Analysis of High-Dimensional Data in the Context of Intrinsic Dimensionality

Dissertation

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

der Technischen Universität Dortmund
an der Fakultät für Informatik

von
Erik Thordsen

Dortmund
2025

Tag der mündlichen Prüfung:	12.05.2025
Dekan:	Prof. Dr. Jens Teubner
Gutachter:	Prof. Dr. Erich Schubert
	Prof. Dr. Martin Aumüller

For my love and my daughter
My anchor and sail
My heart and my home and
Reason to prevail

Abstract

Many modern machine learning applications and data analysis make use of or produce large amounts of high-dimensional data. While it is commonly known, that the performance of many algorithms degrades with increasing dimensionality both in terms of speed and quality, the performance on real-world datasets is often times much better than expected. In fact, real-world datasets tend to occupy only a lower-dimensional manifold in the available observed space, either due to the underlying generative process or the sparsity of the data. To describe that phenomenon, the concept of Intrinsic Dimensionality (ID) was introduced, which describes the minimum number of latent variables to produce the observed manifold. In a pursuit to estimate the ID of non-linear manifolds, small localities of the data are often considered, giving rise to the concept of Local Intrinsic Dimensionality (LID) which aside from estimating a global ID also allows for a spatially resolved analysis of the data. Since the LID eludes direct observation beyond two or three dimensions, it has to be estimated from the data. In this thesis, we explore multiple new approaches to estimate the LID of data in Euclidean spaces, investigate their analytical and empirical properties, and compare them to existing approaches both qualitatively and quantitatively. One of these approaches, the Angle-Based Intrinsic Dimensionality (ABID) estimator, has very useful theoretical properties. We therefore provide an exemplary derivation of ABID to vector fields as a potential control mechanism in algorithms such as Gradient Descent as a showcase of how LID estimation can be useful beyond point clouds. We also investigate if and how LID estimation approaches can be applied to non-Euclidean spaces. While the theory of LID estimation in non-Euclidean spaces remains largely unresolved, we provide a visionary prospect on the future of the field and provide anecdotical evidence for possible future applications of LID.

Acknowledgement

I am extremely grateful for the support, guidance, and cooperation of my supervisor, Prof. Dr. Erich Schubert without whom this thesis would not have been possible. I would like to thank him for his patience, encouragement, and the freedom to explore the research projects of my choice which ultimately led to this thesis. In addition to the academic guidance, I am also thankful for the personal support and the many interesting discussions we had. I am further grateful for the advice and constructive critique of my committee members, Prof. Dr. Christian Janiesch, Prof. Dr. Martin Aumüller, and Prof. Dr. Emmanual Müller. Special thanks to Prof. Dr. Mario Botsch for his mentoring.

I would like to extend my sincere thanks to my current and former colleagues Gloria Feher, Andreas Lang, and Lars Lenssen at the Data Mining Group, as well as the colleagues "next door" Dr. Jan Mühlig, Maximilian Berens, and Roland Kühn from the chair of Databases and Information Systems. The many discussions and conversations gave me lots of necessary support and inspiration. This gratitude extends to all the faculty members and the university, both in research and administration, who have contributed to my academic and personal development. I would like to express my gratitude to Dr. Andre Droschinsky and Prof. Dr. Petra Mutzel for their cooperation during my entry into academic research and for advising me on the path to a Ph.D. Thanks should also go to the Collaborative Research Center 876 at TU Dortmund and the Lamarr Institute for partially funding my research. I am further grateful for the many inspiring and insightful discussions with my international colleagues at conferences and during their visits to Dortmund. They are too many to name, but I am grateful for their fellowship and the many new perspectives they have provided.

I express my deepest appreciation to my friends and family for their continuous support and encouragement. I am grateful to my sports team, the "Rasende Krücken", which included a beginner like me in their ranks, for the many hours of distraction and fun, especially during the COVID years. I am also thankful to my friends from school, university, and beyond, who have been a constant source of support and joy. I want to give special thanks to my family, especially my parents, who have always been there for me and endured my many "when it's done" promises. This thanks extends to my partner's family, who have become my family as well in the past years.

Finally, I could not have undertaken this journey without the support of my partner, who has been a constant source of encouragement, love, and understanding. This thesis has loomed over us for most of our relationship, and I am grateful for her patience, support, and understanding, as well as for our little daughter who has been a constant source of joy and distraction.

Preface

Before delving into the depths of intrinsic dimensionality (ID), I would like to direct a few words to anyone inclined and reflect on the past years. After finishing my master in computer science, I was eager to be a part of academic research - almost independent of the specific field. Since the supervisor of my master's thesis and prospective doctoral supervisor at the time moved to a different university, far outside the area reachable by daily commute, I was recommended (or rather redirected) to Prof. Dr. Erich Schubert. At the time, he was just about to fill the first open positions in his group and I was about to be the first. Luckily enough for me, he proposed one topic closely related to graph theory which previously was my field of choice. Then and even more so in hindsight, joining Erich's group was a great decision for me. I was given the time to compress my master's thesis into a conference paper and partake in my first conference. Although this work did not contribute to my doctoral thesis, it allowed me a first glance at the full scope of academic research, at a time when the prospects of my doctoral topic were just developing. Initially, I was supposed to research topic modeling using graph-based approaches, more specifically using word cooccurrences. Investigating the data, I came to recognize, that e. g. homonyms make segmentation and clustering on the resulting graphs incredibly hard. The difficult words, however, appeared to strongly correlate with an empirically derived measure. Unfortunately, I did not even have the slightest theoretical explanation and inquired about some feedback from Erich. Based on my description, he recommended to take a look at ID and I started climbing down a rabbit hole, that I could not escape within the scope of this thesis. Before running, I first had to learn to walk, though. The field of ID felt rather barren when working with measures of similarity as word cooccurrences happen to be. And deriving a measure for arbitrary similarity spaces seemed too much of a step for a single publication. Thus, my work on similarity-based ID in Euclidean spaces began. The first work, whose derivation, evaluation, and properties make up the majority of this thesis, was well received in the SISAP (Conference on Similarity Search and Applications) community, whose participants produced a decent chunk of the literature on ID. The approach was a straight-forward measure based on simple linear algebra and the simple formulation allows for extensive theoretical derivations not possible with previous related approaches. But it was limited to Euclidean spaces nonetheless and I

still had no theoretical foundation to truly move beyond Euclidean spaces and back to topic modelling. Multiple attempts at reducing non-Euclidean measures to measures of hypothetical manifold geodesics ended up rather fruitless, so I instead tried to generalize the concept of "angle" used in the previous estimator. One way to generalize on that property is to replace the second vector with the linear subspace created by a set of vectors. This produced a set of inequalities that - in the limit - describe "Euclideanness" in a quantifiable way. However, the computational cost of testing this property remained impractical, although some interesting properties emerged from the analysis instead. The second attempt to generalize "angles" was to consider the angles within triangles created by vectors around a point of interest rather than the angle at the point of interest itself. While I had no hopes of pushing beyond Euclideanness in that fashion, I had hopes to find a more robust measure and also a sufficiently orthogonal estimator to pursue non-Euclideanness in terms of an ensemble of estimators. By analyzing a Euclidean proxy space rather than the true non-Euclidean space, a reasonable measure of spatial complexity akin to ID should occur, so I hoped. Such proxy spaces can be shown to exist, but finding a "reasonable" proxy space is rather challenging. This remains an unresolved issue in this thesis, although I hope to make a strong case for the involved methodology in the corresponding chapter. The whole theory of inspecting non-Euclidean spaces is quite fragile for now and I can merely provide a visionary prospect on what might lie in the future of the field. While yet mostly speculation with a slight tint of empirical evidence, ID might just be a missing puzzle piece fitting in the analysis of language data and into the world of algorithmic control heuristics, linking spatial complexity to computational complexity. Especially the latter is almost philosophical, as the general concept of "complexity" just *feels* familiar wherever it is encountered but any attempt to formalize the relationship just slips through one's fingers. But this almost esoteric perspective in a sense fits the domain of ID. After all, the mathematical models do not fit discrete samples, the desired behavior at the boundaries of a dataset or in the transition between two dimensionalities is unclear, and practically relevant datasets have no true ID. These are all shortcomings of attempting to practically measure a property that lives in the infinitesimal, best explained by "counting to infinity with your fingers". But there *are* practical approaches to do so and they *do* yield meaningful values. ID may feel esoteric sometimes, but, I think, it may also just be our best shot at some theoretical problems. I hope this thesis condenses my work and insights on the topic in a most "digestible" way and inspires new ways of applying the concept of ID in yet unforeseen ways. More years than I initially would have liked to invest have culminated in this thesis, prolonged by the occasional more-or-less-related side project and some fortunate and unfortunate personal incidents. Alas, the years between 2019 and 2023 were quite special for most of us.

Erik Thordsen
Dortmund, February 7, 2025

Contents

Abstract i

Acknowledgement iii

Preface v

List of Symbols ix

List of Figures xiii

List of Acronyms xv

Glossary xvii

1 Introduction 1

2 Intrinsic Dimensionality 5

2.1 Latent Spaces and Intrinsic Dimensionality 6

2.2 Relation to Other Machine Learning Tasks 8

2.3 Global Linearity 10

2.4 Local Linearity 12

3 Angle-Based Intrinsic Dimensionality 17

3.1 Derivation of $\mathrm{LID}_{\mathrm{ABID}}$ 18

3.2 About the choice of moments 25

3.3 $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{PCA}}$ 28

3.4 $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$ 34

3.5 Analytical Comparison of $\mathrm{ID}_{\mathrm{ABID}}$ and $\mathrm{ID}_{\mathrm{MLE}}$ 38

3.6 Empirical Evaluation 54

3.7 Gradient Field LID 65

3.8 Conclusion 70

4 Random In-Data Projections and LID **73**
4.1 Pivotal Bounds In Euclidean Spaces 74
4.2 Expected Variance Of Random Projections 78
4.3 Random Projections and ID Estimation 83
4.4 Pivot Filtering Linear Scan 84
4.5 Conclusion 87

5 Chordal Angles-Based Intrinsic Dimensionality **89**
5.1 Derivation of LID_{ALC} 90
5.2 Empirical Evaluation 103
5.3 Conclusion 115

6 Beyond Euclidean **117**
6.1 Some Anecdotal Evidence 117
6.2 Properties of Euclidean Space and Distance 122
6.3 Is (Very) Non-Euclidean Useful? 126
6.4 LID in non-Euclidean space 130
6.5 Conclusion 137

7 Quo Vadis LID? **139**

List of Publications **145**

Bibliography **152**

List of Symbols

Notation	Description
	Basic variables
$\alpha, \beta, \ldots$	Scalar variables
$i, j, \ldots$	Integer indexing variables
$a, b, \ldots$	Vector variables
$A, B, \ldots$	Matrix variables
$X, X_1, X_2, \ldots$	Datasets consisting of vectors
	Geometry
d	Representation dimension of data / vectors
δ	Intrinsic dimensionality of data / manifolds
$\mathbf{0}_d$ / $\mathbf{1}_d$	The all-zeros / all-ones vector in d dimensions (d ommitted when clear)
$\mathbb{0}_d$ / $\mathbb{1}_d$	The all-zeros / all-ones matrix with d rows and columns (d ommitted when clear)
$\mathbb{I}_d$	Identity matrix with d rows and columns (d ommitted when clear)
$\mathscr{V}_d(r)$	Volume of the d-ball with radius r
$\mathscr{A}_d(r)$	Surface "area" of the d-ball with radius r
	Algebra and Analysis
$\|x\|_p$	The Minkowski p-norm of a vector x
$\|x\|$	The norm of a vector x ($\|x\|_2$ in Euclidean space)
$\langle x, y \rangle$	Inner product of two vectors x and y (dot product in Euclidean space)
$\mathrm{tr}(A)$	Trace of a (square) matrix A
∇f / $\nabla^{(1)} f$	Jacobian of a function f
$\nabla^{(2)} f$	Hessian of a function f
$\nabla^{(k)} f$	Tensor of k-th-order partial derivatives of f

Notation	Description
$\lambda(A)$ / $\lambda_i(A)$	Vector of eigenvalues / i-th eigenvalue of a matrix M
$v_i(A)$	i-th eigenvector of a matrix A corresponding to $\lambda_i(A)$
$\Lambda(A)$	Diagonal matrix with $\lambda_i(A)$ as i-th diagonal entry
$V(A)$	Matrix with $v_i(A)$ as i-th column
A^α	α-th (spectral) power of A, i. e. $V(A^\alpha) = V(A)$ and $\lambda_i(A^\alpha) = \lambda_i(A)^\alpha$
$A^{\circ\alpha}$	α-th Hadamard power of A, i. e. $A_{i,j}^{\circ\alpha} = (A_{i,j})^\alpha$
	Statistics
$\mathbb{E}[\ldots]$	Expected value of a term
$\mathbb{E}^{(k)}[\ldots]$	k-th central moment of a term
$\mathbb{E}^{*(k)}[\ldots]$	k-th non-central moment of a term
$\mathrm{Var}[\ldots] = \mathbb{E}^{(1)}[\ldots]$	Variance of a term over reals
$\mathrm{Cov}[\ldots] = \mathbb{E}^{(2)}[\ldots]$	Covariance matrix of a term over vectors
$\mathrm{Cov}^*[\ldots] = \mathbb{E}^{*(2)}[\ldots]$	Non-central covariance matrix of a term over vectors

List of Figures

2.1 Non-convex set in parameter space and its embedding 7
2.2 Examples for different intrinsic dimensionalities 7
2.3 A dataset with two different local intrinsic dimensionalities. 8

3.1 t-th moments of absolute cosines on the sphere 26
3.2 Empirical error for Local Intrinsic Dimensionality (LID) estimation with the t-th moment of absolute cosines on the sphere 27
3.3 Isolines of Angle-Based Intrinsic Dimensionality ($\mathrm{LID}_{\mathrm{ABID}}$) and Principal Component Analysis (PCA) in the eigenvalue space 29
3.4 Probability density of the eigenvalues of random covariance matrices with $\mathrm{LID}_{\mathrm{ABID}}$ isolines 30
3.5 Empirical probability density of the eigenvalues of random matrices with "true" LID shading and Raw Angle-Based Intrinsic Dimensionality ($\mathrm{LID}_{\mathrm{RABID}}$) isolines 32
3.6 Empirical derivation of radians-based LID on the sphere 35
3.7 Empirical derivation of chord length-based LID on the sphere 36
3.8 Empirical convergence of $\mathrm{LID}_{\mathrm{ABID}}$ over neighborhood size 40
3.9 Empirical required neighborhood size for convergence of $\mathrm{LID}_{\mathrm{ABID}}$ over LID 41
3.10 Empirical required neighborhood size for convergence of $\mathrm{LID}_{\mathrm{ABID}}$ over representation dimension 42
3.11 Correlation of covariance eigenvalue estimation error and magnitude . . . 42
3.12 Asymptotic error percentiles of Maximum Likelihood LID Estimator ($\mathrm{LID}_{\mathrm{MLE}}$) over neighborhood size 44
3.13 Asymptotic error percentiles of $\mathrm{LID}_{\mathrm{MLE}}$ over LID 44
3.14 Examples of synthetic non-linear embeddings 47
3.15 Effect of local non-linearity on $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$ with varying representation dimension 48
3.16 Comparison of theoretical and observed error of $\mathrm{LID}_{\mathrm{MLE}}$ for assumption confirmation for non-linear embeddings over varying embedding dimension 49
3.17 Effect of local non-linearity on $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$ with varying LID . . . 50

3.18 Comparison of theoretical and observed error of $\mathrm{LID}_{\mathrm{MLE}}$ for assumption confirmation for non-linear embeddings over varying intrinsicd dimension 51
3.19 Effect of full-dimensional embedding noise on $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$ with varying LID 52
3.20 Comparison of theoretical and observed error of $\mathrm{LID}_{\mathrm{MLE}}$ for assumption confirmation for additive noise 53
3.21 LID comparison on Koch snowflake 55
3.22 LID comparison on M6 dataset 56
3.23 LID stability comparison on M6 dataset 56
3.24 LID comparison on M10c dataset 57
3.25 LID comparison on 8-dimensional noisy lattice 58
3.26 LID comparison on nested hypercubes 58
3.27 LID comparison on adversarially perturbed 5-ball 60
3.28 LID comparison on adversarially perturbed swiss roll 60
3.29 LID comparison on transition from one to two dimensions using a diamond shape 61
3.30 LID comparison on transition from one to two dimensions using a rectangle shape 62
3.31 $\mathrm{LID}_{\mathrm{ABID}}$ vs Tight LID Estimator ($\mathrm{LID}_{\mathrm{TLE}}$) on lollipop dataset 62
3.32 $\mathrm{LID}_{\mathrm{ABID}}$ vs $\mathrm{LID}_{\mathrm{TLE}}$ on Lorenz system 63
3.33 LID comparison on MNIST and Gisette datasets 64
3.34 Comparison of sampled and approximated Gradient Field Local Intrinsic Dimensionality ($\mathrm{LID}_{\mathrm{GFL}}$) 67
3.35 Comparison of different neighborhood definitions for Normalized Gradient Field Local Intrinsic Dimensionality ($\mathrm{LID}_{\mathrm{NGFL}}$) 68
3.36 Convergence of gradient descent with $\mathrm{LID}_{\mathrm{NGFL}}$ controlled learning rate . . 68
3.37 $\mathrm{LID}_{\mathrm{GFL}}$ and $\mathrm{LID}_{\mathrm{NGFL}}$ estimates on the Lorenz system 69

4.1 Eligible search space after pivot filtering with combined vs intersected bounds 77
4.2 Empirical running times of the pivot filtering linear scan index and estimated optimal number of pivots on MNIST 86
4.3 Running times of the pivot filtering linear scan index on ALOI projections of varying dimensionality 86

5.1 Concept figure for the $\mathrm{LID}_{\mathrm{ALC}}$ proof 90
5.2 Theoretical vs Beta fit moments of chordal triangle cosines 97
5.3 Theoretical vs linearly parameterized Beta fit moments of chordal triangle cosines 97
5.4 Relative error of the linearly parameterized Beta fit for LID estimation . . 99
5.5 Absolute error of the linearly parameterized Beta fit for LID estimation . . 99

5.6 Beta fit comparison for cosines in the d-ball 101
5.7 Relative error of LID estimation with $\mathrm{LID}_{\mathrm{BALL}}$ and $\mathrm{LID}_{\mathrm{BALE}}$ 102
5.8 Convergence behavior of the Approximately Linearly-parameterized Chordal Angles LID ($\mathrm{LID}_{\mathrm{ALC}}$) estimator over the number of sampled cosines . . . 104
5.9 Convergence behavior of the $\mathrm{LID}_{\mathrm{ALC}}$ estimator over the number of neighbors 105
5.10 Comparison of LID errors in ideal neighborhoods of varying dimensions . 106
5.11 Comparison of LID errors in biased neighborhoods 107
5.12 Comparison of LID errors in skewed neighborhoods 108
5.13 LID comparison on primitive high-dimensional datasets 110
5.14 LID comparison on synthetic high-dimensional datasets 111
5.15 LID comparison for class separation on the MNIST dataset 113
5.16 LID comparison for class separation on the Fashion-MNIST dataset 113

6.1 $\mathrm{LID}_{\mathrm{ABID}}$ estimates on Levenshtein distances and approximate Euclidean embeddings 118
6.2 $\mathrm{LID}_{\mathrm{ALC}}$ estimates on Levenshtein distances and approximate Euclidean embeddings 118
6.3 Showcase of local "diversity" in MNIST images using $\mathrm{LID}_{\mathrm{ABID}}$ estimates on normalized conditional mutual information 120
6.4 Critical α values for an extremely non-Euclidean distance matrix 128
6.5 Critical α values for the Levenshtein distance 129
6.6 Critical α values for the Manhattan distance 129
6.7 Critical α values for the Great-Circle distance 129
6.8 Distribution of critical α and corresponding non-Euclideanness score values for generated distance matrices 132
6.9 Correlation of canonical and non-Euclidean LID estimates with polynomial fit 132
6.10 Critical α prediction using a combination of LID estimates 133
6.11 Correlation of canonical and non-Euclidean LID estimates for $\mathrm{LID}_{\mathrm{RABID}}$. 134
6.12 Approximation error of power transform and Gramian-based approximation to non-Euclidean distances 135
6.13 Correlation of canonical and Gramian-based approximation LID estimates 135
6.14 Correlation of canonical and Gramian-based approximation $\mathrm{LID}_{\mathrm{RABID}}$ estimates 135

List of Acronyms

Notation	Description
PCA	Principal Component Analysis [66]
EDM	Euclidean Distance Matrix [1]
ID	Intrinsic Dimensionality
LID	Local Intrinsic Dimensionality
LID_{PCA}	PCA Intrinsic Dimensionality [31]
LID_{ABID}	Angle-Based Intrinsic Dimensionality [85]
$\text{LID}_{\text{RABID}}$	Raw Angle-Based Intrinsic Dimensionality [85]
LID_{GFL}	Gradient Field Local Intrinsic Dimensionality
LID_{NGFL}	Normalized Gradient Field Local Intrinsic Dimensionality
LID_{TRIP}	Thresholded Random In-distribution Projections Intrinsic Dimensionality [87]
LID_{ALC}	Approximately Linearly-parameterized Chordal Angles LID
LID_{BALL}	Hyperball LID_{ALC} estimator with linear parameters
LID_{BALE}	Hyperball LID_{ALC} estimator with exponentially corrected parameters
LID_{MLE}	Maximum Likelihood LID Estimator [42, 54]
LID_{MOM}	Measure-of-moments LID estimator based on distances [2]
LID_{GED}	Generalized Expansion Dimension LID estimator [47]
LID_{ALID}	Augmented LID Estimator [19]
LID_{TLE}	Tight LID Estimator [5]
LID_{NC}	"Natural Community"-LID [72]
LID_{MiND}	Minimum Neighbor Distance Intrinsic Dimensionality [71]
LID_{IDEA}	Intrinsic Dimensionality Estimation Algorithm [71]

Notation	Description
$\mathrm{LID}_{\mathrm{DANCo}}$	Dimensionality from Angle and Norm Concentration LID Estimator [17]
$\mathrm{LID}_{\mathrm{ANOVA}}$	Analysis of Variance LID Estimator [26]
$\mathrm{LID}_{\mathrm{FCI}}$	Full Correlation Integral LID Estimator [28]
$\mathrm{LID}_{\mathrm{LIDL}}$	LID estimator using approximate Likelihood [82]

Glossary

Curse of Dimensionality The Curse of Dimensionality refers to the different problems that arise in high-dimensional Euclidean space, such as the increased computational cost of calculating with more coordinates or the vanishing variance of distance that lead to a decreased discriminability.

Principal Component Analysis A method to obtain the directions of closest fit to the data which in return equals the directions of largest spread of the observed data.

Euclidean Distance Matrix The matrix of *squared* pairwise Euclidean distances of a finite sample or a matrix for which an according Euclidean point set exists.

Intrinsic Dimensionality An estimated value of the minimum required amount of dimensions to represent given data.

Local Intrinsic Dimensionality A "localized" version of Intrinsic Dimensionality typically evaluated on the k-nearest neighbors of some point.

Chapter 1

Introduction

The availability of and necessity for large datasets is continuously increasing. Both the economical gain assumed by accumulating, selling, and mining large datasets, and the hunger for more data of current developments like large language models, deep learning-based vector embeddings, and many more contribute to this development. The introduction of vector embeddings further allows to consider almost any data domain in terms of numerical (Euclidean) vectors, expanding applications based on Euclidean space to near arbitrary domains. Such embeddings, however, are typically very high-dimensional and difficult to interpret or understand. Additionally, the vanishing variance of distances in higher dimensions as well as the increased number of coordinates lead to worse performance or quality of algorithms and data structures, which is known as the Curse of Dimensionality. Yet, not all dimensions are equally important as e. g. repeating all dimensions does not change almost any of a dataset's characteristics. This led to the introduction of the term Intrinsic Dimensionality (ID), which describes the number of dimensions minimally required to represent all information of a dataset – up to a relatively small amount which one might consider as "measuring noise". The first practical ways to estimate this number were based on PCA [66, 31] and considered everything below a threshold of the total variance as noise. Considering, that datasets may as well be a union of multiple subspaces of different characteristics, the concept of ID was later generalized to LID, which produces an estimate for each sample in the dataset rather than one for the entire dataset. (Local) ID has thus expanded the set of analytical tools to investigate properties of a dataset at hand such as summary statistics, a popular first step that can be misleading [59], or other tools such as OPTICS [6] that provide insights on the cluster structure of a dataset. Aside from simply informing a practitioner about a dataset, it has practical applications in dimension-reduction to choose appropriate parameters for algorithms such as Isomap [83] or t-SNE [57], control flow to over- or under-fit on high-dimensional parts of a dataset [72], adversarial attack detection [4], and Generative Adversarial Network quality verification [9] to name a few. This thesis is focused on exploring both the esti-

mation of LID in terms of newly introduced estimators as well as some potential future use-cases, that are mostly unexplored as of now. While there previously were multiple estimators available to estimate the LID, none of them appeared sufficiently applicable to our initial problem of detecting anomalies (homonyms/synonyms/...) in word cooccurrences. The data at hand encoded a general similarity measure, for which no LID estimation method was available at the time – aside from first mostly arbitrarily translating the similarities to distances and using these for LID estimation. We thus focused on introducing LID estimators based on similarity values, complementing a vast literature of distance-based estimators. The field of LID estimation as well as practical applications thereof are a currently developing topic and this thesis aims to both introduce the reader to the general concept and our contributions and inspire them to new applications.

The structure of this thesis mostly follows the chronological order of our discoveries, with individual sections added posteriorly. Chapter 2 introduces the concepts relevant to ID, discusses its relation to other fields, and gives an overview of preceding and simultaneously developed estimators. Chapters 3, 4, and 5 then introduce a series of intrinsic dimensionality estimators, and give further insights on their properties and theory derived thereof. The estimator in Chapter 3 was our first contribution to the field of intrinsic dimensionality and has the generally most useful properties as it is closely related to the local covariance of samples. It thus lends itself to thorough analysis and allows for potential applications in vector fields where at least numerical differentiation is possible. The estimator in Chapter 4 emerged from a reflection on the boundaries of "Euclideanness" and is related to the expected variance of samples in subspaces created from directional vectors in the dataset. Asserting these directional vectors to be aligned with geodesics relates the expected values to ID and allows us to predict the dimensionality required for sufficient accuracy in downstream tasks. The estimator family in Chapter 5 is the latest contribution and extends the approach of Chapter 3 towards a different set of angles. The number of these chordal angles is cubic in the neighborhood size rather than quadratic and, thus, they lead to hypothetically more robust estimates. While the domain change also changes the interpretation of the estimated dimensions, the estimates are more reliable in various settings. After introducing multiple estimators, Chapter 6 discusses the possible application of intrinsic dimensionality estimators to non-Euclidean spaces, even in the absence of a fully satisfying theory behind it. We first give a few motivating anecdotes highlighting the efficacy of our estimators in settings, where they are not proven to work. After that, we discuss the meaning of "Euclideanness" and propose an argument for why practically useful metrics are likely "near-Euclidean" – at least locally, but potentially also globally. We proceed by empirically analyzing how intrinsic dimensionality estimates in non-Euclidean spaces relate to a certain type of Euclidean proxy spaces. This analysis displays, that the tested estimators give strongly correlated estimates on non-Euclidean and a canonicalized Euclidean proxy spaces. While these results are still some-

what premature and require further theoretical inquiry, we hope to convince the reader of the usefulness of intrinsic dimensionality in general metric spaces and potentially even in non-metric spaces. Chapter 7 closes this thesis by recapitulating the contents of the thesis and giving an outlook of possible future applications of intrinsic dimensionality. These potential prospects are mostly without sufficient theoretical backing and should merely be understood as an optimistic vision of avenues to explore.

Chapter 2

Intrinsic Dimensionality

As previously motivated, the concepts of Intrinsic Dimensionality (ID) and Local Intrinsic Dimensionality (LID) are aimed at evaluating how many independent features, i. e. dimensions, are minimally required to properly describe a given set of observations. However, we have not yet explained what "properly describe observations" means. The intuitions are almost as numerous as approaches to ID estimation in the literature, yet the approaches can be grouped by their induced semantics. Here, we only discuss the intuitions behind the two largest groups of ID estimators: Correlation-/covariance-based and expansion-/distance-based estimators.

Expansion-based estimators such as the $\mathrm{LID}_{\mathrm{MLE}}$ [54] or the Augmented LID Estimator ($\mathrm{LID}_{\mathrm{ALID}}$) [19] are based on the idea that data should be described by a number of features that induce a space "expanding" as rapidly as the data itself. The expansion rate of these features is linked to the expansion rate of bodies in Euclidean geometry like hyperballs and hypercubes, which expand exponentially in the number of dimensions. In either case, the distances of uniform distributions in these bodies are considered. The concept of expansion-based ID has been linked to the indiscriminability of the distance measure [44]. While this link theoretically lifts distance-based ID beyond the limitations of Euclidean space, this connection has only been proven in an infinitesimal setting that is practically unattainable.

Covariance-based estimators such as $\mathrm{LID}_{\mathrm{ABID}}$ [85] or PCA Intrinsic Dimensionality ($\mathrm{LID}_{\mathrm{PCA}}$) [31] rather focus on the geometrical spread of observations. The ID is deduced from the number of dimensions of a local linear embedding that describes the correlations inside the data well enough. Angle-based approaches like $\mathrm{LID}_{\mathrm{ABID}}$ or $\mathrm{LID}_{\mathrm{ALC}}$ consider the distribution of different types of triangles inside a neighborhood while e. g. $\mathrm{LID}_{\mathrm{PCA}}$ fits a multivariate normal distribution to the data. As these estimators are based on the dot product, their domain is limited to Euclidean spaces – or at least isometrically Euclidean embeddable spaces.

Although the motivations diverge, the goal of both estimator groups – and the numerous other estimators like the nearest-neighbor-graph-based estimator by Costa and Hero [22] – is, therefore, to find the (minimum) number of dimensions of a distribution (in Euclidean space) that behaves most or sufficiently similar to the observations. When considering observations from Euclidean space, i. e. Euclidean vectors, this naturally induces an appropriate linear subspace on which to project observations. The linearity of these subspaces is not obvious, yet, required by the basic geometrical motivations of all previously mentioned estimators. The uniform distributions within hyperballs and on hyperspheres are only comparable to the observed data if both are considered in the same "flat", i. e. not non-linearly deformed, Euclidean space. Using the idea of sufficiently precise subspaces, denoted as latent or parameter spaces, allows for an analytical description of the observations as samples from a generative process. The concept of latent spaces also immediately links ID to dimension reduction and other tasks, which will be discussed in detail in Section 2.1. The importance of LID to other Machine Learning tasks is discussed in Section 2.2. Section 2.3 gives an overview of existing approaches to ID introduced in the literature, which is expanded by a list of LID methods in Section 2.4.

2.1 Latent Spaces and Intrinsic Dimensionality

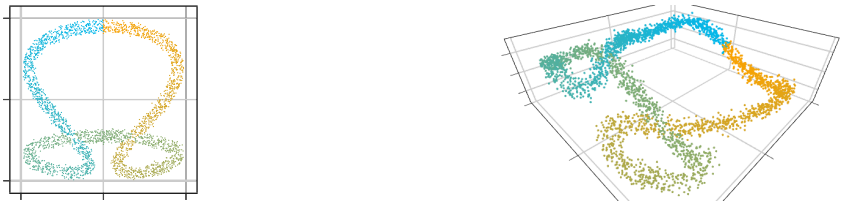
The concept of latent spaces [5, 17, 69] is that observed samples stem from some (possibly lower-dimensional) latent space and are embedded via some embedding function into the observed space. Assuming a smooth embedding function f and a latent vector $y \in \mathbb{R}^\delta$ we would, therefore, observe $x \in \mathbb{R}^d$ given by

$$x = f(y) + \xi \tag{2.1}$$

where ξ is a (possibly 0) error term. Since the latent vector y parametrizes the embedding function f, the latent space is also referred to as the parameter (or parameterization) space [22]. This induces a probability density function in our observed space P_o given by

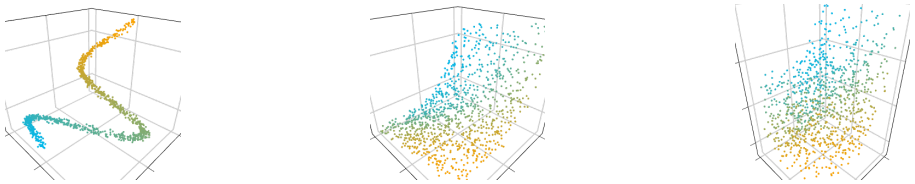
$$P_o(x) = \left(\sum_{y \in f^{-1}(x)} P_l(y) c(f, y) \right) + \Xi(x) \tag{2.2}$$

where P_l is some probability density function in latent space, $\Xi(x)$ is the error term resulting from ξ in the previous equation, and $c(f, y)$ is a correcting factor accounting for compression or expansion of f at y, akin to the surface element in surface integrals. Since the embedding function does not need to be bijective, the preimage of x under f can consist of less or more than one point. Whenever f is bijective on $\{y \mid P_l(y) \geq \varepsilon \geq 0\}$ it projects onto a manifold in the observed space, i. e. a subspace which locally resembles a flat and possibly lower-dimensional Euclidean space. Non-bijective embedding functions lead to self-intersections (singularities) of the varieties – generalizations of manifolds that allow



(**a**) Parameter Space (**b**) Embedding Space

Figure 2.1: A non-convex sample in the parameter space (a) leading to a non-linear embedding in the embedding space (b). The colors indicate a potential one-dimensional parameterization with self-intersection that is locally indistinguishable from the two-dimensional space given the samples.



(**a**) Intrinsically 1d (**b**) Intrinsically 2d (**c**) Intrinsically 3d

Figure 2.2: Different intrinsically dimensional manifolds embedded in three dimensions. Up to noise, the manifolds are 1-dimensional a, 2-dimensional b, and 3-dimensional c, respectively.

for these singularities – onto which the support of P_I is projected. Around singularities, however, the varieties are locally indistinguishable from a higher-dimensional manifold. The entire variety might then be considered a higher-dimensional manifold created from a latent space with a non-convex support as displayed in Fig. 2.1. Approaches to ID estimation, therefore, only regard the bijective case, in opposition to the field of Manifold Learning in which detection and elimination of singularities prior or during their main tasks has been considered [24]. The objective of ID estimation is to find the number of latent dimensions, i. e. the dimensionality of the support of P_I. When assuming that the distribution given by P_I has a variance of $\gg 0$ in every direction, this is identical to the dimensionality of the domain of P_I. The intuition of ID estimation on bijective embedding functions is visually easy to grasp: If the data looks like a curve, it is intrinsically 1-dimensional, if it looks like a deformed plane, it is intrinsically 2-dimensional, and so on (see Fig. 2.2). Since practical approaches, however, do not work on distributions but rather observations sampled from these distributions, we only ever see a limited part of the support of P_I. This might lead to a set of observations with locally quite different dimensionalities as displayed in Fig. 2.3. Due to this locally varying behavior, the concept of ID had been generalized to Local Intrinsic Dimensionality (LID). Instead of inspecting an entire dataset, we rather inspect "localities", e. g. k-nearest-neighborhoods, around any

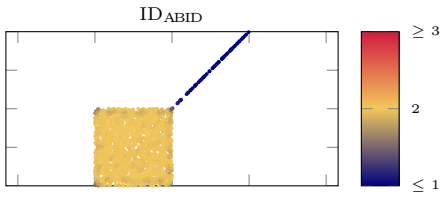


Figure 2.3: A dataset with two different local intrinsic dimensionalities.

observation in the dataset, and assign a LID value to each considered observation. The overall ID of the dataset can then be obtained by inspection of the histograms of these LID estimates or by the mean of all LID values [69]. In general, any ID estimator can be used as a LID estimator by applying it to each locality independently and vice versa. Estimators in the literature, however, are designated as either ID or LID estimators since LID estimators introduce assumptions on the embedding function and the latent distribution that are sound on a localized scale, yet, will likely not hold on an entire dataset.

To account for a mixture of distributions, the model can further be expanded to consider multiple latent distributions and embedding functions as

$$P_o(x) = \left(\sum_i \sum_{y \in f_i^{-1}(x)} P_{l,i}(y) c(f_i, y) \right) + \Xi(x) \tag{2.3}$$

The overlap of the observed manifolds can be an issue since they cannot easily be separated, potentially leading to even larger LID values than the dimensionality of the individual latent spaces. Everywhere else, the LID values indicate the dimensionality of the latent space of the $P_{l,i}$ whose support maps to this part of space. In this case, global ID values or mean LID values lose their meaning. Global ID estimators and aggregated LID values (mean, median, ...) should, therefore, only be used when the dataset is assumed to consist of exactly one manifold, e.g. in (linear) dimensionality reduction.

The models introduced in this section are solely used to explain the idea behind LID estimation and its relation to other problems. They will, however, not be used in the derivation of the estimators. While e.g. the angle-based estimator can be expressed in terms of the Jacobian of the embedding function [86], the derivations of the estimators rather consider ideal neighborhoods.

2.2 Relation to Other Machine Learning Tasks

After introducing ID and LID we can now use the models (2.2) and (2.3) to explain the relation of ID estimation to other machine learning tasks and consequentially to potential application fields. Whenever the dimensionality of the latent space δ is (much) smaller than the dimensionality of the observed space d, the number of features can be reduced

within reasonable loss of information on local relations between observations by approximating the preimage of the observations under the embedding function. This task is known as dimensionality reduction with approaches like PCA [66] for linear embedding functions and Autoencoders [51], Locally Linear Embedding (LLE) [70], or t-SNE [57] for locally linear embedding functions.

Approaches for locally linear embedding functions are also called Manifold Learning since they approximate the latent spaces of manifolds. These approaches do not propose a selection method for the proper amount of latent dimensions and rather require them as a user-specified parameter. Methods for LID estimation can be used to estimate this parameter. Tasks in Manifold Learning include e. g. functional approximations of the data manifold and distance computations on the data manifold, known as geodesics [83]. Since singularities of the manifolds result in errors in the geodesics, a part of the Manifold Learning literature focuses on eliminating these singularities [24]. LID approaches do not explicitly consider singularities, yet their estimated LID values tend to diverge from nearby estimates. This observation has led to the application of ID estimators to detect singularities as anomalous LID values comparable in efficacy to approaches not based on LID [68]. Partially, the goals of Manifold Learning and LID estimation are similar. LID estimation is more focused on estimating the dimensionality of the latent space. Manifold Learning instead focuses on the approximation of the latent space itself. This results in some similar tasks and methods, yet the objectives and applicability of the results differ.

When considering noise or outlier observations as an additional full-dimensional manifold or a set of additional 0-dimensional manifolds, i. e. single points, depending on their relative density, the concept of LID becomes applicable to tasks regarding noise-detection. E.g. the task of Outlier Detection can be described as isolating all observations belonging to a distinctly higher- or lower-dimensional latent space. That is, we are again interested in isolating a set of observations based on their LID. In practice, using LID approaches to identify outliers is obstructed by the problem of choosing a proper locality. Since outliers are much fewer than inliers, selecting k-nearest-neighborhoods around noise observations leads to over-proportionally selecting neighbors from the lower-dimensional manifolds likely to lie in a narrow cone viewed from the noise observations. Depending on the specific LID estimator, the LID estimate of the noise point is very likely distinctly different from that of its neighbors, yet, not necessarily equal to the dimensionality of the observed space. Whilst it is possible to use LID estimators for outlier detection, practical results are not yet reaching the state-of-the-art [48]. Similarly, LID estimators have been proposed to detect adversarial attacks in deep neural networks. Adversarial attacks are considered specific inputs, that are specifically designed to cause a misclassification of a deep neural network, potentially whilst appearing to be in-distribution of the training data. An example could be a picture of a cat with a small amount of noise added, which is then

classified as a dog. The LID estimates of latent vectors using the nearest latent vectors of the training data as localities, were successfully evaluated for anomalies [56, 4].

The mixture of manifolds model (2.3) is further interesting in any setting where the different manifolds have different LIDs like clustering or classification of differently intrinsically dimensional subspaces. Each cluster or class can be considered as an independent latent space and embedding function resulting in the observed dataset. Whenever the corresponding manifolds differ in dimensionality, LID values can be used to identify these separate subspaces. This approach can be used as a clustering algorithm and can be expanded to other spaces such as image segmentation [16].

LID-motivated measures like the "Natural Community"-LID ($\mathrm{LID}_{\mathrm{NC}}$) [72], intend to give a notion of the complexity of neighborhoods within graphs. They can be used to impact the control flow of graph embedding algorithms to improve the quality of the results. The concept of using LID values not only to parameterize algorithms but effectively controlling the behavior of algorithms dependent on high-/low-LID localities is a promising advance, yet, barely explored. Similarly, we proposed a method for the opposite direction – adopting LID methods to better suit specific downstream applications. We created a LID estimator that accounts for an expected observation error named $\mathrm{LID}_{\mathrm{TRIP}}$. This estimator is intended to obtain better estimates in methods for which small errors of Euclidean distance values are acceptable, such as k-nearest neighbor indexing where the binary decision of "being a neighbor" is invariant under small errors of distances. Although derived under the assumption of a multivariate normal distribution, the approach performed well on other distributions as well, advancing the applicability of LID estimators to a broader range of tasks [85].

Lastly, the Jacobian of the embedding function(s) can be approximated using the local covariance of samples in the embedded space. This property is based on a locally uniform distribution in parameter space, i. e. a locally linear probability density function, which approximately holds in small enough neighborhoods. Using that observation, we introduced an oversampling technique that does not change the LID and gives additional samples closer to the sampled manifold than e. g. linear interpolation [86]. Since this work does not fit the topic of this thesis, it is not further expanded upon here.

2.3 Global Linearity

For the described manifold models, we only assumed that the embedding function f is smooth. This assumption alone is not useful for the task of ID estimation, since an entirely arbitrary f could project e. g. a line onto a 20-dimensional hyperball as a space-filling curve or vice-versa by collapsing the variance along all but one dimension. Any meaningful approach to ID estimation, therefore, must involve assumptions on at least f and potentially on P_I and c. On a global scale, an intuitive assumption is for f to be a linear

function, whereby $c(f, \bullet)$ is constant, and for ξ to be "sufficiently" small. The distribution in observed space then has a variance "relatively close" to 0 in any direction independent of the column vectors of the Jacobian ∇f of f as induced by ξ. The variance in dependent directions, i. e. along the manifold, should be much larger. These assumptions make the model very similar to the objectives of both PCA [66] and Bayesian PCA [13] which are well-known techniques for dimension reduction by eliminating correlated features within the data. The close-to-zero-variance directions are meaningless to any machine learning method applied to this data since they do not contain any information on the latent space and, therefore, the generating process. Instead, they increase the computational cost of machine learning methods and potentially induce highly correlated features which negatively impact the efficacy of methods like naïve Bayesian Classification, where the redundancy of a feature increases the importance exponentially. Accordingly, PCA [66] is used to obtain an orthogonal set of directions and variances of the observations in these directions. These vectors correspond to the eigenvectors of the covariance matrix of the observations and, when ordered by the corresponding eigenvalues, i. e. the variance in this direction, each eigenvalue is the maximum variance in any direction after eliminating the directions corresponding to any larger eigenvalue. In other words, the largest and smallest non-zero eigenvalues of the covariance matrix correspond to the maximum and minimum variance of the observations in any direction. Removing the directions corresponding to the smallest eigenvalues, i. e. projecting onto the directions with the largest eigenvalues, yields a linear embedding that conserves most of the observed variance. To remove all close-to-zero-variance directions one "simply" has to remove all eigenvalues that fail some predicate. Since the eigendecomposition of the covariance matrix is already provably optimal in terms of explained cumulative variance [66], PCA and its variants are the preferred methods for globally linear ID estimation. The task, thus, reduces to finding the "correct" number of dimensions to keep, for which many heuristics have been developed. A popular method is to retain the largest eigenvalues up to a certain percentage of the total variance [31] which e. g. is one of the standard methods implemented in libraries like the Python library Scikit-learn [67]. However, many other heuristics like removing all eigenvalues after the first one below a certain percentage of the largest eigenvalue [30] and much more sophisticated methods [63] have been developed. These heuristics are in principle ID estimation techniques [31, 30] – although not necessarily developed for that purpose – and will here be denoted as $\mathrm{LID}_{\mathrm{PCA}}$ where the specific heuristic is mentioned when used.

Although other ID estimation techniques have been specifically designed to estimate one global ID value, they do not share the global linear assumption. Instead, they are evaluated over a local region of the data and then aggregated to a global ID estimate. Since they share the local assumptions of estimators designed for local ID estimation, they will be discussed jointly in Section 2.4.

2.4 Local Linearity

In the previous section, we constrained the embedding function f to be (globally) linear. Resulting of that constraint, PCA is – in terms of remaining variance – the optimal choice. However, the assumption of a linear embedding function is very strong and does not allow for the modeling of non-linear manifolds. To relax the constraint, we can consider the embedding function to be locally linear, whereby c is locally almost constant. We will still assume ξ to be small, though. Local linearity might still appear a significant assumption, although it is practically almost non-existent. In practice, we are confronted with a finite sample of the manifold. Even if the underlying embedding function were to be locally non-linear, the observable data would be indistinguishable from a linear embedding function with either an additional dimension or a high local curvature, since a locally linear interpolation trivially exists. The assumption of local linearity is, therefore, a very reasonable one and is the basis for the concept of LID estimation. Since there are many different approaches to LID estimation, not all estimators can be discussed in this section, but we intend to give the reader a good overview of the historic progress of LID estimation.

The concept of LID estimation is founded on the functional representation of the manifold model and dates back to concepts like the Hausdorff Dimension (1918) [37] – the fundamental work of fractal dimensions. In this function-oriented work, the dimensionality is defined in terms of the expansion rate of a curve compared to its parameters in infinitesimal neighborhoods. These properties, when reparameterizing the embedding function to map the origin of the latent space to the considered point in the embedding space, properly describe the local behavior of the underlying manifold, but are not directly applicable to finite datasets. Practical approaches for LID estimation were, therefore, based on the assumption that the local behavior of the manifold can be approximated by the behavior of the manifold in "small" neighborhoods, typically defined by the nearest neighbors. The earliest data-centric approach to LID estimation is the Correlation Dimension by Grassberger and Procaccia (1983) [33]. This approach considers the number of pairwise distances less than a certain threshold relative to that threshold value. Since geometric bodies in Euclidean space expand exponentially in the number of dimensions, that number can be expected to grow exponentially with the number of dimensions. The discrete nature of that measure, however, makes it very sensitive to the choice of the threshold and the number of observations. The first "school of thought", attempted to improve on the numerical issues and empirical precision of that estimator. These approaches started by fitting a power-law to the distances of the k-nearest neighbors divided by the distance of the $(k{+}1)$-nearest neighbor. The first author to propose such an approach was Takens in 1985 [80]. He calculated the maximum likelihood estimate to be the inverse mean of the logarithm of the distances within some fixed radius normalized to $[0, 1]$. The choice of

this radius, however, is non-trivial and Levina and Bickel proposed the LID_{MLE} estimator (2004) [54] which replaces the cut-off radius with the $(k{+}1)$-nearest neighbor distance. It was later observed to be an application of the Hill estimator (1975) [42] for power-law tails of distributions. The LID_{MLE} estimator defined as

$$\text{ID}_{\text{MLE}}(x, k+1) = \left(\frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{d(x, x_{k+1})}{d(x, x_i)} \right) \right)^{-1} \tag{2.4}$$

where x_i denotes the i-nearest neighbor, is a popular choice for LID estimation in theory and practice to this day[1]. It is easy to implement, fast to compute, and simple enough to derive theoretical properties. In an attempt to get more concise results for high-dimensional data, Rozza et al. expanded on the LID_{MLE} estimator in 2012 with their Minimum Neighbor Distance Intrinsic Dimensionality (LID_{MiND}) and Intrinsic Dimensionality Estimation Algorithm (LID_{IDEA}) estimators [71]. Their LID_{MiND} approaches – these are multiples, but their concept is unison – are aimed at evaluating the optimum δ (the estimated LID) such that the k-nearest neighbor is most likely observed. To that aim they propose to either try out a series of integer values and choose the most likely, evaluate a type of gradient descent on the likelihood function – where LID_{MLE} happens to be a special case – or evaluate for the lowest Kullback-Leibler divergence towards an empirical distribution of distances. While these estimates can be better than LID_{MLE} their computational cost is significantly higher and their theoretical properties are not as well understood. The LID_{IDEA} estimator is rather an approach to fix the susceptibility of the aggregated LID_{MLE} estimates for under-estimation by applying a correction. It, however, requires taking the mean of estimates over all samples and thus does not qualify for LID estimates. The LID_{ALID} proposed by Chelly et al. in 2016 [19] is a further expansion of the LID_{MLE} estimator that also involves pairwise distances of k-nearest neighbors below the $(k{+}1)$-nearest neighbor distance. It allows for slightly smaller neighborhood sizes at an increased computational cost. Its cumbersome formulation makes it unattractive for theoretical work, though. An even more involved approach was proposed by Amsaleg et al. (including the LID_{ALID} authors) in 2019 with their LID_{TLE} estimator [5], which considers all pairwise distances of the neighbors of x as well as their reflections relative to x.

Related to these estimators is the k-nearest neighbor graph-based estimator by Costa and Hero (2004) [22]. They proposed to use the k-nearest neighbor graph, which had previously been used in Manifold Learning approaches like ISOMAP [83] and LLE [70], to estimate the ID of the data. To do so, they investigated the expansion of geodetic distances approximated by distances in the graph, which links it to the Correlation Dimension [33]. Although it was proposed as a global ID estimator, its algorithm takes the mean of LID estimates over all samples and is thus based on localities as well.

[1] Levina and Bickel further proposed to take the mean estimate over multiple values of k, which is oftentimes omitted in theoretical derivations and implementations.

The second "school of thought" expanded the Correlation Dimension on a theoretical level rather than fixing the practical shortcomings. Karger and Ruhl proposed the Expansion Rate (2002) [49] – its $\log_2$ being called Expansion Dimension – to describe the performance of spatial indexes given general metrics. They used a power-law to investigate the number of neighbors within a radius relative to double the radius and observed that the exponent directly affects the performance of spatial indexing structures. The Expansion Rate follows the same power-law as Euclidean bodies, directly connecting it to the Correlation Dimension, however, the approach in its definition is strictly not limited to Euclidean distances. That approach was generalized to the Doubling Dimension by Gupta et al. [35] in 2003. In 2006, Cole and Gottlieb [21] gave a skip list-inspired algorithm linking the algorithmic complexity of spatial indexes to the Doubling Dimension, further supporting the results of Karger and Ruhl. While both Correlation and Doubling Dimension apply to non-Euclidean spaces, the intuition of exponentially expanding bodies is fundamentally based on a Euclidean perspective. The results only imply that if a metric "behaves like" a Euclidean metric, the algorithmic complexity is bounded by the dimensionality of that Euclidean space. That claim is entirely disjoint from the parameterization of a hypothetical latent space. The link between the claim to "non-Euclideanness" and an underlying Euclidean perspective is a point of debate that will be discussed in Chapter 6. Nonetheless, these works were the first immediate link between LID and the performance of algorithms on the data. In 2012, Houle et al. proposed the Generalized Expansion Dimension [47] which provided a more sophisticated theoretical foundation for the Expansion Rate and ultimately introduced the fundamental LID definition

$$\mathrm{LID}(x, \alpha, \beta) = \frac{\log(\phi(x, \alpha)) - \log(\phi(x, \beta))}{\log(\alpha) - \log(\beta)} \tag{2.5}$$

where ϕ is any functional. This generalized form was not included in the 2012 work [47] but was involved in every subsequent work of the authors on the subject [2, 44, 45, 46, 3, 5] in one form or another, typically substituting $\alpha = 1 + \varepsilon$ and $\beta = 1$ and employing a cumulative distribution function for ϕ. It has been shown to converge to the indiscriminability of the inspected measure in the infinitesimal [44]. The generalized definition, however, is so broad, that it principally allows us to define most previous estimators as special cases. By defining ϕ such that it is the β-nearest neighbor distance of x where $\alpha = 1$, we obtain $\mathrm{LID}_{\mathrm{MLE}}$. When defining ϕ as the number of neighbors of x within a radius β relative to double the radius α, we obtain the Expansion Dimension. Since that generalization allows to simulate the purely Euclidean-based estimators, the "non-Euclideanness" inherited from the Expansion Dimension is questionable – or reversely, it may imply that Euclidean distance is not as fundamentally different to other metrics as sometimes assumed.

A related recent advance was the use of deep learning density models employed by Tempczyk et al. [82]. After training predictive models to generate a univariate Gaussian

smoothed density of the dataset, they evaluate the density at varying standard deviations. From the drop-off in density for larger standard deviations, the Expansion Rate into the "unoccupied" space can be evaluated. The resulting dimension estimate, thus, describes the local intrinsic codimensionality of the manifold, i. e. the difference between the LID and the representation dimension. By subtracting the result from the representation dimension, a LID estimate is obtained. While the results presented by Tempczyk et al. are outstandingly good, especially in high dimensions, basing the approach on hard-to-verify deep learning models makes the result rather unreliable. Further, it is unclear how to choose appropriate standard deviations in practice. In our evaluation in Chapter 5, we were not able to produce reasonable estimates on real datasets using a hyperparameter scheme that achieved good results on artificial data.

Lastly, a third "school of thought" investigated the behavior of manifolds without the power-law fit introduced by the Correlation Dimension. The first such approach was the Dimensionality from Angle and Norm Concentration LID Estimator ($\mathrm{LID}_{\mathrm{DANCo}}$) by Ceruti et al. in 2014 [17]. They proposed to use the Kullback-Leibler divergence between the distribution of distances and angles, and predefined empirical distributions, similar to the Kullback-Leibler divergence of the $\mathrm{LID}_{\mathrm{MiND}}$ estimator. Similarly, the computational cost of the $\mathrm{LID}_{\mathrm{DANCo}}$ estimator is significantly higher than that of the $\mathrm{LID}_{\mathrm{MLE}}$ estimator, although a proposed Fast $\mathrm{LID}_{\mathrm{DANCo}}$ algorithm [17] reduces the computational cost to a comparable order of magnitude. However, the involved process makes this approach rather difficult to analyze as well. In 2019, Erba et al. proposed the Full Correlation Integral LID Estimator ($\mathrm{LID}_{\mathrm{FCI}}$) estimator [28] which is based on the distribution of points on a unit-sphere by normalizing the points in each locality relative to their center. In opposition to all other LID estimators, they proposed to center each locality at its mean, which is a significant deviation from the typical approach of using the inspected point as the center. They did not explain why they mean-centered the localities and we would advise against doing so since it skews the distribution especially around outliers and in localities where the embedding function has a high curvature. Instead of computing a closed-form estimate, the $\mathrm{LID}_{\mathrm{FCI}}$ estimator fits an a priori function against a distribution sampled from the data, similar to the $\mathrm{LID}_{\mathrm{DANCo}}$ and $\mathrm{LID}_{\mathrm{MiND}}$ estimators. That process is again very laborious and the formulation does not lend to theoretical analysis. In the same year, Díaz et al. proposed the Analysis of Variance LID Estimator ($\mathrm{LID}_{\mathrm{ANOVA}}$) estimator [26], which computes the variance of angles in a locality relative to the center and chooses the number of dimensions that best matches the variance of angles between uniform samples from the unit sphere. The matching variance values are constants and can be precomputed, which made the $\mathrm{LID}_{\mathrm{ANOVA}}$ estimator the first LID estimator based on angles to be computationally efficient. The authors included variants of the estimator to fit the angle variance with a correction term or a kernel, however, the estimates are limited to integer values. That makes $\mathrm{LID}_{\mathrm{ANOVA}}$ just as difficult to analyze as the other angle-based estimators. In

2020 we independently proposed the Angle-Based Intrinsic Dimensionality (LID_{ABID}) estimator [84] which employs the distribution of squared cosines of angles between points in a locality relative to the inspected point. We provided closed-form equations for the expected squared cosine on a unit-sphere, which can be inverted to obtain the intrinsic dimensionality as

$$\text{LID}_{\text{ABID}}(x,k) = \left(\frac{1}{k^2} \sum_{i,j=1}^{k} \left\langle x_i - x, x_j - x \right\rangle^2 \right)^{-1} \tag{2.6}$$

The LID_{ABID} estimator is the first LID estimator based on angles to be computationally efficient and to provide a closed-form fractional estimate. We also proposed the $\text{LID}_{\text{RABID}}$ estimator [84] which does not include the regularization of LID_{ABID}, which enforces estimates to be below the true intrinsic dimensionality. These estimators provide both an optimistic and a realistic estimate, respectively, and can be translated using only a few basic arithmetic operations. This allows us to evaluate both estimators basically at the cost of just one. The simple formulation in terms of the squared cosines makes the LID_{ABID} and $\text{LID}_{\text{RABID}}$ estimators handy for analysis and in our extended work [85] we showed multiple properties, e.g. an interpretation in terms of the spectrum of the covariance matrix and the intimate relation to PCA. It was after publication, that we noticed that LID_{ABID} computes the same estimate as the LID_{FCI} estimator in the limit when mean-centering the localities. Yet, the simpler formulation, lower computational cost, and forthcoming analytical properties make LID_{ABID} the superior choice. Following LID_{ABID} we proposed the Thresholded Random In-distribution Projections Intrinsic Dimensionality (LID_{TRIP}) estimator [87] which uses random projections to estimate the LID given an amount of acceptable error in variance. This estimator follows the same principles as LID_{ABID} and inspects the expected variance minus a squared truncated Euclidean distance – a generalization of squared cosines to multiple dimensions. Aside from being the first task-aware LID estimator, LID_{TRIP} also generalizes on LID_{ABID}, which is a special case given specific parameters. Lastly, in yet unpublished work, we expanded the concept of LID_{ABID} to consider the distribution of cosines in chordal triangles, i.e. triangles formed by three points on the unit sphere. For this estimator, which we call LID_{ALC}, we again provide closed-form equations for the moments, from which the LID can be computed efficiently. Similarly to the extension of LID_{MLE} to LID_{TLE}, we hoped to obtain a more accurate estimate by aggregating over a cubic number of values rather than a squared number of values, achieving stable estimates at smaller neighborhood sizes.

Chapter 3

Angle-Based Intrinsic Dimensionality

This chapter will focus on our first contribution to the field of Local Intrinsic Dimensionality (LID) estimation, which is the introduction of LID_{ABID} and $\text{LID}_{\text{RABID}}$ as novel methods to estimate the LID of a dataset using angles. We intended to provide a method to evaluate the local complexity in non-Euclidean data, specifically on word-cooccurrences observed in natural text. For this purpose, we required a method purely based on a similarity measure. Yet, deriving such a method for general similarities appeared far out of reach at the time, and we instead focused on the special case of cosines in Euclidean space. Using a squared amount of values per neighborhood compared to the linear amount of distances in LID_{MLE} and related estimators, further made us hopeful for a faster converging method in terms of neighborhood size. Unbeknownst to us at the time, other approaches to angle-based LID estimation had been published [17, 28, 26] – mostly during the development of the theory – yet none of these approaches are efficient, accurate, fractional, and allow for rigorous analytical discussion. The latter is primarily important to expand beyond the Euclidean case. Therefore, we consider our approach to be a significant contribution to the field of LID estimation. Further, it was the basis for all subsequent works on Euclidean data [86, 87] and the implicit boundaries of Euclidean space, i.e. inequalities beyond the triangle inequality and Ptolemy's inequality that follow from the Euclidean axioms [87].

In this chapter, we will first provide the derivation and definition of the LID_{ABID} and $\text{LID}_{\text{RABID}}$ estimators in Section 3.1. Section 3.2 will discuss why the second non-central moment used in the definition of LID_{ABID} is a good choice. We will then discuss the relation of LID_{ABID} to other LID estimators in Section 3.3 and Section 3.4. Lastly, we will compare LID_{ABID} to LID_{MLE} in terms of analytical measures backed by empirical data in Section 3.5 and conclude with the empirical evaluation of LID_{ABID} in Section 3.6.

Multiple sections in this chapter are heavily oriented along the already published works on the topic, "ABID: Angle Based Intrinsic Dimensionality" [84] and the extended

journal version "ABID: Angle Based Intrinsic Dimensionality - Theory and analysis" [85], denoted by a citation on the section titles. Not every adopted sentence is marked as a quote, but the reader should assume that the majority of the content in these sections is based on previous publications. The remaining sections are novel content for this thesis.

3.1 Derivation of $\mathrm{LID}_{\mathrm{ABID}}$ [85]

Approaches to LID estimation are generally focused on distributions of measurable quantities. The immediately measurable similarities in Euclidean space are both inner products and their normalized "cousins" the cosines. Since the focus of LID estimation is on as small localities as possible, the inner products are more vulnerable to potential noise and are therefore less suitable for this purpose. Consequentially we focus on the cosine similarity or inner product of normalized data, i.e. the pairwise angle distribution on the sphere. To visualize the effect this has on the resulting LID estimates, consider the following example: The Milky Way galaxy is a spiral galaxy, i.e. one would be inclined to abstract it to a sort of disc with a visible but not dominant height. In terms of intrinsic dimensionality, it would intuitively be a "2 and a bit"-dimensional manifold. Our solar system is in the outskirts of the galaxy, and we can observe the stars in the night sky, which for all intents and purposes appears to us as a sphere. And with the naked eye, we can only observe the angles between the stars – we inspect the Milky Way stars as normalized points. The collection of stars that make the majority of the Milky Way, appear as a band on the night sky, with a visible but not dominant thickness. From our perspective, we therefore perceive the Milky Way as a "2 and a bit"-dimensional manifold. This is the intuition behind the $\mathrm{LID}_{\mathrm{ABID}}$ estimator: estimating the LID of the manifold from the shape the manifold projects onto the sphere. That intuition leads to a varying behavior dependent on the scale that we apply. If we choose a scale, such that the "locality" is a spec of sand, then the Milky Way appears in nearly 3 dimensions, as the solar system itself is all around it. If the locality spans a good portion of the Milky Way, we obtain the abstracted disc interpretation of "2 and a bit". If the locality spans multiple galaxies, a single galaxy can be abstracted to a single point of 0 dimensions. Accordingly, the proper answer to "How many dimensions does a collection of samples have?" is dependent on the scale of the locality we are considering. In LID estimation, the choice is often to be as local as possible, since we assume to have way too few samples to take larger scales into account, but if a more abstract view is required for a downstream task, the locality must be increased.

For the angle-based estimators discussed in this chapter, we employ the same local linearity assumption as $\mathrm{LID}_{\mathrm{MLE}}$ that was described in Section 2.4. Thereby, we assume that the manifold projects onto a $(\delta-1)$-dimensional subsphere of the full dimensional sphere placed around each point of interest. That point of interest is the point x in the

dataset for which we desire to compute the LID estimate. We choose the neighborhood as the intersection of a d-ball centered at x and the dataset at the k-neighbor distance, i.e. the k-nearest neighbors of x. The neighborhood then gets projected to unit distance around x. We are then interested in the distribution of angles sampled uniformly at random from the unit sphere. That distribution has been provided by Cai et al. [14].

3.1.1 Theorem (Distribution of random angles in a $(d-1)$-sphere [14]). *The distribution of angles φ between two random points sampled independently and uniformly from a $(d-1)$-sphere converges as the number of samples goes to infinity to*

$$P(\varphi) = \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{d-1}{2})} \cdot \sin(\varphi)^{d-2} \tag{3.1}$$

where Γ is the gamma function and φ is defined on $[0, \pi]$.

Since the distribution of angles is independent of the length of the vectors, that distribution is viable for all spherically symmetric distributions like uniform hyperballs or univariate normal distributions. As popularly known from the Curse of Dimensionality, all angles tend to become approximately orthogonal as dimensionality approaches infinity. This causes (3.1) to concentrate around $\frac{\pi}{2}$ as shown by Cai et al. [14]. The distribution above is unwieldy and expensive to compute (as we need to compute the arcus cosines). We, therefore, prefer to work directly on the cosines. By applying the Legendre duplication formula and doing a change of variables, we obtain the distribution of cosines.

3.1.2 Theorem (Distribution of cosine similarities of points in a $(d-1)$-sphere). *The distribution of pairwise cosine similarities C between random points sampled independently and uniformly from a $(d-1)$-sphere is*

$$P(C) = \tfrac{1}{2} B(\tfrac{1+C}{2}; \tfrac{d-1}{2}, \tfrac{d-1}{2}) \tag{3.2}$$

where $B(x; a, b)$ is the Beta distribution probability density function and C is defined on $[-1, 1]$.

Proof. For this proof, we modify the well-known Legendre duplication formula:

$$\Gamma(x)\Gamma(x + \tfrac{1}{2}) = 2^{1-2x}\Gamma(\tfrac{1}{2})\Gamma(2x) \tag{3.3}$$

$$\Rightarrow \qquad \frac{\Gamma(x + \frac{1}{2})}{\Gamma(x)\Gamma(\frac{1}{2})} = \frac{2^{1-2x}\Gamma(2x)}{\Gamma(x)^2} = \frac{1}{B(x, x)} \cdot \tfrac{1}{2}^{2x-1} \tag{3.4}$$

where $B(\cdot, \cdot)$ is the beta function. By using (3.4) in (3.1) for $x = \frac{d-1}{2}$, we obtain

$$P(\varphi) = \frac{1}{B(\frac{d-1}{2}, \frac{d-1}{2})} \cdot \left(\tfrac{1}{2}\sin(\varphi)\right)^{d-2} \tag{3.5}$$

We can now substitute φ with $\arccos(C)$ by a change of variable:

$$P(C) = \frac{1}{B(\frac{d-1}{2}, \frac{d-1}{2})} \cdot \left(\tfrac{1}{2}\sin(\arccos(C))\right)^{d-2} \cdot \left|\tfrac{\partial}{\partial C} \arccos(C)\right| \tag{3.6}$$

$$= \frac{1}{B(\frac{d-1}{2}, \frac{d-1}{2})} \cdot (\tfrac{1}{2}\sqrt{1-C^2})^{d-2} \cdot \frac{1}{\sqrt{1-C^2}} \tag{3.7}$$

$$= \frac{1}{B(\frac{d-1}{2}, \frac{d-1}{2})} \cdot (\tfrac{(1-C)(1+C)}{2 \cdot 2})^{\frac{d-2}{2}} \cdot ((1-C)(1+C))^{-\frac{1}{2}} \tag{3.8}$$

$$= \frac{1}{B(\frac{d-1}{2}, \frac{d-1}{2})} \cdot (1 - \tfrac{1+C}{2})^{\frac{d-1}{2}-1} \cdot (\tfrac{1+C}{2})^{\frac{d-1}{2}-1} \cdot \tfrac{1}{2} \tag{3.9}$$

$$= \tfrac{1}{2} B\left(\tfrac{1+C}{2}; \tfrac{d-1}{2}, \tfrac{d-1}{2}\right) \tag{3.10}$$

which is a beta distribution rescaled to the interval $[-1, 1]$, on which C is defined. □

Based on this, we can easily obtain the following helpful Corollary:

3.1.3 Corollary. *The average cosine similarity of two random points sampled independently and uniformly from a d-ball is given by*

$$\mathbb{E}[C] = 0\,. \tag{3.11}$$

The variance and the non-central second moment are given by

$$\operatorname{Var}[C] = \mathbb{E}\left[C^2\right] = \tfrac{1}{d}\,. \tag{3.12}$$

Proof. This follows immediately from the central moments of Beta distributions. By Theorem 3.1.2 we have $\frac{1+C}{2} \sim B(\bullet; \frac{d-1}{2}, \frac{d-1}{2})$. This *symmetric* beta distribution has a mean of $\frac{1}{2}$, and hence $\mathbb{E}[C] = 0$. The variance of this beta distribution given $a = b = \frac{d-1}{2}$ is $\operatorname{Var}\left[\frac{1+C}{2}\right] = \frac{1}{4d}$, and hence $\mathbb{E}\left[\left(\frac{1+C}{2} - \frac{1}{2}\right)^2\right] = \mathbb{E}\left[\frac{C^2}{4}\right] = \frac{1}{4d}$, which multiplied by 4 gives $\mathbb{E}\left[C^2\right] = \frac{1}{d}$. Because the mean is 0, the variance and the second non-central moment agree trivially. □

The $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ estimators are immediate consequences of Theorem 3.1.2 and Corollary 3.1.3. They employ a method of moments approach based on the cosine distribution to estimate the intrinsic dimensionality of the data. The first moment of Corollary 3.1.3 cannot be used for estimation because it does not depend on d. Both the variance and the second non-central moment, however, are suitable for estimating intrinsic dimensionality as they depend inversely on d. This simple dependency stands in contrast to the expansion-rate-based approaches, which generally obtain an exponential relation to the dimensionality as the volume of a d-ball has d in its exponent. In concordance with the squared number of values obtained from pairwise cosines, we hope to obtain a more robust measure even with smaller neighborhood sizes (i.e., fewer samples); and as we do not need to compute logarithms it can be computed more efficiently. But we still have two choices: we can either estimate using the variance $\hat{d} = 1/\operatorname{Var}[C]$ or using the non-central

second moment $\hat{d} = 1/\mathbb{E}\left[C^2\right]$, which only agrees if $\mathbb{E}[C] = 0$ as expected for a uniform ball.

Consider the scenario of many points sampled from a hyperplane but the point of interest is not on this hyperplane. The local neighborhood will then consist of samples in a circular region on this plane. If we move the point of interest away from the plane the average cosine tends to 1, and the variance to 0. The variance-based estimate would hence tend to infinity, while the second non-central moment estimate will tend to 1. We argue that this is the more appropriate estimate, as the data concentrates in a single far away area, collapsing to a point in the limit. In that case, we would consider the point of interest as a far outlier, best considered as a singular point, resulting in a LID of 0, which coincides with the limit of the second non-central moment estimate.

Inspired by the $\mathrm{LID}_{\mathrm{TLE}}$ approach [5], we investigated the idea of considering the reflections of all points with respect to the point of interest. Such a reflection causes the average cosine in this example to be 0 as every pair of points can be matched to the pair with the second point reflected. In the above example, we would obtain two opposite discs of points and the resulting variance would tend to 1. The estimates of the variance-based estimator would thus agree with the non-central moment. When adding reflected points, the variance and the non-central second moment become equivalent (which could serve as additional justification for $\mathrm{LID}_{\mathrm{TLE}}$): Since $\cos(x_i, -x_j) = -\cos(x_i, x_j) = \cos(-x_i, x_j)$, adding reflections yields two positive and two negative copies of each cosine. The resulting average then is 0, and hence $\mathrm{Var}[C'] = \mathbb{E}\left[C'^2\right] - \mathbb{E}[C']^2 = \mathbb{E}\left[C'^2\right] = \mathbb{E}\left[C^2\right]$. Since the reflected points are implicitly represented by the second non-central moment, they do not require an explicit consideration.

Instead of discussing the limit cases of distributions, we will now work with a fixed data sample of k points centered around a point of interest. For simplicity, we assume that the data has been translated such that the point of interest is always at the origin, and that this point and any duplicates of it have been removed from the sample. We now use *all pairwise* cosine similarities in a $k \times k$ matrix denoted C. The diagonal of this matrix is usually excluded from computations, as required by Theorem 3.1.2. We use the term C_1 when the ones on the diagonal are to be included. By C^2, we denote the individual squaring of cosines. The next theorem will use both a fixed sample and the matrix C_1 with the diagonal included.

3.1.4 Theorem (Upper bound). *Let $X = \{x_1, \ldots, x_k\} \subset \mathbb{R}^d$ be a sample from a δ-dimensional subspace embedded in $\mathbb{R}^d$ for some $\delta \leq d$. Formally, let X contain at least δ linearly independent vectors and let all x_i be linear combinations of a given set of δ orthonormal basis vectors. Then the following inequality holds*

$$\mathbb{E}\left[C_1^2\right]^{-1} \leq \delta\,. \tag{3.13}$$

Proof. Let $\widetilde{X}$ be the $k \times \delta$ matrix obtained from X by first performing a change of basis to the given orthonormal basis of size δ, then normalizing each vector to unit length to produce $\tilde{x}_i$. Neither the change of basis (which is a rotation) nor the posterior normalization affects the cosine similarities, and we hence have

$$\cos(\tilde{x}_i, \tilde{x}_j) = \cos(x_i, x_j) \ . \tag{3.14}$$

It immediately follows that $\widetilde{X}$ has a rank of δ, as we still have δ linearly independent vectors. The matrix $\widetilde{C}_1 = \widetilde{X}\widetilde{X}^T$ then contains entries of the form $\langle \tilde{x}_i, \tilde{x}_j \rangle$. As all $\tilde{x}_i$ are normalized, $\widetilde{C}_1$ is equal to the cosine similarities. Per (3.14) it then follows that $\widetilde{C}_1$ is exactly C_1. Because C_1 is a cosine similarity matrix, the diagonal entries are all 1 and we have $\operatorname{tr}(C_1) = k$. Since $\widetilde{X}$ is a $k \times \delta$ matrix with rank δ, we know that the rank of C_1 is δ as well. Therefore C_1 has δ eigenvalues $\lambda_1, \dots, \lambda_\delta$ with $\sum_{i=1}^{\delta} \lambda_i = \operatorname{tr}(C_1) = k$. The sum of squared entries $\|C_1\|^2$ equals the sum of squared eigenvalues $\sum_{i=1}^{\delta} \lambda_i^2$ and is minimized if every eigenvalue equals $\frac{k}{\delta}$, which means we have the following lower bound:

$$\|C_1\|^2 = \textstyle\sum_{i=1}^{\delta} \lambda_i^2 \geq \delta \cdot \left(\frac{k}{\delta}\right)^2 = \frac{k^2}{\delta} \tag{3.15}$$

and by taking the inverse we obtain the upper bound $\mathbb{E}\left[C_1^2\right]^{-1} \leq \delta$. □

This is an upper bound for estimating the intrinsic dimensionality using C_1, and we can use this to also obtain an upper bound for C.

3.1.5 Corollary. *Let $X = \{x_1, \dots, x_k\} \subset \mathbb{R}^d$ be a sample from a δ-dimensional subspace embedded in $\mathbb{R}^d$ as defined in Theorem 3.1.4. If $k > \delta$, then*

$$\mathbb{E}\left[C^2\right]^{-1} \leq \tfrac{k-1}{k-\delta} \cdot \delta \ . \tag{3.16}$$

Proof. As the difference between C and C_1 is the diagonal of ones, (3.15) yields

$$\|C\|^2 = \|C_1\|^2 - k \geq \tfrac{k^2}{\delta} - k = \tfrac{k(k-\delta)}{\delta} \tag{3.17}$$

and hence the average of the remaining $k^2 - k$ cells is

$$\mathbb{E}\left[C^2\right] \geq \tfrac{k-\delta}{k-1} \cdot \tfrac{1}{\delta} \ , \tag{3.18}$$

which is equivalent to the inequality above. For $k = \delta$ we obtain a trivial bound. □

Accordingly, the difference of including the diagonal or not vanishes for large enough k. One could attempt to regularize $\mathbb{E}\left[C^2\right]$ with $\frac{k-1}{k-\delta}$. The major problem therein is that we do not know δ in advance. To control the maximal overestimation of δ, a sufficiently large neighborhood can be used to lower the margin of error. For example, to bound $\mathbb{E}\left[C^2\right]^{-1} \leq \delta + c$, at least $k \geq \frac{1}{c}\delta^2 + (1 - \frac{1}{c})\delta$ neighbors are required. For the practically relevant bound of δ+1 (c=1) this means, that we require $k \geq \delta^2$ samples.

From an engineering perspective, an iterative refinement of the estimate of δ might be tempting. Further, an a priori approach to find the best fitting δ, that is a fixed point of the regularization, might be a good fit as well. The following Theorem shows that both of these estimates are equal to using the estimate of Theorem 3.1.4, i.e. to use C_1 instead of C.

3.1.6 Theorem. *Let X and C be defined as in Corollary 3.1.5, C_1 as in Theorem 3.1.4. Further let k be at least $\delta+2$ and $(\delta_i)_{i=0,\dots}$ be an infinite sequence with*

$$\delta_0 = \mathbb{E}\left[C^2\right]^{-1} \text{ and } \delta_i = \mathbb{E}\left[C^2\right]^{-1} \cdot \tfrac{k-\delta_{i-1}}{k-1} \text{ for all } i \geq 1. \tag{3.19}$$

Then the following equations hold:

$$\lim_{i\to\infty} \delta_i = \lim_{i\to\infty} \mathbb{E}\left[C^2\right]^{-1} \cdot \tfrac{k-\delta_i}{k-1} = \mathbb{E}\left[C_1^2\right]^{-1} \tag{3.20}$$

Proof. For the first equation we will analyze the series of δ_i and show that it has a closed form solution. We will use $c = \mathbb{E}\left[C^2\right]$ as a shorthand notation. If we recursively develop d_i by insertion, we can easily see that it is equal to

$$\delta_i = -k \cdot \left(\sum_{j=1}^{i} \left(\tfrac{-1}{c\cdot(k-1)}\right)^j\right) + \delta_0 \cdot \left(\tfrac{-1}{c\cdot(k-1)}\right)^i \tag{3.21}$$

As per Corollary 3.1.5, we know that $c \geq \frac{1}{\delta}$. Hence $c \cdot (k-1)$ has to be truly greater than 1. By that constraint, as i goes to infinity, we obtain

$$\lim_{i\to\infty} \delta_i = \lim_{i\to\infty} -k \cdot \left(\sum_{j=1}^{i} \left(\tfrac{-1}{c\cdot(k-1)}\right)^j\right) + \delta_0 \cdot \left(\tfrac{-1}{c\cdot(k-1)}\right)^i = k \cdot \tfrac{1}{c\cdot(k-1)+1} \tag{3.22}$$

As we insert (3.22) into (3.20), we obtain

$$k \cdot \tfrac{1}{c\cdot(k-1)+1} = \tfrac{k-k/(c\cdot(k-1)+1)}{c\cdot(k-1)} \tag{3.23}$$

$$\Leftrightarrow k(c\cdot(k-1)) = k\cdot(c\cdot(k-1)+1) - k \tag{3.24}$$

$$\Leftrightarrow k^2c - kc = k^2c - kc + k - k \tag{3.25}$$

The last equation obviously holds. The second equality of (3.20) follows immediately from inserting $(k^2-k)\,\mathbb{E}\left[C^2\right] = k^2\,\mathbb{E}\left[C_1^2\right] - k$ (slight modification of (3.17)) into (3.22).□

Therefore, the relaxed estimate $\mathbb{E}\left[C_1^2\right]$ is a properly regularized intrinsic dimensionality estimator based on the pairwise cosines of points in local neighborhoods.

Based on these theoretic considerations, we can now introduce the angle-based estimators of intrinsic dimensionality.

3.1.7 Definition ($\mathbf{LID_{ABID}}$). Given a dataset $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$, the *regularized* angle-based intrinsic dimensionality estimator for a point x_i is:

$$\mathrm{LID}_{\mathrm{ABID}}(x_i;k) := \mathbb{E}\left[C_1\left(B_k(x_i)\right)^2\right]^{-1} \tag{3.26}$$

where $B_k(x_i)$ are the directional vectors from x_i to the k nearest neighbors of x_i and $C_1\left(B_k(x_i)\right)$ are the pairwise cosine similarities within $B_k(x_i)$.

By choosing the neighborhood of any point in the specified set as the k nearest neighbors, the measure is invariant under scaling and varying local densities. Analogously, one can instead define the neighborhood by a maximum distance to the central point. The sole restriction thereby is that the size of the neighborhood has to be greater or equal to $\delta + 2$ as for any smaller neighborhood, the estimator does not need to be properly regularized. Since the error of the non-regularized estimate is limited for any neighborhood size $\propto \delta^2$, we further introduce a non-regularized version for comparative analysis.

3.1.8 Definition ($\mathbf{LID_{RABID}}$). Given a dataset $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$, the *raw* angle-based intrinsic dimensionality estimator for a point x_i is defined as

$$\mathrm{LID}_{\mathrm{RABID}}(x_i; k) := \mathbb{E}\left[C\left(B_k(x_i)\right)^2\right]^{-1} \tag{3.27}$$

where $B_k(x_i)$ are the directional vectors from x_i to the k nearest neighbors of x_i and $C\left(B_k(x_i)\right)$ are the pairwise cosine similarities of *different* vectors in $B_k(x_i)$.

Beware that this estimator can cause a division by zero if all k vectors are pairwise orthogonal, and can return values larger than k. In the prior case, there simply is no reason to assert a dimensionality, since no correlation has been observed. For the latter case, one has to remember, that the upper bound in Corollary 3.1.5 only holds for $k>\delta$. If $\delta>k$, the estimates of $\mathrm{LID}_{\mathrm{ABID}}$ are necessarily upper-bound δ, whilst $\mathrm{LID}_{\mathrm{RABID}}$ *can* still estimate the correct LID. When "undersampling" the intrinsic dimensionality, the purely distribution-based estimator $\mathrm{LID}_{\mathrm{RABID}}$ can, therefore, still operate and remains theoretically sound. One might though want to confirm the estimates by repeated evaluation on a larger neighborhood or with another method. For large k both estimators converge to the same value, providing a consistency check for the estimates.

To interpret the estimates by these methods, it is important to consider the domain they operate on. The angle-based measure is bounded by the spanning dimensionality of the point set. While distributions of angles are usually distorted by non-linear transformations, many transformations such as rotations will retain this bound. Hence, the bound may nevertheless apply – at least approximately – for many projections of lower-dimensional manifolds in higher dimensional embeddings. It is easy to see that angle-preserving transformations do not affect our measure while distance-preserving transformations will not affect distance-based estimators. These measures are less affected by local non-linear contractions and expansions such as the decreasing density on the outer parts of Gaussian distributions but they tend to estimate higher dimensionality than distance-based approaches when the transformations are non-linear. We do not consider this to be a flaw, just a different design that may or may not have advantages: a common assumption in many methods and applications like manifold learning is to have locally linear transformations, that preserve small neighborhoods which will then affect neither angles nor densities. The above estimators, which can be seen as estimating how many dimensions

such a locally linear embedding needs to have, is arguably very close to the idea of such applications. Similarly, by implicitly mirroring all points at the point of interest, $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ employ the same notion of "continuity" in the data as $\mathrm{LID}_{\mathrm{TLE}}$ does, whilst $\mathrm{LID}_{\mathrm{MLE}}$ does not employ that strategy. Especially on outlying points, e. g. those on the surface of a manifold, that mirroring stabilizes the estimates by continuing the manifold distribution to its "outside". This is not necessarily better, but again a different design concept. One also has to keep in mind, that LID estimates may, in general, be lower than a minimal "reasonable" embedding space as, e.g., a Möbius strip is locally intrinsically two dimensional everywhere yet requires three dimensions to be embedded in Euclidean space.

3.2 About the choice of moments [85]

In this section, we will discuss the angle-based approach in terms of different moments of the cosine distribution. As both $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ can be expressed in terms of either the regularized or immediate second non-central moment of the distribution of cosines, it is worth considering if other moments might be comparable or even better estimators for the LID. Every odd (non)-central moment of the cosine distribution, is obviously 0 as the distribution is symmetric around 0. Here, we want to investigate the idea of using higher-order non-central moments of the distribution of absolute cosines, whose odd non-central moments do not vanish. We found a function properly describing the t^{th} moment of the absolute cosine distribution and validated it using symbolic integration in the SymPy Python package [60]. As the pattern matching-based integration algorithm gives a rather lengthy series of integration steps we omit most of the integration but give an outline of the general approach.

3.2.1 Theorem (Moment-generating function). *The moment-generating function for the non-central moments of the absolute cosine distribution between pairs of different points in $(d-1)$-spheres is*

$$M_{|C|}(t) = \frac{\Gamma\left(\frac{d}{2}\right)\Gamma\left(\frac{1+t}{2}\right)}{\Gamma\left(\frac{d+t}{2}\right)\Gamma\left(\frac{1}{2}\right)} \tag{3.28}$$

Proof. In the proof, we will replace the random variable of cosines C with a scaled and translated variable $X = \frac{C+1}{2}$ that is defined on $[0, 1]$ and can be directly inserted into the beta distribution. Instead of $\mathbb{E}\left[|C|^t\right]$ we, hence, need to compute $\mathbb{E}\left[|2X-1|^t\right]$. The expected value can be computed via the definite integral

$$\mathbb{E}\left[|2X-1|^t\right] = \int_0^1 |2x-1|^t B\left(x; \tfrac{d-1}{2}, \tfrac{d-1}{2}\right)\, \mathrm{d}x\,. \tag{3.29}$$

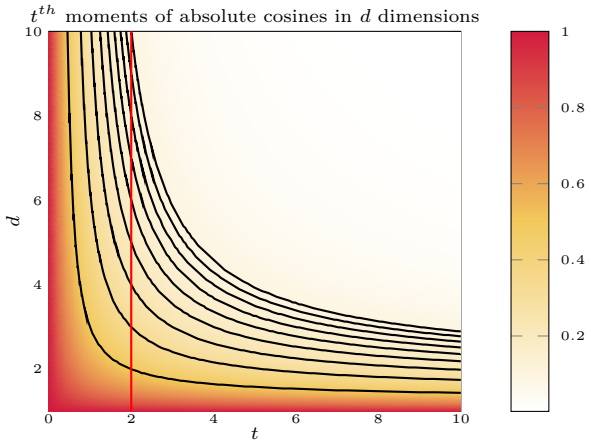


Figure 3.1: Values of the theoretical t-th moment of the absolute cosine distribution at d dimensions. The black contour lines are at the values $\frac{1}{2}$ through $\frac{1}{10}$ from bottom left to top right. The red line is at the second moment used by $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$.

Due to the symmetry of the beta distribution, the integrals from 0 to $\frac{1}{2}$ and from $\frac{1}{2}$ to 1 are equal. Replacing the full integral with twice the upper half gives

$$\mathbb{E}\left[|2X-1|^t\right] = \int_{\frac{1}{2}}^{1} 2(2x-1)^t B\left(x; \tfrac{d-1}{2}, \tfrac{d-1}{2}\right)\, \mathrm{d}x. \tag{3.30}$$

By expanding the beta distribution with its definition and substituting $x(1-x)$ with $\frac{1}{4} - (x-\frac{1}{2})^2$ we can further reduce this expression to

$$\mathbb{E}\left[|2X-1|^t\right] = \int_{\frac{1}{2}}^{1} 2^{t+1}\left(x-\tfrac{1}{2}\right)^t \left(\tfrac{1}{4} - \left(x-\tfrac{1}{2}\right)^2\right)^{\frac{d-3}{2}} B\left(\tfrac{d-1}{2}, \tfrac{d-1}{2}\right)^{-1}\, \mathrm{d}x. \tag{3.31}$$

This formulation allows substituting $(x-\frac{1}{2})$ in the following steps which we do not provide here. The integral, however, has been verified with symbolic integration using the SymPy Python package [60]. The indefinite integral, as well as the definite integral from $\frac{1}{2}$ to 1 with ${}_2F_1$ being the ordinary hypergeometric function, is

$$\mathbb{E}\left[|2X-1|^t\right] = \left[\frac{2^{3-d+t}\left(x-\frac{1}{2}\right)^{t+1}\Gamma\left(\frac{1+t}{2}\right)}{B\left(\frac{d-1}{2}, \frac{d-1}{2}\right)\Gamma\left(\frac{3+t}{2}\right)}\, {}_2F_1\left(\begin{matrix}\frac{3-d}{2}, \frac{1+t}{2} \\ \frac{3+t}{2}\end{matrix}\middle| 4\left(x-\tfrac{1}{2}\right)^2\right)\right]_{\frac{1}{2}}^{1} = \frac{\Gamma\left(\frac{d}{2}\right)\Gamma\left(\frac{1+t}{2}\right)}{\Gamma\left(\frac{d+t}{2}\right)\Gamma\left(\frac{1}{2}\right)}. \tag{3.32}$$

□

This moment-generating function is applicable to any positive t, even non-integers. As the cosine distribution still only accounts for cosines between different points, the second non-central moment $M_{|C|}(2)$ is equivalent to $\mathrm{LID}_{\mathrm{RABID}}$. In the case of $t{\neq}2$ the equation for $M_{|C|}(t)$ is not easily solvable for d. If $t{=}2m$ with $m{\in}\mathbb{N}$ the inverse of $M_{|C|}(t)$ is a polynomial in d of degree m and for all other values of t the gamma functions do not easily cancel out. We, therefore, do not provide a closed-form for any moment besides $t{=}2$.

Fig. 3.1 displays the values of the moments dependent on t and d. The isolines appear to have a hyperbolic shape with a center at $t{=}0$ and $d{=}1$. All of the moments are

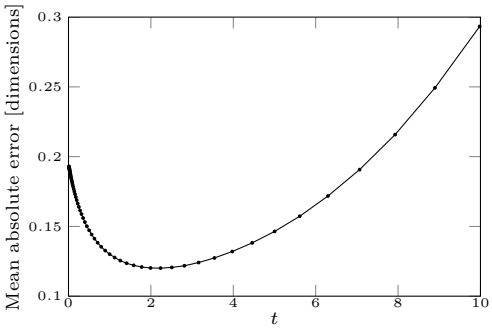


Figure 3.2: The mean absolute estimate error for varying t-th moments on 5000 32-balls with 400 samples each. The smallest error is observed at $t \approx 2$. Similar results were obtained on repeated runs and different parameterization for the number of dimensions and samples.

monotonously decreasing in both t and d. The equations can therefore be approximately solved to arbitrary precision by binary search on an appropriate interval or via selecting the closest fit of a sufficiently densely sampled lookup table of precomputed moment values for varying d. In many cases, values for d between 1 and the representation dimension will be suitable for this task. As a measure of sensitivity to d we propose the relative gradient magnitude, that is $|\frac{\partial}{\partial d} M_{|C|}(t)| / M_{|C|}(t)$ which equates to $\frac{1}{2}(\psi(\frac{d+t}{2}) - \psi(\frac{d}{2}))$ where ψ is the digamma function. At exactly $t=2$ lies $\mathrm{LID}_{\mathrm{RABID}}$ with an estimate of $\frac{1}{d}$ and a relative gradient magnitude of $\frac{1}{d}$. For $t<2$ the relative gradient magnitude decreases and approaches 0 as t approaches 0. This introduces numerical stability problems as the derivative for d gets orders of magnitudes smaller than the value of the moments, even faster than for $t=2$. For $t>2$ the relative gradient magnitude increases, which would suggest improved estimates as the numerical stability problems from the relative magnitude of the gradient are not as pronounced. However, the moment values are vanishing, as is their variance $M_{|C|}(2t) - (M_{|C|}(t))^2$ making estimates based on higher moments much more sensitive to rounding errors, which in turn increase due to growing exponents. We conclude that both smaller and larger values for t are prone to numerical stability issues. A choice of $t=2$ could be a valuable middle ground that even allows avoiding the gamma function altogether, which in praxis is only approximate.

To validate these claims we tested the different moments on a large set of randomly generated d-balls. For each of the balls, we computed the absolute difference between d and the moment-based estimate and computed a mean absolute error. An example of the results is displayed in Fig. 3.2. The minimum mean absolute error was achieved for $t \approx 2$. The estimates for arbitrary moments were computed with binary search from 1 to $d+10$ up to a precision of ten decimals on the moment values. To rule out faulty results due to the vanishing moment values for higher moments, we also performed the same experiment with lookup tables for moments on the same interval of dimensions with 10^{-3} increments that gave equivalent results. Besides being a lot easier to compute due to the closed form,

the second non-central moment seemingly gives the most reliable values of all possible moments. As mentioned before, lower numerical stability for small and large values of t is to be expected. However, the smooth curve through integer and non-integer values and around 2 suggests an additional inherent quality of $t{\approx}2$ that goes beyond considerations of numerical stability. What inherent quality this might be remains unclear.

To summarize, the estimates of the second non-central moment seemingly produce the highest quality of estimates. In addition to that, they can be computed with basic mathematical operations. The possibility for regularization for an upper bound by d further induces a connection to eigen decomposition-based approaches such as PCA.

3.3 $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{PCA}}$ [85]

In the proof of Theorem 3.1.4 we used the eigenvalues of the cosine matrix for a sample of normalized vectors $\widetilde{X}$. It is a known fact that the sum of squared eigenvalues of $\widetilde{X}\widetilde{X}^T$ is equal to the sum of squared eigenvalues of $\widetilde{X}^T\widetilde{X}$. Since $\mathrm{LID}_{\mathrm{ABID}}$ is equal to the inverse sum of squared eigenvalues of the cosine matrix divided by k^2 it, hence, is also equal to the inverse sum of squared eigenvalues of $\frac{1}{k}\widetilde{X}^T\widetilde{X}$ which is the "non-central covariance matrix" of $\widetilde{X}$, i. e. the covariance matrix assuming a mean of 0 in all dimensions. This allows to compute $\mathrm{LID}_{\mathrm{ABID}}$ in $\mathcal{O}\left(kd^2\right)$ instead of the naive $\mathcal{O}\left(k^2d\right)$ complexity resulting from the cosine matrix. Dependent on whether k or d is larger, either approach can be faster, resulting in an overall complexity of $\mathcal{O}\left(\max(k,d)\min(k,d)^2\right)$ for the $\mathrm{LID}_{\mathrm{ABID}}$ estimator for inputs given as Euclidean vectors. If only the cosines or distances are given, it is still faster to use the cosine matrix with a complexity of $\mathcal{O}\left(k^2\right)$.

The intimate relation to the covariance matrix of normalized vectors, however, also implies a strong relationship to local Principal Component Analysis (PCA). A common heuristic with (local) PCA used, e.g., in cluster analysis [52], is to choose the number of eigenvalues whose sum amounts to at least a threshold fraction of the total sum of eigenvalues of either the covariance or correlation matrix, which has been proposed as an integer estimate of LID [31]. When applied to the covariance matrix this corresponds to the number of components to explain at least x% of the total variance in the dataset. In LID estimation, the local PCA is generally applied to the covariance matrix and threshold values of, e.g., 95% allowing to drop up to 5% of the total variance as embedding noise. A central observation of using local PCA for LID estimation is that the resulting estimate is strongly influenced by this threshold value, which is mostly arbitrarily chosen and has an unfortunate direct dependency on the estimate. If, for example, a threshold of less than $1-\frac{1}{\delta}$ is chosen, it is almost certain to underestimate the LID on point distributions uniformly distributed along with the largest δ principal components.

Since $\mathrm{LID}_{\mathrm{ABID}}$ is a measure on the covariance eigenvalues for a normalized dataset, and PCA Intrinsic Dimensionality ($\mathrm{LID}_{\mathrm{PCA}}$) is a measure on the covariance eigenvalues

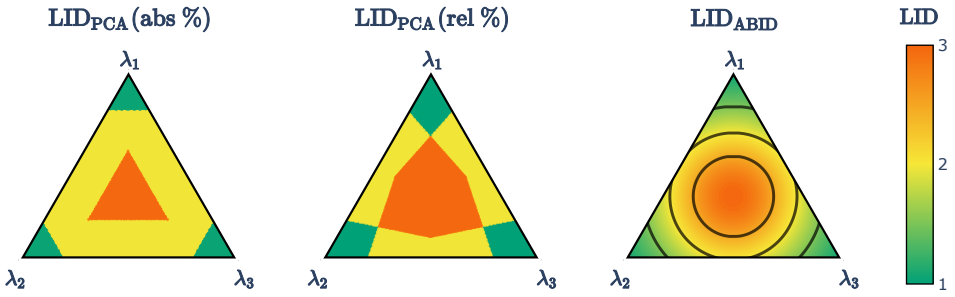


Figure 3.3: LID estimates of two $\mathrm{LID}_{\mathrm{PCA}}$ variants and $\mathrm{LID}_{\mathrm{ABID}}$ dependent on the relative size of eigenvalues of the covariance matrix. The "abs %" $\mathrm{LID}_{\mathrm{PCA}}$ variant counts the number of largest eigenvalues whose sum explains 78.87% of the total variance. The "rel %" $\mathrm{LID}_{\mathrm{PCA}}$ variant counts the number of eigenvalues that are at least 26.79% of the largest eigenvalue. These thresholds were chosen to provide the same "corners" (estimate = 1) as $\mathrm{LID}_{\mathrm{ABID}}$ (estimates < 1.5). Circles in the $\mathrm{LID}_{\mathrm{ABID}}$ plot are isolines at the estimate values 2.5, 2, and 1.5 from inside to outside.

for a non-normalized dataset, we can compare their behavior in the space of eigenvalues. Fig. 3.3 displays the LID estimates based on the relative magnitudes of three eigenvalues. The $\mathrm{LID}_{\mathrm{ABID}}$ estimates are radially symmetric and smoothly distributed such that an equilibrium of i eigenvalues with relative magnitude $\frac{1}{i}$ corresponds to a LID estimate of i. The $\mathrm{LID}_{\mathrm{PCA}}$ estimates are not radially symmetrically distributed and are integer only. Even though both estimators can be described in the same domain, the radial symmetry states a very dominant difference. The isolines of $\mathrm{LID}_{\mathrm{ABID}}$ are circles (1-spheres) and the isoplanes of $\mathrm{LID}_{\mathrm{PCA}}$ explaining x% of the total variance are triangles (2-simplices), intersected with the viable eigenvalue space simplex ($\sum_i \lambda_i = 1,\ 0 \leq \lambda_i \leq 1$). When generalizing these plots to d eigenvalues, the isovolumes of $\mathrm{LID}_{\mathrm{ABID}}$ generalize to concentric $(d-2)$-spheres. The isovolumes of $\mathrm{LID}_{\mathrm{PCA}}$ for the values ≥ 2 and $\geq d$ generalize to differently scaled dual $(d-1)$-simplices. The isovolumes for the values ≥ 3 through $\geq d-1$, however, assume shapes not as easily described. For example, for d=4 the isovolume of $\mathrm{LID}_{\mathrm{PCA}}$ for an estimate of ≥ 3 is a cube with corners facing the same direction as the corners of both the 3-simplices representing the values ≥ 2 and ≥ 4. As a hypothesis, it appears that the isovolumes from ≥ 2 to $\geq d$ describe a continuous dualization of the $(d-1)$-simplices for ≥ 2 and $\geq d$ combined with an increase in scale. In that case, the isovolumes have disjoint shapes symmetric under permutation of coordinates but not radially symmetric, whereas $\mathrm{LID}_{\mathrm{ABID}}$ promotes isovolumes of identical shape. The isoplanes provided by the PCA with a relative threshold value follow a far more intricate geometric shape. As a key difference to the absolute threshold value, a single smaller eigenvalue does not decrease the estimate as quickly (outward bending of the inner triangle sides), but a single larger eigenvalue increases the estimate much more quickly (inward bending of the outer triangle sides).

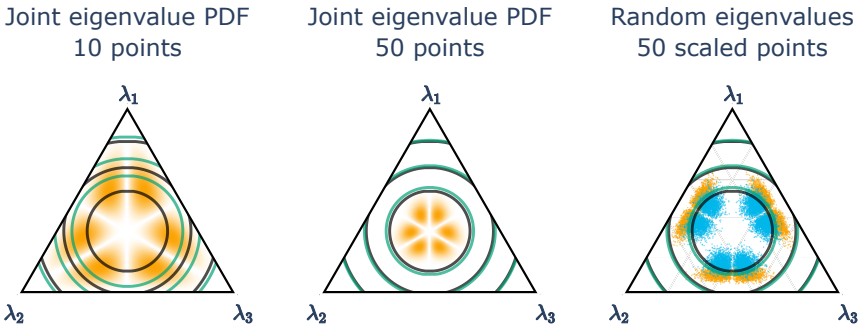


Figure 3.4: A priori probability density and empirical distribution for random covariance matrix eigenvalue vectors. The two left plots display the "tightening" impact of a larger dataset. The right plot displays the impact of linearly scaling one of the dimensions down by a factor of $\frac{1}{2}$. Blue points correspond to normalized and orange to not normalized datasets. All plots are overlayed with $\mathrm{LID}_{\mathrm{ABID}}$ isolines in black and $\mathrm{LID}_{\mathrm{RABID}}$ isolines in green. The isolines represent estimated LIDs of 2.5, 2, and 1.5 from inside to outside.

As for propositions towards the qualitative advantage of either of the approaches, radially symmetrical or simplicial, we make use of results from random matrix theory. In the field of random matrix theory Wishart matrices are used as a model for sample covariance matrices of random samples from multivariate normal distributions (see, e.g., Muirhead [64]). In the simplest case, which we will discuss here, the dimensions of the dataset are assumed to be independent, uncorrelated, and have the same standard deviation σ. The probability density function for the joint distribution of eigenvalues of the covariance matrix then reduces to

$$\mathrm{P}_{k,d,\sigma}(\lambda_1,\dots,\lambda_d) = c_{k,d,\sigma} \exp\left(-\tfrac{1}{2\sigma^2}\sum_{i=1}^{d}\lambda_i\right)\prod_{i=1}^{d}\lambda_i^{\frac{k-d-1}{2}}\prod_{i<j}^{d}|\lambda_i-\lambda_j| \tag{3.33}$$

where $c_{k,d,\sigma}$ is a normalization constant dependent on the number of points k, the number of features d, and σ, and λ_i are the eigenvalues of the covariance matrix [64, p. 107]. Most of the probability mass is concentrated in a finite number of blobs whereas the variance of these blobs decreases as the number of observed points increases. For estimates from $\mathrm{LID}_{\mathrm{PCA}}$ the eigenvalues are considered after dividing them by their sum or the largest value. The probability density function for eigenvalues with a sum equal to 1 is identical to the above function up to an adjusted normalization constant. This function can, therefore, be used to analyze the qualitative performance of $\mathrm{LID}_{\mathrm{PCA}}$ estimates based on an absolute threshold. For the relative threshold, we claim, that it is useful nonetheless, since in the inspected localities of LID estimation, we assume the data to form a hyperball, whose sum of eigenvalues solely depends on the dimensionality and can thus be considered constant. For $\mathrm{LID}_{\mathrm{ABID}}$ we normalize the neighborhood vectors prior to computing the eigenvalues, which transforms the shape of the probability density function towards a more circular shape. Thereby, the above probability density function is not immediately

applicable for $\mathrm{LID}_{\mathrm{ABID}}$, yet the tightening effect of a larger dataset still applies. Since we here condition the distribution on the number of samples leading to the covariance matrix, the estimates of $\mathrm{LID}_{\mathrm{RABID}}$ are different to those of $\mathrm{LID}_{\mathrm{ABID}}$. The isolines of $\mathrm{LID}_{\mathrm{RABID}}$ have a larger radius, which compensates for the increased uncertainty of the eigenvalues. The $\mathrm{LID}_{\mathrm{RABID}}$ estimate at the center ($\lambda_1{=}\lambda_2{=}\lambda_3$), which is almost never observed according to the probability distribution, is ≈ 3.86. This further highlights the difference between $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ estimates on smaller sample sizes. The $\mathrm{LID}_{\mathrm{RABID}}$ estimates are better fitting the probability density function, i. e. they are more "realistic", yet the $\mathrm{LID}_{\mathrm{ABID}}$ estimates can not overestimate the LID at the cost of occasionally being too "optimistic" about the required number of dimensions. In convergence, i. e. for large enough sample sizes, both estimators converge, giving an estimate that combines "realistic" and "optimistic".

Aside from providing visualization for the probability density, we generated eigenvalue vectors from random point sets to showcase the effect of scaling individual dimensions in Fig. 3.4. Scaling a single feature down forces the probability mass to shift towards the sides of the plot. In that case, the quality of estimates of $\mathrm{LID}_{\mathrm{PCA}}$ is extremely sensitive to the threshold parameter, as it controls the scale of the region for estimates of 3. Dependent on the threshold value almost all of the eigenvalue spectra will be identified as either intrinsically 2- or 3-dimensional. The choice of the threshold value, hence, results in extreme outcomes that are at best difficult to foresee yet very likely to encounter in practical ID estimation applications. Choosing a threshold value that positions this line right in the middle of these blobs even results in an error rate of $\approx 50\%$ no matter what behavior the user intended. As mentioned before, it does not suffice to choose a threshold value for deliberate results on the feature dimension but one has to consider the intrinsic dimension, as it is intimately connected to the margins in eigenvalue space. If the dataset consists of a mixture of differently dimensional manifolds this generally cannot be solved with a single threshold value. A non-discrete version of $\mathrm{LID}_{\mathrm{PCA}}$ could circumvent these effects, yet the complex shape of isovolumes makes defining a soft $\mathrm{LID}_{\mathrm{PCA}}$ non-trivial. Whilst the $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ estimates are of course also affected by differently scaled dimensions, the fractional estimates allow for a continuous transition between different intrinsic dimensions. That results in a smaller numerical error. As the centers of mass of the blobs approximately lie on a circle around the center of the eigenvalue space, the radially symmetric shape of $\mathrm{LID}_{\mathrm{ABID}}$'s isolines results in decent estimates on average. Normalizing the dataset further directs the shape of the probability towards a circular shape. Still, the eigenvalue distribution is not oriented solely along the isolines of $\mathrm{LID}_{\mathrm{ABID}}$, whereby a sufficiently large number of observed points is required to tighten the blobs. In terms of LID estimation, this means that the variance of eigenvalue vectors is too large on smaller neighborhoods to obtain reliable results with $\mathrm{LID}_{\mathrm{ABID}}$. These experiments also showcase that a radially symmetrical approach is less suited for non-normalized datasets. The in-

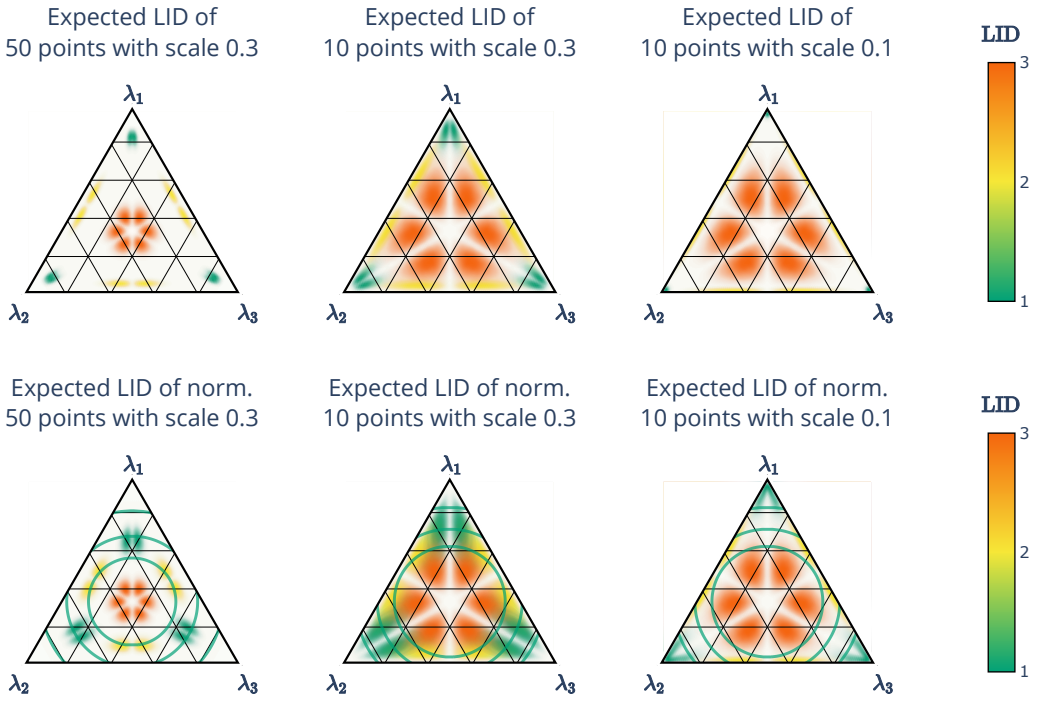


Figure 3.5: Empirical "true" LID values for random covariance matrix eigenvalue vectors. The top row considers the covariance on the samples as is and the bottom row considers the covariance on the normalized samples. Each plot is an RBF-smoothed scatter plot of the eigenvalues of random samples from uniform balls in 3 dimensions. For each ball, either 0, 1 or 2 dimensions were scaled down by the specified factor. Each point is colored by the number of non-scaled dimensions – the "true" LID under varying amounts of full-dimensional univariate noise. The opacity is controlled by the sampling density, i. e. white areas are areas where no corresponding spectra were observed. Each plot contains the same amount of samples for each "true" LID, yet especially in the upper row, the sampling density is concentrated to the sides and corners of the plot, leading to barely visible distributions. Overlaid lines in the lower plots are $\text{LID}_{\text{RABID}}$ isolines at the estimate values 2.5, 2, and 1.5 from inside to outside.

creased radius of $\text{LID}_{\text{RABID}}$ can be used to compensate for the increased uncertainty of the eigenvalues.

As for the comparison of different LID_{PCA} variants, they to some degree characterize different interpretations of LID. As can be seen in Fig. 3.5, the amount of noise in "non-intrinsic dimensions" as well as the number of samples affect the shape of the eigenvalue distribution that should yield a specific LID value. The distributions given with a larger amount of noise (larger scale values) exhibit the "denting in" of the corners, that can be observed in the isoplanes of the "rel %" variant, yet the sides especially in the lower noise regime appear rather flat like in the "abs %" variant. The choice of variant can, therefore, be deliberately used to cancel out high noise or rather consider noise part of the signal. The choice of the threshold value, however, is still an open issue and the number of samples used in computation must be considered when choosing the threshold.

As for $\mathrm{LID}_{\mathrm{RABID}}$, the errors of the estimator occur primarily in regions with low sample density or high noise. Errors in the low sample density areas obviously is not a big issue, yet under high noise, the continuous characteristic of the $\mathrm{LID}_{\mathrm{RABID}}$ estimate requires the collapse of different true LID values to a single estimate. If two dimensions have a noise term that accounts for 30% of the variance in the "signal dimensions", it appears like two additional "somewhere between 0 and 1"-dimensions. Evaluating these as, say "half a dimension" results in observing one additional dimension – similar to observing one more dimension without noise. That constraint is due to the real-valued nature of the estimator and extends to any fractional estimator in some way. In high-noise regimes, the $\mathrm{LID}_{\mathrm{RABID}}$ estimator, and by extension the $\mathrm{LID}_{\mathrm{ABID}}$ estimator, tends to overestimate the true LID by summing fractional observed dimensions. The "catastrophic" example in Fig. 3.5 is, however, a very extreme case. For robust LID estimation, having 30% additional full-dimensional noise quite severely violates the assumptions of the LID estimation problem. To some extend, a high local curvature yields the same increment in variance of non-intrinsic dimensions, yet, the total noise due to uncertainty in the samples and curvature of the manifold is assumed much less significant. For the more appropriate noise regime $\leq 10\%$, the $\mathrm{LID}_{\mathrm{RABID}}$ estimator visually gives the correct LID estimate for almost all samples. In that regime, the specific choice of the $\mathrm{LID}_{\mathrm{PCA}}$ variant also is less relevant but promotes the "abs %" variant. Hypothetically, an estimator based on the empirical eigenvalue distributions could be used to fit the shapes of the distributions and give a series of estimates over a varying amount of assumed noise, yet the complexity of the problem is not to be underestimated.

To conclude, the simplicial shape of the eigenvalue distribution on non-normalized data is as akin to $\mathrm{LID}_{\mathrm{PCA}}$ as the mostly spherical shape on normalized data is to $\mathrm{LID}_{\mathrm{ABID}}$. Both estimators do not fully characterize the distribution of eigenvalues, but certain variants do fit these distributions on average quite good. This hints towards comparable results of the two estimator types, which requires further research of joint eigenvalue distributions on normalized datasets as well as the effective mapping on eigenvalues due to normalizing the data for confirmation. The major difference remains the necessity of a threshold value for $\mathrm{LID}_{\mathrm{PCA}}$ and the discrete estimates, although $\mathrm{LID}_{\mathrm{PCA}}$ could possibly be generalized to soft estimates. As for $\mathrm{LID}_{\mathrm{PCA}}$ on correlation matrices, the connection towards $\mathrm{LID}_{\mathrm{ABID}}$ is only weak. The eigenvalues of the correlation matrix and the covariance matrix only agree on large enough standardized datasets, which is generally not the case in LID estimation. Eigenvalue distributions on these matrices have been researched in the field of random matrix theory as well, with similar results, yet we cannot clearly connect the results in terms of LID estimation. Normalization alone does not suffice to guarantee similar eigenvalues, which makes $\mathrm{LID}_{\mathrm{ABID}}$ mostly incomparable to $\mathrm{LID}_{\mathrm{PCA}}$ on correlation matrices of normalized vectors. $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{PCA}}$ on correlation matrices, in general, do not share a common domain.

3.4 $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$

In this section, we will inspect the relation of $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$. The two estimators are based on independently developed approaches, yet, we provide a somewhat continuous transition between the two. For these considerations, we will assume the idealized setting of an infinite dataset and a neighborhood radius approaching 0. These arguments, thereby, do not extend to realistic settings but demonstrate the intricate relation between the two estimators. Both estimators can be considered as placeholders for their corresponding family of estimators, since the theoretical objective of say $\mathrm{LID}_{\mathrm{MLE}}$, $\mathrm{LID}_{\mathrm{ALID}}$, and $\mathrm{LID}_{\mathrm{TLE}}$ is the same but with different practical considerations. The $\mathrm{LID}_{\mathrm{MLE}}$ estimator is specifically concerned with the expansion rate of (geodetic) distances in the vicinity of a point. The $\mathrm{LID}_{\mathrm{ABID}}$ estimator considers cosines of points in the vicinity of a point projected onto a unit-sphere. That sphere surface locally approaches a 1-codimensional subspace of the original neighborhood as the neighborhood radius approaches 0. In these localities on the sphere, the sines of angles are approximations to the geodetic distances on the sphere and thereby an approximation of the geodetic distances in the original space with one less dimension. Evaluating $\mathrm{LID}_{\mathrm{MLE}}$ on a locality of the projected points on the sphere should hence produce a LID estimate with exactly one dimension less than the original space. The $\mathrm{LID}_{\mathrm{ABID}}$ estimator now extends that concept by considering the limit of the $\mathrm{LID}_{\mathrm{MLE}}$ estimate over the entire space, i. e. the sphere, under the assumption of a specific distribution. A similar concept was previously proposed by Erba et al. by extending the Correlation Dimension over the sphere, yet they did not connect their result to the distribution of angles on a sphere [28]. Their estimator, hence, results in very similar estimates to $\mathrm{LID}_{\mathrm{RABID}}$ – in fact given some transformations, like mean-centering the neighborhood, they provide identical estimates in the limit – yet it requires the very costly approximation of the LID by sampling from a cumulative distribution function. Following that transition lends to intermediate steps as considerations for potential LID estimators, which we will sketch below. We will start with the distribution of geodetic distances, i. e. great circle distances, on the sphere within a specific radius on the sphere surface. From there we can continue with the distribution of pairwise geodetic distances over the entire unit sphere which coincides with the distributions of angles and leads to the distribution of chord lengths. Lastly, we arrive at the distribution of cosines considered by $\mathrm{LID}_{\mathrm{ABID}}$ and derive an approximation that results in the $\mathrm{LID}_{\mathrm{RABID}}$ estimator.

That continuous transition highlights some key observations when considering non-Euclidean distances since the great circle distance is non-Euclidean. This can be easily seen by the fact, that the geodetic "triangle" $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ has three right angles on the surface, which is not possible in a Euclidean space. For Euclidean spaces, extending the neighborhood radius of the $\mathrm{LID}_{\mathrm{MLE}}$ estimator over a uniform full-dimensional distribution should not affect the resulting LID estimate. When expanding the radius over

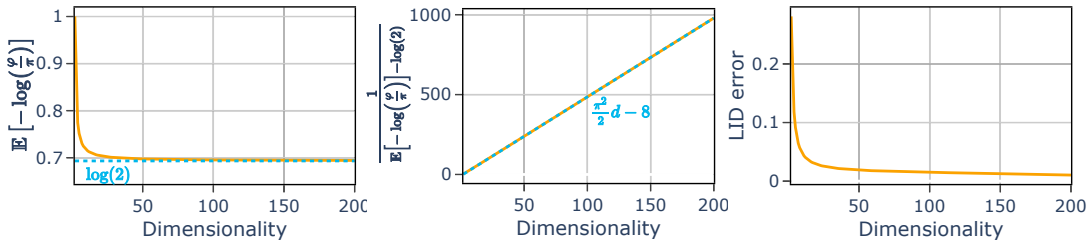


Figure 3.6: Derivation of an estimator for LID based on the expected log ratio of angles on a uniform sphere. The left plot displays the expected log ratio of angles on the sphere approaching $\log(2)$. The reciprocal of the difference to $\log(2)$ is displayed in the middle and can be approximated by a linear function. The right plot displays the error in estimated dimensionality using the linear fit.

a non-Euclidean space such as the unit sphere with the great circle distance, however, we need to accommodate for the non-Euclidean nature of the space and in the limit arrive at an entirely different estimator. For the $\mathrm{LID}_{\mathrm{MLE}}$ estimator to give correct results, we therefore also require the assumption of a locally Euclidean space, which limits the original work on the Expansion Dimension, where the space can be arbitrarily non-Euclidean [49]. Local Euclideanness holds for geodetic distances on manifolds but not for arbitrary metric spaces. To amend that "error", the Generalized Expansion Dimension [47] involves the Lebesgue measure of balls in the specific metric space, which alleviates the issue but is not applicable without rigorous analysis of the metric. Further, the links between practical complexity, i. e. in the context of indexing, and the LID have primarily been investigated for the immediate application of estimators [18, 49, 21], not for any generalized variants. In that regard, the $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$ estimator both rely on the local Euclideanness of the space they operate on.

When investigating the immediate vicinity of points on the sphere, we can in the limit of the number of samples to infinity employ the $\mathrm{LID}_{\mathrm{MLE}}$ estimator. The resulting function we are then interested in can be described in terms of an expected value:

$$\delta = \mathop{\mathbb{E}}_{\varphi \in [0, \varphi_0]} \left[-\log\left(\frac{\varphi}{\varphi_0}\right) \right]^{-1} + 1 = \left(\int_0^{\varphi_0} -\log\left(\frac{\varphi}{\varphi_0}\right) \mathrm{P}(\varphi) \ \mathrm{d}\varphi \right)^{-1} + 1 \tag{3.34}$$

Here φ_0 describes the neighborhood radius on the sphere and $\mathrm{P}(\varphi)$ is the probability density function of angles. For uniform distributions on the sphere, that probability density function is given by Cai et al. [14] as

$$\mathrm{P}(\varphi) = \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d-1}{2}\right)} (\sin(\varphi))^{d-2} \tag{3.35}$$

As $\varphi_0 \to \pi$ that estimator covers the entire sphere. Solving the integral to evaluate the expected value, however, is highly non-trivial and is not of interest in this work, especially since there is no hope for the resulting estimator to outperform $\mathrm{LID}_{\mathrm{ABID}}$ or $\mathrm{LID}_{\mathrm{RABID}}$. We

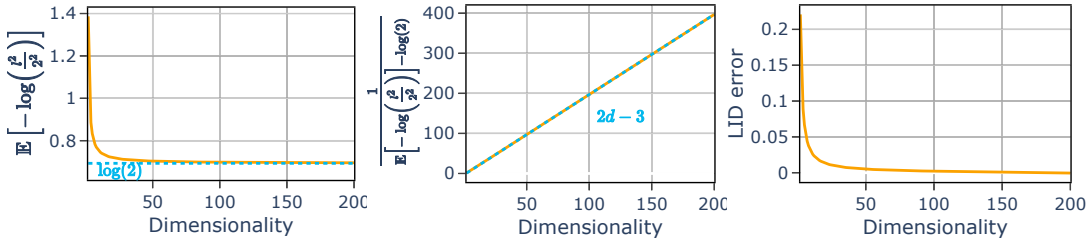


Figure 3.7: Derivation of an estimator for LID based on the expected log ratio of squared chordal lengths on a uniform sphere. The left plot displays the expected log ratio of squared chordal lengths on the sphere approaching $\log(2)$. The reciprocal of the difference to $\log(2)$ is displayed in the middle and can be approximated by a linear function. The right plot displays the error in estimated dimensionality using the linear fit.

can instead numerically evaluate the integral and attempt to fit a function to the result. Fig. 3.6 displays the process of numerically finding an appropriate fit function. The first observation is, that the expected log ratio approaches $\log(2)$ for increasing dimensionality and that the differential to $\log(2)$ drops approximately proportional to $\frac{1}{x}$. Plotting the reciprocal of the difference to $\log(2)$, we can approximate the result by a linear function, whose slope and intercept are close to $\frac{\pi^2}{2}$ and 8. We chose these values as our linear fit since the values are closely matched and simple. Equating the linear fit with the reciprocal of the difference to $\log(2)$ and solving for d yields the empirical great circle distance estimator

$$\delta = \frac{2}{\pi^2 \left(\mathbb{E}_{\varphi \in [0, \varphi_0]} \left[-\log\left(\frac{\varphi}{\varphi_0}\right) \right] - \log(2) \right)} + \frac{16}{\pi^2} \,. \tag{3.36}$$

Comparing that estimator to the plain reciprocal of the estimated log ratio as per $\mathrm{LID}_{\mathrm{MLE}}$, we observe that the functions are quite similar, except that both additive components are linearly scaled. The dominant term is still the reciprocal of the expected value – corrected by its value in the suspected limit of $d \to \infty$. In a similar fashion, we can derive an estimator based on the expected log ratio of chord lengths. The chord lengths are given as $(2 \sin\left(\frac{\varphi}{2}\right))$ and thus merely a change of variables away from the angles. The probability density function of chord lengths is given by Sidiropoulos [79] as

$$\mathrm{P}(l) = \frac{l}{\mathrm{B}\left(\frac{d-1}{2}, \frac{1}{2}\right)} \left(l^2 - \frac{l^4}{4} \right)^{\frac{d-3}{2}} \tag{3.37}$$

where $\mathrm{B}(\cdot, \cdot)$ is the beta function. Using that probability density function, we can now evaluate the expected value of the log ratio of squared chordal lengths and fit a function to the result. Fig. 3.7 displays the process of numerically finding an appropriate fit function in the same manner as we did for the log ratio of angles. This time, the maximum radius

is 2, leading to a different scaling factor inside the expected value. We end up with the empirical squared chordal length estimator

$$\delta = \frac{1}{2\left(\mathbb{E}_{l\in[0,2]}\left[-\log\left(\frac{l^2}{2^2}\right)\right]-\log(2)\right)}+\frac{3}{2} \tag{3.38}$$

$$= \frac{1}{2\left(2\,\mathbb{E}_{l\in[0,2]}\left[-\log\left(\frac{l}{2}\right)\right]-\log(2)\right)}+\frac{3}{2}\,. \tag{3.39}$$

The resulting estimator is again quite similar to the plain reciprocal of the expected value as per $\mathrm{LID}_{\mathrm{MLE}}$. Further, the estimator is also applicable to non-squared chord lengths by simply pulling the power out of the logarithm. We expect these estimators to work in practice, though since we do not provide a more theoretical derivation, we do not assign them names or further investigate them. In theory, they may show different behavior for e. g. outliers and might be a useful addition to ensemble methods, nonetheless. Here, we are rather interested in their usefulness for a "transition" between $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{ABID}}$. The squared chord lengths are immediately related to cosines since the squared chord length to an angle φ is equal to $2-2\cos(\varphi)$. Where we previously were interested in the expansion of chord lengths from 0 to 2, we are now interested in the expansion of cosines from 1 to -1. By substituting the ratio $\frac{l^2}{2^2}$ with $\frac{1-\cos(\varphi)}{2}$ and changing the limits of the expected value from $l\in[0,2]$ to $\cos(\varphi)\in[-1,1]$, we obtain the analogous estimator for the expected log ratio of cosines. To move from the logarithm over cosines to the squared cosines, we can employ the Taylor expansion-based approximation of the expected logarithm by Teh, Newman, and Welling [81] given as $\mathbb{E}\left[\log(x)\right]\approx\log(\mathbb{E}\left[x\right])-\frac{\mathrm{Var}[x]}{2\,\mathbb{E}[x]^2}$. We then obtain

$$\delta = \frac{1}{2\left(\mathbb{E}_{\cos(\varphi)\in[-1,1]}\left[-\log\left(\frac{1-\cos(\varphi)}{2}\right)\right]-\log(2)\right)}+\frac{3}{2} \tag{3.40}$$

$$\approx \frac{1}{2\left(-\left(\log\left(\mathbb{E}\left[\frac{1-\cos(\varphi)}{2}\right]\right)-\frac{\mathbb{E}\left[\left(\frac{1-\cos(\varphi)}{2}\right)^2\right]-\mathbb{E}\left[\frac{1-\cos(\varphi)}{2}\right]^2}{2\,\mathbb{E}\left[\frac{1-\cos(\varphi)}{2}\right]^2}\right)-\log(2)\right)}+\frac{3}{2} \tag{3.41}$$

$$= \frac{1}{2\left(-\left(\log\left(\frac{1}{2}\right)-\frac{\mathbb{E}\left[\left(\frac{1-2\cos(\varphi)+\cos(\varphi)^2}{4}\right)^2\right]-\left(\frac{1}{2}\right)^2}{2\left(\frac{1}{2}\right)^2}\right)-\log(2)\right)}+\frac{3}{2} \tag{3.42}$$

$$= \frac{1}{2\left(-\left(-\log(2)-\frac{\mathbb{E}[\cos(\varphi)^2]}{2}\right)-\log(2)\right)}+\frac{3}{2} \tag{3.43}$$

$$= \frac{1}{\mathbb{E}\left[\cos(\varphi)^2\right]}+\frac{3}{2} = \mathrm{LID}_{\mathrm{RABID}}+\frac{3}{2} \tag{3.44}$$

The additional value of $\frac{3}{2}$ is likely a result of the Taylor expansion-based approximation of the expected value of a logarithm since the estimates produced by the empirically fit

estimator for cosines are quite precise. In any case, we are not concerned with an exact result but rather with the theoretical link between the $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{ABID}}$ estimator, which is stated sufficiently clearly by the result. We obtained a continuous transition between the two estimators, whether the intermediate steps are exact in value or not, and thereby demonstrated the intricate relation between the two estimators.

In conclusion, the $\mathrm{LID}_{\mathrm{MLE}}$ estimator evaluated on geodetic distances on a sphere can be expanded to an estimator over the sphere by increasing the neighborhood radius to its maximum value, similar to the approach proposed by Erba, Gherardi, and Rotondo [28]. Similarly, the $\mathrm{LID}_{\mathrm{MLE}}$ could be expanded over any bounded geometry. The $\mathrm{LID}_{\mathrm{RABID}}$ estimator emerges from that approach after a series of intermediate steps. These are very similar to the $\mathrm{LID}_{\mathrm{MLE}}$ formula, yet differ due to the non-Euclideanness of their domain. This implies, that the $\mathrm{LID}_{\mathrm{MLE}}$ estimator is only applicable to non-Euclidean metrics under the assumption of local Euclideanness and within small localities. We have not provided a rigorous derivation for the intermediate steps but have provided empirical evidence, that they approximate the LID. These intermediate steps may be useful for ensemble methods but lack the theoretical foundation of other LID estimators as of now.

3.5 Analytical Comparison of $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$

A general problem in applying LID estimators is the definition of a locality. Commonly, the k-nearest neighbors surrounding a point of interest are considered. That is a natural choice following the preceding assumptions of a locally linear embedding and a smooth distribution density. In small k-nearest neighborhoods, the observed samples should then approach a uniform distribution in some linearly embedded δ-ball as the neighborhood radius approaches 0. Since the estimates are statistical measures, their result converges onto δ with increasing sample sizes. Combining these two observations, a very large number of neighbors in an infinitesimal neighborhood would be ideal. Yet, the number of samples in any given radius is limited in practice. Consequentially, the choice of neighborhood sizes is a balancing act between the estimate accuracy and the compliance with our preceding assumptions. Breaking the fundamental assumptions again impacts the quality of our estimates. For too small k the estimates do not necessarily converge on the ground truth result while for too large k the neighborhood may resemble a subspace with a LID other than δ.

In this section, we will investigate the convergence rates of both $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$ in Subsection 3.5.1. In Subsection 3.5.2 we will then investigate how either high-dimensional additive noise or non-linear embeddings affect these two estimators by disturbing their assumptions. We chose these two LID estimators as proxies for the quality of covariance- and expansion-based estimators as their estimates highly correlate with any estimator in their respective group and due to their rather simple definition which allows for analyt-

ical inspection. The observations presented here, therefore, are assumed to extend to all estimators in these two groups.

3.5.1 Convergence Rates and Sample Sizes

To evaluate how many samples are required in a locality to evaluate $\mathrm{LID}_{\mathrm{ABID}}$ within a certain error margin, we can take a look at the converge rate of covariance matrices. As per Cai et al. [15], the sample covariance matrix $\widehat{\Sigma}$ given N samples in $\mathbb{R}^d$ converges to the true covariance matrix Σ in expectancy at worst proportional to $\sqrt{d}$ and inversely proportional to $\sqrt{N}$, that is

$$\mathbb{E}\left[\left\|\Sigma - \widehat{\Sigma}\right\|\right] \in \mathcal{O}\left(\left(\frac{d}{N}\right)^{1/2}\right) \tag{3.45}$$

As the operator norm equals the maximum absolute eigenvalue of a matrix this induces that the maximum absolute difference of the i-th largest eigenvalues of Σ and $\widehat{\Sigma}$ respectively must also obey this asymptotical law. That is because we can assume the eigenvectors to align for large enough N as $\Sigma - \widehat{\Sigma}$ would otherwise include eigenvalues proportional to some $\left|\lambda_i(\Sigma) - \lambda_j(\widehat{\Sigma})\right|$ which does not decrease with growing N. To inspect the convergence rate of $\mathrm{LID}_{\mathrm{ABID}}$ we can consider it in terms of the eigenvalues of the covariance matrix of normalized and mirrored neighbors. The mirroring is necessary for the non-central covariance matrix and the regular covariance matrix to align. It technically doubles N, but we will not consider that scalar factor in the following, as it does not affect the asymptotic convergence rate. Since the convergence rate given by Cai et al. [15] does not depend on any specific distribution, it does hold for the normalized data as well. We can write

$$\Delta_{\mathrm{ID}} = \left|\mathrm{LID}_{\mathrm{ABID}}(\Sigma) - \mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})\right| = \left|\frac{1}{\sum_i \lambda_i(\Sigma)^2} - \frac{1}{\sum_i \lambda_i(\widehat{\Sigma})^2}\right| \tag{3.46}$$

Since every eigenvalue of $\widehat{\Sigma}$ converges to an eigenvalue in Σ, they can be bound asymptotically with $\lambda_i(\widehat{\Sigma}) - \mathcal{O}\left(\left(\frac{d}{N}\right)^{1/2}\right) \leq \lambda_i(\Sigma) \leq \lambda_i(\widehat{\Sigma}) + \mathcal{O}\left(\left(\frac{d}{N}\right)^{1/2}\right)$. Using $s \in [-1,1]^d$:

$$\Delta_{\mathrm{ID}} \in \left[\left|\frac{1}{\sum_i \left(\lambda_i(\widehat{\Sigma}) + s_i \mathcal{O}\left(\left(\frac{d}{N}\right)^{1/2}\right)\right)^2} - \frac{1}{\sum_i \lambda_i(\widehat{\Sigma})^2}\right|\right]_{s\in[-1,1]^d} \tag{3.47}$$

where $[f(x)]_{x\in X}$ is the interval $[\min_{x\in X} f(x), \max_{x\in X} f(x)]$. We can increase the interval size by replacing the $\mathcal{O}(\cdot)$ with a scalar multiple $c\left(\frac{d}{N}\right)^{1/2}$:

$$\subseteq \left[\left|\frac{1}{\sum_i \lambda_i(\widehat{\Sigma})^2 + c^2\frac{d}{N} + 2s_i\lambda_i(\widehat{\Sigma})\, c\left(\frac{d}{N}\right)^{1/2}} - \mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})\right|\right]_{s\in[-1,1]^d} \tag{3.48}$$

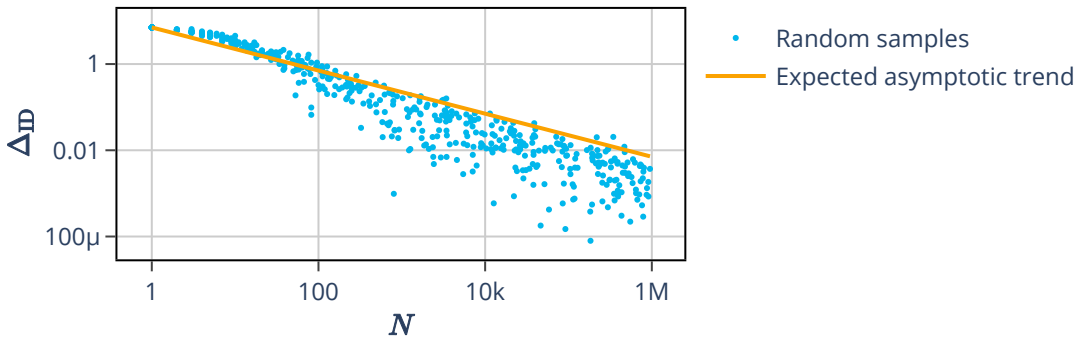


Figure 3.8: Empirical analysis of the expected asymptotic worst-case convergence behavior of $\mathrm{LID}_{\mathrm{ABID}}$ over varying sample size. Samples have 15 dimensions and follow a distribution with a "true" $\mathrm{LID}_{\mathrm{ABID}}$ value of ≈ 8.07 estimated on 10M samples.

Since the sums of eigenvalues of Σ and $\widehat{\Sigma}$ must both be 1, the s_i must be a mix of both positive and negative values, as long as the error is not 0. We can increase the interval by lowering the lower bound to 0. We further increase the interval by allowing all signs to be either positive or negative 1, whichever affects the denominator the most:

$$\subseteq \left[0, \max_{\hat{s}\in\{-1,1\}} \left| \frac{1}{\mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})^{-1} + c^2\frac{d^2}{N} + 2\hat{s}c\left(\frac{d}{N}\right)^{1/2}\sum_i \lambda_i(\widehat{\Sigma})} - \mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma}) \right| \right] \tag{3.49}$$

$$= \left[0, \max_{\hat{s}\in\{-1,1\}} \mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma}) \left| \frac{1}{1 + \mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})\left(c^2\frac{d^2}{N} + 2\hat{s}c\left(\frac{d}{N}\right)^{1/2}\right)} - 1 \right| \right] \tag{3.50}$$

$$= \left[0, \max_{\hat{s}\in\{-1,1\}} \mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma}) \left| \frac{c\mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})\frac{cd^2+2\hat{s}\sqrt{dN}}{N}}{1 + c\mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})\frac{cd^2+2\hat{s}\sqrt{dN}}{N}} \right| \right] \tag{3.51}$$

$$= \left[0, \max_{\hat{s}\in\{-1,1\}} \left| \frac{c\mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})^2\left(cd^2 + 2\hat{s}\sqrt{dN}\right)}{N + c\mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})\left(cd^2 + 2\hat{s}\sqrt{dN}\right)} \right| \right] \tag{3.52}$$

The convergence behavior is then dependent on what values are fixed and what values are growing. For fixed $N, \mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})$ and growing d, the expected worst-case error Δ_{ID} grows asymptotically linear – same as for fixed N, d and growing $\mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})$. For fixed $d, \mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})$ and growing N, Δ_{ID} converges to 0 as $\Delta_{\mathrm{ID}} \in \mathcal{O}\left(N^{-1/2}\right)$.

Fig. 3.8 empirically supports this theoretical asymptotic behavior. To obtain random covariance matrices, we sampled d samples from $\mathcal{N}(\mathbf{0}_d, \mathbb{I}_d)$ and computed their non-central covariance matrix Σ^*. Afterward, 10 million samples were drawn from $\mathcal{N}(\mathbf{0}_d, \Sigma^*)$ and then normalized. Using this normalized distribution, a "true" non-central covariance matrix Σ was computed. Afterward, samples of varying sizes were drawn and normalized, and their non-central covariance matrix $\widehat{\Sigma}$ was computed. We then evaluated the difference in $\mathrm{LID}_{\mathrm{ABID}}$ estimates using either $\widehat{\Sigma}$ or Σ. Fig. 3.8 includes the empirical values

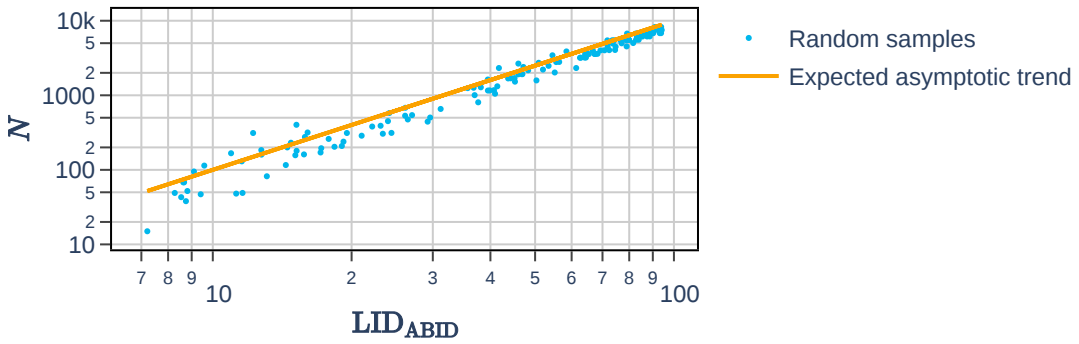


Figure 3.9: Empirical analysis of the expected asymptotic worst-case convergence behavior of $\mathrm{LID}_{\mathrm{ABID}}$ over varying covariances. Samples have 100 dimensions and sample sizes N have been chosen such that the "true" $\mathrm{LID}_{\mathrm{ABID}}$ (evaluated on 100k samples) is approximated with $\Delta_{\mathrm{ID}} \in [0.9, 1.1]$.

and a line corresponding to $(\mathrm{LID}_{\mathrm{ABID}}(\Sigma) - 1)N^{-1/2}$, since for $N = 1$ the $\mathrm{LID}_{\mathrm{ABID}}$ estimate via $\widehat{\Sigma}$ will be exactly 1 and therefore have an error of $\Delta_{\mathrm{ID}} = (\mathrm{LID}_{\mathrm{ABID}}(\Sigma) - 1)$. Note that the highest points in the plot only need to follow the trend of the line and not necessarily lie below it. Firstly, we do not know the exact scalar implied by the $\mathcal{O}(\cdot)$. Secondly, the bound only implies asymptotic behavior, and the bound, therefore, does not need to hold for any N below an unknown threshold value. Thirdly, the evaluation of the non-central covariance matrix over a huge number of samples suffers from numerical instability, which could explain the increased variation in the log-log-plot for large N.

Similarly we can use above equation to evaluate how fast the required sample size for some fixed Δ_{ID} needs to increase for varying d and $\mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})$. Since Δ_{ID} asymptotically grows linear in $\mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})$ it is implied that the sample size N needs to grow quadratic in $\mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})$ to obtain a constant error margin. And again, empirical analysis displayed in Fig. 3.9, where the drawn line equates to $\mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})^2$, suggests that this relation does hold. For that experiment, the covariance matrices were created in the same manner as before, yet, a random power in $[e^{-2}, e^2]$ was applied to all eigenvalues in Σ^* to increase or decrease the intrinsic dimensionality, which would otherwise cluster around $d/2$. We could in analogy deduce that the sample size is invariant towards the number of representation dimensions as well as Δ_{ID} is invariant towards d. Again, empirical evaluation displayed in Fig. 3.10 supports this claim. The drawn line equating to the constant $\mathrm{LID}_{\mathrm{ABID}}(\widehat{\Sigma})^2$ aligns with the trend of the empirical observations. A sample size independent of d might at first glance appear counterintuitive. The error on each of the eigenvalues, however, is very likely not uniform as suggested by the worst-case above and will in probability be proportional to the magnitude of each eigenvalue as can be seen in Fig. 3.11. Since we can expect about as many eigenvalues to be "much" larger than the rest – approximately equal to the intrinsic dimensionality – it would be intuitive, that any correlation between

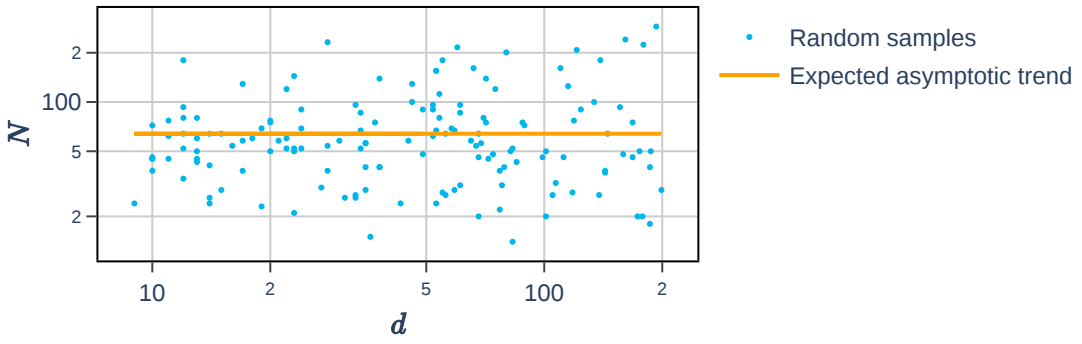


Figure 3.10: Empirical analysis of the expected asymptotic worst-case convergence behavior of $\mathrm{LID}_{\mathrm{ABID}}$ over varying representation dimensionalities. Samples have 8 to 200 dimensions and distributions are chosen such that the "true" $\mathrm{LID}_{\mathrm{ABID}}$ (evaluated on 100k samples) is in $[7.9, 8.1]$. Sample sizes N have been chosen such that the "true" $\mathrm{LID}_{\mathrm{ABID}}$ is approximated with $\Delta_{\mathrm{ID}} \in [0.9, 1.1]$.

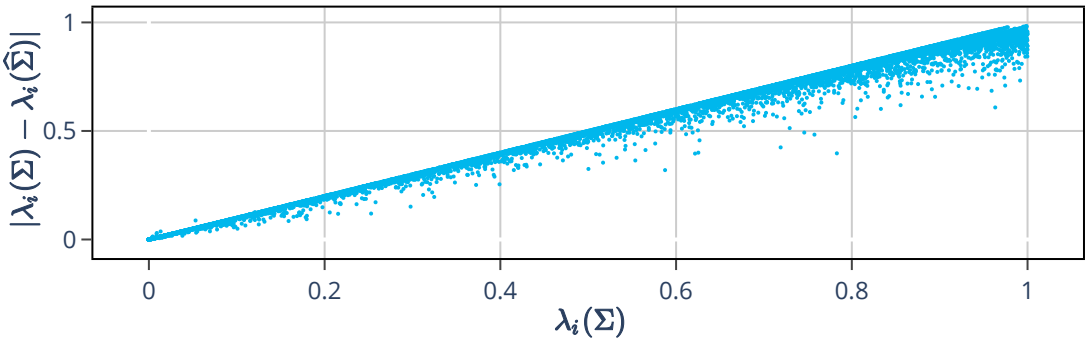


Figure 3.11: The absolute error of each eigenvalue of 150 random covariance matrices. Dimensions were randomly chosen between 10 and 200 and sample sizes between 10 and 1000.

required sample size and representation dimensionality disappears once we control for intrinsic dimensionality. Accordingly, the required sample size should rather be dictated by the intrinsic than by the representational dimensions. To enforce an approximately constant "true" $\mathrm{LID}_{\mathrm{ABID}}$, eigenvalues were iteratively chosen at random from the admissible interval to obtain said $\mathrm{LID}_{\mathrm{ABID}}$ value. For that, we use that the sum of eigenvalues S must be 1 and the sum of squared eigenvalues T must be $1/\mathrm{LID}_{\mathrm{ABID}}$. For one or two eigenvalues, these values are well-defined due to these constraints. For more than two eigenvalues, we can iteratively choose eigenvalues by considering the possible eigenvalue space induced by the constraints as the intersection of a hyperplane and a hypersphere, which again forms a hypersphere. By limiting the hypersphere to non-negative coordinates, we can deduce the possible values for the next eigenvalue to necessarily be within

$$\left[\max\left\{0, \frac{S-\sqrt{(d-1)(dT-S^2)}}{d}\right\}, \frac{S+\sqrt{(d-1)(dT-S^2)}}{d}, \right] \tag{3.53}$$

whenever the entire hypersphere of possible eigenvalues has non-negative coordinates, i.e. whenever $2T - S^2 < 0$, and otherwise

$$\left[0, \frac{S-\sqrt{2T-S^2}}{2}\right] \cup \left[\frac{S+\sqrt{2T-S^2}}{2}, \frac{S+\sqrt{(d-1)(dT-S^2)}}{d}\right] \tag{3.54}$$

For the recursion, S can be diminished by the chosen eigenvalue and T by its square whereas d must be decreased by 1.

After choosing the eigenvalues for the initial covariance matrix we adjust the eigenvalues via a power to obtain the desired $\mathrm{LID}_{\mathrm{ABID}}$ value on the covariance matrix of multivariate normal samples after normalization. The required exponents were computed using numerical optimization. This process was designed to produce as diverse covariance matrices as possible within a reasonable time since choosing truly at random from some distribution and rejecting based on $\mathrm{LID}_{\mathrm{ABID}}$ is not feasible for larger values of d. Nonetheless, the observed invariance of N towards d, when controlling for $\mathrm{LID}_{\mathrm{ABID}}$, has been observed for all evaluated generation schemes for covariance matrices as well, including a generation process similar to the one used in Fig. 3.9.

All of these relations, however, only apply to the asymptotic worst-case behavior and do not necessarily reflect on the average behavior, which should center on the "true" $\mathrm{LID}_{\mathrm{ABID}}$ value as the "true" covariance matrix is the mode of all sampled covariance matrices. Since every locality might follow a different local distribution, this argument does not necessarily extend to the histogram of $\mathrm{LID}_{\mathrm{ABID}}$ estimates. Neither can we deduce any argument on the skewness of the $\mathrm{LID}_{\mathrm{ABID}}$ estimates around their mean, i.e. the sidedness of the error. It does, however, support the colloquially shared (and apparently never written down) rule-of-thumb that the number of samples should be chosen proportional to the squared number of (intrinsic) dimensions, as in that case, the expected worst-case error Δ_{ID} remains invariant.

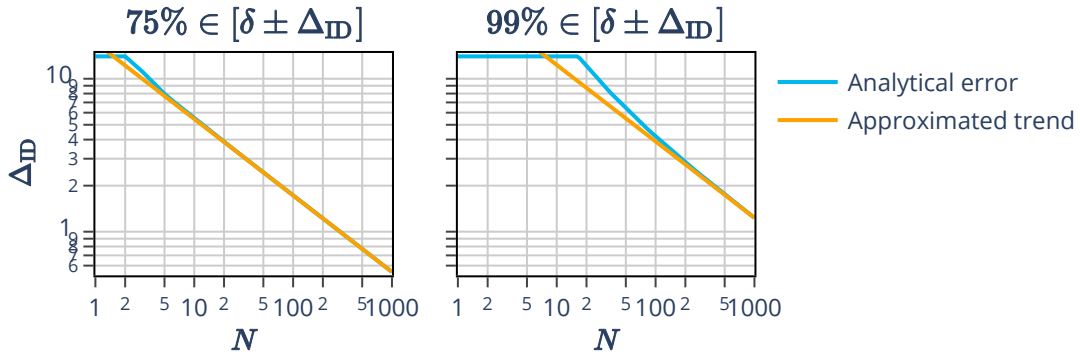


Figure 3.12: Error bound on $\mathrm{LID}_{\mathrm{MLE}}$ estimates in probability as indicated by the respective title. The ground truth ID used in these plots is $\delta = 15$ and the approximation is $f(x) = x^{-1/2}$ fitted through the last point of the analytical function.

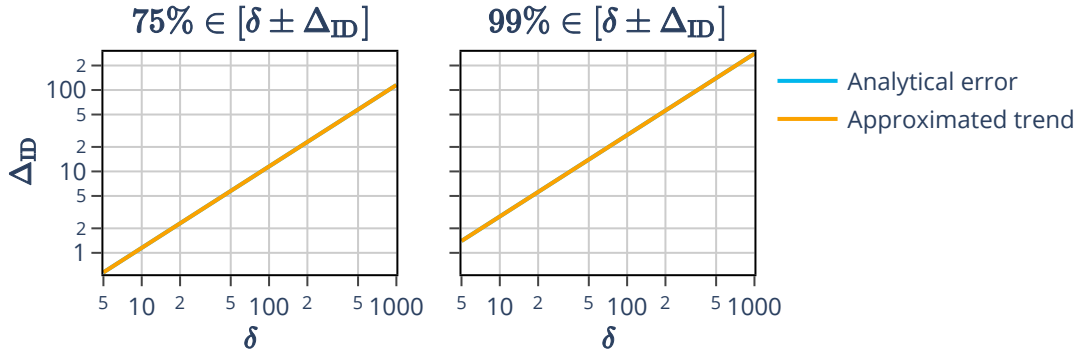


Figure 3.13: Error bound on $\mathrm{LID}_{\mathrm{MLE}}$ estimates in probability as indicated by the respective title. The sample size used in these plots is $N = 100$ and the approximation is $f(x) = x$ fitted through the last point of the analytical function.

To analyze the convergence rates of the $\mathrm{LID}_{\mathrm{MLE}}$ estimator we assume the data to lie uniformly in some unit δ-ball. In that way, we can drop the largest sampled distance from the definition of $\mathrm{LID}_{\mathrm{MLE}}$ which then becomes the negative reciprocal of the mean logarithm of observed distances. The distribution of this mean value can be derived analytically starting at the probability to observe any specific distance

$$\mathrm{P}_r(x) = \frac{\mathscr{A}_\delta(x)}{\mathscr{V}_\delta(1)} = \delta x^{\delta-1} \tag{3.55}$$

where $\mathscr{V}_\delta(x)$ and $\mathscr{A}_\delta(x)$ are the volume and surface of the δ-ball with radius x, respectively. By change of variable, we obtain the probability of the negative logarithm of any such distance

$$\mathrm{P}_{-\log(r)}(x) = \mathrm{P}_r(e^{-x}) \left| \frac{\partial}{\partial x} e^{-x} \right| = \delta e^{-\delta x} \tag{3.56}$$

To obtain the probability distribution of the mean of N negative logarithms we first investigate the sum of 2 of such values

$$\mathrm{P}^{(2)}_{-\log(r)}(x) = \int_0^x \mathrm{P}_{-\log(r)}(x-y)\, \mathrm{P}_{-\log(r)}(y)\, \mathrm{d}y = \delta^2 x e^{-\delta x} \tag{3.57}$$

Similarly we can generalize this distribution to the sum of N values by replacing the left distribution in the integral and obtain

$$\mathrm{P}^{(N)}_{-\log(r)}(x) = \int_0^x \mathrm{P}^{(N-1)}_{-\log(r)}(x-y)\, \mathrm{P}_{-\log(r)}(y)\, \mathrm{d}y = \frac{\delta^N x^{N-1} e^{-\delta x}}{(N-1)!} \tag{3.58}$$

By change of variable we then get the ideal distribution of $\mathrm{LID}_{\mathrm{MLE}}$ estimates as

$$\mathrm{P}_{\mathrm{LID}_{\mathrm{MLE}}}(x) = \mathrm{P}^{(N)}_{-\log(r)}\left(\frac{N}{x}\right) \left| \frac{\partial}{\partial x} \frac{N}{x} \right| = \frac{(\delta N)^N e^{-\frac{\delta N}{x}}}{x^{N+1}(N-1)!} \tag{3.59}$$

The expected value of this distribution is $\delta N/(N-1)$. The $\mathrm{LID}_{\mathrm{MLE}}$ estimator is slightly biased which can be fixed by adding one degree of freedom to the mean. This observation has previously been made by Levina and Bickel in their paper introducing the $\mathrm{LID}_{\mathrm{MLE}}$ estimator [54]. Below the mean, the function is dominated by the exponential part $e^{-\frac{\delta N}{x}}$ and the polynomial part $x^{-(N+1)}$ dictates the behavior above the mean. Consequentially, the distribution is asymmetric with a heavy upper tail and the potential error of $\mathrm{LID}_{\mathrm{MLE}}$ is unbounded. Although Levina and Bickel already analyzed the convergence behavior of $\mathrm{LID}_{\mathrm{MLE}}$, their analysis was limited to the variance of estimates [54]. To evaluate the convergence behavior similar to the analysis for $\mathrm{LID}_{\mathrm{ABID}}$ given above, we instead focus on an error bound in probability. We start with the probability that the $\mathrm{LID}_{\mathrm{MLE}}$ estimate deviates from δ by less than some threshold value Δ_{ID} by integrating over the distribution in (3.59)

$$\mathrm{P}_{err}(\Delta_{\mathrm{ID}}) = \int_{\delta-\Delta_{\mathrm{ID}}}^{\delta+\Delta_{\mathrm{ID}}} \mathrm{P}_{\mathrm{LID}_{\mathrm{MLE}}}(x)\, \mathrm{d}x = P\left(N, \frac{\delta N}{\delta - \Delta_{\mathrm{ID}}}\right) - P\left(N, \frac{\delta N}{\delta + \Delta_{\mathrm{ID}}}\right) \tag{3.60}$$

where $P(a, b)$ is the regularized incomplete lower gamma function. To obtain the error bound Δ_{ID} which is held with probability p, we would need to invert this equation for which we do not know an analytic solution. Yet, since $\text{P}_{err}(\Delta_{\text{ID}})$ is monotonously increasing, we can numerically invert the equation. By doing so we can plot the error bound in probability p both over the sample size as shown in Fig. 3.12 and over δ as shown in Fig. 3.13. As can be seen in the plots, the error bound appears to be asymptotically proportional to both δ and $N^{-1/2}$ in probability. Since, by definition, LID_{MLE} is invariant to additional zero-variance dimensions, Δ_{ID} must be invariant to d. These are the same scaling factors as in the previous derivation for LID_{ABID}, whereby the same arguments apply here. That is, for expansion-based approaches, a quadratic amount of observed samples relative to the LID should suffice.

A quadratic number of samples might appear quite large, since e. g. 100^2 points can barely be considered a "locality" in any practical dataset and, hence, properly evaluating the intrinsic dimensionality of datasets with a very high intrinsic dimensionality remains an open problem in many practical cases due to an insufficient number of observations. Yet, requiring "only" a number of samples proportional to the square of intrinsic dimensions remains in stark contrast to the potentially exponential amount required in count-based estimators like the Correlation Dimension [33] as has been argued by Eckmann and Ruelle [27].

Aside from the impact of sample sizes on LID estimates, we can, further, derive an argument on whether or not the assumptions of LID_{MLE} (and by extension any of the expansion-based estimators) are satisfied from (3.60). When observing a distribution of LID_{MLE} estimates from some manifold, an interval of $2\Delta_{\text{ID}}$ around the mean should cover about $\text{P}_{err}(\Delta_{\text{ID}})$ of estimates. If the error margin covers much more or less estimates, then any of the assumptions of LID_{MLE} (local uniformity, local linearity) may not hold. In the next section, we will, therefore, investigate the impact of broken assumptions on the two estimators discussed here and further inspect whether or not (3.60) can be used to test for the assumptions of LID_{MLE}.

3.5.2 Impact of Broken Assumptions

Under the assumption of a locally linear embedding function and a locally linear distribution density, the assumptions of both LID_{MLE} and LID_{ABID} can be reduced to observing small enough localities and a neglectable amount of high-dimensional noise. The local uniformity of LID_{MLE}, for example, is induced by the locally linear distribution density, which produces the same distances from some point of reference as a locally uniform distribution, as the increase of probability in one direction counteracts the decrease of probability in the opposite direction. Similarly, the spherical uniformity assumption of

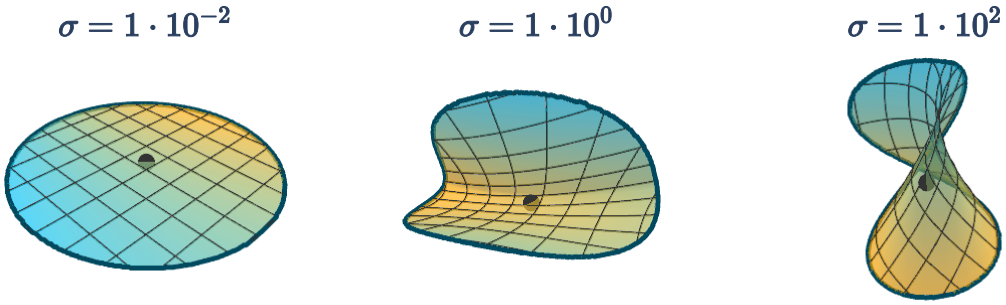


Figure 3.14: Examples of random non-linear embedding functions from two into three dimensions. The standard deviation σ controls the size of random supersymmetric tensors used in a Taylor series-like fashion. The dot represents the origin and the outline shows points at a fixed distance to the origin, i. e. the bounds of a neighborhood in embedding space.

$\mathrm{LID}_{\mathrm{ABID}}$ is satisfied, as the square of cosines implicitly introduces mirrored instances to even out the distribution density.

Both locally linear embedding function and distribution density, in turn, are sound assumptions, since we do not work on analytical representations but on discrete samples on which a locally non-linear embedding function is virtually indistinguishable from one with locally high curvature. That is, one practically can not differentiate between these assumptions being unsatisfied and the inspected localities being too large. We, therefore, consider only the case of "localities" being chosen too large or having a non-neglectable amount of high-dimensional noise.

To evaluate how $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{ABID}}$ behave when the locality assumption is not satisfied, we propose to investigate random embedding functions with increasing local non-linearity. To obtain random embedding functions that are as general as possible, we consider random polynomials with a weighting scheme similar in structure to the Taylor series. The derivatives of higher dimensional functions take the form of (super-)symmetrical tensors, i. e. tensors that are invariant under permutation of their vector arguments. By fixing the first $(k-1)$ vector arguments of an order k supersymmetric tensor T to some vector v, we obtain a vector that contains all possible polynomials over coefficients in v with degree $(k-1)$. This vector then simulates the $(k-1)$-th derivative of some implicit function evaluated for the argument vector v. Since we assume supersymmetric tensors, fixing the vector arguments can be done by applying $(k-1)$ dot products with the given vector, denoted as $T \cdot^{(k-1)} v = T \cdot v \cdot \ldots \cdot v$. By interpreting the resulting vectors as "derivatives" of an argument vector v and hence as coefficients of a Taylor series, we can simulate arbitrary (infinitely differentiable) functions from $\mathbb{R}^d$ to $\mathbb{R}^d$. We, hence, propose the embedding function

$$E(v, k_{\max}) = v + \sum_{i=2}^{k_{\max}} \frac{T_i \cdot^{(i-1)} v}{(i-1)!} \tag{3.61}$$

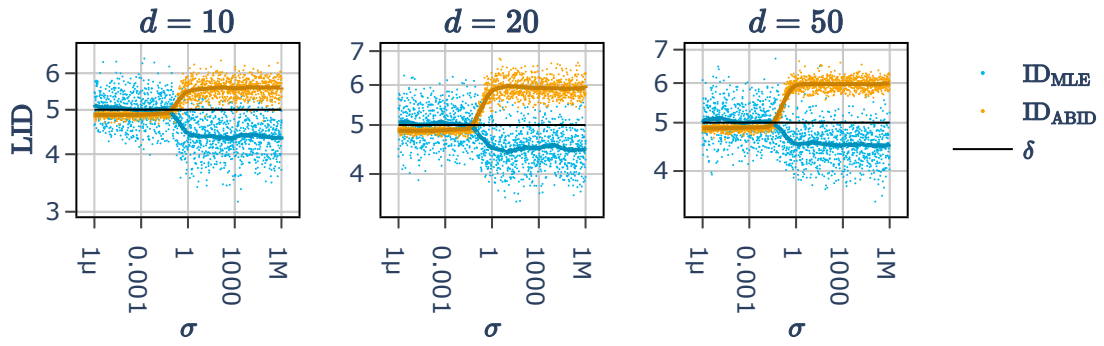


Figure 3.15: $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{ABID}}$ estimates for non-linear embeddings from $\delta = 5$ into d dimensions over a varying amount of non-linearity in the embedding function with $N = 150$ chosen from 1500 embedded points. The overlaid lines are mean values over a sliding window of 160 samples over LID values sorted by the corresponding σ.

where T_i are (random) order i tensors. To obtain an embedding function from $\mathbb{R}^\delta$ to $\mathbb{R}^d$, we can pad the δ-dimensional vectors with $(d - \delta)$ zeros. The maximum degree of this embedding function is dictated by $k_{\max}$ which, in practice, can not be chosen too large, since the computational cost of the embedding functions grows exponentially in $k_{\max}$. In the following experiments, we used $k_{\max} = 4$. The coefficients of the random tensors were sampled from a normal distribution under varying standard deviations. The closer these coefficients are to 0, the closer the embedding function is to the identity, whereas larger coefficients amplify the non-linear parts of the embedding function. Fig. 3.14 displays some examples of random embedding functions from 2 to 3 dimensions for varying standard deviations of the coefficient distribution. As the neighborhoods, in practice, are chosen in the embedded space, Fig. 3.14 displays the intersection of the produced variety with a full-dimensional ball. In the following experiments, this behavior is obtained by producing a larger set of samples in the latent space and selecting the k-nearest neighbors to the origin after embedding. As can be seen from the examples in Fig. 3.14, the non-linearity increases with σ. Due to the probabilistic nature of the process, not all generated functions strictly adhere to this rule and the expected asymptotic behavior for very large σ is uncertain, yet, we can expect a regime change from almost linear functions to very non-linear functions.

We practically evaluated the effect of increased non-linearity for varying δ and d. As displayed in Fig. 3.15, changing d did not change the effect of increased non-linearity on either $\mathrm{LID}_{\mathrm{MLE}}$ or $\mathrm{LID}_{\mathrm{ABID}}$ a lot. The $\mathrm{LID}_{\mathrm{ABID}}$ estimates have a lower variance than those by $\mathrm{LID}_{\mathrm{MLE}}$ and both estimators experience a regime change roughly between $\sigma = 0.1$ and $\sigma = 1$. The increased estimates in $\mathrm{LID}_{\mathrm{ABID}}$ can be explained by the increased geometrical complexity due to higher curvature. The $\mathrm{LID}_{\mathrm{MLE}}$ estimates instead become smaller due to the non-uniform data distribution after embedding. Why that is, however, remains

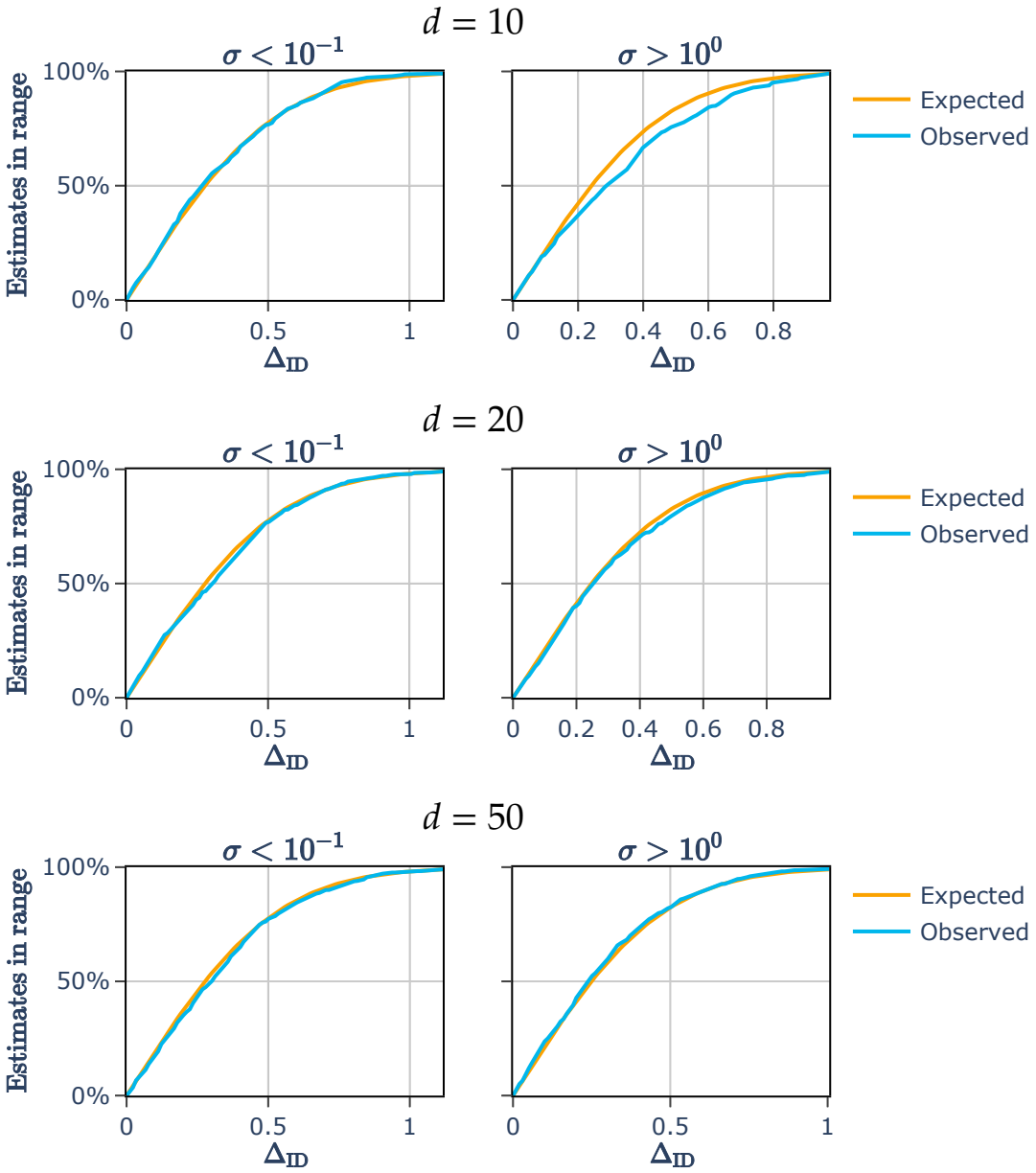


Figure 3.16: The expected fraction of $\mathrm{LID}_{\mathrm{MLE}}$ estimates within a prescribed estimation error to the mean estimate Δ_{ID} of $\delta = 5$ intrinsically dimensional vectors non-linearly embedded into $d = 10, 20, 50$ dimensions over a varying amount of non-linearity with $N = 150$ chosen from 1500 initial samples.

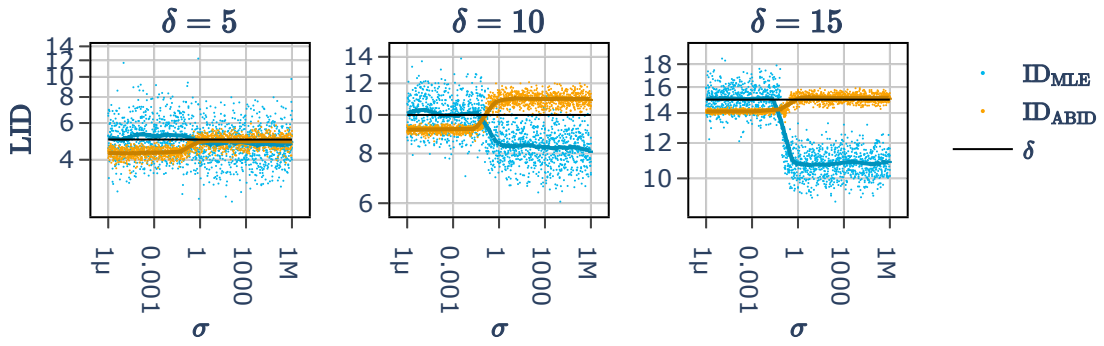


Figure 3.17: $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{ABID}}$ estimates for non-linear embeddings from δ into $d = 20$ dimensions over a varying amount of non-linearity in the embedding function with $N = \delta^2$ chosen from $5\delta^3$ embedded points. The overlaid lines are mean values over a sliding window of 160 samples over LID values sorted by the corresponding σ.

unclear. It may be due to the polynomial distortion of the sampling density leading to a modified expansion rate. The effect itself then is likely dependent on the generative process of the functions and not easily generalizable. We can not explain, why the estimates converge on some fixed value but propose that it happens due to convergent behavior of the generative process of embedding functions. Similar behavior, however, also appeared when replacing the normal coefficient distribution with a uniform distribution of growing intervals around 0. A better (understood) generative process for arbitrary non-linear embedding functions would be required to further investigate this behavior but none is known to the author. The non-linearity of the embedding function nonetheless appears to have an adversarial effect on the $\mathrm{LID}_{\mathrm{MLE}}$ estimator. However, this effect can not be explained by the hypothesized assumption test using (3.60). As displayed in Fig. 3.16, the observed error margins for $\delta = 5, d = 10$ do not fit the expected error margins, suggesting broken assumptions. Comparing the area under the curves, the $\delta = 5, d = 10$ case is off by ≈6.3%, whereas the error on all other curves is below ≈2%. Thus, the assumption test does not necessarily raise any concerns about the validity of the $\mathrm{LID}_{\mathrm{MLE}}$ estimator in the non-linear regime but may be an indicator.

Fig. 3.17 displays the same experiment as Fig. 3.15 but with fixed d and varying δ. We again experience a regime change between $\sigma = 0.1$ and $\sigma = 1$ after which the $\mathrm{LID}_{\mathrm{MLE}}$ estimates are smaller and the $\mathrm{LID}_{\mathrm{ABID}}$ estimates are slightly larger. Considering the proposed test for adherence to the $\mathrm{LID}_{\mathrm{MLE}}$ assumptions, no clear indication of broken assumptions can be observed. According results are displayed in Fig. 3.18. The area under the curves does not diverge from the expected values by more than 1.5% for any of the curves. Whether or not that is due to the chosen generative process for the non-linear embedding functions is unclear, though.

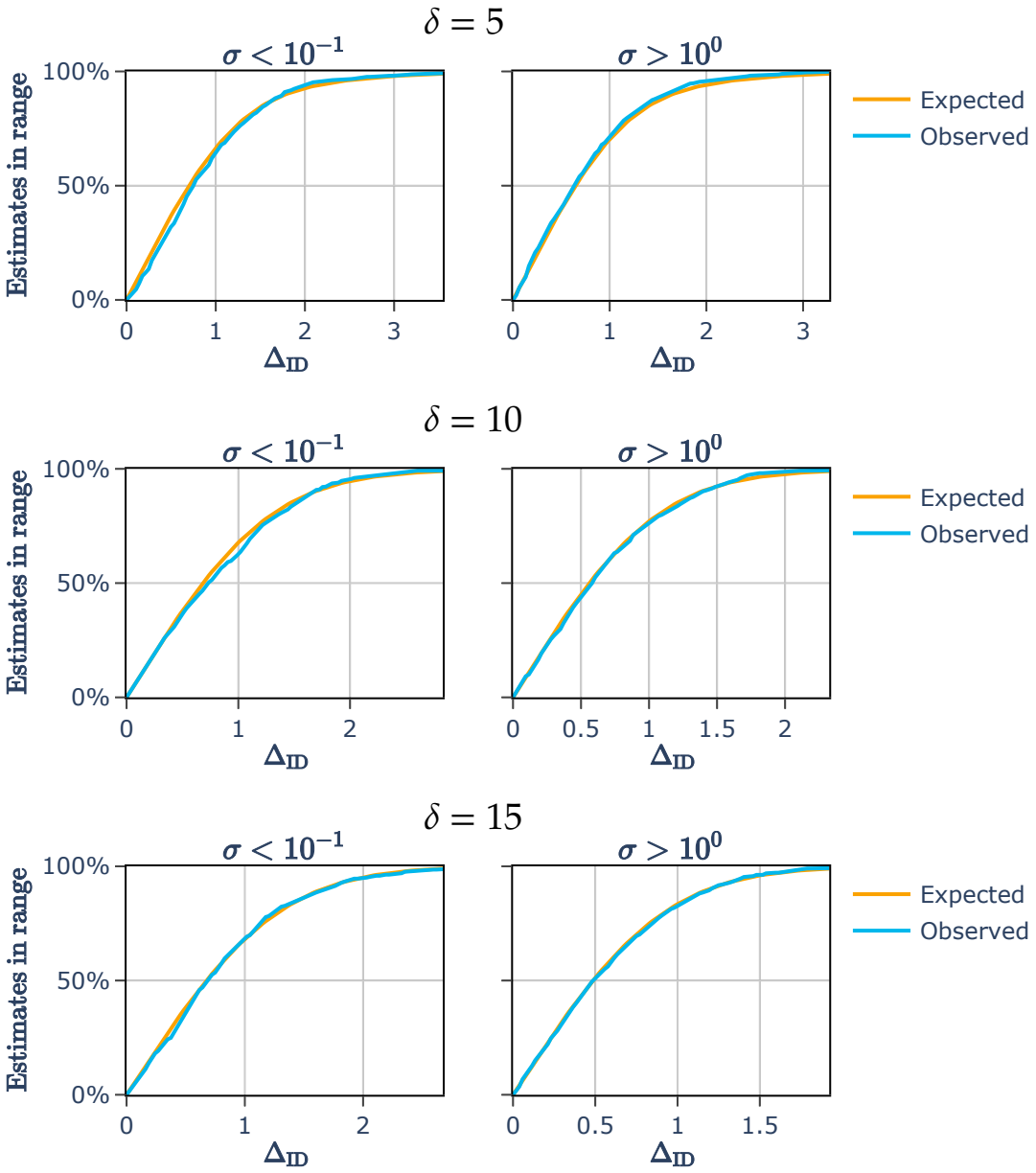


Figure 3.18: The expected fraction of $\mathrm{LID}_{\mathrm{MLE}}$ estimates within a prescribed estimation error to the mean estimate Δ_{ID} of $\delta = 5, 10, 15$ intrinsically dimensional vectors non-linearly embedded into $d = 20$ dimensions over a varying amount of non-linearity with $N = \delta^2$ chosen from $5\delta^3$ initial samples.

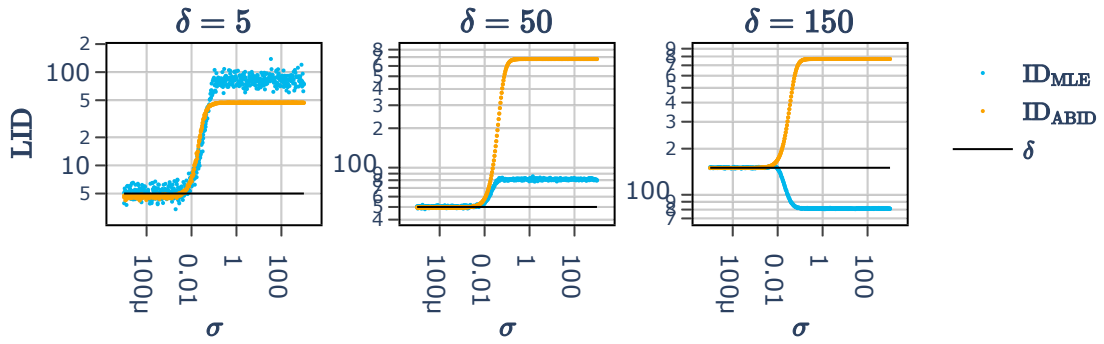


Figure 3.19: $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{ABID}}$ estimates for linear embeddings with added full-dimensional noise from δ into $d = 784$ dimensions over a varying amount of noise with $N = 2\delta^2$ chosen from $20\delta^2$ initial samples.

We further observe, that the $\mathrm{LID}_{\mathrm{MLE}}$ estimates can be much smaller than the true ID when the non-linearity is very pronounced and the error increases with δ, whereas the $\mathrm{LID}_{\mathrm{ABID}}$ estimates at least stay in the same order of magnitude. The evidence provided here is, however, still rather anecdotal, since the distribution of random functions proposed here has not been thoroughly studied and the reason for the surprisingly convergent behavior for large σ is unclear. The limitation of the tensor order to $k_{\max} = 4$ might limit the occupied geometry by the generated variety, yet, the example in Fig. 3.14 visually displays a rather extreme outcome. On varieties as folded up as those for $\sigma = 1$ and $\sigma = 10^2$ in Fig. 3.14, the considered neighborhood is well beyond what one could consider locally-linear. In these cases, one would need to choose either a smaller neighborhood to adhere to the local-linearity assumption or a much larger neighborhood to analyze the "macro structure" of the variety and consider the non-linearities as measuring noise.

After inspecting the effect of non-linearity we will now discuss the impact of d-dimensional additive noise, i. e. a d-dimensional random shift like measuring noise, on $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{ABID}}$. For that, we padded δ-dimensional vectors from a uniform ball distribution with $(d-\delta)$ zeros and added noise to each component sampled from a univariate normal distribution. By changing the standard deviation of the noise distribution, we can control the amount of noise added to the datasets. Fig. 3.19 displays the effect of this kind of noise on the estimates of $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{ABID}}$. We again observe a regime change when the noise becomes so strong, that the data appears indistinguishable from full-dimensional at similar σ values for both estimators. While the $\mathrm{LID}_{\mathrm{ABID}}$ estimator properly recognizes the high-dimensional space, the $\mathrm{LID}_{\mathrm{MLE}}$ estimates may even decrease, likely due to the non-uniformity of the additive noise. Fig. 3.20 displays the expected and measured error margin curves for the experiment displayed in Fig. 3.19. For the $\sigma < 0.001$ regime, the measured probabilities are quite noisy but follow the general trend of the expected probabilities. For $\sigma > 0.1$, the measured probabilities for any Δ_{ID} are higher than

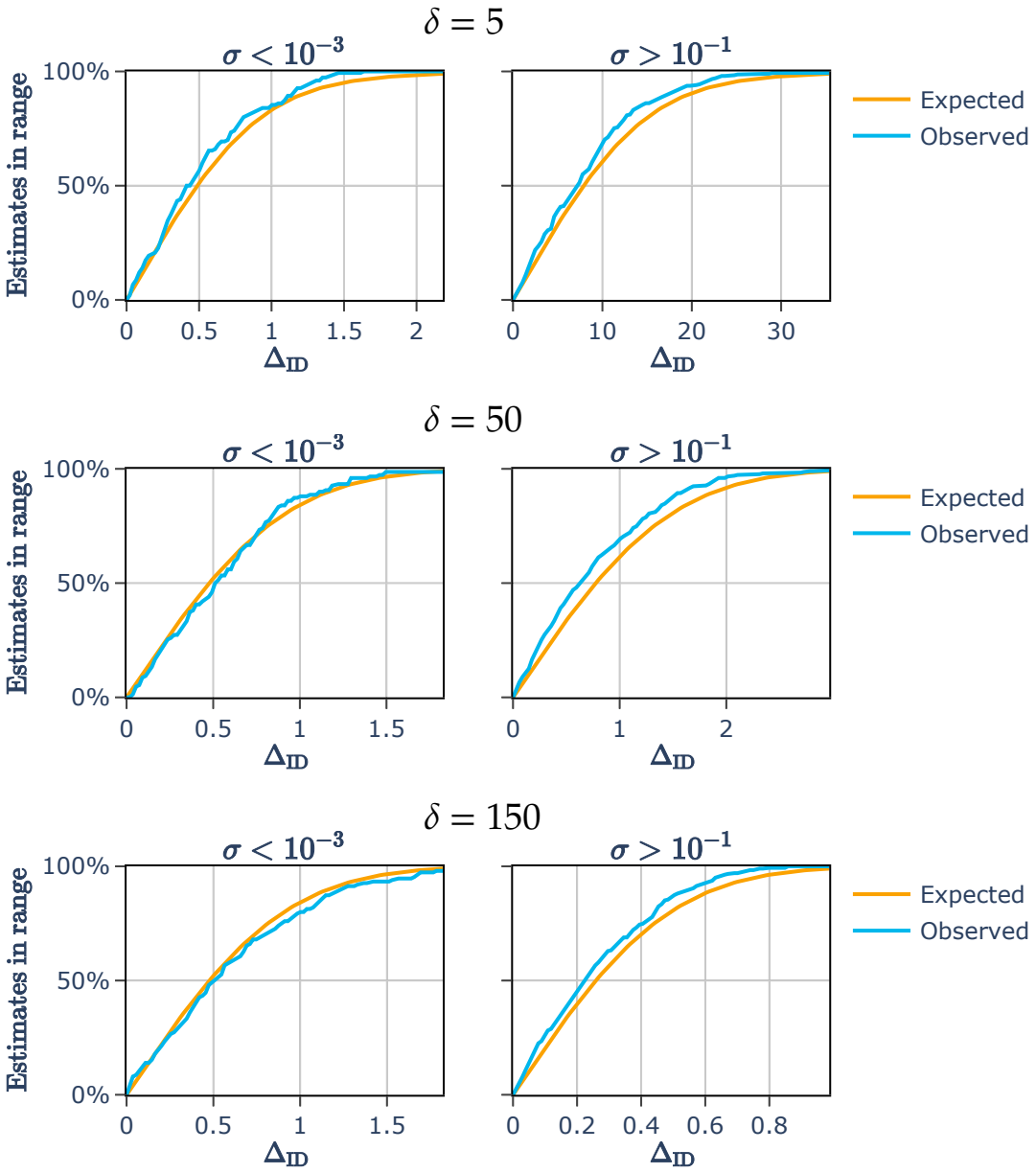


Figure 3.20: The expected fraction of $\mathrm{LID}_{\mathrm{MLE}}$ estimates within a prescribed estimation error to the mean estimate Δ_{ID} of $\delta = 5, 50, 150$ intrinsically dimensional vectors with added full-dimensional noise in $d = 784$ dimensions over a varying amount of noise with $N = 2\delta^2$ chosen from $20\delta^2$ initial samples.

the expected probabilities. These curves were calculated on < 250 $\mathrm{LID}_{\mathrm{MLE}}$ estimates, and corresponding curves may be much smoother for larger sample sizes. This time, considering the area under the curve hints at broken assumptions in the $\sigma > 0.1$ regime. The maximum divergence from the expected area under the curve in the $\sigma < 0.001$ regime is $\approx$4.5% and the minimum divergence in the $\sigma > 0.1$ regime is $\approx$5%. Here, a threshold of 5% would separate the two regimes. These observations would suggest that the probability-based criterion may be more reliable for the noise case than for the non-linearity case. As per these results, a "suspicious" relative error of the area under the curve beyond 5% may be a good indicator, that the $\mathrm{LID}_{\mathrm{MLE}}$ estimates may not be reliable. In conclusion, the test using the area under the curve can not be used to confirm the assumptions of $\mathrm{LID}_{\mathrm{MLE}}$ but may be used to reject them. As per visual inspection, the curves in the case of broken assumptions appear to be systematically smaller or larger than the expected values, whereas non-systematic deviations from the expected curves align with the non-broken assumptions. Further research into the topic is still mandatory for a reliable test, yet, the possibility of such a test is promising. A similar test for the $\mathrm{LID}_{\mathrm{ABID}}$ or $\mathrm{LID}_{\mathrm{RABID}}$ estimator would require an in-depth analysis of the distribution of eigenvalues of the non-central covariance matrix of normalized samples, which is beyond the scope of this work. Yet, in the absence of a better test, $\mathrm{LID}_{\mathrm{ABID}}$ estimates over less than $\mathcal{O}(\delta^2)$ samples are likely unreliable. The same, however, does not hold for $\mathrm{LID}_{\mathrm{RABID}}$ which is much less prone to underestimating the LID due to not being regularized to be below it.

3.6 Empirical Evaluation [85]

In this section, we empirically evaluate the performance of the proposed methods. We consider several LID estimators on many standard evaluation datasets of both artificial and natural origin. As measures of quality, we analyze the estimated LID's consistency both with expected values (for synthetic data) and with each other (for natural data with no true value). We will further inspect the stability of LID estimates for varying neighborhood sizes. Depending on the density of datasets, approaches that require a large neighborhood to stabilize tend to be inapplicable.

The histograms shown in this section are limited to a region of interest in both x and y direction for improved readability. Outside of the presented range along the x-axis, the distributions always show a smooth drop to zero with no further peaks but may have a long tail.

3.6.1 Reference Estimators

We compare $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ to $\mathrm{LID}_{\mathrm{PCA}}$ estimates with a threshold value of 95% explained variance, the $\mathrm{LID}_{\mathrm{MLE}}$ estimator [54], the Measure-of-moments LID estimator

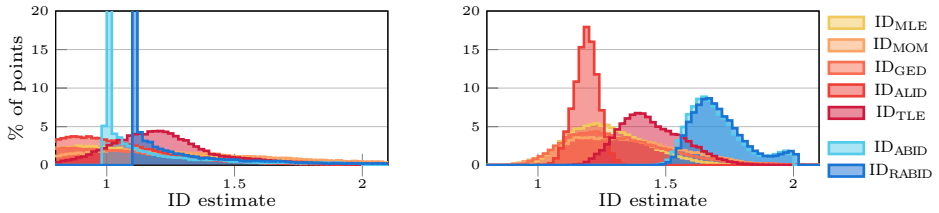


Figure 3.21: Histograms of ID estimates of points sampled from a Koch snowflake using the 10 (left) and 200 (right) nearest neighbors for ID estimate computation.

based on distances (LID_{MOM}) similar to LID_{MLE} [2], the Generalized Expansion Dimension LID estimator (LID_{GED}) [47], the Augmented LID Estimator (LID_{ALID}) [19], and its successor the Tight LID Estimator (LID_{TLE}) [5] using the implementations in the ELKI framework [76]. All of these alternative estimators are based on the expansion rate. The LID_{TLE} is supposed to reduce the necessary sample size in the neighborhood to acquire a good estimate, yet in our experiments tends to give higher estimates than the other distance-based approaches.

3.6.2 Dimensionality of Fractal Curves

In line with the theoretical foundation of this work and to demonstrate the different semantics of angle-based and distance-based LID estimation, we analyze the estimated LID of a well-known fractal, the Koch snowflake. The Koch snowflake is a popular example of fractal dimension. It can be constructed by starting with an equilateral triangle and iteratively adding an equilateral triangle three times smaller to the middle of each line segment. The resulting shape looks roughly like the outline of a snowflake, which would intuitively be considered one-dimensional, yet, the infinite small triangles on every line segment give it a certain "width". It is, hence, neither exactly one- nor two-dimensional but something in between. Structures that can not be properly assigned an integer dimension are called fractals and their dimensionality can be computed in multiple ways. The first definition was given by Hausdorff in 1918 [37] and describes the minimum dimension of a countable set of d-balls covering a given set. For simple examples like the Koch snowflake, the Hausdorff dimension can be evaluated as the logarithm of the increase in the number of line segments divided by the scale of new line segments per iteration. Each iteration multiplies the number of line segments by 4 with each line segment being 3 times shorter than the previous one resulting in a Hausdorff dimension of $\frac{\log 4}{\log 3} \approx 1.26$. As seen in Fig. 3.21, most distance-based approaches estimate a dimensionality roughly around the Hausdorff dimension of the Koch snowflake when we consider enough neighbors (k=200). This result is not surprising, as the distance-based approaches are conceptually closely related to the Hausdorff dimensions. Our angle-based estimates, however, estimate a dimensionality of ≈1.6 for larger neighborhoods, which is larger than the fractal dimension, yet smaller than the representation dimension. The difference can

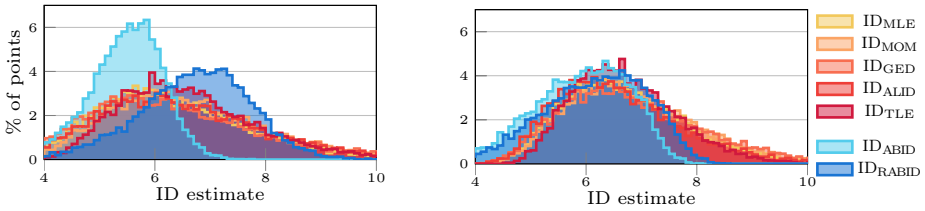


Figure 3.22: Histograms of LID estimates of the M6 set with neighborhood sizes of 30 (left) and 150 (right) points.

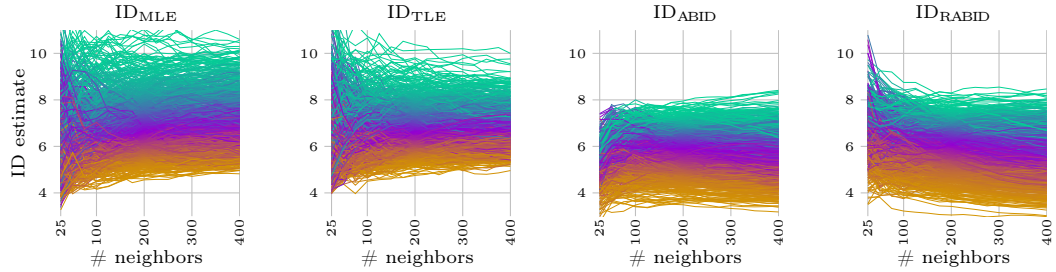


Figure 3.23: Trails of estimates of 1000 points for varying neighborhood sizes on the M6 set. Trail colors are assigned in order of LID estimates at 200 neighbors.

be explained by the highly non-linear shape of the snowflake, as two consecutive line segments are overlapping in a singularity. Whilst points sampled from a finite recursion Koch snowflake lie on a curve, they might be locally indistinguishable from samples from $\mathbb{R}^2$. A higher LID estimate may turn out to be more robust for downstream applications such as manifold learning. The results on further fractals, such as the outline of n-flakes, were similar.

It is worth mentioning that the scale of the neighborhood has a large impact on the estimates. When choosing a neighborhood small enough to mostly stay within a line segment of the fractals (here k=10), the LID estimates approximate 1, as most neighborhoods lie on straight lines. For larger neighborhoods (e.g. k=200), the estimates approach a proper representation of the manifold space. For too large neighborhoods (e.g. k=2000), however, boundaries of the point set as well as observing points distant on the manifold, yet close in the embedding space tend to corrupt the estimates.

3.6.3 Synthetic Data

Rozza et al. introduced a large collection of synthetic datasets [71] mostly based on generators initially proposed by Hein and Audibert [39] to evaluate ID estimators. The M6 dataset consists of points sampled from a 6-dimensional manifold non-linearly embedded in a 36-dimensional space. As can be seen in Fig. 3.22, for k=150 all estimators agree on the dataset to be inherently 6-dimensional at most points. Where distance-based estimators tend to have a long tail towards higher dimensions, the angle-based approaches have an

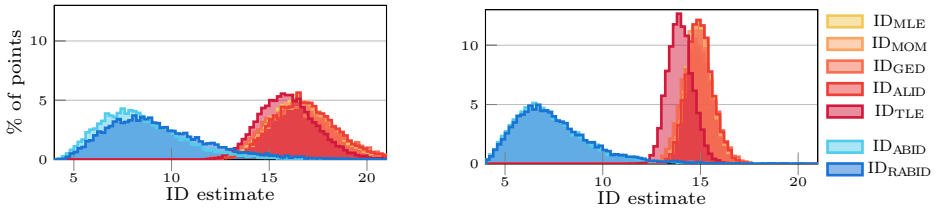


Figure 3.24: Histograms of ID estimates of the M10c data for neighborhood sizes of 100 (left) and 500 (right) points.

upper bound. Even though this non-linearity shifts the upper bound beyond 6, the angle-based approaches tend to have a shorter upper tail and drop off to zero faster. To obtain an order-based score for the stability of LID estimators, we computed both Spearman's and Pearson's correlation between the orderings of data points by their LID estimates. For that, we selected each two adjacent neighborhood sizes from our neighborhood size grid (e.g., 150 and 175) and compared the LID estimates at these neighborhood sizes. Between 25 and 400, the angle-based approaches achieve higher scores than the distance-based approaches on both Spearman's and Pearson's correlation of estimates.

In this sense, the angle-based approaches are more stable both in value as well as in the order when varying neighborhood size. In more extreme neighborhoods ($<$ 30 and $>$ 300), artifacts from having too few samples for a reliable estimate and reaching the boundaries of the dataset, respectively, cause results to become less stable. The stability is visualized in Fig. 3.23 using trails of LID estimates for individual points when varying the neighborhood size. In a perfectly stable result, all lines would be parallel; instability causes lines that cross outside their own color range (which represents the order at k=200) and hence the mixing of the colors. The improved stability of the angle-based estimates is shown by a fairly stable plot from 125 to 300 neighbors, whereas the distance-based estimates deviate much more at $\leq$150 and $\geq$250 neighbors, respectively, already. This can be seen from the more volatile mixing of trails below and above these values. Additionally, we can see in this plot that the average (the purple region) of the distance-based estimates tends to increase with growing neighborhood size, whereas the distribution of LID_{ABID} and $\text{LID}_{\text{RABID}}$ appear stable upwards of 100 neighbors. The higher stability means that smaller neighborhoods suffice for estimation and that the neighborhood parameter is easier to choose. We can further observe the upper bound property of LID_{ABID} compared to $\text{LID}_{\text{RABID}}$. The LID_{ABID} estimator approaches a stable LID distribution from below whereas the $\text{LID}_{\text{RABID}}$ estimates for smaller neighborhood sizes can overestimate the true LID, yet allow for an on-average correct estimation even in the $k < \delta^2$ regime.

The dataset of Rozza et al. [71] with the highest intrinsic dimension, M10c, is a 24-dimensional, uniformly sampled hypercube embedded in 25 dimensions. On that dataset, we observed a larger discrepancy between the angle- and the distance-based approaches, shown in Fig. 3.24. However, M10c consists of only 10000 points, which is the number

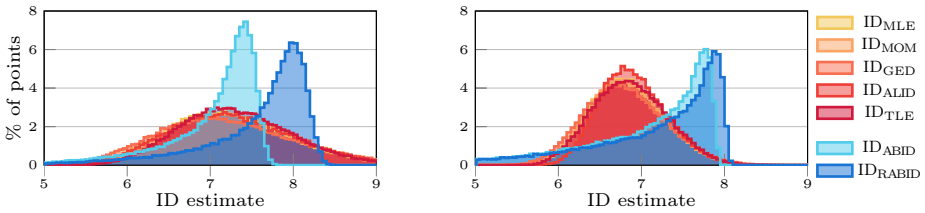


Figure 3.25: Histograms of ID estimates of points on an 8-dimensional noisy lattice using 100 (left) and 500 (right) nearest neighbors for LID estimation, respectively.

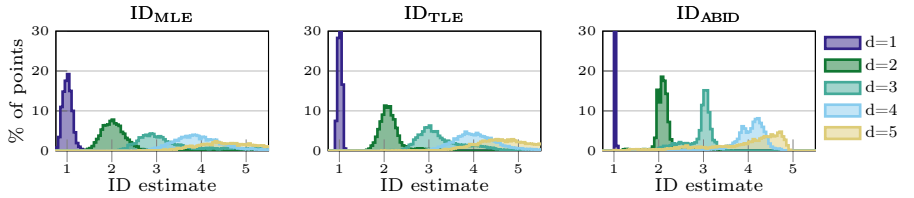


Figure 3.26: Histograms of LID estimates of nested hypercubes with a neighborhood size of 100 colored by the hypercube from which they were sampled.

of corners of a $\log_2(10000) \approx 13$ dimensional hypercube. Hence, we doubt that this small sample can reliably represent a full 24-dimensional manifold, so the data is likely of much lower dimensionality. The estimates of the distance-based approaches move towards this value as the neighborhood size increases. Each of these 10000 points then is essentially the corner of a 13-dimensional hypercube and will see the other data points as forming a hypercone, producing smaller angles than if the data evenly surrounded the point. We believe it is because of this effect (essentially a variant of the curse of dimensionality) that the angle-based approaches estimate a far lower LID. To support this theory, we created a dataset consisting of one point sampled from each cell of an 8-dimensional grid of 4^8=65536 cells. This yields a dataset consisting of 4^8 points that is more evenly distributed than uniform random sampling and truly spans an 8-dimensional space. On this dataset, only the angle-based approaches were able to estimate the correct dimension for most points as can be seen in Fig. 3.25. The points where LID_{ABID} and $\text{LID}_{\text{RABID}}$ estimate lower values are likely the *many* points at the corners, edges, and sides of this lattice.

To test the reliability of estimators in a mixture of manifolds, we created instances of 1- through 5-dimensional hypercubes linearly projected into the same 5-hypercube. The projection was chosen such that every d_i-dimensional hypercube intersects every d_j-dimensional hypercube in a $\min(d_i, d_j)$-dimensional subspace. For every hypercube, we sampled 5000 points uniformly at random and computed LID estimates using a neighborhood of different sizes. In all experiments, the angle-based approaches were visibly more capable of differentiating between the different dimensional subspaces, which can be seen from the sharper spikes in Fig. 3.26. When the dataset contains noise points independent of the generative mechanism of the data we want to analyze, these noise points can be

higher-dimensional up to the feature dimension of the data. This high-dimensional noise manifold then contains the lower-dimensional manifold of interest. Being capable of separating overlapping differently-dimensional subspaces is, hence, an important feat, and we believe this new LID estimate may help subspace discovery approaches that are based on intrinsic dimensionality (e.g., [10]). We observe that the angle-based approaches are more robust against noise and in the presence of overlapping subspaces.

3.6.4 Adversarial Datasets

To evaluate the robustness of the different estimators we further looked into the impact of non-uniform distributions in parameter space (by either noise or sampling) and by additive noise in the embedded space. To get a repeatable process that is capable of showing limit cases for point distributions, we adopted the random perturbation approach introduced by Matejka and Fitzmaurice [59] for their "Datasaurus" dataset. The basic principle is to randomly attempt to perturb a part of the dataset and check for an acceptance condition. If the new perturbation of a part is accepted, it is integrated into the dataset. This process is repeated for a fixed number of iterations or until some stopping criterion is fulfilled.

Because we aimed to generate "adversarial" datasets, where estimators misestimate the "true" intrinsic dimensionality, our acceptance condition hence was to change the mean LID estimate of the perturbed part for the selected target LID estimator (as different estimators might need different perturbation). We distinguish two cases in the following: parameter space and embedded space. In the first case, arbitrary perturbation is allowed to make the dataset as difficult as possible. Decreasing the LID estimates in this setting is too easy to be informative because the dataset can simply be squeezed to a lower-dimensional manifold; increasing is much more interesting. In embedded space, we enforce similarity to the original data by constraining the perturbation to have an absolute mean below 0.001 and the standard deviation to be at most 0.3. Here, we did not register any successful perturbations for lowering the mean LID estimate, and hence, in the following, we will only discuss increasing the LID. We started in parameter space with 10 000 points uniformly sampled from a 5-ball and in embedding space with 10 000 points uniformly sampled from a swiss roll (2d manifold in 3d) without noise. We ran 100 000 iterations on each run where each iteration was allowed to make 50 attempts at increasing the mean LID estimate by perturbing 1% of the dataset. To improve the running time of the algorithm, we only updated the LID estimates of the perturbed points upon success and recomputed all LID estimates for every 20 successful perturbations. All LID estimates were computed using 150 neighbors.

An excerpt of the results of this randomized approach is displayed in Fig. 3.27 and Fig. 3.28. The experiments have been conducted for other estimators as well, yet the re-

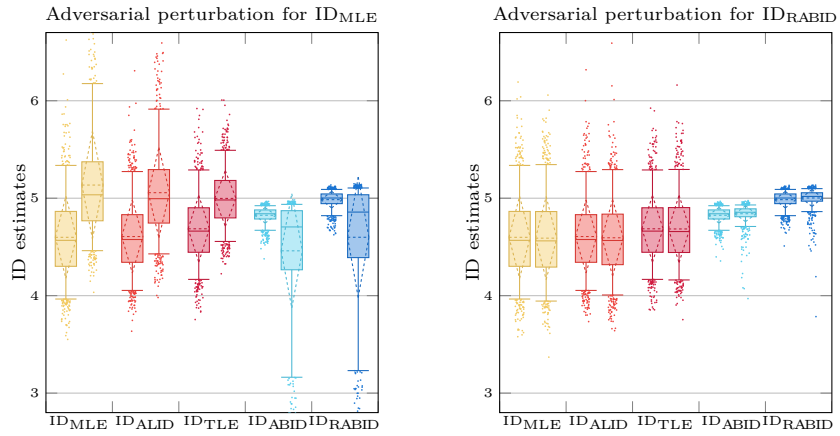


Figure 3.27: LID estimate distributions for the 5-ball (left in group) and the adversarially perturbed 5-ball (right in group) for multiple LID estimates. The dashed diamonds describe mean and standard deviation, whiskers are at 5- and 95-percentiles, and only 10% of outliers are displayed.

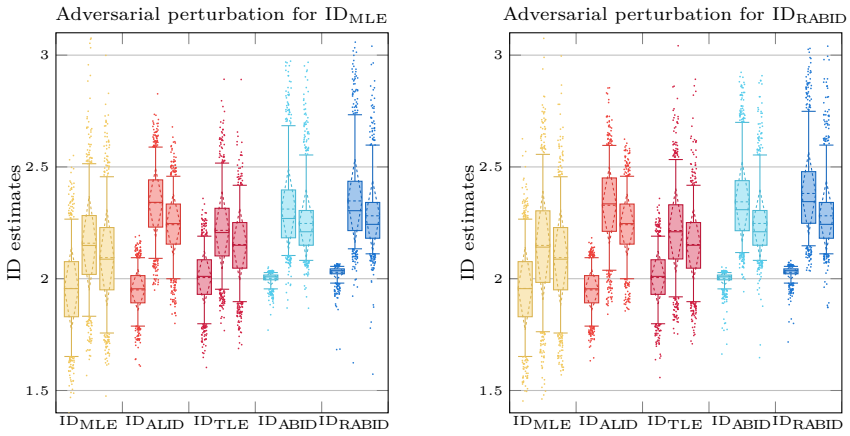


Figure 3.28: LID estimate distributions for the swiss roll (left in group), the adversarially perturbed swiss roll (middle in group), and a swiss roll with normally distributed noise (right in group) for multiple LID estimates. The dashed diamonds describe mean and standard deviation, whiskers are at 5- and 95-percentiles, and only 10% of outliers are displayed.

sults among datasets optimized for angle-based and distance-based estimators strongly correlate. In Fig. 3.27, the dataset on the left was optimized to be misestimated by LID_{MLE}, while the right plot was optimized against LID_{RABID}. For each method, we show two bars: the first for the original 5-ball, the second for the adversarially perturbed version. As the left dataset was optimized against LID_{MLE}, it is to be expected that it overestimates the most, but we can observe that the other distance-based methods are also affected. The heavily skewed and much lower estimates for the angle-based approaches suggest that these bad cases have a decreasing density towards the border as outliers have a lower LID in angle-based approaches. In the right dataset, none of the methods changes much. Even

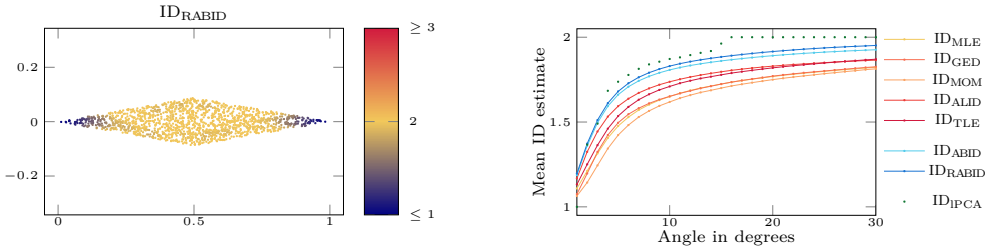


Figure 3.29: Mean LID estimates for 150 neighbors on diamond datasets with varying sharp angles. Each dataset has 1500 points uniformly sampled from a diamond with width 1 and a prescribed sharp angle. An example with an angle of 20° colored by $\text{LID}_{\text{RABID}}$ estimates is displayed on the left.

for the target estimator $\text{LID}_{\text{RABID}}$, the mean estimate was only pushed up very slightly. We presume that angle-based approaches are not systematically prone to overestimating for any distributions in the full-dimensional parameter space and the data perturbation attack did not succeed to find a suitable perturbation. Fig. 3.28 uses the embedded swiss-roll datasets. Here we added a third bar for each method for a reference dataset consisting of a swiss roll with random normal distributed noise with mean 0 and standard deviation 0.3 (without adversarial optimization). This experiment shows that difficult cases due to additive noise in the embedded space affect all types of estimators to a similar extent and a case optimized to be difficult for one method is also difficult for the others. Comparing the second and third bars of each group, we can see that the adversarial optimization produced higher LID estimates than the added random noise with the same summary statistics, demonstrating that the perturbation procedure itself succeeded at finding a difficult instance. We deduce from the similarity between the two subfigures that the same types of instances result in overestimated LIDs for both types of estimators. Unlike the first 5-ball example, the difficulty in additive noise in feature space is, hence, rather a matter of geometry than of distribution density. On real data, both effects can be present simultaneously, of course. The angle-based estimates, however, are again skewed with a tendency towards the noise-free LID, whereas the distance-based estimates are distributed mostly symmetric around the mean.

3.6.5 Experiments on Toy Datasets

We evaluated a number of toy datasets to obtain further qualitative results. Firstly, we wanted to see the transition from one- to two-dimensional estimates on different scaling approaches of a line to a surface. In these cases, the ground truth LID is ambiguous as the transition from "noisy line" to "surface" is fuzzy. The results, hence, do not assign a better or worse quality but describe the behavior of the estimators and can help interpret estimates. We created datasets of diamonds and rectangles in two dimensions with

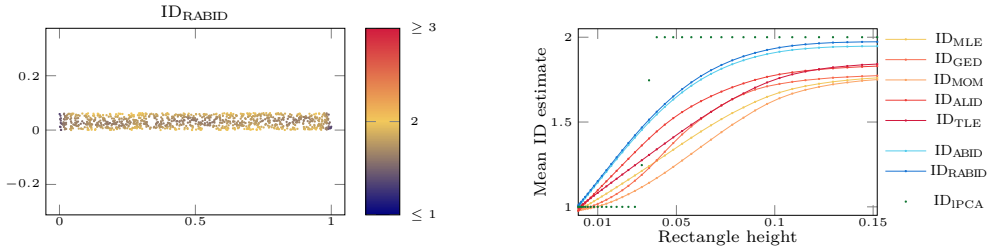


Figure 3.30: Mean LID estimates for 150 neighbors on rectangle datasets with varying heights. Each dataset has 1000 points uniformly sampled from a rectangle with width 1 and a prescribed height. An example with a height of 0.625 colored by $\text{LID}_{\text{RABID}}$ estimates is displayed on the left.

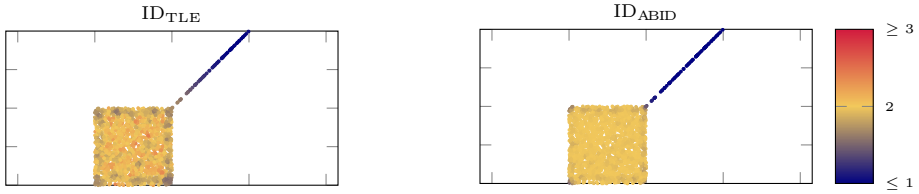


Figure 3.31: Estimates on points sampled from a square with an adjoint line starting at one of the square corners (lollipop) with a neighborhood size of 100.

varying proportions. Examples of such datasets as well as the mean LID estimates on these instances in relation to the scaling factor are shown in Fig. 3.29 and Fig. 3.30. From these experiments, we conclude that angle-based approaches are likely more sensitive to geometrically orthogonal components, as the mean LID estimates increase faster even for small scaling factors. Surprisingly, the mean estimates for angle-based approaches as well as for LID_{PCA} pass the mean ID estimate of 1.5 at about the same scaling factor in both experiments. Whether this is by chance or generalizes to different shapes and higher dimensions is yet unclear, but cannot be ruled out due to the relationship between LID_{ABID} and LID_{PCA}.

In another experiment to analyze the transition from one-dimensional to two-dimensional estimates in the same dataset rather than between different datasets, we used a square with an adjoint line (lollipop). As can be seen in Fig. 3.31, the angle-based estimates provide a more rapid transition from the two-dimensional square to the one-dimensional line. The estimates for the angle-based approach drop to about 1 right at the corner of the square, whereas the distance-based estimates and even the LID_{PCA} estimates remain high until the line dominates the neighborhood.

Lastly, we sampled random points from the Lorenz system, which is a set of three differential equations introduced by Lorenz [55]. The equations represent the spatial gradient of an object in 3d space with respect to time. By starting from some initial point and iteratively adding a small fraction of the gradient, one can approximate a trail through time that complies with the Lorenz system. It is well known that these trails converge to

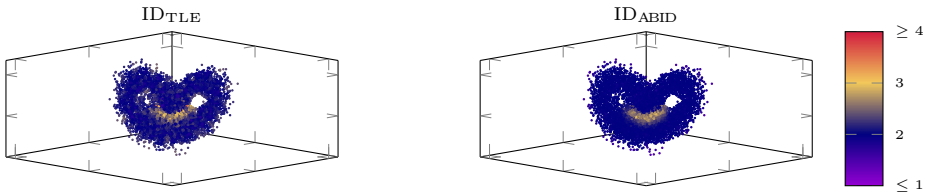


Figure 3.32: Estimates on points sampled from the Lorenz system with a neighborhood size of 100.

the so-called Lorenz attractor where the set of states over time tend towards two overlapping disks with the default parameters.[1] We, hence, started at some arbitrary position and computed a trail of 1 million points in small increments of which we chose 4000 random points as our dataset. By taking a random part of the trail samples we ensured not to have any periodic behavior in our sample due to the fixed time increments, which would yield strongly correlating neighborhoods. As ground truth on this dataset, we assume a LID of 2 for all points in the disk-shaped portions of the Lorenz attractor and 3 for the points in the intersection of these disks. The estimates on one of these samplings are displayed in Fig. 3.32. Similar to the lollipop dataset, this experiment showcases a transition between different dimensional estimates within the same dataset. As for the diamond and rectangle datasets, the exact point at which the LID should transition from one value to the other is fuzzy and has no single true answer. All estimators provide estimates in line with our fuzzy ground truth assumption with outlier points having a slightly lower LID for the angle-based approaches and slightly higher for the distance-based approaches.

The adversarial datasets suggest that the sampling density largely affects distance-based estimators, possibly leading to overestimation of the true LID. Angle-based estimators are solely prone to non-symmetric neighbor distributions that result in outlying lower LID estimates and a skewed distribution with a long lower tail. Additive noise in feature space can in turn result in higher estimates for both approaches, also giving a skewed angle-based estimate distribution, yet with a long upper tail. The disturbance of estimates by noise in feature space seems to be a result of geometric deformation affecting both approaches similarly.

On the toy datasets, we observed that angle-based estimates are more sensitive even to small orthogonal components and thereby generally "react faster" to an increased geometric deformation. This sensitivity leads to faster increasing mean LID estimates comparable to $\mathrm{LID}_{\mathrm{PCA}}$, which we also observed on the adversarial datasets in Fig. 3.28. Besides the rapid transition between different LID values, the angle-based estimates generally have a lower local variance than distance-based approaches even for rather large neighborhoods. This can be seen by the locally smooth colors in Fig. 3.31 and Fig. 3.32. Where the $\mathrm{LID}_{\mathrm{TLE}}$

[1] This behavior is dependent on the parameters of the Lorenz system. Here, we assume the parameters proposed by [55].

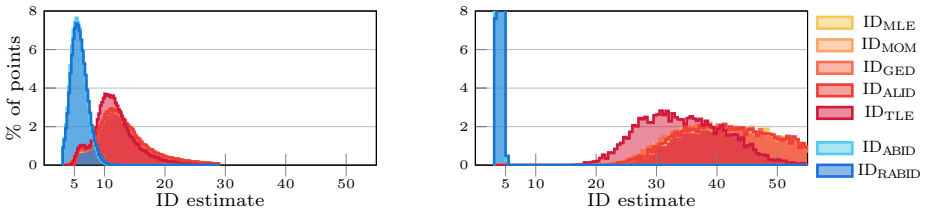


Figure 3.33: Histograms of LID estimates for MNIST and Gisette with 300 neighbors.

estimates for single points in the square in Fig. 3.31 deviate from their neighbors as much as 0.5, the local differences of LID_{ABID} do not exceed 0.2. The locally smooth angle-based estimates have been noticeable in all of our experiments and appear to be an inherent property of angle-based estimators. The combination of locally smooth estimates with a more abrupt transition between different LID values can be useful in the separation of differently dimensional subspaces which supports our results on the nested hypercube dataset displayed in Fig. 3.26.

3.6.6 Real Data

We also analyzed the LID estimates on real data such as the MNIST and Gisette dataset that have been previously used to evaluate LID estimates, e.g., by Hein and Audibert [39] and Amsaleg et al. [5]. On the MNIST dataset, both of our proposed approaches estimate a LID of about 6 for most points, whereas the distance-based approaches peak around 10 to 11. From neighborhood sizes of 100 upwards, the distance-based approaches, however, start forming a second peak at the same LID as the angle-based approaches (around a LID of 5.5), visible in Fig. 3.33. A possible explanation could be that the MNIST dataset is not uniformly random on the manifold, whereby small environments are too noisy for distance-based approaches. On the 5000-dimensional variant of this dataset, Gisette, the added noise harshly increased the estimates of the distance-based approaches. The angle-based approaches, however, estimate a slightly lower LID of about 4 for most points. We consider the smaller change of the angle-based estimates as more plausible, even though the proposed estimates for the Gisette dataset might be slightly too low, as the high-dimensional noise might have sparsened the local neighborhoods too much. Nevertheless, we observe that the angle-based approach can be more robust against noise in such a semi-real scenario.

3.6.7 Estimator Interactions

These experiments can also give some insight into the differences and interactions of the different estimators. As expected from theory, LID_{ABID} and LID_{RABID} converge towards the same value for sufficiently large neighborhood sizes. Because it is trivial to compute

both estimators at once – the estimates can even be converted in posterior if the producing neighborhood size is known –, we can use the difference of the estimates to assess the quality; if they differ much we may need larger sample sizes, if they are close the sample size should be sufficient for this dimensionality. Otherwise, they provide a realistic and an optimistic estimate of the number of required dimensions. Because the angle-based estimators appear to require fewer samples than the distance-based approaches to produce stable results (cf. Section 3.6.3 and Fig. 3.23, but also throughout the other experiments), this may also help to choose neighborhood sizes for these methods. Secondly, if the angle-based estimates are much smaller than the distance-based estimates, the dataset might not be sufficiently densely sampled for this dimensionality; if the angle-based estimates are much larger than the distance-based estimates, the embedding may be highly non-linear (as in the Koch snowflake example), or may not preserve local density.

3.7 Gradient Field LID

In this section, we will expand $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ towards a measure of the gradient field of a dataset. It primarily showcases the analytical qualities of $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ and how new estimators in other domains could be derived. The methods derived here are proof of concepts and have not been evaluated thoroughly in practice. By transferring LID estimation to the gradient field we intend to measure the local complexity of the function landscape. The gradient field of a function is most important in gradient descent and related topics such as the training of neural networks or the study of dynamic systems – if we are not to consider the prior a subset of the latter. Assuming a differentiable function f, the gradient field is defined as the set of all gradients of f at all points in the domain of f. To then investigate the local complexity of that vector field, we can limit our view to the gradients of f at all points of the neighborhood of a point. If all of these gradients are aligned in the same direction, we assume the local complexity of the function to be low. In terms of gradient descent, a low gradient field complexity would mean that the exact position of the point is not too important for the direction of the gradient, likely allowing for less numerical precision, i. e. a larger step size. If the gradients point in different directions, the local complexity is high, and even small "missteps" might lead to significantly different results. We would claim, that the $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ estimators are specifically useful for this case since we are primarily interested in the LID of directional vectors. Note, that near or at a (local) optimum the Gradient Field Local Intrinsic Dimensionality ($\mathrm{LID}_{\mathrm{GFL}}$) can even be fully dimensional (i. e. have the same number of dimensions as f

has arguments) since any direction potentially yields an equal amount of change in f. To evaluate the $\mathrm{LID}_{\mathrm{GFL}}$ using $\mathrm{LID}_{\mathrm{ABID}}$, we are thus interested in

$$\mathrm{LID}_{\mathrm{GFL}}(f,p) = \mathop{\mathbb{E}}_{x,y\in N(p)}\left[\frac{\langle \nabla f(x)-\nabla f(p), \nabla f(y)-\nabla f(p)\rangle^2}{\|\nabla f(x)-\nabla f(p)\|^2 \|\nabla f(y)-\nabla f(p)\|^2}\right]^{-1} \tag{3.62}$$

where $N(p)$ is any spherically symmetric neighborhood of p. We could evaluate $\mathrm{LID}_{\mathrm{GFL}}$ by sampling a very small neighborhood around p, yet that would require the evaluation of a significant number of gradients. Instead, we propose to approximate the gradient function by a locally linear function, i.e. a hyperplane in the tangent space of p. This allows us to reformulate the $\mathrm{LID}_{\mathrm{GFL}}$ analytically as a first-order (in the Taylor expansion of ∇f) approximation since the differences of gradients can be expressed in terms of the Hessian $H(p)$ of f evaluated at p

$$\mathrm{LID}_{\mathrm{GFL}}(f,p) \approx \mathop{\mathbb{E}}_{x,y\in N(p)}\left[\frac{\langle H(p)x, H(p)y\rangle^2}{\langle H(p)x, H(p)x\rangle \langle H(p)y, H(p)y\rangle}\right]^{-1} \tag{3.63}$$

$$= \mathop{\mathbb{E}}_{x,y\in N(p)}\left[\frac{(x^T H(p)H(p)y)^2}{(x^T H(p)H(p)x)(y^T H(p)H(p)y)}\right]^{-1} \tag{3.64}$$

Since the expected value is invariant under the distribution of $N(p)$, as long as it is spherically symmetric, we can assert a normal distribution. The expected value is then given by the sum of the squared eigenvalues of the covariance matrix of normalized samples drawn from $\mathcal{N}(0, H(p)H(p))$

$$\mathrm{LID}_{\mathrm{GFL}}(f,p) \approx \left(\sum_i \lambda_i\left(\widetilde{\Sigma}_{H(p)H(p)}\right)^2\right)^{-1} \tag{3.65}$$

where $\widetilde{\Sigma}_{H(p)H(p)}$ is the covariance matrix of normalized samples drawn from $\mathcal{N}(0, H(p)H(p))$ and λ_i denotes the i-th eigenvalue of a matrix. Using an approximation that will be detailed in Chapter 4, we can approximate the eigenvalues of $\widetilde{\Sigma}_{H(p)H(p)}$ by the eigenvalues of $H(p)H(p)$ as

$$\lambda_i\left(\widetilde{\Sigma}_{H(p)H(p)}\right) \propto \lambda_i\left(H(p)H(p)\right)^{\frac{d}{d+2}} = \lambda_i\left(H(p)\right)^{\frac{2d}{d+2}} \tag{3.66}$$

with the additional constraint, that the sum of eigenvalues after normalization is equal to 1. That gives a simplified expression for the $\mathrm{LID}_{\mathrm{GFL}}$ as

$$\mathrm{LID}_{\mathrm{GFL}}(f,p) \approx \left(\sum_i \left(\frac{\lambda_i\left(H(p)\right)^{\frac{2d}{d+2}}}{\sum_j \lambda_j\left(H(p)\right)^{\frac{2d}{d+2}}}\right)^2\right)^{-1}. \tag{3.67}$$

That is, we can evaluate the $\mathrm{LID}_{\mathrm{GFL}}$ of a function f by evaluating the eigenvalues of the Hessian matrix. Computing both the Hessian matrix and its eigenvalues comes at a significant cost, yet it is likely cheaper than evaluating the $\mathrm{LID}_{\mathrm{GFL}}$ by sampling. This approximation

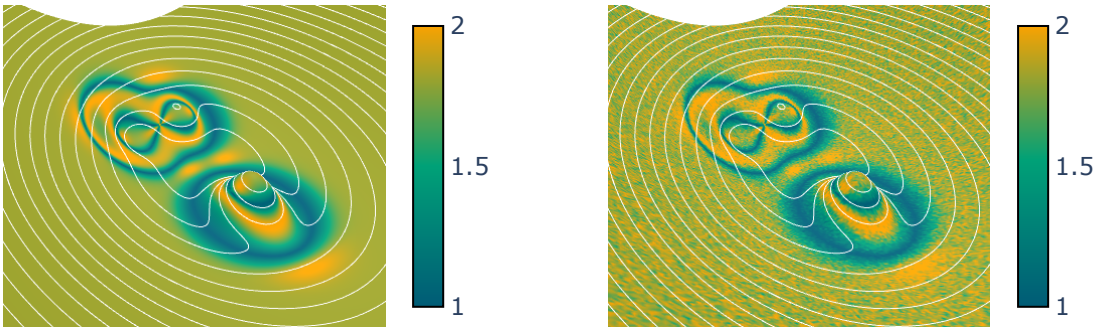


Figure 3.34: Comparison of the analytical approximation of the $\mathrm{LID}_{\mathrm{GFL}}$ of a function f using the Hessian (left) compared to a sampling-based estimate using a spherical neighborhood with radius 10^{-8} (right). The x and y coordinates are in parameter space of f and the z axis in the 3d plot corresponds to $f(x, y)$.

is more accurate for higher dimensions but gives good results for lower dimensions as well, as can be seen in Fig. 3.34. The approximation allows for a smooth evaluation of the $\mathrm{LID}_{\mathrm{GFL}}$ over the landscape whose error in value is made up for by not inducing the error from sampling the local gradient distribution.

Fig. 3.35 displays the resulting $\mathrm{LID}_{\mathrm{GFL}}$ estimates over the landscape of a function f. The estimates cover the local geometric complexity, where locally mostly flat surfaces are estimated to have a low $\mathrm{LID}_{\mathrm{GFL}}$ and locally strongly curved surfaces are estimated to have a high $\mathrm{LID}_{\mathrm{GFL}}$. However, since we do not solely consider the direction of the gradients, but also their magnitude, the estimates are not solely sensitive to converging and diverging gradient flows, but also to "acceleration and deceleration" of the gradient flow as induced by curvature on the landscape. While being functionally interesting, this makes it less useful for the case of controlling gradient descent, where the step size should be controlled to be lower in regions of diverging gradient flow and higher in regions of parallel gradient flow. To accommodate for that, we propose the $\mathrm{LID}_{\mathrm{NGFL}}$ as a normalized version of the $\mathrm{LID}_{\mathrm{GFL}}$ that normalizes the gradients in the neighborhood of p. The $\mathrm{LID}_{\mathrm{NGFL}}$ is then equivalent to (3.62) except that the gradients are normalized. As a side effect, the choice of the neighborhood distribution now matters, since the difference in normalized gradients is not invariant under scaling. For infinitesimal neighborhoods, these differences vanish almost everywhere, so a proper and truly larger-than-zero choice for the neighborhood radius is recommended. As a viable starting point for e. g. gradient descent, we propose to use a uniform spherical distribution around p with a radius equal to the norm of the gradient at p. In the context of gradient fields, we expect the gradient norm to be a naturally interesting scale of the neighborhood. Fig. 3.35 displays the resulting $\mathrm{LID}_{\mathrm{NGFL}}$ estimates over the landscape of a function f, where either a spherical neighborhood or a normal distribution along the gradient direction, i. e. points of the form $p + a\nabla f(p)$ with $a \sim \mathcal{N}(0, \|\nabla f(p)\|)$, was used. We can see, that the $\mathrm{LID}_{\mathrm{NGFL}}$ estimates are sensitive to di-

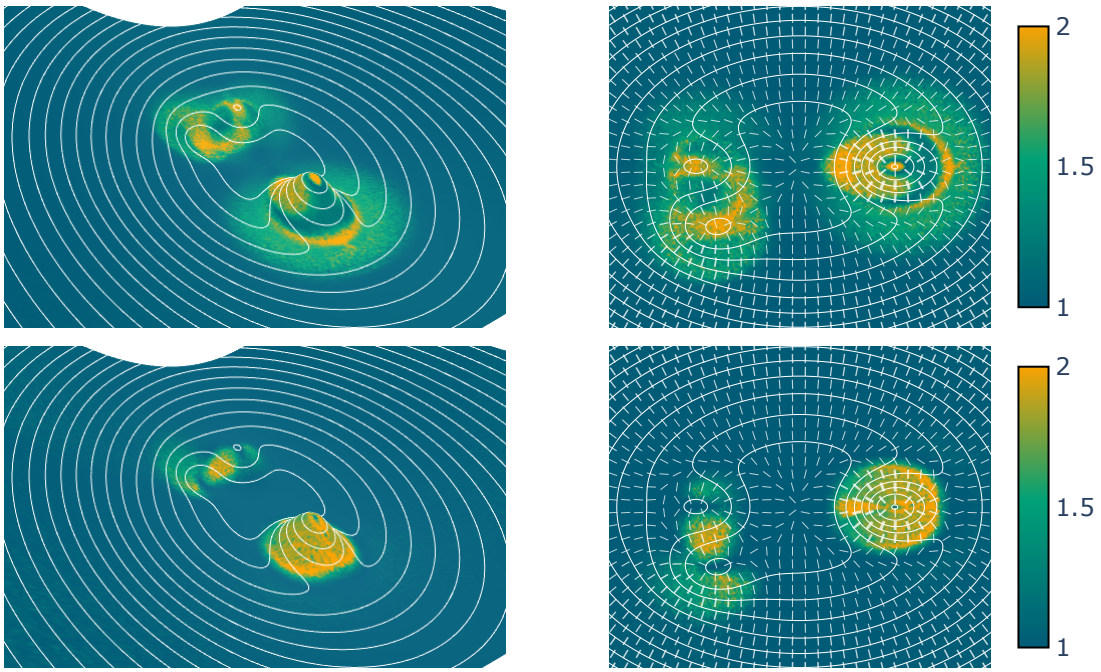


Figure 3.35: Estimates of the $\mathrm{LID}_{\mathrm{NGFL}}$ of a function f over its landscape. The $\mathrm{LID}_{\mathrm{NGFL}}$ estimates in the top row were computed on a spherical neighborhood with a radius equal to the norm of the gradient at the respective point. The $\mathrm{LID}_{\mathrm{NGFL}}$ estimates in the bottom row were computed on a normal line distribution in the direction of the gradient with a standard deviation equal to the norm of the gradient at the respective point. The x and y coordinates are in parameter space of f and the z axis in the 3d plot corresponds to $f(x, y)$. The right plot indicates the gradient field of f with overlaid contours of $f(x, y)$. The color of the surfaces is given by the $\mathrm{LID}_{\mathrm{NGFL}}$.

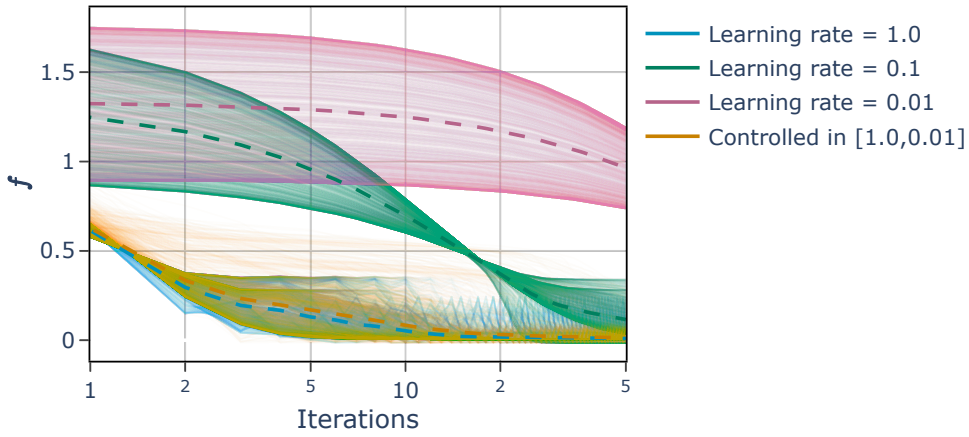


Figure 3.36: Gradient descent on the function displayed in Fig. 3.34 using either fixed learning rates or a learning rate controlled by the $\mathrm{LID}_{\mathrm{NGFL}}$ estimator. To translate a LID estimate into a learning rate, we used exponential interpolation between the maximum learning rate 1.0 for $\mathrm{LID}_{\mathrm{NGFL}}$=1 and the minimum learning rate 0.01 for $\mathrm{LID}_{\mathrm{NGFL}}$=2. Each trace displays the value of f along the number of iterations of the gradient descent and the dashed lines display the mean value of f over all traces of each group. Each run started at a random position on the outskirts of the landscape.

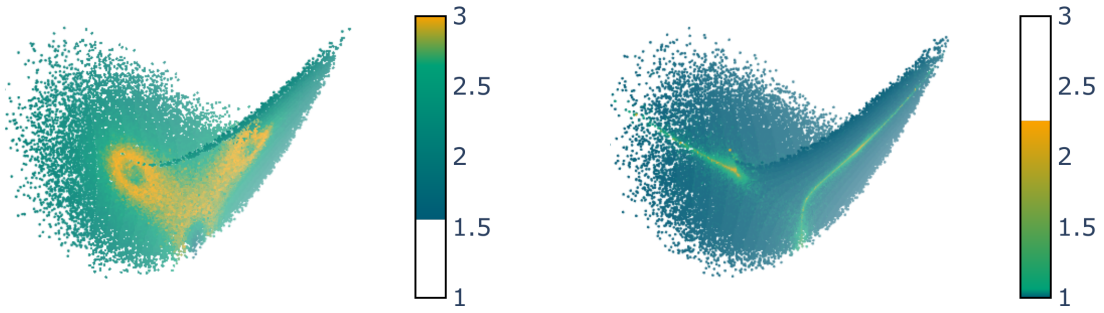


Figure 3.37: Estimates of the $\mathrm{LID}_{\mathrm{GFL}}$ and $\mathrm{LID}_{\mathrm{NGFL}}$ of a function f evaluated on the Lorenz system. The $\mathrm{LID}_{\mathrm{GFL}}$ estimates were computed on a spherical neighborhood with a radius equal to the norm of the gradient at the respective point. The $\mathrm{LID}_{\mathrm{NGFL}}$ estimates were computed on a normal line distribution in the direction of the gradient with a standard deviation equal to the norm of the gradient at the respective point. Shading was added to the color values to indicate distance from the camera for improved spatial perception.

verging and converging paths, in principle allowing to control the step size of gradient descent. The spherical distribution focuses on the entire gradient field whereas the normal line distribution is focused on single gradient de-/ascents. For functional analysis, the prior is likely more useful while for gradient descent the latter is more interesting since it allows to recognize areas, where the step size has to be diminished to not "overshoot" a de-/ascending trace. Unfortunately, the definition of $\mathrm{LID}_{\mathrm{NGFL}}$ does not allow for an analytically reduced expression, making sampling the only viable option to evaluate it for now. To confirm that the $\mathrm{LID}_{\mathrm{NGFL}}$ can be used for controlling gradient descent, we performed a simple experiment whose results are displayed in Fig. 3.36. As can be seen, the controlled learning rate based on the $\mathrm{LID}_{\mathrm{NGFL}}$ estimator on average allows for a convergence with similar performance to the maximum learning rate but without the risk of non-converging oscillation near the optimum, which occurs for a static maximum learning rate. In practice, heuristics for learning rate scheduling are often used, to prevent oscillations like these. While it is clear, that reducing the learning rate allows one to fall "deeper" into local minima, the proper choice of the schedule is often a matter of trial and error. The $\mathrm{LID}_{\mathrm{NGFL}}$ estimator could potentially be used to derive a more principled approach in the future, for now unbeknownst whether the observed results generalize well to other functions and higher dimensional spaces. In this work, we are satisfied with the proof of concept that the $\mathrm{LID}_{\mathrm{NGFL}}$ can be used to control gradient descent in a meaningful way and leave the further exploration of this concept to future research.

The $\mathrm{LID}_{\mathrm{GFL}}$ concept is of course also applicable to other vector fields, such as dynamic systems. Fig. 3.37 displays the $\mathrm{LID}_{\mathrm{GFL}}$ and $\mathrm{LID}_{\mathrm{NGFL}}$ estimates evaluated at positions of orbits in the Lorenz system, previously mentioned in Section 3.6. The $\mathrm{LID}_{\mathrm{GFL}}$ estimates computed by sampling from a spherical distribution around the respective point, assign

higher values to the regions of the attractor where the flow is more chaotic while also being sensitive to the curvature of the flow. That produces an average LID estimate on the attractor of around 2, with higher values of up to 3 at the most chaotic or curved regions. For the $\mathrm{LID}_{\mathrm{NGFL}}$ estimates, we sampled from a normal distribution along the gradient direction. The intuition is here to evaluate the local complexity of individual orbits rather than the attractor as a whole. In the result, most LID estimates are around 1, indicating that the orbits are almost insensitive to small perturbations in the direction of the gradient, i. e. the time domain of the dynamic system. Solely the chaotic regions of the attractor show higher LID estimates of around 2, indicating that the system is sensitive to perturbations at this point of the orbit.

The results displayed in this section show, that adaptations of $\mathrm{LID}_{\mathrm{ABID}}$ can potentially yield interesting insights into vector fields, whether they are gradient fields of functions or flow fields of dynamic systems. The results also extend to other vector fields, such as fluid dynamics. The proposed variants ($\mathrm{LID}_{\mathrm{GFL}}$ and $\mathrm{LID}_{\mathrm{NGFL}}$, sphere and normal line neighborhood, scale of neighborhoods) can give vastly different results and are not yet fully understood, but show potential for future research. For analytical work, especially the reduced approximation for $\mathrm{LID}_{\mathrm{GFL}}$ is a promising starting point for further investigations. The discussion in this section primarily focused on the concept of LID can be extended to other domains beyond point clouds, such as functional analysis, and how the analytical qualities of $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ can be used to derive new estimators in these fields. The effectiveness of the proposed estimators for practical applications is yet to be evaluated. Practically, an efficient realization of the estimators remains an open problem.

3.8 Conclusion

To conclude this chapter, we have introduced $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ as novel methods to estimate the LID of a dataset using angles. This approach is intimately related to $\mathrm{LID}_{\mathrm{MLE}}$, for which it is an efficient extension of the method over a bounded distribution, the uniform unit sphere, and by extension to the family of expansion-based estimators. Compared to $\mathrm{LID}_{\mathrm{MLE}}$ the $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ estimators further provide a squared amount of values per neighborhood compared to the linear amount of distances in $\mathrm{LID}_{\mathrm{MLE}}$ and related estimators. That does not decrease the asymptotic convergence rate, which requires a number of neighbors squared in the number of latent dimensions for all estimators, but our empirical studies allowed for faster-stabilizing estimates. $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ are also related to $\mathrm{LID}_{\mathrm{PCA}}$ since we can express the estimators in terms of eigenvalues of the covariance matrix of normalized data. Investigations on the distribution of random matrices suggest that $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ are an appropriate fractional extension of $\mathrm{LID}_{\mathrm{PCA}}$ to normalized samples which also allows to drop the manual choice of threshold values. Since arbitrary threshold values can result in individually arbitrary LID estimates, this is

a significant advantage. Lastly, the $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ estimators are related to other angle-based estimators such as $\mathrm{LID}_{\mathrm{DANCo}}$, $\mathrm{LID}_{\mathrm{ANOVA}}$, and $\mathrm{LID}_{\mathrm{FCI}}$. Especially the relation to $\mathrm{LID}_{\mathrm{FCI}}$ is important, since it attempts to compute the same quantity as $\mathrm{LID}_{\mathrm{RABID}}$ but in a computationally much less efficient way. As a consequence, the $\mathrm{LID}_{\mathrm{FCI}}$ estimator does not lend itself to further analytical investigations and is practically and theoretically of less use. Specifically the efficiency of $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ combined with the fractional nature and analytical qualities elevate these estimators above other angle-based estimators. These are significantly too complex, inaccurate, or computationally expensive to be used in practice or for further analytical investigations such as the ones we provided in this chapter. Most notably, the extension of $\mathrm{LID}_{\mathrm{ABID}}$ to $\mathrm{LID}_{\mathrm{GFL}}$ and $\mathrm{LID}_{\mathrm{NGFL}}$ – measures to evaluate the local complexity of a vector field – are not as easily derived for other LID estimators and may provide new insights into gradient descent and dynamic system analysis.

Chapter 4

Random In-Data Projections and LID

In this chapter, we will discuss our second contribution to the field of LID estimation, which is the introduction of Thresholded Random In-distribution Projections Intrinsic Dimensionality (LID_{TRIP}) [87]. Since this chapter is closely oriented along the paper, "On Projections to Linear Subspaces" [87], larger portions are directly adopted from the paper, and corresponding sections are annotated accordingly. The initial motivation for this estimator was to generalize the notion of "angle" to a measure on more than two points, and thus provide a larger number of values per neighborhood, stabilizing the estimation at smaller neighborhood sizes. A second motivation was to investigate the "boundaries" of Euclidean space and search for sufficient inequalities that help identify Euclidean spaces. The proposed angle generalization can in short be described as the minimum angle between a point and a linear subspace, i.e. a non-affine hyperplane, spanned by k other points. The number of possible angle computations then grows exponentially with k. That measure is related to the truncated Euclidean distance on linear subspaces sampled randomly from the data distribution and provides inequalities over k points for both inner products and Euclidean distances. In the limit of $k \to \infty$ (or at least to the ID of the dataset), this implies Euclideanness of the data – albeit to little surprise since it induces a parameterization of the vector space. Regarding the intended goals of this work, both motivations were thus mostly successful. Yet, in an immediate application, the estimated LID also grows with k, yielding the same "self-fulfilling prophecy" as the threshold choice in LID_{PCA}. To compensate for that effect, we proposed to parameterize the estimator with an acceptable loss in variance, which e.g. in similarity search arises naturally due to the discrete nature of the data. This allows us to find the balancing point between the number of points and the computational cost of Euclidean distance boundary calculations. In an empirical study, we could thus deduce the number of dimensions for which a truncated Euclidean distance is a "good enough" approximation to the true Euclidean distance in similarity search. The quality of LID_{TRIP} as a LID estimator was not derived directly, yet the LID_{ABID} estimator is a special case of LID_{TRIP}. Rather than a general LID estimator, we

consider LID_{TRIP} a practically oriented parameter estimator for the truncated Euclidean distance in applications with a known tolerance on the distance accuracy, i. e. a task-aware LID estimator. In the pursuit of practical applications of LID theory beyond plain dimensionality estimation, the approach of task-aware LID was a new concept complementing the LID-aware algorithms [72, 50]. The derivation of LID_{TRIP} further provides approximate expected values for explained variance in projections to random subspaces of the data distribution, similar to PCA, and an approximation for the covariance matrix of data after vector normalization given the covariance matrix before vector normalization. These results are not only of interest to LID estimation but also to the understanding of data distributions in Euclidean space in general.

We denote the i-th eigenvalue of some matrix M with $\lambda_i(M)$ and the corresponding eigenvector with $v_i(M)$. Whenever the matrix is clear from the context, we will omit the argument and simply write λ_i and v_i. In this chapter, we do not care about the specific order of eigenvalues but assume that the corresponding eigenvalues of matrices that admit the same eigenvectors are in the same order. We write M^c as an abbreviation for the spectral power $V(M)\,\Lambda(M)^{\circ c}\,V(M)^T$ where $V(M)$ is the matrix containing the eigenvectors of M as columns and $\Lambda(M)^{\circ c}$ is the diagonal matrix containing λ_i^c on the diagonal. We write $C(X)$ for the covariance matrix of datasets X where we assume X to be origin-centered unless otherwise specified. We denote the normalizations of vectors x and datasets X with $\tilde{x}$ and $\widetilde{X}$, respectively. Whenever Euclidean spaces and distances are discussed, the dot product is implied by the inner product.

The chapter is structured as follows: In Section 4.1, we will derive the bounds for the Euclidean distance and inner products between a point and a linear subspace spanned by k other points. These bounds are a generalization of the triangle inequality and the triangle inequality for cosine similarity [74]. In Section 4.2, we will investigate the expected values of the variance after projection to the k pivots used in the bounds. This analysis allows us to estimate the bound quality of a variable number of pivots. Section 4.3 covers the derivation of the LID_{TRIP} estimator and its parameterization by starting from the special case LID_{ABID}. In Section 4.4, we will evaluate how the LID_{TRIP} estimator can be used to predict the number of pivots required for a given task. Finally, in Section 4.5, we will summarize the findings in the context of this thesis.

4.1 Pivotal Bounds In Euclidean Spaces [87]

We consider linear subspace projections of *query points* onto the linear subspace spanned by (not necessarily orthogonal) *pivots* or *reference points* $\{r_1, \dots, r_k\}$, $k \leq d$ drawn from the same distribution as the analyzed dataset, e.g., by choosing them from the dataset itself. In the case of affine subspace projections, both the query and reference points are shifted by a *center* point c. We assume all (shifted) reference points to be linearly in-

dependent. Otherwise, we discard reference points until linear independence holds. The projection $\pi(x-c; r_1-c, \dots, r_k-c)$ of some shifted query point $x-c$ onto the affine subspace (shortened to $\pi(x-c)$ whenever the choice of reference points is clear) is then given by

$$\pi(x-c) = \sum_{i=1}^{k} \langle x-c, \hat{r}_i \rangle \hat{r}_i \tag{4.1}$$

where the $\hat{r}_i$ are the normalized orthogonal vectors obtained from the Gram-Schmidt process applied to the r_i-c. These can be recursively computed from

$$\hat{r}_1 = \frac{r_1-c}{\|r_1-c\|} \qquad \hat{r}_i = \frac{(r_i-c) - \sum_{j=1}^{i-1} \langle r_i-c, \hat{r}_j \rangle \hat{r}_j}{\left\|(r_i-c) - \sum_{j=1}^{i-1} \langle r_i-c, \hat{r}_j \rangle \hat{r}_j\right\|} \tag{4.2}$$

where $\|x\|$ is shorthand for $\langle x,x\rangle^{1/2}$. In the following, we will repeatedly require the evaluation of $\langle \cdot, \hat{r}_i \rangle$ and $\|\pi(\cdot;\cdot)\|$. Although (4.1) and (4.2) can be evaluated explicitly every time, it can be more convenient to represent the (squared) norm after projection in terms of inner products (especially in kernel spaces):

$$\|\pi(x-c)\|^2 = \sum_{i=1}^{k} \langle x-c, \hat{r}_i \rangle^2 \tag{4.3}$$

since all $\hat{r}_i$ are normalized and pairwise orthogonal. We can reduce $\langle \cdot, \hat{r}_i \rangle$ to

$$\langle x-c, \hat{r}_i \rangle = \frac{\langle c,c\rangle - \langle c,x\rangle - \langle c,r_i\rangle + \langle x,r_i\rangle - \sum_{j=1}^{i-1} \langle x-c,\hat{r}_j\rangle \langle r_i-c,\hat{r}_j\rangle}{\left(\langle c,c\rangle - 2\langle c,r_i\rangle + \langle r_i,r_i\rangle - \sum_{j=1}^{i-1} \langle r_i-c,\hat{r}_j\rangle^2\right)^{1/2}} \tag{4.4}$$

which can also be used recursively to compute the $\langle r_i-c, \hat{r}_j \rangle$ in (4.4). In the non-affine case, $c = \mathbf{0}$, (4.4) simplifies to

$$\langle x, \hat{r}_i \rangle = \frac{\langle x,r_i\rangle - \sum_{j=1}^{i-1} \langle x,\hat{r}_j\rangle \langle r_i,\hat{r}_j\rangle}{\left(\langle r_i,r_i\rangle - \sum_{j=1}^{i-1} \langle r_i,\hat{r}_j\rangle^2\right)^{1/2}} \tag{4.5}$$

Note that the denominator and parts of the nominator need to be computed just once. Further, we omit the explicit computation of any $\hat{r}_i$ which would be infeasible in, e.g., RBF kernel and general inner product spaces. With dynamic programming, $\|\pi(x-c)\|^2$ can be computed in $\Theta(pk^2)$ time, where p is the effort required to compute an inner product.

In spatial indexing, pivots have been successfully used to bound distances via the triangle inequality [61, 65]. We propose to bound distances in terms of a decomposition of the squared Euclidean norm into dot products given by

$$d_{Euc}(x,y)^2 = \|x-y\|^2 = \langle x-y, x-y\rangle = \langle x,x\rangle + \langle y,y\rangle - 2\langle x,y\rangle \tag{4.6}$$

From this we can derive bounds for the Euclidean distance between two points given a bound on the dot product $\langle x,y\rangle$, assuming $\langle x,x\rangle$ and $\langle y,y\rangle$ are known. Let $\hat{r}_1, \dots, \hat{r}_k$

be pivot points previously orthogonalized by the Gram-Schmidt process as defined in Section 4.2. We can decompose $x-c$ and $y-c$ into k components aligned along the $\hat{r}_i$ and one orthogonal remainder. We will call this $(k+1)$-th component $x_\perp$ and $y_\perp$, respectively. It then follows that

$$\langle x-c, y-c\rangle = \langle x_\perp, y_\perp\rangle + \sum_{i=1}^{k} \langle \langle x-c, \hat{r}_i\rangle \hat{r}_i, \langle y-c, \hat{r}_i\rangle \hat{r}_i\rangle \tag{4.7}$$

Because the $\hat{r}_i$ are pairwise orthogonal, this decomposition is uniquely defined. Since all $\hat{r}_i$ have a unit norm, we can rewrite this equation to

$$\langle x, y\rangle = \langle x_\perp, y_\perp\rangle + \langle c, x\rangle + \langle c, y\rangle - \langle c, c\rangle + \sum_{i=1}^{k} \langle x-c, \hat{r}_i\rangle \langle y-c, \hat{r}_i\rangle \tag{4.8}$$

All of the terms on the right-hand side then either depend on x or y, but not on both, except for $\langle x_\perp, y_\perp\rangle$. In the semantics of Euclidean spaces, both $x_\perp$ and $y_\perp$ lie in the same $(d-k)$-dimensional linear subspace. We can compute both as

$$x_\perp = (x-c) - \pi(x-c) \quad \text{and} \quad y_\perp = (y-c) - \pi(y-c) \tag{4.9}$$

but do not know their relative orientation. Yet, we can bound their inner product using the Cauchy-Schwarz inequality resulting in the bounds $\pm\left(\langle x_\perp, x_\perp\rangle \cdot \langle y_\perp, y_\perp\rangle\right)^{1/2}$. By orthogonality of $x_\perp$ and $\pi(x-c)$ we know $\|x_\perp\|^2 = \|x-c\|^2 - \|\pi(x-c)\|^2$. The bounds for the inner product $\langle x-c, y-c\rangle$ then follow as

$$\begin{aligned} &\langle c, x\rangle + \langle c, y\rangle - \langle c, c\rangle + \textstyle\sum_{i=1}^{k} \langle x-c, \hat{r}_i\rangle \langle y-c, \hat{r}_i\rangle \\ &\pm \left(\begin{array}{l} \left(\langle x, x\rangle + \langle c, c\rangle - 2\langle c, x\rangle - \textstyle\sum_{i=1}^{k} \langle x-c, \hat{r}_i\rangle^2\right) \\ \cdot \left(\langle y, y\rangle + \langle c, c\rangle - 2\langle c, y\rangle - \textstyle\sum_{i=1}^{k} \langle y-c, \hat{r}_i\rangle^2\right) \end{array} \right)^{1/2} \end{aligned} \tag{4.10}$$

which in the non-affine case, $c = \mathbf{0}$, becomes

$$\sum_{i=1}^{k} \langle x, \hat{r}_i\rangle \langle y, \hat{r}_i\rangle \pm \left(\left(\langle x, x\rangle - \sum_{i=1}^{k} \langle x, \hat{r}_i\rangle^2 \right) \cdot \left(\langle y, y\rangle - \sum_{i=1}^{k} \langle y, \hat{r}_i\rangle^2 \right) \right)^{1/2} \tag{4.11}$$

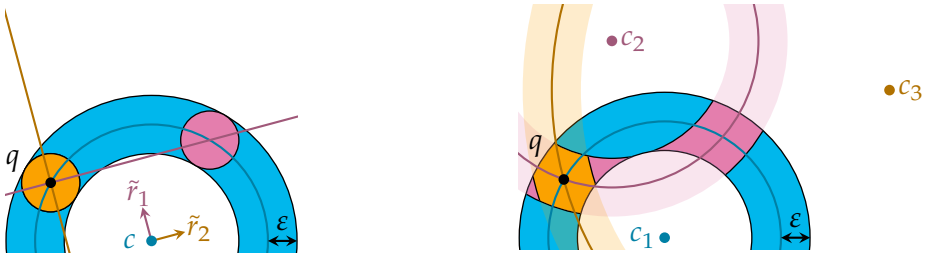
Inserting both of these values into (4.6) gives bounds on the squared Euclidean distance and, consequentially, on the Euclidean distance. These bounds are a generalization of at least two bounds known from the literature. When we assume the affine case and $k=0$ pivots, the bounds derived from (4.6) and (4.11) reduce to

$$\langle x, x\rangle + \langle y, y\rangle - 2\langle c, x\rangle - 2\langle c, y\rangle + 2\langle c, c\rangle \pm 2\|x-c\|\left\|y-c\right\| \tag{4.12}$$

$$= \left(\|x-c\| \pm \left\|y-c\right\|\right)^2 \tag{4.13}$$

which are the bounds easily derivable from the triangle inequality. For the non-affine case with $k=1$ pivots and normalized x and y, the inner product bounds (4.11) reduce to

$$\langle x, \hat{r}_1\rangle \langle y, \hat{r}_1\rangle \pm \left(\left(1 - \langle x, \hat{r}_1\rangle^2\right) \left(1 - \langle y, \hat{r}_1\rangle^2\right) \right)^{1/2} \tag{4.14}$$



(**a**) Combined center and pivot (**b**) Intersection of triangle inequality

Figure 4.1: Eligible search spaces around a query point q after filtering with the lower bounds obtained from one, two, or three centers and/or pivots.

which is the triangle inequality for cosines introduced in [74]. Triangle-inequality-based bounds have been used in spatial indexing in methods like, e.g., LAESA [61]. For multiple pivots, these approaches take the minimum or maximum of the bounds obtained separately for each pivot. In our terminology, we refer to such pivots as centers c. Those are fundamentally different from the term pivots introduced here: When performing an ε-range query for a query point y, the eligible search space for vectors x according to the upper bound in (4.13) is a hyperspherical shell centered at c. This geometric shape can be described as the sumset (the set of all sums of pairs in the cartesian product) of a $(d{-}1)$-sphere of radius $\|y{-}c\|$ centered at c and a d-ball of radius ε. When using pivots as per our definition, each pivot induces a hyperplane orthogonal to the $\hat{r}_i$ which intersects with the hypersphere. Consequentially, the resulting eligible search space is the sumset of a $(d{-}1{-}k)$-sphere of radius $(\|y{-}c\|^2 - \|\pi(y{-}c)\|^2)^{1/2}$ and a d-ball of radius ε. This is illustrated in two dimensions in Fig. 4.1. Each of the pivots eliminates an entire dimension from the sphere-part of the search space whereas the minimum lower bounds obtained from multiple centers produce an intersection of multiple hyperspherical shells. While $d{-}1$ pivots can reduce the search space to the sumset of at most 2 points and an ε-ball, the intersection of even d hyperspherical shells in the best case produces a volume that can be roughly described as a distorted hypercube with an "edge length" of about 2ε. The resulting volume can be exponentially larger in d than the search volume using $d{-}1$ pivots. As the volumes of regular shapes in Euclidean space expand exponentially in dimensions, one would expect an approximately exponential reduction in search space over an increasing number of pivots, whereas using the minimum upper bound over multiple centers does not induce such a reduction in search space volume. It is, therefore, of little surprise that the cosine bounds introduced in [74] ($k{=}1$), produced tighter bounds empirically than the triangle inequality ($k{=}0$), and were successfully applied to improve the performance of spherical k-means clustering [75]. Qualitatively, there is a clear argument for using a larger amount of pivots. However, the reduction in search space comes at the price of increased computational cost as the evaluation of $\langle y, \hat{r}_i \rangle$ is quadratic and

the evaluation of the bounds is linear in k. Blindly increasing k is not universally advantageous for the computational cost of spatial indexing queries or other algorithms. But how many pivots tighten the bounds enough to counterweigh the overhead? More precisely, how much more of a point's squared norm does the k-th randomly drawn pivot cover on average? Although the answer does not refer to an optimal pivot choice, by arguing over expectations of underlying distributions, this conservative argument likely holds for previously unknown query points.

4.2 Expected Variance Of Random Projections [87]

The analysis of squared norms after projection is closely related to spectral analysis. If we chose any normalized vector v, $\mathbb{E}_{x\in X}\left[\left\|\pi(x-\mathbb{E}_{y\in X}[y];v)\right\|^2\right]$ is simply the variance of X in direction v. Consequentially, for any pair of a normalized eigenvector $v_i(C(X))$ and its corresponding eigenvalue $\lambda_i(C(X))$, we know that $\mathbb{E}_{x\in X}\left[\|\pi(x;v_i(C(X)))\|^2\right]=\lambda_i(C(X))$ for any origin-centered X. By orthogonality of the eigenvectors, this argument can be extended to any number of eigenvectors $v_1(C(X)),\dots,v_n(C(X))$ as

$$\mathop{\mathbb{E}}_{x\in X}\left[\|\pi(x;v_1(C(X)),\dots,v_n(C(X)))\|^2\right]=\sum_{i=1}^{n}\lambda_i(C(X)) \tag{4.15}$$

Pearson [66] showed that the eigenvectors of the covariance matrix are precisely the maximizers of this term, i. e. they are the solution to

$$\mathop{\arg\max}_{v_1(C(X)),\dots,v_n(C(X))}\ \mathop{\mathbb{E}}_{x\in X}\left[\|\pi(x;v_1(C(X)),\dots,v_n(C(X)))\|^2\right] \tag{4.16}$$

If one intended to evaluate how much of the squared norm of any point is remaining after the projection onto k directions maximally, the answer immediately follows from the sum of the k largest eigenvalues. Employing the corresponding eigenvectors as $\hat{r}_i$ would then be a reasonable approach. Yet, both eigenvectors and eigenvalues can be sensitive to noise in limited datasets [29]. They may not be an optimal choice when new and unknown data arises. We, hence, focus on the expectation of these values for a random set of reference points drawn from the data. More precisely we inspect

$$E_k^{\Sigma}(X) \quad := \quad \mathop{\mathbb{E}}_{\substack{r_1,\dots,r_k\in X\\ \forall i\neq j: r_i\neq r_j}}\left[\mathop{\mathbb{E}}_{x\in X}\left[\|\pi(x-c;r_1-c,\dots,r_n-c)\|^2\right]\right] \tag{4.17}$$

As with the eigenvectors and eigenvalues of the covariance matrix, this expected value is the sum of components introduced by each additional reference point taken into consideration. This naturally sums up the total variance of the dataset for $k=d$. Through varying k we can obtain a cumulative description of how much variance an arbitrary linear projection within the dataset can explain and the difference of neighboring values gives the amount of variance explained at random by the k-th component. We will write this difference as $E_k(X) := E_k^{\Sigma}(X)-E_{k-1}^{\Sigma}(X)$ where $E_0^{\Sigma}(X)=0$. It follows that $E_k^{\Sigma}(X)=\sum_{i=1}^{k}E_k(X)$.

Practically evaluating the expected value from any dataset X for any $k \gg 1$ is infeasible, as it involves $\binom{|X|}{k}$ possible sets of reference points. It is much easier to estimate the value by the Monte Carlo method (i.e. choosing a fixed number of random sets of reference points) or to approximate it from the covariance matrix if it well describes the dataset's distribution.

We will only consider the non-affine case of $c = \mathbf{0}$, as the affine case is analogous and introduces numerous subtractions hindering readability. We will also omit the constraint that the reference points must not be linearly dependent to improve readability. Starting from (4.17) we can deduce

$$E_k(X) = E_k^{\Sigma}(X) - E_{k-1}^{\Sigma}(X) = \mathop{\mathbb{E}}_{\substack{x\in X,\\ r_1,\dots,r_k\in\widetilde{X}}} \left[\left\langle x, \frac{r_k - \pi(r_k; r_1,\dots,r_{k-1})}{\|r_k - \pi(r_k; r_1,\dots,r_{k-1})\|} \right\rangle^2 \right] \tag{4.18}$$

Here the term $r_k - \pi(r_k; r_1, \dots, r_{k-1})$ is the projection of r_k onto the linear subspace orthogonal to all $r_1, \dots, r_{k-1}$. We can represent this projection by a matrix multiplication with a matrix, which we will call A_{k-1}.

$$= \mathop{\mathbb{E}}_{\substack{x\in X,\\ r_1,\dots,r_k\in\widetilde{X}}} \left[\frac{\langle x, A_{k-1} r_k\rangle^2}{\langle A_{k-1} r_k, A_{k-1} r_k\rangle} \right] = \mathop{\mathbb{E}}_{\substack{x\in X,\\ r_1,\dots,r_k\in\widetilde{X}}} \left[x^T \frac{A_{k-1} r_k r_k^T A_{k-1}^T}{\operatorname{tr}(A_{k-1} r_k r_k^T A_{k-1}^T)} x \right] \tag{4.19}$$

By rewriting $r_i r_i^T$ as R_i this further simplifies to

$$= \mathop{\mathbb{E}}_{\substack{x\in X,\\ r_1,\dots,r_k\in\widetilde{X}}} \left[x^T \frac{A_{k-1} R_k A_{k-1}^T}{\operatorname{tr}(A_{k-1} R_k A_{k-1}^T)} x \right] \tag{4.20}$$

$$= \operatorname{tr}\left(\mathop{\mathbb{E}}_{r_1,\dots,r_{k-1}\in\widetilde{X}} \left[\mathop{\mathbb{E}}_{r_k\in X} \left[\frac{A_{k-1} R_k A_{k-1}^T}{\operatorname{tr}(A_{k-1} R_k A_{k-1}^T)} \right] \right] \mathop{\mathbb{E}}_{x\in X} \left[x x^T \right] \right) \tag{4.21}$$

By replacing $\mathbb{E}_{x\in X}\left[xx^T\right]$ with the covariance matrix $C(X)$ and renaming the innermost expected value to $C_k(X)$ we then obtain

$$E_k(X) = \mathop{\mathbb{E}}_{r_1,\dots,r_{k-1}\in\widetilde{X}} \left[\operatorname{tr}\left(C_k(X) C(X) \right) \right] \tag{4.22}$$

A_0 is the identity matrix $\mathbb{I}_d$, as the linear subspace orthogonal to an empty set of vectors is the entire space. Consequentially, we can define A_k recursively as

$$A_k = A_{k-1} - \frac{A_{k-1} R_k A_{k-1}^T}{\operatorname{tr}(A_{k-1} R_k A_{k-1}^T)} = A_{k-1} - \frac{A_{k-1} R_k A_{k-1}}{\operatorname{tr}(A_{k-1} R_k A_{k-1})} \tag{4.23}$$

As all R_i are symmetric, all A_i are symmetric as well. The expected value over r_k of $\frac{A_{k-1} R_k A_{k-1}}{\operatorname{tr}(A_{k-1} R_k A_{k-1})}$ now (approximately) equals the covariance matrix of X after being projected to the linear subspace orthogonal to $r_1, \dots, r_{k-1}$ and normalized. It follows immediately that $C_1(X) = C(\widetilde{X})$ and thereby $E_1(X) = \operatorname{tr}\left(C(\widetilde{X}) C(X)\right)$. However, $E_k(X)$ for $k > 1$ is much less easily defined because the A_i are dependent on the effective values of all r_j, $j \leq i$, and not only on r_i. To circumvent the problem we assume that all A_i are aggregate matrices just like $C(X)$ and sufficiently independent of each other to evaluate the $C_k(X)$ recursively. To

highlight this assumption we will denote the approximated A_i as a function of X as $A_i(X)$. We further assume that all $A_i(X)$, $C_i(X)$, and $C(X)$ admit the same eigenvectors, whereby

$$E_k(X) = \mathop{\mathbb{E}}_{r_1,\dots,r_{k-1}\in\widetilde{X}}[\operatorname{tr}(C_k(X)C(X))] = \sum_{i=1}^{d}\lambda_i(C_k(X))\,\lambda_i(C(X)) \tag{4.24}$$

We will hereafter omit the argument (X) to the matrix argument of the eigenvalues and omit the matrix argument to eigenvectors altogether, since all relevant matrices admit the same eigenvectors. Although the resulting values are no longer exact due to these two assumptions, they allow us to approximate the expected value by deriving the value of $\lambda_i(C_k)$. Assuming that X is multivariate normally distributed, we can extract this value from the definition of C_k using the corresponding eigenvector v_i:

$$\lambda_i(C_k) = v_i^T C_k(X) v_i = \operatorname{tr}\left(v_i v_i^T C_k(X)\right) \tag{4.25}$$

$$= \mathop{\mathbb{E}}_{r_k\in X}\left[\frac{r_k^T A_{k-1}(X) v_i v_i^T A_{k-1}(X) r_k}{r_k^T A_{k-1}(X)^2 r_k}\right] \tag{4.26}$$

$$= \mathop{\mathbb{E}}_{r_k\in\mathcal{N}_{0_d,\mathbb{I}_d}}\left[\frac{r_k^T C(X)^{1/2} A_{k-1}(X) v_i v_i^T A_{k-1}(X) C(X)^{1/2} r_k}{r_k^T C(X)^{1/2} A_{k-1}(X)^2 C(X)^{1/2} r_k}\right] \tag{4.27}$$

$$= \mathop{\mathbb{E}}_{r_k\in\mathcal{N}_{0_d,\mathbb{I}_d}}\left[\frac{r_k^T v_i v_i^T C(X) A_{k-1}(X)^2 r_k}{r_k^T C(X) A_{k-1}(X)^2 r_k}\right] \tag{4.28}$$

We now substitute $C(X)A_{k-1}(X)^2$ with $D_{k-1}(X)$ which entails $\lambda_j(C)\left(\lambda_j(A_{k-1})\right)^2$ is equal to $\lambda_j(D_{k-1})$. In favor of brevity we will assume the matrix argument (D_{k-1}) for eigenvalues when none is present from here on. As per Proposition 2 in Kan and Bao [8], $\lambda_i(C_k)$ then equals

$$\lambda_i(C_k) = \int_0^\infty \frac{\operatorname{tr}\left(v_i v_i^T D_{k-1}(X)(\mathbb{I}_d + 2tD_{k-1}(X))^{-1}\right)}{|\mathbb{I}_d + 2tD_{k-1}(X)|^{1/2}}\,\mathrm{d}t \tag{4.29}$$

$$= \int_0^\infty \frac{\lambda_i}{(1+2t\lambda_i)^{1/2}\prod_{j=1}^{d}\left(1+2t\lambda_j\right)^{1/2}}\,\mathrm{d}t \tag{4.30}$$

This integral is closely related to elliptic integrals and we do not provide a simple and closed-form solution. Solving the integral numerically would again involve too much computational effort. We instead propose to substitute the λ_j in the denominator with $(\lambda_i^2\prod_{j=1}^{d}\lambda_j)^{1/(d+2)}$ whereby the integral takes the form of a scaled beta prime distribution:

$$\lambda_i(C_k) \approx \lambda_i \mathrm{B}(\alpha,\beta)\int_0^\infty \frac{t^{\alpha-1}(1+2\left(\lambda_i^2\prod_{j=1}^{d}\lambda_j\right)^{\frac{1}{d+2}}t)^{-\alpha-\beta}}{\mathrm{B}(\alpha,\beta)}\,\mathrm{d}t \tag{4.31}$$

where $\alpha = 1$, $\beta = \frac{d}{2}$, and $\mathrm{B}(\alpha, \beta)$ is the beta function. The integral over the scaled beta distribution is known to equal the scaling factor, whereby

$$\lambda_i(C_k) \approx \frac{\lambda_i \mathrm{B}(\alpha, \beta)}{2\left(\lambda_i^2 \prod_{j=1}^{d} \lambda_j\right)^{\frac{1}{d+2}}} \propto \lambda_i^{\frac{d}{d+2}} \tag{4.32}$$

As the $\lambda_i(C_k)$ are eigenvalues of a normalized distribution, their sum must equal 1. Using this constraint, we can drop all factors independent of λ_i and derive

$$\lambda_i(C_k) \approx \lambda_i^{\frac{d}{d+2}} \Big/ \sum_{j=1}^{d} \lambda_j^{\frac{d}{d+2}} \tag{4.33}$$

As the λ_j are dependent on $\lambda_j(C)$ and $\lambda_j(A_{k-1})$, this leads to the recursive definition

$$\lambda_i(C_k) \approx \frac{(\lambda_i(C)\,(\lambda_i(A_{k-1}))^2)^{\frac{d}{d+2}}}{\sum_{j=1}^{d}(\lambda_j(C)\,(\lambda_j(A_{k-1}))^2)^{\frac{d}{d+2}}} \qquad \lambda_i(A_k) \approx \lambda_i(A_{k-1}) - \lambda_i(C_k) \tag{4.34}$$

This recursion terminates at $\lambda_i(A_0) = 1$ and $\lambda_i(C_0) = 0$. These approximations can be computed efficiently in $\Theta(dk)$ and inserted in (4.24) to give an approximation of $E_k(X)$. Since the approximations are based on the assumption that X is distributed according to some multivariate normal distribution they need not be accurate. Since all occurrences of any r_k in the formulae involve some sort of normalization, this approximation extends to any distribution of X for which $\{C(X)^{-1/2}x \mid x \in X\}$ is spherically symmetrically distributed, which includes cases like, e.g., d-balls. We also did not compensate for the requirement that all r_k must be pairwise different, as these arguments are based on distributions rather than point sets. The biggest issue with this approximation is the fact, that while the A_i as variables in r_1 through r_i must have eigenvalues in $\{0, 1\}$, the approximated eigenvalues $\lambda_i(A_k)$ can become negative whereby latter E_k can be vastly overestimated. As we know that the $E_k^{\Sigma}(X)$ must sum to the total variance of X, we propose to cut off any excess in $E_k^{\Sigma}(X)$ and determine the $E_k(X)$ based on these cut values. To summarize, the approximation proceeds as follows: For all $1 \leq k \leq d$ compute the $\lambda_i(C_k)$ values using the recursive formulations (4.34). Use these values to compute $E_i(X)$ values and reduce $E_i(X)$ values for larger k to not have their sum exceed the total variance of X, which compensates for negative $\lambda_i(A_k)$. Even though this approximation from a theoretical point makes the wrong assumptions that the r_k are pairwise different and that the $C_i(X)$ are statistically independent, the approximation in our experiments gave close enough results to have it worth considering, especially as the exact computation of values has an enormous computational cost. The approximation via the Monte Carlo method is known to converge on the exact values, yet, might require enormous samples.

As an additional result, that is somewhat hidden between the lines, we can derive the approximate eigenvalues of the covariance matrix of the dataset after normalization from $C_1(X)$ as

$$\lambda_i\big(C(\widetilde{X})\big) \approx \frac{(\lambda_i(C)\,)^{\frac{d}{d+2}}}{\sum_{j=1}^{d}(\lambda_j(C)\,)^{\frac{d}{d+2}}} \tag{4.35}$$

This result, which is not explicitly stated in the original paper [87], has to our knowledge not been published before and is useful for the understanding of the distribution of normalized datasets.

While (4.24) requires the covariance matrix of a mean-centered dataset, the approach via Monte Carlo sampling applies directly to inner product values and, hence, to kernel spaces. The approximation in (4.24) can then be used in black-box optimization to obtain an approximate spectral analysis of the kernel space. The obtained spectrum is neglecting the scale of the eigenvalues of the covariance matrix as the $E_i(X)$ are invariant under the scaling of these values. In this manner, we can perform approximate spectral analysis even in spaces that do not allow for a direct approach, such as the RBF kernel space which has infinitely many dimensions. Naturally, the method must be applied in a truncated fashion for infinite dimensions, for which we here propose two solutions: Firstly, one can estimate $E_1(X)$ through $E_k(X)$ for some fixed k using the Monte Carlo method and rescale these values to sum to 1. This implies neglecting the remaining $d-k$ dimensions and assuming the data to have 0 variance along with these directions. The $d-k$ smallest eigenvalues of the covariance of such a dataset must then be 0, too. Finding any set of k eigenvalues that leads to these $E_1(X)$ through $E_k(X)$ values then solves the truncated case. Secondly, one can assume that the remaining variance not explained by $E_k^{\Sigma}(X)$ is distributed over the remaining $d-k$ values according to some user-defined distribution. Assuming a uniform distribution, for example, would explain the remaining variance as noise in the embedding space which might be a reasonable assumption.

A special case can further be made on the evaluation of $E_k(X)$ values on normalized data. When working on $\widetilde{X}$ instead of X, which can be achieved in kernel space by dividing the occurrences of x in the formulae by $\langle x, x\rangle^{1/2}$, we immediately obtain that $E_1(\widetilde{X})$ equals the sum of squared eigenvalues of $C(\widetilde{X})$. While this equality does not hold for the approximation via eigenvalues of $C(\widetilde{X})$, it is approximately obtained from the Monte Carlo method or precisely for an exhaustive evaluation of $E_1(\widetilde{X})$. Just as the constraint of the sum of eigenvalues of $C(\widetilde{X})$ equalling 1, this additional constraint can be used in the black-box optimization for retrieving the original eigenvalues from $E_k(\widetilde{X})$ values. Using (4.32), these eigenvalues can be approximately translated into the relative eigenvalues of the non-normalized data whenever the data can be assumed to obey the distributional constraints of the approximation.

4.3 Random Projections and ID Estimation [87]

As stated in the previous section, $E_1(\widetilde{X})$ equals the sum of squared eigenvalues of $C(\widetilde{X})$. The reciprocal of this specific value, considering $\widetilde{X}$ as the "local neighborhood", is just the $\mathrm{LID}_{\mathrm{ABID}}$ estimator introduced in Section 3.1, i. e.

$$\mathrm{LID}_{\mathrm{ABID}}(X) = E_1(\widetilde{X})^{-1} = E_1^{\Sigma}(\widetilde{X})^{-1} \,. \tag{4.36}$$

This observation adds additional semantics to the meaning of $\mathrm{LID}_{\mathrm{ABID}}$ as the number of basis vectors of a random projection to fully explain the variance in a dataset. Yet, it also implies the applicability of the E_k values in the realm of ID estimation. Although E_1 gives the part of total variance a random projection based on in-distribution basis vectors can explain, not all E_k values are necessarily equal. That is, the projection onto two random directions does not necessarily cover twice the variance covered by projecting onto one random direction. This linearity is exclusively true for spherically symmetrical distributions such as d-balls and for all other distributions we would certainly expect $E_2^{\Sigma}(X) < 2E_1^{\Sigma}(X)$. Ultimately, we are looking for the smallest k such that $E_k^{\Sigma}(X) \geq \mathrm{tr}\,(C(X))$, that is, the number of random projections required to explain the entire variance of X. Unfortunately, we only have formulae for integer k but we can generalize the approach of $\mathrm{LID}_{\mathrm{ABID}}$ in the sense of extrapolating from a fixed E_k which results in a parameterized ID estimator which we name the Thresholded Random In-distribution Projections Intrinsic Dimensionality ($\mathrm{LID}_{\mathrm{TRIP}}$) Estimator:

$$\mathrm{LID}_{\mathrm{TRIP}}(X, k, \eta) = k + \frac{(1-\eta)\,\mathrm{tr}\,(C(X)) - E_k^{\Sigma}(X)}{E_k(X)} \tag{4.37}$$

where k is the number of considered projections and $\eta \in [0, 1]$ is a fraction describing how much of the variance we attribute to noise. Semantically this answers the question *"How many random projections are required to explain* $(1-\eta)$ *of the total variance if every further projection covers as much variance as the last one?"*. In the linear case of spherically symmetrical distributions as above, this estimator is ideally constant for $\eta = 0$ and all $1 \leq k \leq d$. On other distributions with $\eta = 0$, we would expect a curve that starts at (approximately, dependent on implementation) $\mathrm{LID}_{\mathrm{ABID}}(X)$ for $k = 1$ and approaches k for increasing k as the $E_i(X)$ are monotonically falling. Equality is likely only reached for $k = d$, as this requires zero variance after k projections, which is unlikely in presence of high-dimensional noise. The factor η is intended to compensate for this. For $\eta > 0$, the curve again starts at approximately $\mathrm{LID}_{\mathrm{ABID}}(X)$, approaches k, and after some k drops below it. As for parameter choice, η is application dependent whereas k can either be chosen empirically, or we can inspect values $1 \leq k \leq d$ to find the k at which $\mathrm{LID}_{\mathrm{TRIP}}(X, k, \eta)$ is closest to k. The latter is likely not feasible in a LID fashion when using the Monte Carlo or exhaustive methods but can be done when using the approximation introduced in Section 4.2. When using a fixed k, obtaining an ID below this k is a strong indicator of having chosen k too large. In

addition, the curve of $\mathrm{LID}_{\mathrm{TRIP}}(X, k, \eta)$ over varying k, just like the curve of $E_i(X)$, gives insights into the local distribution characteristics of the dataset that goes beyond ID estimation. These curves can theoretically help distinguish different subspaces, even when they share similar LID.

Referring back to the discussions of indexing with linear projections in Section 4.1, we can now state a clear connection between indexing with random in-distribution pivots and intrinsic dimensionality measures. The $E_k^{\Sigma}(X)$ values answer how much variance on average is covered by a set of k random pivots. The expected covered variance is – in an idealized case of, e.g., uniformly distributed hyperballs – reciprocally related to intrinsic dimensionality. This is most explicitly stated in the relation to $\mathrm{LID}_{\mathrm{ABID}}$ and gives rise to the $\mathrm{LID}_{\mathrm{TRIP}}$ estimator above. Using this geometric concept of ID estimation, we can argue on an on-average appropriate number of pivots in spatial indexing. In Section 4.1 we observed that the eligible search space for range queries when using k pivots is the sumset of a $(d-1-k)$-sphere and an ε-ball. The radius of the hypersphere is equal to the norm of the component orthogonal to all pivots, and roughly describes how close the bounds derived in Section 4.1 are to the true distances. But there is a clear limit as to how much precision one needs in a finite dataset. If this radius drops below the distance between nearest points, removing this slack from the distance estimates does not improve the discriminability. By choosing $\eta = \delta^2 / \operatorname{tr}(C(X))$ where δ is the, e.g., mean/median/p-percentile of nearest neighbor distances, we can use the $\mathrm{LID}_{\mathrm{TRIP}}$ estimator to evaluate just how many random projections exhaust the discriminative potential of pivoted indexing on average.

4.4 Pivot Filtering Linear Scan [87]

For quality evaluation of the bounds as well as to validate the theoretical claims, we embed the bounds in a simple and easy-to-implement index. During the initialization, we choose k random pivots. As mentioned in Section 4.1, we pre-compute all parts of the equations that are independent of query points such as $\langle x, \hat{r}_i \rangle$ or the denominators in (4.4). Range and n-nearest neighbor queries were then implemented according to Algorithms 4.1 and 4.2. The algorithms are quite similar to LAESA [61] but do not require aggregation of multiple bounds as discussed in Section 4.1. Both algorithms are at least linear in $|X|$, which should be accounted for when comparing the performance with tree-based indices. Integrating the bounds into a tree-based index is a nearby extension but out of the scope of this chapter. Both Algorithms 4.1 and 4.2 are trivially adaptable to search for the largest instead of the smallest distances. This index is also trivially adaptable to work on inner products instead of distances by exchanging the bounds. For our experiments, we implemented the index in the Rust language and called the functions from a Python wrapper to compare them to the cKDTree and BallTree implementations

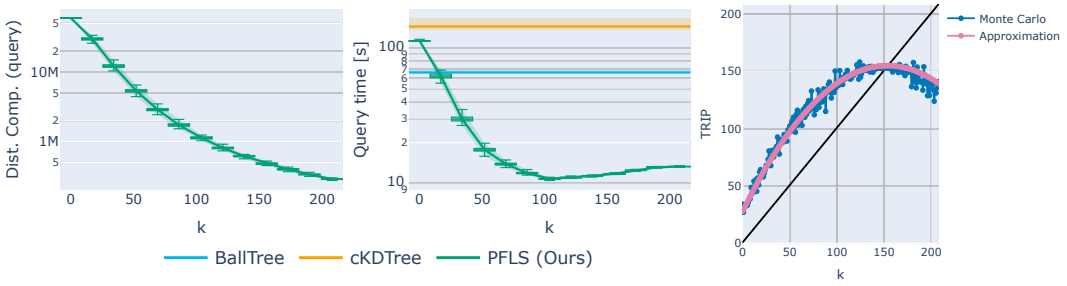
function QUERY($y \in \mathbb{R}^d, n \geq 1$)
  $ls \leftarrow$ lower bounds of $d(x, y)$ for all $x \in X$ as per (4.6) and (4.11)
  $h \leftarrow$ empty max heap
  sort X by ascending $ls[x]$
  for $x \in X$ **do**
    if $|h| < n$ **or** $(ls[x] < h.max.key$ **and** $d(x, y) < h.max.key)$ **then**
      push x onto h with key $d(x, y)$
      if $|h| > n$ **then** remove entry with largest key from h
      else if $ls[x] \geq h.max.key$ **then break**
  return h as array/list

Algorithm 4.1: n-nearest neighbor query for distances

function QUERY-RANGE($y \in \mathbb{R}^d, \varepsilon \in \mathbb{R}$)
  $ls, hs \leftarrow$ lower and upper bounds of $d(x, y)$ for all $x \in X$ as per (4.6) and (4.11)
  $v \leftarrow$ empty list
  for $x \in X$ **do**
    if $ls[x] < \varepsilon$ **and** $(hs[x] < \varepsilon$ **or** $d(x, y) < \varepsilon)$ **then** push x into v
  return v

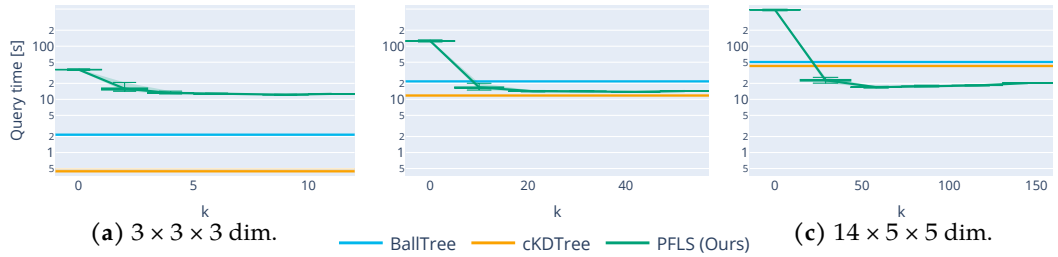
Algorithm 4.2: Range query for distances

of SciPy [89]. The source code is publicly available at `https://github.com/eth42/pfls`. Using this very simple index, we investigated the theoretical claims and the quality of the bounds. Fig. 4.2 displays the results of applying the index to the MNIST training dataset. All queries were 100-nearest-neighbor queries for 1000 query points drawn from the same dataset. We performed 100 queries for each set of parameters and instantiated a new index for each query. As seen in Fig. 4.2a, the number of distance computations initially drops exponentially as we increase the number of pivots, which supports the theoretical claim that each pivot effectively eliminates one dimension from the dataset and reduces the remaining search space exponentially. For increasing k, the descent in distance computations diminishes as the bounds become tight enough to sufficiently discriminate on neighboring points, and the query time eventually increases due to the cost of computing the bounds. In Section 4.3, we argued that the bounds only need to be as tight as to differentiate between nearest neighbors. To validate this claim, we investigated the $\mathrm{LID}_{\mathrm{TRIP}}$ values using an η equal to the 10-percentile of squared 1-nearest-neighbor distances divided by the total variance of the distribution. The smallest k for which $\mathrm{LID}_{\mathrm{TRIP}}(X, k, \eta) \leq k$ is around 150 as can be seen in Fig. 4.2c. The minimum computation time in Fig. 4.2a is around $k = 100$ but the query time at $k = 150$ is not that much larger. The exact percentile is an educated guess and could be supported by inspecting the histogram of nearest-



(**a**) Distance comp. (**b**) Computation times (**c**) $\mathrm{LID}_{\mathrm{TRIP}}$ with $\eta > 0$

Figure 4.2: Experimental results on varying numbers of pivots. Additional pivots exponentially reduce the distance computations, but the query time stagnates once the average discriminative power of the bounds has been exploited. A suitable number of pivots is suggested at the crossing point of $\mathrm{LID}_{\mathrm{TRIP}}$ with the diagonal. Lines are average values, shaded area indicates the minimum and maximum.



(**a**) $3 \times 3 \times 3$ dim. (**c**) $14 \times 5 \times 5$ dim.

(**b**) $14 \times 3 \times 3$ dim.

Figure 4.3: Query times for ALOI color histograms with varying dimensionality.

neighbor distances. Yet, the region of k that provides low query times is wide enough that rough estimates and educated guesses are likely to give good results. We conclude that $\mathrm{LID}_{\mathrm{TRIP}}$ can be used to estimate a proper value for k by deriving η from a percentile of 1-nearest neighbor distances. To estimate a proper k efficiently, the approximation introduced in Section 4.2 can be used, which practically is sufficiently similar to the values obtained from Monte Carlo sampling as displayed in Fig. 4.2c. Lastly, we compared query times on HSV color histograms of the ALOI dataset with varying numbers of dimensions [77]. The considered variants consist of 110250 instances with 27, 126, and 350 dimensions, respectively. As can be seen in Fig. 4.3 the query performance of our index is much less affected by increasing dimensionality. Due to our index using a linear scan, the tree-based reference implementations were faster on low dimensionality. For sufficiently high dimensional or small enough datasets, our index can outperform these reference implementations. For larger datasets, extending the approach to a tree-based structure appears promising.

4.5 Conclusion

This chapter was focused on an attempt to expand the $\mathrm{LID}_{\mathrm{ABID}}$ estimator to a measure based on more than just pairs of points. To achieve that, we generalized the concept of "angle" to the angle between a point and a linear subspace spanned by k other points. That generalization resulted in inequalities for the Euclidean distance and inner products, that can also be used in e. g. spatial indexing. In the limit of $k \to \infty$, the inequalities imply Euclideanness of the data, which is not surprising as they imply a parameterization of the vector space. After all, the inequalities are mere reformulations of the truncated Euclidean distance. By analyzing the expected value when the projection directions are sampled from the data itself, we could derive the $\mathrm{LID}_{\mathrm{TRIP}}$ estimator. As we intended to generalize on the concept of $\mathrm{LID}_{\mathrm{ABID}}$, it comes as little surprise, that $\mathrm{LID}_{\mathrm{ABID}}$ is a special case of $\mathrm{LID}_{\mathrm{TRIP}}$. The $\mathrm{LID}_{\mathrm{TRIP}}$ estimator can be parameterized with an acceptable loss in variance, yielding an estimate for an optimal parameter choice to balance computational cost and pruning efficiency. That claim was supported by the given empirical evidence. The task-specific parameterization makes $\mathrm{LID}_{\mathrm{TRIP}}$ the first task-aware LID estimator, complementing previous LID-aware algorithms from the literature [72, 50].

Chapter 5

Chordal Angles-Based Intrinsic Dimensionality

Having access to a square number of values per neighborhood made the $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ estimators already intuitively attractive compared to the distance-based estimators. Nonetheless, the $\mathrm{LID}_{\mathrm{ABID}}$ estimator turned out to have the same asymptotic dependency on the neighborhood size as $\mathrm{LID}_{\mathrm{MLE}}$. The $\mathrm{LID}_{\mathrm{TRIP}}$ approach even went beyond and resulted in an exponential number of values per neighborhood. The resulting computational effort and many approximative shortcuts necessary to even obtain a result, however, made the $\mathrm{LID}_{\mathrm{TRIP}}$ approach less attractive for practical use in LID estimation and rather for the analysis and parameterization of e.g. spatial indexes. In an attempt to go beyond a squared number of values while still obtaining a sound closed form solution, we developed the Approximately Linearly-parameterized Chordal Angles LID ($\mathrm{LID}_{\mathrm{ALC}}$) estimator, which will be the topic of this chapter. The $\mathrm{LID}_{\mathrm{ALC}}$ estimator is based on the chordal angles of samples projected onto the unit sphere. Chordal angles are the angles inside the triangles of chords connecting pairs of points on the unit sphere. Consequentially, we obtain a cubic number of values. As we will only use the first two moments of these angles, a sub-cubic sample is sufficient for approximate results. This allows for the same practical benefits (fractional, fairly efficient) as the $\mathrm{LID}_{\mathrm{ABID}}$ estimator. Ultimately we aim to decrease the required number of neighbors, resulting in smaller localities and thereby arriving closer to the "linear within locality"-assumption of the underlying manifold. The resulting estimator further extends to inner angles of triangles of samples in uniformly distributed balls, albeit only with an empirical and not with an analytical derivation. In this chapter we will first provide the derivation and definition of the $\mathrm{LID}_{\mathrm{ALC}}$ estimator family in Section 5.1. We will then provide a short empirical evaluation of the estimator compared to other approaches in Section 5.2. The chapter closes in Section 5.3 with a summary of the contributions and results.

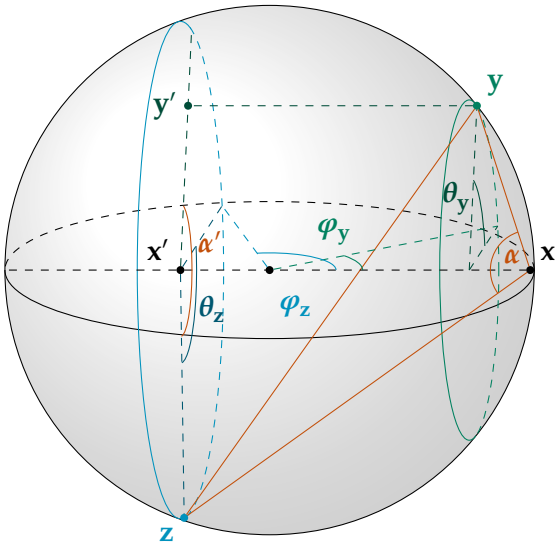


Figure 5.1: Three dimensional showcase of the relevant points and angles in the derivation for the distribution of $\cos(\alpha)$.

5.1 Derivation of $\mathrm{LID}_{\mathrm{ALC}}$

Similar to the derivation of the $\mathrm{LID}_{\mathrm{ABID}}$ estimator, we will inspect the moments of chordal angles to derive the $\mathrm{LID}_{\mathrm{ALC}}$ estimator. To do so, we will need a way to evaluate the probability density function of a single chordal angle. As a first step, we can derive the probability density function of a chordal angle, given the adjacent side lengths, in terms of the probability density function of pairwise cosines on the sphere in Theorem 3.1.2. Moving from this conditional probability to the general probability of a chordal angle would require integrating the conditional probability over all possible side lengths. This integral, however, is a generalized elliptic integral and does not lend to a closed-form solution. Integrating each moment individually is feasible, though. We therefore derive the moments of the conditional probability and then integrate them to obtain the moments of the general probability. Given the moments of the general probability, we show that the distribution rapidly approaches a Beta distribution as the dimensionality of the sphere increases. Ultimately, the obtained formulae allow for three LID estimators of varying complexity: Estimating with the a priori moments, estimating with the moments obtained from the Beta distribution approximation, and lastly estimating with a linear approximation of the Beta distribution parameters. The computational cost of these variants decreases while the parameter error remains below 2.5% for $\delta \geq 3$. We consequentially propose to use the cheapest variant in practice, which we name $\mathrm{LID}_{\mathrm{ALC}}$. Fig. 5.1 provides a three-dimensional showcase of the relevant points and angles to guide through the derivation. The points x, y, z are the vertices of a random chordal triangle and α is the angle whose cosine we want to inspect. Without loss of generality, we assume

$x = (1, 0, \dots, 0), y = (\cos(\varphi_y), \sin(\varphi_y)\cos(\theta_y), \sin(\varphi_y)\sin(\theta_y), 0, \dots, 0)$, and z similar to y. The vectors x, y, and z thus only have non-zero entries in the first three dimensions. Any triplet of vectors x, y, and z can be rotated to this configuration without changing the angles between them. The derivations in this chapter all assume $d > 2$, which is necessary for most of the steps and will not be explicitly stated everywhere. The resulting estimators are then numerically applicable to d or $\delta \leq 2$ albeit without theoretical justification. We will give empirical evidence, that the error in that domain is negligible though.

First, we need to remember the distribution of pairwise cosines for uniformly random points on the sphere introduced in Theorem 3.1.2.

5.1.1 Theorem (Cosines on the sphere [84]). *The probability density function of cosines of angles φ between two points uniformly at random from the $(d-1)$-sphere is*

$$\mathrm{P}\left(\cos(\varphi)\right) \quad = \quad \tfrac{1}{2}\mathrm{B}\left(\tfrac{\cos(\varphi)+1}{2}; \tfrac{d}{2} - \tfrac{1}{2}, \tfrac{d}{2} - \tfrac{1}{2}\right) \tag{5.1}$$

where $\mathrm{B}(x; a, b)$ is the Beta distribution.

Using that distribution, we can derive the conditioned probability density function of $\cos(\alpha)$ given φ_y and φ_z. The proof is a construction of $\cos(\alpha)$ in terms of $\cos(\alpha')$, which is by definition a pairwise cosine of uniformly random samples from a $(d-2)$-subsphere.

5.1.2 Lemma. *Let x, y, z be points uniformly at random from the $(d-1)$-sphere. When $\cos(\varphi_y) := \langle x, y\rangle$ and $\cos(\varphi_z) := \langle x, z\rangle$ are fixed, then*

$$\begin{aligned}&\mathrm{P}\left(\cos(\alpha) \mid \varphi_y, \varphi_z\right)\\ &\sim \mathrm{B}\left(\frac{\cos(\alpha)}{\sqrt{1+\cos(\varphi_y)}\sqrt{1+\cos(\varphi_z)}} - \frac{\sqrt{1-\cos(\varphi_y)}\sqrt{1-\cos(\varphi_z)}}{2\sqrt{1+\cos(\varphi_y)}\sqrt{1+\cos(\varphi_z)}} + \tfrac{1}{2}; \tfrac{d}{2} - 1, \tfrac{d}{2} - 1\right)\end{aligned} \tag{5.2}$$

where α is the chordal angle $\angle_{yxz}$.

Proof. Without loss of generality, we choose

$$x = (1,\ 0, \dots, 0) \tag{5.3}$$

$$y = (\cos(\varphi_y),\ \sin(\varphi_y)\cos(\theta_y),\ \sin(\varphi_y)\sin(\theta_y),\ 0, \dots, 0) \tag{5.4}$$

$$z = (\cos(\varphi_z),\ \sin(\varphi_z)\cos(\theta_z),\ \sin(\varphi_z)\sin(\theta_z),\ 0, \dots, 0) \tag{5.5}$$

This assumption is viable since α, φ_y, and φ_z are invariant under rotation of the sphere and we can rotate any three points such that all but their first three coordinates are 0 and that one point is aligned with the first axis. The chosen representation then represents x, y, and z in spherical coordinates of one such rotation. Instead of deriving the distribution of α directly, we use the result of Theorem 5.1.1 on the angle between y and z in the $(d-2)$-subsphere orthogonal to x, i.e. starting at the second coordinate in the representation above. If we projected the chordal triangle of x, y, and z on that subspace, we obtain

a distorted angle α' as displayed in Figure 5.1. Since y and z are chosen uniformly at random from the $(d-1)$-sphere, they are also uniformly at random from any subsphere. Accordingly, the probability of $\cos(\alpha')$ is given as

$$\mathrm{P}\left(\cos(\alpha')\right) = \tfrac{1}{2}\mathrm{B}\left(\tfrac{\cos(\alpha')+1}{2}; \tfrac{d}{2}-1, \tfrac{d}{2}-1\right) \tag{5.6}$$

We can then apply the law of cosines to derive $\cos(\alpha)$ from $\cos(\alpha')$ by first expressing $\|z-y\|$ in terms of α' – i.e. without θ_y and θ_z. For that, we use the projections x' and y' of x and y onto the $(d-2)$-subsphere orthogonal to x that contains z. This allows us to use the Pythagorean theorem to decompose the distance between y and z into the distances between y' and y, and between y' and z. Using the law of cosines, we obtain the distance between y' and z as

$$\cos(\alpha') = \frac{\|y'-x'\|^2 + \|z-x'\|^2 - \|y'-z\|^2}{2\|y'-x'\|\|z-x'\|} \tag{5.7}$$

$$\Rightarrow \quad \cos(\alpha') = \frac{\sin(\varphi_y)^2 + \sin(\varphi_z)^2 - \|y'-z\|^2}{2\sin(\varphi_y)\sin(\varphi_z)} \tag{5.8}$$

$$\Rightarrow \quad \|y'-z\|^2 = \sin(\varphi_y)^2 + \sin(\varphi_z)^2 - 2\cos(\alpha')\sin(\varphi_y)\sin(\varphi_z) \tag{5.9}$$

We can then apply the Pythagorean theorem to obtain the distance between y and z as

$$\|y-z\|^2 = \|y'-y\|^2 + \|y'-z\|^2 \tag{5.10}$$

$$\Rightarrow \quad \|y-z\|^2 = \left(\cos(\varphi_y) - \cos(\varphi_z)\right)^2 + \sin(\varphi_y)^2 + \sin(\varphi_z)^2 - 2\cos(\alpha')\sin(\varphi_y)\sin(\varphi_z) \tag{5.11}$$

$$\Rightarrow \quad \|y-z\|^2 = 2 - 2\cos(\varphi_y)\cos(\varphi_z) - 2\cos(\alpha')\sin(\varphi_y)\sin(\varphi_z) \tag{5.12}$$

Finally using the law of cosines we get

$$\cos(\alpha) = \frac{\|y-x\|^2 + \|z-x\|^2 - \|y-z\|^2}{2\|y-x\|\|z-x\|} \tag{5.13}$$

$$\Rightarrow \quad \cos(\alpha) = \frac{\begin{pmatrix} (1-\cos(\varphi_y))^2 + \sin(\varphi_y)^2 + (1-\cos(\varphi_z))^2 + \sin(\varphi_z)^2 \\ -2 + 2\cos(\varphi_y)\cos(\varphi_z) + 2\cos(\alpha')\sin(\varphi_y)\sin(\varphi_z) \end{pmatrix}}{2\sqrt{(1-\cos(\varphi_y))^2+\sin(\varphi_y)^2}\sqrt{(1-\cos(\varphi_z))^2+\sin(\varphi_z)^2}} \tag{5.14}$$

$$\Rightarrow \quad \cos(\alpha) = \frac{(1-\cos(\varphi_y))(1-\cos(\varphi_z)) + \cos(\alpha')\sin(\varphi_y)\sin(\varphi_z)}{2\sqrt{1-\cos(\varphi_y)}\sqrt{1-\cos(\varphi_z)}} \tag{5.15}$$

$$\Rightarrow \quad \cos(\alpha') = \frac{2\cos(\alpha)\sqrt{1-\cos(\varphi_y)}\sqrt{1-\cos(\varphi_z)} - (1-\cos(\varphi_y))(1-\cos(\varphi_z))}{\sin(\varphi_y)\sin(\varphi_z)} \tag{5.16}$$

$$\Rightarrow \quad \cos(\alpha') = \frac{2\cos(\alpha)}{\sqrt{1+\cos(\varphi_y)}\sqrt{1+\cos(\varphi_z)}} - \frac{\sqrt{1-\cos(\varphi_y)}\sqrt{1-\cos(\varphi_z)}}{\sqrt{1+\cos(\varphi_y)}\sqrt{1+\cos(\varphi_z)}} \tag{5.17}$$

By inserting (5.17) into (5.1), we obtain (5.2). □

To generalize the above distribution to the unconditioned probability density function of $\cos(\alpha)$, we would need to integrate (5.2) multiplied with a correction for the relative covered volume over φ_y and φ_z. That integral, however, is a generalized form of the elliptic integral and is not known to be tractable. We can, however, derive the moments of the conditional probability and then integrate them to obtain the moments of the general probability. For that, we start by deriving the moments of the conditioned probability.

5.1.3 Hypothesis. *Let x, y, z be points uniformly at random from the $(d-1)$-sphere where $\cos(\varphi_y) := \langle x, y\rangle$ and $\cos(\varphi_z) := \langle x, z\rangle$ are fixed, then the k-th non-central moment of the cosines of angles $\alpha := \angle_{yxz}$ is*

$$\begin{aligned}
&\mathbb{E}\left[\cos(\alpha)^k \mid \varphi_y, \varphi_z\right] \\
&= \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} T(k,i) \frac{2^{-k-i}\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d-1+2i}{2}\right)} \\
&\left((1-\cos(\varphi_y))(1-\cos(\varphi_z))\right)^{\frac{k}{2}-i} \left((1+\cos(\varphi_y))(1+\cos(\varphi_z))\right)^{i}
\end{aligned} \tag{5.18}$$

where $T(k,i)$ is the Bessel number of the second kind $Bes(k, k-i)$[1] *[20], i.e.*

$$T(k,i) = \frac{2^{-i}\Gamma(k+1)}{\Gamma(k-2i+1)\,\Gamma(i+1)} \tag{5.19}$$

Argument. From Lemma 5.1.2 we know, that a linear affine transformation of the chordal cosines is Beta-distributed. That implies

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\cos(\alpha)}{\sqrt{1+\cos(\varphi_y)}\sqrt{1+\cos(\varphi_z)}} - \frac{\sqrt{1-\cos(\varphi_y)}\sqrt{1-\cos(\varphi_z)}}{2\sqrt{1+\cos(\varphi_y)}\sqrt{1+\cos(\varphi_z)}} + \frac{1}{2}\right)^k \mid \varphi_y, \varphi_z\right] \\
&= \mathbb{E}^{(k)}\left[\mathrm{B}\left(\cdot; \tfrac{d}{2}-1, \tfrac{d}{2}-1\right)\right]
\end{aligned} \tag{5.20}$$

where $\mathbb{E}^{(k)}[\cdot]$ denotes the k-th non-central moment of a distribution. Applying the binomial theorem on the left side of this equation and extracting anything that does not depend on α out of the expected values, we obtain

$$\begin{aligned}
&\mathbb{E}^{(k)}\left[\mathrm{B}\left(\cdot; \tfrac{d}{2}-1, \tfrac{d}{2}-1\right)\right] \\
&= \sum_{i=0}^{k} \binom{k}{i} \frac{\mathbb{E}\left[\cos(\alpha)^i \mid \varphi_y, \varphi_z\right]}{2^{k-i}\left((1+\cos(\varphi_y))(1+\cos(\varphi_z))\right)^{\frac{i}{2}}} \\
&\left(1 - \left(\frac{(1-\cos(\varphi_y))(1-\cos(\varphi_z))}{(1+\cos(\varphi_y))(1+\cos(\varphi_z))}\right)^{\frac{1}{2}}\right)^{k-i}
\end{aligned} \tag{5.21}$$

[1] Bessel numbers of the second kind are typically denoted $B(n,k)$, but to avoid confusion with the Beta function, we here use $Bes(n,k)$.

Solving this equation for $\mathbb{E}\left[\cos(\alpha)^k \mid \varphi_y, \varphi_z\right]$ and inserting the definition of the non-central moments of the Beta distribution, we get the recursive form

$$\begin{aligned}
&\mathbb{E}\left[\cos(\alpha)^k \mid \varphi_y, \varphi_z\right] \\
=&\frac{\Gamma(d-2)\,\Gamma\left(\frac{d}{2}-1+k\right)}{\Gamma\left(\frac{d}{2}-1\right)\Gamma(d-2+k)}\left((1+\cos(\varphi_y))(1+\cos(\varphi_z))\right)^{\frac{k}{2}} \\
&-\sum_{i=0}^{k-1}\binom{k}{i}\frac{\mathbb{E}\left[\cos(\alpha)^i \mid \varphi_y, \varphi_z\right]}{2^{k-i}}\sum_{j=0}^{k-i}\binom{k-i}{j}(-1)^{k-i-j} \\
&\left((1-\cos(\varphi_y))(1-\cos(\varphi_z))\right)^{\frac{k-i-j}{2}}\left((1+\cos(\varphi_y))(1+\cos(\varphi_z))\right)^{\frac{j}{2}}
\end{aligned} \tag{5.22}$$

For $k = 0$ and $k = 1$ we can solve this equation directly, as it does not include the recursion and obtain

$$\mathbb{E}\left[\cos(\alpha)^0 \mid \varphi_y, \varphi_z\right] = 1 \tag{5.23}$$

$$\mathbb{E}\left[\cos(\alpha)^1 \mid \varphi_y, \varphi_z\right] = \frac{1}{2}\left((1-\cos(\varphi_y))(1-\cos(\varphi_z))\right)^{\frac{1}{2}} \tag{5.24}$$

and for $k = 2$ (the highest moment that we will use practically) we obtain

$$\begin{aligned}
\mathbb{E}\left[\cos(\alpha)^2 \mid \varphi_y, \varphi_z\right] = &\frac{1}{4}\left((1-\cos(\varphi_y))(1-\cos(\varphi_z))\right) \\
&+\frac{1}{4(d-1)}\left((1+\cos(\varphi_y))(1+\cos(\varphi_z))\right)
\end{aligned} \tag{5.25}$$

which agrees with (5.18). Similarly, we can unroll the sum for any fixed k. For $k \leq 10$ we manually confirmed, that the moments given in (5.18) are correct. A general proof, however, requires a resolution of arbitrarily large sums over quotients of Gamma functions, for which we could neither deduce a closed form nor find any results in the available literature. We, thus, have to leave the rest of the proof as a hypothesis. As per numerical evaluation, the functions appear to be exact, nonetheless. Further, the coefficients are so specific, that we deem it highly unlikely to not be the correct result if it holds for $k \leq 10$.□

We can integrate the non-central moments of the conditioned probability over possible vectors y and z to obtain the non-central moments of the general probability.

5.1.4 Theorem (Moments of Chordal Cosines). *For every k, for which Hypothesis 5.1.3 is correct, the k-th non-central moment of cosines of angles $\alpha := \angle_{yxz}$ where x, y, and z are independent samples uniformly at random from the $(d-1)$-sphere is*

$$\mathbb{E}\left[\cos(\alpha)^k\right] = \frac{2^{2d-4}\Gamma\left(\frac{d}{2}\right)^2\Gamma(k+1)}{\pi\Gamma\left(\frac{d-1}{2}\right)\Gamma\left(\frac{2d-2+k}{2}\right)^2}\sum_{i=0}^{\left\lfloor\frac{k}{2}\right\rfloor}\frac{2^{-2i}\Gamma\left(\frac{d-1+2i}{2}\right)\Gamma\left(\frac{d-1+k-2i}{2}\right)^2}{\Gamma(i+1)\,\Gamma(k-2i+1)} \tag{5.26}$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left[\cos(\alpha)^k\right] \\
&= \int_0^\pi \int_0^\pi \frac{\mathscr{A}_{d-1}\big(\sin(\varphi_y)\big)\mathscr{A}_{d-1}(\sin(\varphi_z))}{\mathscr{A}_d(1)^2} \mathbb{E}\left[\cos(\alpha)^k \mid \varphi_y, \varphi_z\right] \mathrm{d}\varphi_y \,\mathrm{d}\varphi_z && (5.27)
\end{aligned}$$

where $\mathscr{A}_d(r)$ is the surface area of the d-dimensional sphere with radius r. The quotients $\mathscr{A}_{d-1}(\sin(\cdot))/\mathscr{A}_d(1)$ are the probability densities of y and z, respectively. The integral thus follows the general form of the expectation of a function of two random variables.

$$\begin{aligned}
&= \int_{-1}^1 \int_{-1}^1 \frac{\mathscr{A}_{d-1}\big(\sin(\varphi_y)\big)\mathscr{A}_{d-1}(\sin(\varphi_z))}{\mathscr{A}_d(1)^2 \sin(\varphi_y)\sin(\varphi_z)} \mathbb{E}\left[\cos(\alpha)^k \mid \varphi_y, \varphi_z\right] \mathrm{d}\cos(\varphi_y)\,\mathrm{d}\cos(\varphi_z) && (5.28)\\
&= \int_{-1}^1 \int_{-1}^1 \frac{\Gamma\left(\frac{d}{2}\right)^2 \sin(\varphi_y)^{d-3}\sin(\varphi_z)^{d-3}}{\pi\Gamma\left(\frac{d-1}{2}\right)^2} \mathbb{E}\left[\cos(\alpha)^k \mid \varphi_y, \varphi_z\right] \mathrm{d}\cos(\varphi_y)\,\mathrm{d}\cos(\varphi_z) && (5.29)
\end{aligned}$$

Abbreviating $x = (1+\cos(\varphi_y))(1+\cos(\varphi_z))$ and $y = (1-\cos(\varphi_y))(1-\cos(\varphi_z))$ and using $\varphi_y, \varphi_z \in [0, \pi)$ we get

$$\begin{aligned}
&= \int_{-1}^1 \int_{-1}^1 \frac{\Gamma\left(\frac{d}{2}\right)^2}{\pi\Gamma\left(\frac{d-1}{2}\right)^2} x^{\frac{d-3}{2}} y^{\frac{d-3}{2}} \mathbb{E}\left[\cos(\alpha)^k \mid \varphi_y, \varphi_z\right] \mathrm{d}\cos(\varphi_y)\,\mathrm{d}\cos(\varphi_z) && (5.30)\\
&= \int_{-1}^1 \int_{-1}^1 \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} \frac{2^{-k-2i}\Gamma\left(\frac{d}{2}\right)^2\Gamma(k+1)}{\pi\Gamma\left(\frac{d-1}{2}\right)\Gamma(i+1)\Gamma\left(\frac{d-1+2i}{2}\right)\Gamma(k-2i+1)} x^{\frac{d-3+2i}{2}} y^{\frac{d-3+k-2i}{2}} \mathrm{d}\cos(\varphi_y)\,\mathrm{d}\cos(\varphi_z) && (5.31)\\
&= \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} \frac{2^{-k-2i}\Gamma\left(\frac{d}{2}\right)^2\Gamma(k+1)}{\pi\Gamma\left(\frac{d-1}{2}\right)\Gamma(i+1)\Gamma\left(\frac{d-1+2i}{2}\right)\Gamma(k-2i+1)} \left(\int_{-1}^1 (1+t)^{\frac{d-3+2i}{2}} (1-t)^{\frac{d-3+k-2i}{2}} \mathrm{d}t\right)^2 && (5.32)\\
&= \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} \frac{2^{-k-2i}\Gamma\left(\frac{d}{2}\right)^2\Gamma(k+1)}{\pi\Gamma\left(\frac{d-1}{2}\right)\Gamma(i+1)\Gamma\left(\frac{d-1+2i}{2}\right)\Gamma(k-2i+1)} \cdot \left(\frac{2^{\frac{2d-4+k}{2}}\Gamma\left(\frac{d-1+2i}{2}\right)\Gamma\left(\frac{d-1+k-2i}{2}\right)}{\Gamma\left(\frac{2d-2+k}{2}\right)}\right)^2 && (5.33)\\
&= \frac{2^{2d-4}\Gamma\left(\frac{d}{2}\right)^2\Gamma(k+1)}{\pi\Gamma\left(\frac{d-1}{2}\right)\Gamma\left(\frac{2d-2+k}{2}\right)^2} \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} \frac{2^{-2i}\Gamma\left(\frac{d-1+2i}{2}\right)\Gamma\left(\frac{d-1+k-2i}{2}\right)^2}{\Gamma(i+1)\Gamma(k-2i+1)} && (5.34)
\end{aligned}$$

□

Given that we know the non-central moments of the general probability, we can make an "educated guess" on the type of distribution. From visual inspection, we observe, that the distribution looks familiar to the Beta distribution – which is a nearby result, as the pairwise cosines on the sphere are Beta-distributed as well. We can therefore assume, that the distribution of $\cos(\alpha)$ approximates a Beta distribution and conversely be approximated by one. Using the moments of the distribution of $\cos(\alpha)$, we can estimate the parameters of the Beta distribution.

5.1.5 Theorem (Method of Moments Beta Distribution). *The method of moments fit of the Beta distribution to $\frac{\cos(\alpha)+1}{2}$ where α is an angle inside a chordal triangle of points uniformly at random from the $(d-1)$-sphere is*

$$\mathrm{B}\left(\cdot; \frac{3d-4}{4} \cdot \frac{2p^4q^4 + p^2q^2}{dp^4q^4 - (d-1)}, \frac{3d-4}{4} \cdot \frac{2p^4q^4 - p^2q^2}{dp^4q^4 - (d-1)}\right) \tag{5.35}$$

where $p := \frac{\Gamma\left(\frac{d}{2}-\frac{1}{4}\right)}{\Gamma\left(\frac{d}{2}\right)}$ *and* $q := \frac{\Gamma\left(\frac{d}{2}+\frac{1}{4}\right)}{\Gamma\left(\frac{d}{2}\right)}$.

Proof. We orient the proof along the moments of the above Beta distribution. We can obtain these moments by performing a change of variable on the moments of (5.26).

$$\begin{aligned}
&\mathbb{E}\left[\left(\tfrac{\cos(\alpha)+1}{2}\right)^k\right] \\
&= 2^{-k}\sum_{i=0}^{k}\binom{k}{i}\mathbb{E}\left[\cos(\alpha)^i\right] && (5.36)\\
&= 2^{-k}\sum_{i=0}^{k}\frac{\Gamma(k+1)}{\Gamma(i+1)\Gamma(k-i+1)}\frac{2^{2d-4}\Gamma\left(\frac{d}{2}\right)^2\Gamma(i+1)}{\pi\Gamma\left(\frac{d-1}{2}\right)\Gamma\left(\frac{2d-2+i}{2}\right)^2}\sum_{j=0}^{\lfloor\frac{i}{2}\rfloor}\frac{2^{-2j}\Gamma\left(\frac{d-1+2j}{2}\right)\Gamma\left(\frac{d-1+i-2j}{2}\right)^2}{\Gamma(j+1)\Gamma(i-2j+1)} && (5.37)\\
&= \frac{2^{2d-4-k}\Gamma(k+1)\Gamma\left(\frac{d}{2}\right)^2}{\pi\Gamma\left(\frac{d-1}{2}\right)}\sum_{i=0}^{k}\sum_{j=0}^{\lfloor\frac{i}{2}\rfloor}\frac{2^{-2j}\Gamma\left(\frac{d-1+2j}{2}\right)\Gamma\left(\frac{d-1+i-2j}{2}\right)^2}{\Gamma(k-i+1)\Gamma\left(\frac{2d-2+i}{2}\right)^2\Gamma(j+1)\Gamma(i-2j+1)} && (5.38)
\end{aligned}$$

And more specifically we get

$$\mathbb{E}\left[\left(\tfrac{\cos(\alpha)+1}{2}\right)^1\right] = \frac{2^{2d-5}\Gamma\left(\frac{d}{2}\right)^4}{\pi\Gamma\left(\frac{2d-1}{2}\right)^2} + \frac{1}{2} \tag{5.39}$$

$$\mathbb{E}\left[\left(\tfrac{\cos(\alpha)+1}{2}\right)^2\right] = \frac{2^{2d-5}\Gamma\left(\frac{d}{2}\right)^4}{\pi\Gamma\left(\frac{2d-1}{2}\right)^2} + \frac{5d-4}{16(d-1)} \tag{5.40}$$

Using these moments, we can apply the method of moments to estimate the parameters a and b of the Beta distribution as

$$\begin{aligned}
a &= \frac{\pi d^4(3d-4)\left(4^d\Gamma\left(\frac{d}{2}+1\right)^4+\pi d^4\Gamma\left(d-\frac{1}{2}\right)^2\right)\Gamma\left(d-\frac{1}{2}\right)^2}{2\left(-4\cdot 2^{4d}d\Gamma\left(\frac{d}{2}+1\right)^8+4\cdot 2^{4d}\Gamma\left(\frac{d}{2}+1\right)^8+\pi^2d^9\Gamma\left(d-\frac{1}{2}\right)^4\right)} && (5.41)\\
&= \frac{(3d-4)}{4}\cdot\frac{2\Gamma\left(\frac{d}{2}+\frac{1}{4}\right)^4\Gamma\left(\frac{d}{2}-\frac{1}{4}\right)^4+\Gamma\left(\frac{d}{2}\right)^4\Gamma\left(\frac{d}{2}+\frac{1}{4}\right)^2\Gamma\left(\frac{d}{2}-\frac{1}{4}\right)^2}{d\Gamma\left(\frac{d}{2}+\frac{1}{4}\right)^4\Gamma\left(\frac{d}{2}-\frac{1}{4}\right)^4-(d-1)\Gamma\left(\frac{d}{2}\right)^8} && (5.42)\\
&= \frac{(3d-4)}{4}\cdot\frac{2\left(\frac{\Gamma\left(\frac{d}{2}-\frac{1}{4}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^4\left(\frac{\Gamma\left(\frac{d}{2}+\frac{1}{4}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^4+\left(\frac{\Gamma\left(\frac{d}{2}-\frac{1}{4}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^2\left(\frac{\Gamma\left(\frac{d}{2}+\frac{1}{4}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^2}{d\left(\frac{\Gamma\left(\frac{d}{2}-\frac{1}{4}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^4\left(\frac{\Gamma\left(\frac{d}{2}+\frac{1}{4}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)^4-(d-1)} && (5.43)
\end{aligned}$$

By using the abbreviations

$$p = \frac{\Gamma\left(\frac{d}{2}-\frac{1}{4}\right)}{\Gamma\left(\frac{d}{2}\right)} \quad \text{and} \quad q = \frac{\Gamma\left(\frac{d}{2}+\frac{1}{4}\right)}{\Gamma\left(\frac{d}{2}\right)} \tag{5.44}$$

this further simplifies to

$$a = \frac{(3d-4)}{4}\cdot\frac{2p^4q^4+p^2q^2}{dp^4q^4-(d-1)} \quad \text{and similarly} \quad b = \frac{(3d-4)}{4}\cdot\frac{2p^4q^4-p^2q^2}{dp^4q^4-(d-1)} \tag{5.45}$$

□

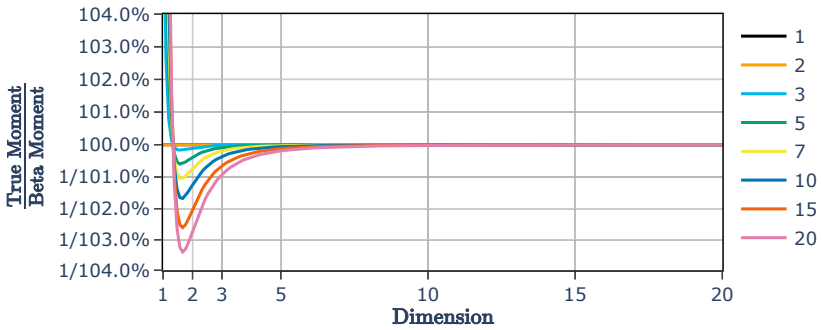


Figure 5.2: The quotient of theoretically derived moments and moments of the Beta distribution fit for cosines in chordal triangles. Values were evaluated using 500 digits of precision and integer and non-integer values of d.

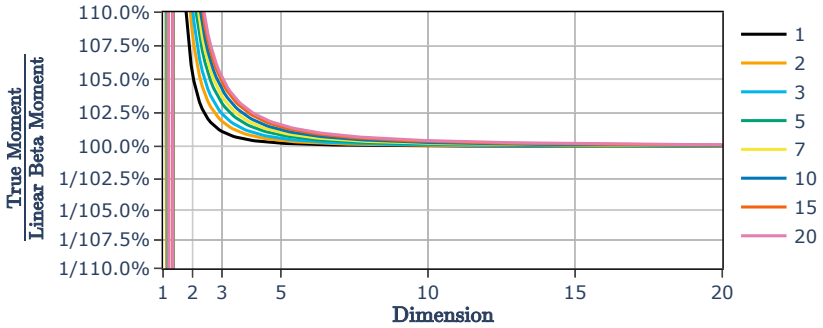


Figure 5.3: The quotient of theoretically derived moments and moments of the linearly parameterized Beta distribution fit for cosines in chordal triangles. Values were evaluated using 500 digits of precision and integer and non-integer values of d.

Notably, Theorem 5.1.5 only uses the first two moments, for which Hypothesis 5.1.3 provably holds. It is, thus, not conditioned on any unproven results. Fig. 5.2 visualizes the quotient of the theoretical moments and the moments of the fitted Beta distribution. As can be seen, the moments match almost perfectly and rapidly converge even for large k. Further, the numerical values for large k approach 0, such that the distribution is much less affected by the relative error. Since the Beta distribution is uniquely determined by its moments[2], fitting all moments closely implies that the distribution of $\cos(\alpha)$ is indeed closely matched by the Beta distribution. Further, as the parameterization of the Beta distribution reduces to linear functions in the limit of $d \to \infty$, we can further approximate the distribution of $\cos(\alpha)$ using a Beta distribution with coefficients linear in d.

[2]Billingsley [11, Theorem 30.1] provides that property for any distribution with a positive radius of convergence, i. e. with a converging moment generating function in a non-zero interval around 0. As the moment generating function of the Beta distribution with parameter t can be bound by 1 and $1 + \sum_{k=1}^{\infty} \frac{t^k}{k!} = e^t$, it trivially converges for any t and thus has an infinite radius of convergence.

5.1.6 Theorem (Approximate linearly-parameterized Beta distribution). *The method of moment fitted distribution for the cosines of angles of chordal triangles on the $(d-1)$-sphere given in Theorem 5.1.5 and*

$$\mathrm{B}\left(\cdot; \frac{3d-4}{2}, \frac{d-1}{2}\right) \tag{5.46}$$

converge as $d \to \infty$.

Proof. For this proof, we will make use of Gautschi's inequality [32] or rather a slightly derived form of it provided by the NIST[3]:

$$x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < (x+1)^{1-s}. \tag{5.47}$$

Using these inequalities, we can bound p and q as follows

$$p \in \left(\left(\frac{2}{d}\right)^{\frac{1}{4}}, \left(\frac{2}{d-2}\right)^{\frac{1}{4}}\right) \qquad \text{using } x = \frac{d}{2} - 1,\ s = \frac{3}{4} \tag{5.48}$$

$$q \in \left(\left(\frac{2d-3}{4}\right)^{\frac{1}{4}}, \left(\frac{2d+1}{4}\right)^{\frac{1}{4}}\right) \qquad \text{using } x = \frac{d}{2} - \frac{3}{4},\ s = \frac{3}{4} \tag{5.49}$$

By inserting the bounds for p and q into the formula for the shape parameters a and b, we can also bound their values as

$$a \in \left(\frac{(d-2)\left(4d+2+\sqrt{(4d+2)d}\right)}{4d}, \frac{(3d-4)\left(4d-6+\sqrt{(4d-6)(d-2)}\right)}{12(d-2)}\right) \tag{5.50}$$

$$\overset{d\to\infty}{=} \left(\frac{d\left(4d+\sqrt{4d^2}\right)}{4d} + o(d), \frac{3d\left(4d+\sqrt{4d^2}\right)}{12d} + o(d)\right) = \left(\frac{3d}{2} + o(d), \frac{3d}{2} + o(d)\right) \tag{5.51}$$

$$b \in \left(\frac{(d-2)\left(4d+2-d\sqrt{\frac{4d-6}{d-2}}\right)}{4d}, \frac{(3d-4)\left(4d-6-(d-2)\sqrt{\frac{4d+2}{d}}\right)}{12(d-2)}\right) \tag{5.52}$$

$$\overset{d\to\infty}{=} \left(\frac{d\left(4d-d\sqrt{4}\right)}{4d} + o(d), \frac{3d\left(4d-d\sqrt{4}\right)}{12d} + o(d)\right) = \left(\frac{d}{2} + o(d), \frac{d}{2} + o(d)\right) \tag{5.53}$$

One can easily see, that a scales asymptotically linear with a slope of $\frac{3}{2}$ and b scales asymptotically linear with a slope of $\frac{1}{2}$ since it applies for both relevant bounds, respectively. As $d \to \infty$, the distributions, thus, converge. The intercepts of the linear functions are not necessary for the proof since their impact on the moments vanishes for $d\to\infty$. Yet, they are useful for any practical applications to get a good approximation for small d. The bounds above are not tight, and have a difference of 1 in the limit of $d\to\infty$. The proposed intercepts were, thus, chosen empirically and analytically confirmed using the Gruntz algorithm [34] implemented in SymPy [60]. □

[3] http://dlmf.nist.gov/5.6.E4

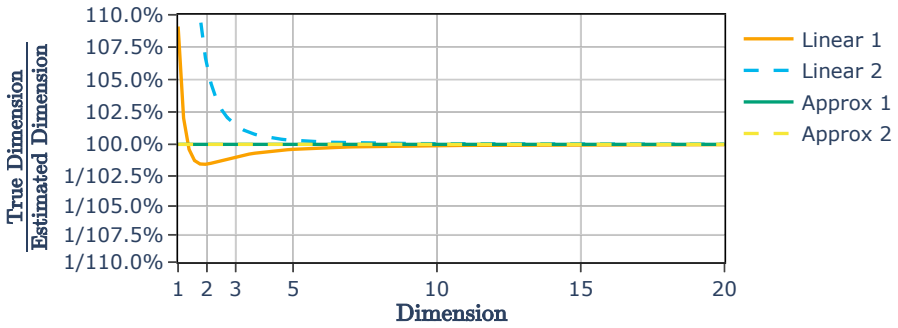


Figure 5.4: The quotient of the true dimension and the estimated dimension based on the theoretical moments. The linear variants are based on Definition 5.1.7 and the approximate variants are based on the moment equations in the proof of Theorem 5.1.5. Values were evaluated using 500 digits of precision and integer and non-integer values of d.

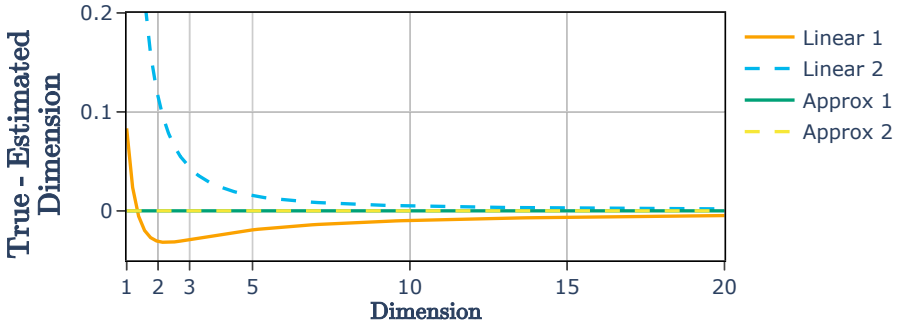


Figure 5.5: The difference between the true dimension and the estimated dimension based on the theoretical moments. The linear variants are based on Definition 5.1.7 and the approximate variants are based on the moment equations in the proof of Theorem 5.1.5. Values were evaluated using 500 digits of precision and integer and non-integer values of d.

An analogous plot to Fig. 5.2 for the linearly parameterized Beta distribution is shown in Fig. 5.3. While the approximation error is larger, the moments still converge rapidly. In practice, the approximation error is negligible for $d \geq 3$. For the first two moments, we can observe a relative error of below 2.5%.

We are now interested in using the moments and distribution parameters for ID estimation. While using the theoretical moments of (5.26) is possible, it would require a (binary) search over possible values of d. The approximate, yet quite complicated, moments of the approximate Beta distribution also do not allow for obtaining a closed-form solution for d by inverting the method of moments. The linear approximation, however, can be used to provide a method of moments estimator for d using the first two non-central moments of the Beta distribution. That approach gives two closed-form solutions for d which we call the $\mathrm{LID}_{\mathrm{ALC}}$ estimators type 1 and 2, one for each parameter of the linearly-parameterized Beta distribution. The formulae are the solutions to equating the

linear parameterization with the method of moment estimates obtained by inverting the equations for the first and second non-central moments of the Beta distribution.

5.1.7 Definition ($\mathbf{LID_{ALC}}$). For normalized samples X uniformly at random from a spherically symmetric distribution, the $\mathrm{LID}_{\mathrm{ALC}}$ estimator types 1 and 2 are

$$\mathrm{LID}_{\mathrm{ALC}1}(X) := \frac{2}{3} \cdot \frac{\mu_1^2 + \mu_1\mu_2 - 2\mu_2}{\mu_1^2 - \mu_2} = \frac{2a+4}{3} \tag{5.54}$$

$$\mathrm{LID}_{\mathrm{ALC}2}(X) := \frac{3\mu_1^2 - 2\mu_1\mu_2 - 2\mu_1 + \mu_2}{\mu_1^2 - \mu_2} = 2b + 1 \tag{5.55}$$

where $a = \frac{\mu_1(\mu_2-\mu_1)}{\mu_1^2-\mu_2}$, $b = \frac{(1-\mu_1)(\mu_2-\mu_1)}{\mu_1^2-\mu_2}$, and μ_1 and μ_2 are the first and second non-central moment of $\frac{\cos(\measuredangle_{yxz})+1}{2}$ over all pairwise different $x, y, z \in X$ respectively.

Similar results can in principle be obtained just by using one of the moments, but we observed, that estimates using more than one moment are more stable. Although these two types visually appear quite different, the resulting estimates are strongly correlated in practice. Fig. 5.4 visualizes the estimated intrinsic dimensionality of the proposed $\mathrm{LID}_{\mathrm{ALC}}$ estimators compared to the estimates obtained from the moments of the approximate Beta distribution (5.35) based on the theoretical moments (5.26). Since the theoretical moments are exact, the estimated dimensionality should be identical to d, if the binary search is successful. As can be seen, the increased cost of the theoretically better moments is barely justified by better estimates, as the relative error of the linear approximation is below 7.5% for $d \geq 2$ and below 1% for $d \geq 4$. For integer parameterization of downstream tasks, any error below $\frac{1}{2d}$ is acceptable as it results in misjudging the true dimensionality by less than 1. Fig. 5.5 thus visualizes the difference in d and estimated dimensionality. While $\mathrm{LID}_{\mathrm{ALC}2}$ has a larger error for $d \leq 3$, it converges faster than $\mathrm{LID}_{\mathrm{ALC}1}$. The error on both variants is practically negligible, though.

In contrast to the $\mathrm{LID}_{\mathrm{ABID}}$ estimator, whose angles are focused on the center of the sphere, the $\mathrm{LID}_{\mathrm{ALC}}$ estimators do not implicitly mirror samples on the sphere. Consequentially, the $\mathrm{LID}_{\mathrm{ALC}}$ estimators can be evaluated both on mirrored and non-mirrored neighborhoods to detect skewed neighborhoods. This can theoretically be used to, e.g. detect outliers by significantly smaller estimates with mirrored neighbors compared to non-mirrored neighbors. Another difference between $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{ALC}}$ is that the vector norms now do matter. That is, the estimator does not trivially extend to other spherically symmetric distributions but is tailored to the uniform distribution on the sphere. A natural question now is, if the $\mathrm{LID}_{\mathrm{ALC}}$ estimator can be extended to uniform distributions in hyperballs. That would allow the usage of the norms and could theoretically increase the accuracy of the estimator. Unfortunately, the required integrals for the angle distribution of triangles sampled uniformly at random from hyperballs, are even more complicated

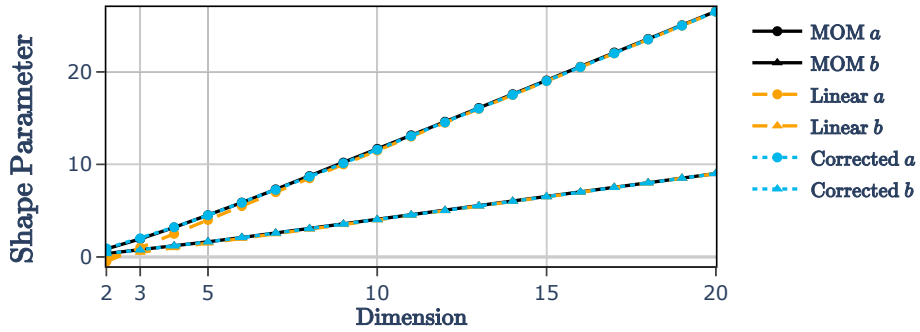


Figure 5.6: The shape parameters for a fitted Beta distribution for cosines of triangles in the d-ball. The method of moment traces are based on 5 000 000 random cosines while the other traces use parameters computed via the definition in Hypothesis 5.1.8 either without or with the exponential correction term.

than the ones for the hypersphere. However, we have a strong intuition, that the distribution should converge to a Beta distribution as well and that the parameters of the Beta distribution should be approximately linear in the dimensionality of the hyperball. By empirical evaluation displayed in Fig. 5.6, we obtained linear parameters and further observed, that for small dimensions, the parameters can be improved by adding an exponential term to the parameters that vanishes for larger d.

5.1.8 Hypothesis (Approximate Beta distribution for balls). *The cosines of angles of triangles whose vertices x, y, z are sampled uniformly at random from the d-ball are approximately Beta distributed, i.e. the distribution of $\frac{\cos(\alpha)+1}{2}$ where $\alpha = \angle_{yxz}$ and*

$$\mathrm{B}\left(\cdot; \frac{3d-7}{2} + 2^{-\frac{d-3}{2}}, \frac{d-2}{2} + 2^{-\frac{d+1}{2}}\right) \tag{5.56}$$

converge as $d \to \infty$. Since the exponential terms vanish rapidly, the distributions also converge on the linearly-parameterized Beta distribution without these terms.

As an interesting side note, the linear parameterization equals the linear approximation for spherical samples in one dimension less. Intuitively, one would expect the opposite to be true, that the ball angles correspond to sphere angles in one dimension more, since the LID within these volumes is equal. Alas, the opposite is true and we do not explain this apparent paradox. The formulaic similarity in our eyes, however, lends some credibility to the method, as we would intuitively expect a similar distribution. By inverting the parameters in Hypothesis 5.1.8 using the method of moment equations for the Beta distribution, we can derive $\mathrm{LID}_{\mathrm{ALC}}$-style estimators for hyperballs.

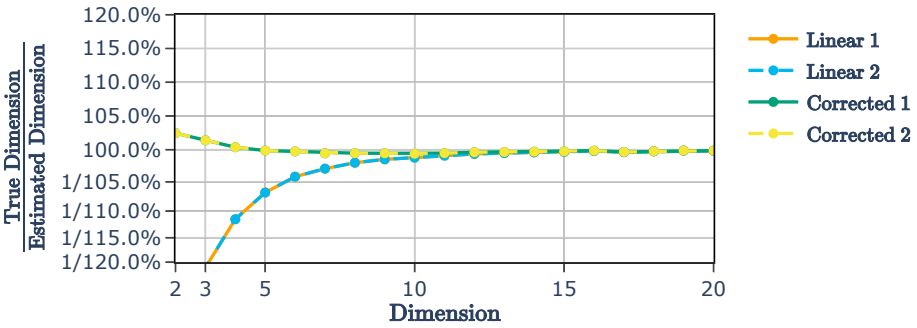


Figure 5.7: The estimation error for the dimensionality using the linear $\mathrm{LID}_{\mathrm{BALL}}$ or the exponentially corrected $\mathrm{LID}_{\mathrm{BALE}}$ estimators. All traces are based on 5 000 000 random cosines.

5.1.9 Definition (Hyperball $\mathbf{LID}_{\mathbf{ALC}}$). For samples X uniformly at random from a hyperball, the Hyperball $\mathrm{LID}_{\mathrm{ALC}}$ estimator with linear parameters ($\mathrm{LID}_{\mathrm{BALL}}$) types 1 and 2 are

$$\mathrm{LID}_{\mathrm{BALL}1}(X) := \frac{5\mu_1^2 + 2\mu_1\mu_2 - 7\mu_2}{3(\mu_1^2 - \mu_2)} = \frac{2a+7}{3} \tag{5.57}$$

$$\mathrm{LID}_{\mathrm{BALL}2}(X) := \frac{2\mu_1(2\mu_1 - \mu_2 - 1)}{\mu_1^2 - \mu_2} = 2b + 2 \tag{5.58}$$

where $a = \frac{\mu_1(\mu_2-\mu_1)}{\mu_1^2-\mu_2}$, $b = \frac{(1-\mu_1)(\mu_2-\mu_1)}{\mu_1^2-\mu_2}$, and μ_1 and μ_2 are the first and second non-central moment of $\frac{\cos(\angle_{yxz})+1}{2}$ over all pairwise different $x, y, z \in X$ respectively.
The Hyperball $\mathrm{LID}_{\mathrm{ALC}}$ estimator with exponentially corrected parameters ($\mathrm{LID}_{\mathrm{BALE}}$) types 1 and 2 are

$$\mathrm{LID}_{\mathrm{BALE}1}(X) := \frac{2a+7}{3} + \frac{2}{\log(2)} W\left(-2^{-\frac{a-1}{3}} \frac{\log(2)}{3}\right) \tag{5.59}$$

$$\mathrm{LID}_{\mathrm{BALE}2}(X) := 2b + 2 + \frac{2}{\log(2)} W\left(-2^{-\frac{2b+3}{2}} \log(2)\right) \tag{5.60}$$

where W is the Lambert W function defined by $x = W(xe^x)$ and a, b, μ_1, and μ_2 are as above. The $\mathrm{LID}_{\mathrm{BALE}}$ equations were derived using SymPy [60] and individual steps are omitted for brevity.

As can be seen in Fig. 5.7, the ball-based estimators both converge to the true dimensionality as d grows. The exponentially corrected $\mathrm{LID}_{\mathrm{BALE}}$ estimators perform better than the "cheaper" $\mathrm{LID}_{\mathrm{BALL}}$ variants, yet, the difference becomes negligible for $d \geq 10$. The error of the ball-based estimators, however, is larger than the theoretically more sound $\mathrm{LID}_{\mathrm{ALC}}$ estimators. Differences in $\mathrm{LID}_{\mathrm{ALC}}$ and $\mathrm{LID}_{\mathrm{BALL}}$ estimates might though be important to detect certain geometric properties of the data, such as too large "localities", where the projection to the sphere discards too much information. Similarly, we can again choose whether or not to mirror neighborhoods before evaluation, which depends on the specific

application and may reveal additional information. Mirroring neighborhoods can be semantically beneficial and reduce the variance of the estimates, although it comes at the risk of perturbing the LID estimate. All of the LID estimator variants proposed here have an increased computational cost compared to $\mathrm{LID}_{\mathrm{ABID}}$ – cubic instead of quadratic – but we believe that it leads to more robust moment estimates and thereby to better LID estimates. This allows us to inspect smaller neighborhoods with a decreased risk of breaking the "locality assumption" of LID estimation on k-nearest neighbors. To validate the quality of these estimators, we will now consider some empirical results beyond the theoretical considerations.

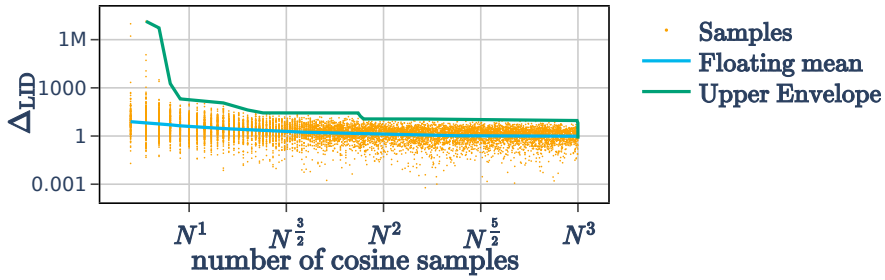
5.2 Empirical Evaluation

In this section, we empirically evaluate the $\mathrm{LID}_{\mathrm{ALC}}$ family of estimators in terms of their convergence behavior and the quality of their estimates compared to other estimators. We will first consider the convergence behavior both in terms of the number of samples and the number of chordal angles used. We will then compare the quality of the estimates in ideal, biased, and skewed settings, and on artificial and real datasets to other estimators such as $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$.

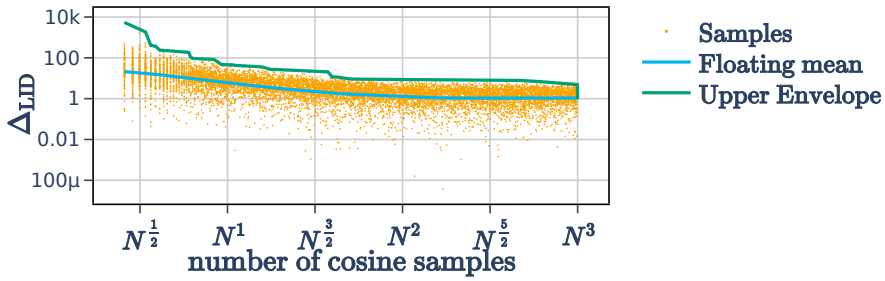
5.2.1 Convergence

As mentioned before, we can at most consider a number of chordal angles cubic in the number of samples in the neighborhood, but this implies an equivalent computational cost. It may thus be more useful to rather use a fixed number of random angles as we are solely interested in the first and second moments of their cosines. These moments should be sufficiently converged on a random subset. We choose these random cosines by selecting indices uniformly at random in triplets and discarding any triplet that includes an index more than once. This possibly includes duplicates, but the implementation is much simpler and faster than a more sophisticated selection process. As can be seen in Fig. 5.8, the error of the $\mathrm{LID}_{\mathrm{ALC}}$ type 1 estimator converges quickly with the number of cosines used. The average error for $N = 3d$ neighbors drops below 1 for approximately $N^{\frac{3}{2}}$ cosines. Further, different neighborhood sizes of the same linear multiplicity of their respective dimensionality behave mostly similarly in terms of this convergence for $N^{\frac{3}{2}}$ and more cosines. In terms of computational complexity, it thus likely suffices to use between $N^{\frac{3}{2}}$ and N^2 random cosines for the $\mathrm{LID}_{\mathrm{ALC}}$ type 1 estimator when having $N = 3d$ neighbors and around $N^{\frac{5}{2}}$ when having $N = 2d$ neighbors. The same result extends to all other estimators of this family.

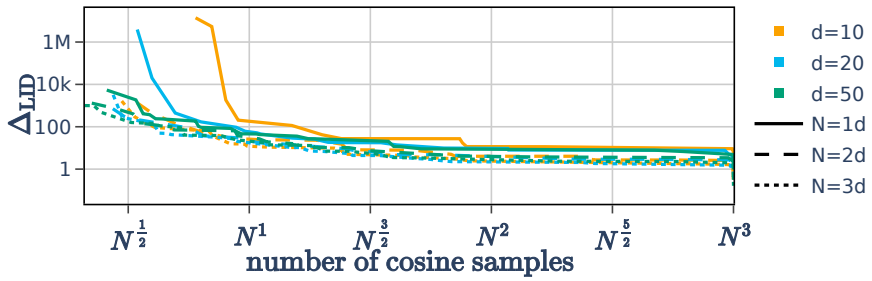
Knowing, how the number of random cosines used in computation affects the estimate quality, we can inspect how the number of neighbors affects the convergence behavior for



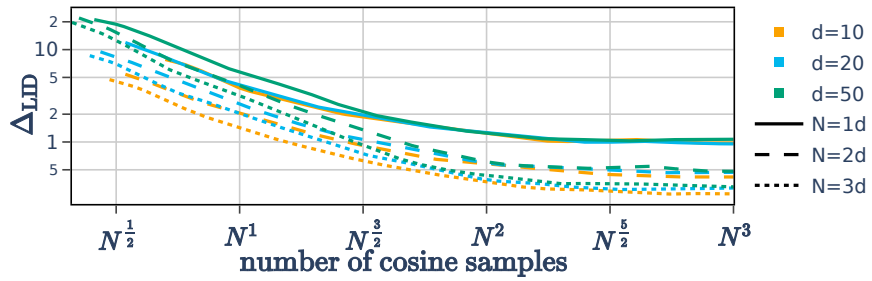
(**a**) $d = 10, N = 10$



(**b**) $d = 50, N = 50$

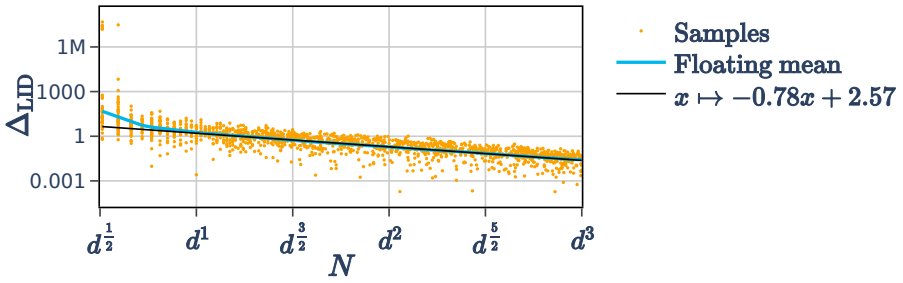


(**c**) Upper envelopes

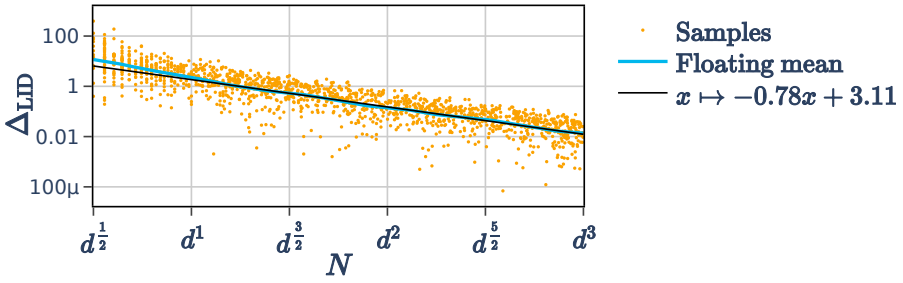


(**d**) Floating means

Figure 5.8: Convergence behavior of the $\mathrm{LID}_{\mathrm{ALC}}$ type 1 estimator for different numbers of random cosines as compared to the $\mathrm{LID}_{\mathrm{ALC}}$ value using all cosines. The underlying distributions were univariate normal distributions with varying dimension d. Each data point in 5.8a and 5.8b represents a single run, i. e. a different random sample of size N. 5.8c and 5.8d show the upper envelopes and floating means of the errors over the number cosines.



(a) $d = 15, N^{\frac{3}{2}}$ cosines



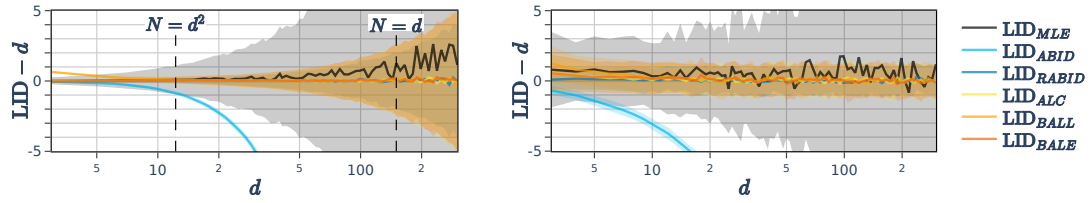
(b) $d = 25, N^{\frac{3}{2}}$ cosines

Figure 5.9: Convergence behavior of the $\mathrm{LID}_{\mathrm{ALC}}$ type 1 estimator for different neighborhood sizes as compared to the true dimensionality. The underlying distributions were univariate normal distributions with varying dimension d. The used number of random cosines for estimation are annotated to the plots. Each data point represents a single run, i.e. a different random sample of size N. Aside from the floating mean, a linear fit in the log-log-space is shown to indicate the empirical power of the convergence.

varying numbers of dimensions and cosines. Fig. 5.9 displays two exemplary plots for the $\mathrm{LID}_{\mathrm{ALC}}$ type 1 estimator. The estimates for $N < d$ are quite unstable but they rapidly converge to an error below 1 for $N \geq d^{\frac{3}{2}}$. Using a linear fit in log-log-space, we obtain empirical powers for the convergence rates by taking the negative reciprocal of the slope. The linear fits were computed for the almost linear part of the error curve beyond $N > d$. The steeper slope left of that point would induce a lower convergence rate, thus the fit provides an upper bound for the convergence rate. For the specific cases in Fig. 5.9, we find that the convergence rate is approximately 1.28 for $d = 15$ and $d = 25$ when using $N^{\frac{3}{2}}$ cosines. This rate is well below the convergence rate of 2 for $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$, which has been empirically supported in Section 3.4. As this rate depends on the number of cosines used and is merely empirical, thus likely worse for higher dimensions, we repeated that process multiple times and collected the results in Table 5.1. The convergence rates for $d = 15$ and $d = 25$ are almost identical, suggesting that any trend towards higher dimensions is likely not significant beyond $d = 25$. Further, the convergence rate for N^2 cosines is near linear, which coincides with the results in Fig. 5.8. When investing

Cosines	$d = 5$	$d = 15$	$d = 25$
N^1	1.47	1.79	1.79
$N^{1.25}$	1.30	1.54	1.59
$N^{1.5}$	1.06	1.28	1.28
$N^{1.75}$	0.97	1.08	1.12
N^2	0.94	1.04	1.03

Table 5.1: Empirical convergence rates for the $\mathrm{LID}_{\mathrm{ALC}}$ type 1 estimator for different numbers of dimensions and random cosines.



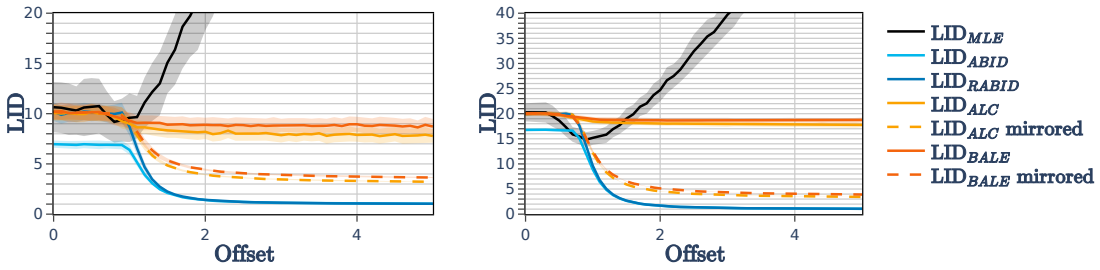
(**a**) $N = 150$ (**b**) $N = 2d$

Figure 5.10: Mean error of different LID estimators on uniform ball distributions in different dimensions. The LID estimates were computed at the center of the balls and we either used a constant number of neighbors or a number linear in the dimensionality. The shaded areas indicate the error range between plus and minus one standard deviation around the mean. All estimators of the $\mathrm{LID}_{\mathrm{ALC}}$ family were evaluated using N^2 random cosines. For each d, we performed 1000 runs with different random samples.

a computational effort squared in the number of neighbors, smaller neighborhoods down to size $N \in \mathcal{O}(d)$ should give accurate estimates for the $\mathrm{LID}_{\mathrm{ALC}}$ type 1 estimator. Again, these results extend to the other estimators of this family. We can so far conclude, that the estimators of the $\mathrm{LID}_{\mathrm{ALC}}$ family converge in much smaller neighborhoods than the $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$ estimators. This does not necessarily speak for the practical quality of the estimates, but any errors are much less likely due to too-small neighborhoods. Since $\mathrm{LID}_{\mathrm{BALE}}$ is a heuristically improved version of $\mathrm{LID}_{\mathrm{BALL}}$, we will from here on omit $\mathrm{LID}_{\mathrm{BALL}}$ in the experiments. We will further only consider the type 1 estimators of the $\mathrm{LID}_{\mathrm{ALC}}$ family, as the type 2 estimators do not significantly diverge from their type 1 counterparts in practice.

5.2.2 Ideal Case

Ideally, each neighborhood is distributed according to an intrinsically dimensional uniform ball distribution. Fig. 5.10 displays the results of this case under varying dimensionality. For a fixed number of neighbors, the $\mathrm{LID}_{\mathrm{ABID}}$ estimator starts underestimating the



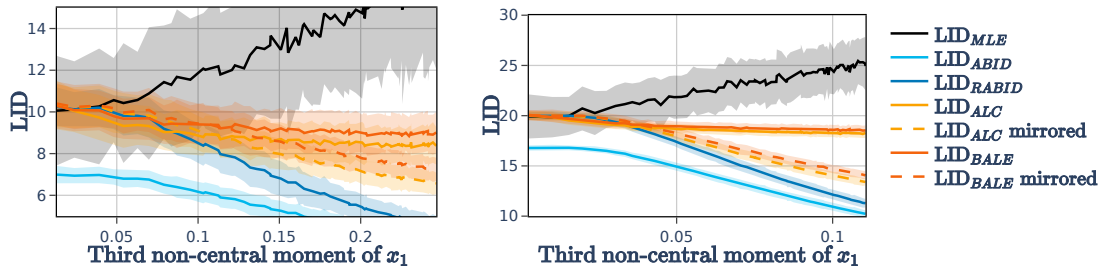
(**a**) $d = 10, N = 20$ (**b**) $d = 20, N{=}100d$

Figure 5.11: Mean LID estimates on uniform unit-ball distributions centered at $(x_1, 0, \ldots, 0)$ for varying x_1. The LID estimates were computed at the origin using the closest N of 10 000 samples drawn from the ball as a neighborhood. The shaded areas indicate the error range between plus and minus one standard deviation around the mean. All estimators of the $\mathrm{LID}_{\mathrm{ALC}}$ family were evaluated using N^2 random cosines. For each x_1, we performed 150 runs with different random samples.

dimensionality due to its bounding property. The $\mathrm{LID}_{\mathrm{MLE}}$ estimator, whilst being mostly correct on average, has an increasing standard deviation in its estimates. The estimators of the $\mathrm{LID}_{\mathrm{ALC}}$ family as well as the $\mathrm{LID}_{\mathrm{RABID}}$ estimator remain near exact in the mean with a much lower, yet also increasing standard deviation. Notably, the $\mathrm{LID}_{\mathrm{BALL}}$ estimator has an increased error for $d < 10$ due to the approximation error for small d in the computation of the $\mathrm{LID}_{\mathrm{BALL}}$ estimator. This error does not exceed 1 and is thus negligible in practice. The very close and slightly better performance of the $\mathrm{LID}_{\mathrm{RABID}}$ estimator is quite surprising, as it is limited to the same square number of cosines available to the $\mathrm{LID}_{\mathrm{ABID}}$ estimator. When using $N = 2d$ neighbors, all estimators of the $\mathrm{LID}_{\mathrm{ALC}}$ family and $\mathrm{LID}_{\mathrm{RABID}}$ have an almost constant standard deviation of approximately 1 whereas the standard deviation of the $\mathrm{LID}_{\mathrm{MLE}}$ estimator still increases for larger d. This is congruent with the better convergence behavior of the $\mathrm{LID}_{\mathrm{ALC}}$ family compared to the $\mathrm{LID}_{\mathrm{MLE}}$ estimator. The $\mathrm{LID}_{\mathrm{RABID}}$ estimator appears to have the same convergence behavior as the $\mathrm{LID}_{\mathrm{ALC}}$ family, opposite to the $\mathrm{LID}_{\mathrm{ABID}}$ estimator. The relaxation of $\mathrm{LID}_{\mathrm{RABID}}$ to $\mathrm{LID}_{\mathrm{ABID}}$ in terms of the spectral norm of the covariance matrix for one allows for a better analytical understanding but also hinders the convergence behavior.

5.2.3 Biased and Skewed Cases

Skewness and bias are two ways for neighborhoods to violate the ideal uniform ball distribution. In the case of bias, the neighborhood is shifted away from the point of interest and thus inspected "from the outside". This case is of special interest with regard to real-world data, which contains many samples "on the outside" due to the natural boundaries of finite data. In the case of skewness, the distribution density is locally non-uniform. While



(**a**) $d = 10, N = 20$ (**b**) $d = 20, N$=100

Figure 5.12: Mean LID estimates on unit-ball distributions uniform in all but the first dimension. Samples were drawn from the unit cube $[0, 1]^d$ and the first coordinate was transformed by $x_1 \mapsto x_1^c$. All samples outside the inscribed hypersphere were discarded. The LID estimates were computed at the center of the inscribed hypersphere. The shaded areas indicate the error range between plus and minus one standard deviation around the mean. All estimators of the LID_{ALC} family were evaluated using N^2 random cosines. We performed 11 250 runs with different random samples in total.

a linear change in density does not affect e. g. LID_{MLE} or LID_{ABID} (the increased density on one side cancels out with the decreased density on the other side), a non-linear change in density does.

Fig. 5.11 shows the mean LID estimates for biased neighborhoods. We investigated different distances between the point of interest and the center of the ball. At an offset of 1, the point of interest lies on the boundary of the ball. In that case, LID_{MLE}, $\text{LID}_{\text{RABID}}$, and the mirrored variants of the LID_{ALC} family agree on a LID lower than d. The mirrored variants use the reflections of neighbors x at the point of interest p, i. e. $p - (x - p)$, as additional neighbors, similar to $\text{LID}_{\text{RABID}}$. Beyond an offset of 1, the estimators diverge, as the expansion-based LID_{MLE} increases, the mirrored estimators and $\text{LID}_{\text{RABID}}$ decrease and the non-mirrored LID_{ALC} variants remain near constant. The LID_{ABID} estimator underestimates the LID for all offsets, as the neighborhood size is too small but also drops for larger offsets. We propose two possible interpretations as to what the "true" LID should be: Either it should approach $d - 1$, as the surface of the ball is $d - 1$ dimensional, or it should approach 1, as a very far neighborhood is indistinguishable from a single point. The non-mirrored LID_{ALC} variants appear to robustly follow the first interpretation, whereas LID_{ABID} and $\text{LID}_{\text{RABID}}$ follow the second. The mirrored LID_{ALC} variants lie somewhere in between and the LID_{MLE} estimator is not interpretable in this case but can be explained by all neighbors "appearing" at an increasingly similar distance. This behavior would be expected from much higher-dimensional data, whereby the LID_{MLE} estimates increase. The unintuitive behavior of LID_{MLE} can in certain tasks be useful though, as the large increase has been used as an indicator for outliers in the past [48]. A combi-

nation with the $\mathrm{LID}_{\mathrm{ALC}}$ variants may improve that approach by differentiating between truly increased LID and outliers.

Fig. 5.12 shows the mean LID estimates for skewed neighborhoods. Here, we produced a non-linear change in density by applying a varying power to the first coordinate of the $[0,1]^d$ cube and retaining only samples inside the inscribed ball. The distribution of this ball is thus non-linearly skewed along one axis. The estimates were computed at the center of this ball. Instead of the power, we display the resulting third non-central moment of the first axis in the plot, which we consider better interpretable. We can observe similar results to the bias test, as $\mathrm{LID}_{\mathrm{MLE}}$ rises, the mirrored estimators and $\mathrm{LID}_{\mathrm{RABID}}$ fall and the non-mirrored $\mathrm{LID}_{\mathrm{ALC}}$ variants slowly approach a slightly lower value. In the limit, all samples aggregate in a small cap where one of the dimensions practically vanishes. That again induces the same interpretations as above but with a different scale and reasoning. Thus, the same conclusions can be drawn from this test as from the bias test: Non-mirrored $\mathrm{LID}_{\mathrm{ALC}}$ variants realize one interpretation, and mirrored $\mathrm{LID}_{\mathrm{ALC}}$, $\mathrm{LID}_{\mathrm{ABID}}$, and $\mathrm{LID}_{\mathrm{RABID}}$ realize the other. As the mirrored variants mimic $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$, we will omit the mirrored variants in the following experiments.

5.2.4 Artificial Datasets

The previous results in this section required the application of the estimators to single neighborhoods in numerous datasets. Whilst giving a good overview of the estimators' behavior, it inhibited a comparison with the recent LID estimator using approximate Likelihood ($\mathrm{LID}_{\mathrm{LIDL}}$) by Tempczyk et al. [82]. $\mathrm{LID}_{\mathrm{LIDL}}$ uses a pre-trained density model of the dataset to estimate the intrinsic dimensionality based on an expansion-based principle familiar to $\mathrm{LID}_{\mathrm{MLE}}$. Tempczyk et al. claim an outstanding performance of the estimator on very high-dimensional datasets and gave empirical evidence for that claim on artificial datasets [82]. It is thus of interest to compare the $\mathrm{LID}_{\mathrm{LIDL}}$ estimator to the $\mathrm{LID}_{\mathrm{ALC}}$ family of estimators, but training one density model per neighborhood as in the previous experiments is not feasible as the training times are often in the order of minutes or higher. We will now instead consider a limited set of random datasets and consider the distribution of LID estimates within each dataset. Thus, not every neighborhood is ideally distributed as some of the samples necessarily lie on the boundary of the dataset but we only need to compute the density model for $\mathrm{LID}_{\mathrm{LIDL}}$ once for all neighborhoods within a dataset. This perspective further gives insight into the robustness of the estimators in practice for bounded known geometries before we continue with real datasets of unknown geometry in the next section. As in Chapter 3, we will use some of the datasets provided by Rozza et al. [71] and also consider high-dimensional balls and univariate normal distributions as done by Tempczyk et al. [82]. Contrary to the other estimators, $\mathrm{LID}_{\mathrm{LIDL}}$ does not require a neighborhood size but several radii for the expansion analy-

50 dimensions

Dataset	$\mathrm{LID}_{\mathrm{ALC}}$	$\mathrm{LID}_{\mathrm{BALE}}$	$\mathrm{LID}_{\mathrm{LIDL}}$
gaussian	38.91 ± 1.64	39.07 ± 1.60	48.97 ± 0.69
ball	47.81 ± 1.15	48.71 ± 1.17	48.99 ± 0.39
cube	43.29 ± 1.26	43.64 ± 1.32	47.85 ± 0.56
Dataset	$\mathrm{LID}_{\mathrm{ABID}}$	$\mathrm{LID}_{\mathrm{RABID}}$	$\mathrm{LID}_{\mathrm{MLE}}$
gaussian	5.56 ± 1.39	5.95 ± 1.58	28.12 ± 2.83
ball	8.24 ± 0.62	8.89 ± 0.75	29.22 ± 2.84
cube	6.97 ± 1.33	7.40 ± 1.55	28.03 ± 2.71

100 dimensions

Dataset	$\mathrm{LID}_{\mathrm{ALC}}$	$\mathrm{LID}_{\mathrm{BALE}}$	$\mathrm{LID}_{\mathrm{LIDL}}$
gaussian	76.11 ± 3.03	77.02 ± 2.98	97.68 ± 1.27
ball	96.78 ± 2.37	98.65 ± 2.35	97.19 ± 0.73
cube	87.06 ± 2.47	86.99 ± 2.57	96.38 ± 0.94
Dataset	$\mathrm{LID}_{\mathrm{ABID}}$	$\mathrm{LID}_{\mathrm{RABID}}$	$\mathrm{LID}_{\mathrm{MLE}}$
gaussian	4.88 ± 0.80	5.15 ± 0.89	42.33 ± 4.18
ball	6.68 ± 0.23	7.09 ± 0.25	45.81 ± 4.37
cube	5.82 ± 0.72	6.14 ± 0.80	43.81 ± 4.22

Figure 5.13: Mean ± standard deviation of various LID estimators on artificial datasets with 50 (left) and 100 (right) dimensions and 10 000 samples based on 100 neighbors.

sis and hyperparameters for the density model. We based the radii on the median distance of 2 000 random samples to their 1- to k-nearest neighbors. The used radii were then $0.001, 0.02575, 0.0505, 0.07525$, and 0.1 times this median distance. Aside from basing the $\mathrm{LID}_{\mathrm{LIDL}}$ estimates on roughly the same scale as the other estimators, they produced reliably good results in our experiments. This approach, however, alleviates the "benefit" of not relying on sample neighborhoods, as claimed by its authors [82]. As for the density model, we used the standard parameters provided by their software package.[4]

At first, we compared the quality of the LID estimates on simple artificial datasets such as univariate normal distributions (`gaussian`), uniform hyperballs (`ball`), and uniform hypercubes (`cube`). The results are displayed in Fig. 5.13. Overall, the $\mathrm{LID}_{\mathrm{LIDL}}$ estimator has an outstanding performance on these datasets in recovering the ID of the distributions. On the `ball` sets, where the LID should agree with the ID in most cases, the $\mathrm{LID}_{\mathrm{BALE}}$ estimator has a near-exact result, closely followed by $\mathrm{LID}_{\mathrm{ALC}}$. The stark underestimation of these estimators on `gaussian` and `cube` datasets is likely due to the multidimensional skewness of the local neighborhoods. It is therefore not necessarily "wrong" but rather a different interpretation of the LID in these cases. Contrary, $\mathrm{LID}_{\mathrm{LIDL}}$ appears to be agnostic of the local geometry, which may or may not be beneficial depending on the task. The estimators $\mathrm{LID}_{\mathrm{ABID}}$, $\mathrm{LID}_{\mathrm{RABID}}$, and $\mathrm{LID}_{\mathrm{MLE}}$ do not recognize the high dimensional distributions. This is most likely due to the insufficient neighborhood sizes. These exact observations were reproducible for multiple other dimensionalities and neighborhood sizes omitted here for brevity.

Next, we considered the artificial datasets provided by Rozza et al. [71]. These datasets are samples from different dimensional manifolds embedded in different dimensional spaces. Some of these cases are more complicated manifolds compared to the simple ones

[4] https://github.com/opium-sh/lidl

Dataset	ID	d	LID_{ALC}		LID_{BALE}		LID_{LIDL}	
			mean	**log mean**	**mean**	**log mean**	**mean**	**log mean**
m1	10	11	1.61%	1.62%	3.59%	3.57%	8.59%	8.57%
m2	3	5	5.49%	6.08%	4.42%	4.71%	65.84%	65.45%
m3	4	6	9.59%	10.88%	8.04%	8.65%	42.47%	42.19%
m4	4	8	7.89%	8.02%	9.92%	9.72%	95.04%	94.85%
m5	2	3	4.74%	5.00%	2.87%	2.99%	38.13%	37.20%
m6	6	36	13.11%	12.86%	20.34%	19.93%	242.01%	241.18%
m7	2	3	5.39%	5.55%	3.16%	3.26%	45.50%	44.77%
m8	12	72	22.00%	21.82%	27.71%	27.54%	171.74%	171.46%
m9	20	20	13.90%	16.20%	11.42%	12.95%	4.10%	4.29%
m10a	10	11	14.65%	17.36%	10.83%	12.25%	1.81%	1.83%
m10b	17	18	14.63%	17.20%	11.77%	13.40%	1.67%	1.70%
m10c	24	25	14.03%	16.38%	11.84%	13.48%	1.41%	1.42%
m11	2	3	8.50%	8.99%	2.97%	3.06%	49.96%	49.80%
m12	20	20	20.75%	26.38%	17.74%	21.70%	1.78%	1.82%
m13	1	13	93.60%	93.55%	1.16%	1.15%	1073.52%	1070.82%
overall			16.66%	16.40%	9.85%	10.30%	122.90%	64.77%
Dataset	**ID**	**d**	LID_{ABID}		LID_{RABID}		LID_{MLE}	
			mean	**log mean**	**mean**	**log mean**	**mean**	**log mean**
m1	10	11	2.73%	2.82%	6.74%	6.72%	11.24%	12.93%
m2	3	5	2.97%	3.12%	1.79%	1.87%	9.44%	10.34%
m3	4	6	8.97%	10.04%	9.34%	10.18%	14.51%	16.26%
m4	4	8	11.11%	11.18%	14.07%	14.01%	10.72%	11.12%
m5	2	3	1.03%	1.06%	1.16%	1.17%	8.54%	8.99%
m6	6	36	10.93%	11.97%	13.01%	13.65%	16.54%	16.20%
m7	2	3	1.94%	1.92%	2.39%	2.37%	9.69%	10.49%
m8	12	72	45.15%	86.47%	41.80%	76.16%	17.25%	16.74%
m9	20	20	56.10%	135.23%	52.03%	116.68%	27.79%	39.17%
m10a	10	11	24.69%	34.77%	19.22%	25.91%	18.21%	22.87%
m10b	17	18	48.69%	100.86%	44.00%	85.14%	25.15%	34.26%
m10c	24	25	64.76%	192.79%	61.64%	170.66%	30.57%	44.72%
m11	2	3	2.13%	2.13%	2.83%	2.82%	8.81%	9.25%
m12	20	20	66.93%	219.56%	64.64%	201.36%	22.65%	30.02%
m13	1	13	1.22%	1.21%	1.24%	1.22%	7.88%	8.01%
overall			23.29%	40.93%	22.39%	37.70%	15.93%	18.89%

Figure 5.14: Mean and log mean error of various LID estimators on artificial datasets with 10 000 samples each evaluated using 100 neighbors each.

considered before. Since the ID of these datasets varies, we aggregated the results as the mean absolute relative error of estimates x_i and true ID δ:

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_i-\delta}{\delta}\right|. \tag{5.61}$$

We also considered the log mean error, which evaluates the mean error in log-space:

$$\exp\left(\frac{1}{N}\sum_{i=1}^{N}\left|\log\left(\frac{x_i}{\delta}\right)\right|\right)-1. \tag{5.62}$$

The log mean error is less sensitive to outliers and rather gives the average error in orders of magnitude. Both of these error scores are displayed in Fig. 5.14. On the datasets that resemble the primitive geometries considered before (`m1`, `m2`, `m9-m10c`, `m12`), the $\mathrm{LID}_{\mathrm{LIDL}}$ estimator again tends to outperform all other estimators except for the affine space in `m2`. On the non-linear manifolds (`m3-m8`, `m11`) $\mathrm{LID}_{\mathrm{LIDL}}$ vastly overestimates the ID. The $\mathrm{LID}_{\mathrm{ALC}}$, $\mathrm{LID}_{\mathrm{BALE}}$, and $\mathrm{LID}_{\mathrm{MLE}}$ estimators also overestimate the ID but to a lesser extent. $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ also tend to slightly overestimate the ID but drop off for higher dimensions (`m8`). On the curve (`m13`) set, all but $\mathrm{LID}_{\mathrm{ALC}}$ and $\mathrm{LID}_{\mathrm{LIDL}}$ estimate the true ID well. The $\mathrm{LID}_{\mathrm{ALC}}$ estimator appears to recognize the curvature of the manifold as an additional dimension. The $\mathrm{LID}_{\mathrm{LIDL}}$ estimator instead assumes a full-dimensional space. Considering the overall performance, the $\mathrm{LID}_{\mathrm{BALE}}$ estimator produces the most reliable estimates on these datasets closely followed by $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{ALC}}$. This is due to $\mathrm{LID}_{\mathrm{ALC}}$ being more sensitive to the local geometry and $\mathrm{LID}_{\mathrm{MLE}}$ having a higher variance in its estimates. The $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ estimators, whilst giving good results in lower dimensions, drop off in higher dimensions in part due to insufficient neighborhood sizes. The $\mathrm{LID}_{\mathrm{LIDL}}$ estimator ends up with the worst overall performance despite its outstanding performance on simple artificial datasets. This may be due to a bad choice of radii, but the choice of radii is not clear in practice. The parameter regime derived on the simple cases appears to fail in more complex cases.

5.2.5 Real Datasets

Lastly, we will consider the application of the $\mathrm{LID}_{\mathrm{ALC}}$ family of estimators to real datasets. Real datasets are typically of unknown intrinsic dimensionality as the geometry of the manifold is unclear. Both varying sample densities and noise tend to further complicate the estimation of the intrinsic dimensionality. Accordingly, there often is no clear-cut answer as to which LID is the correct value, despite Rozza et al. [71] claiming true ID values for some of these datasets. Nonetheless, LID estimation can be beneficial in the design of machine learning models, as the ID can be used to determine the complexity of the model. Aside from the numerical LID values themselves, their relative order may be of interest to compare the complexity of different localities in the dataset. We will here consider

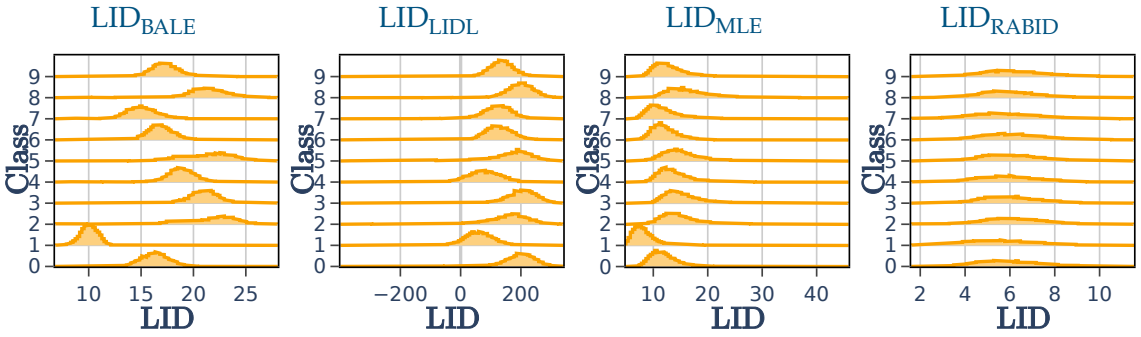


Figure 5.15: Distribution of LID estimates for each class of the MNIST dataset using 100 neighbors. The LID estimates were computed on the full dataset and afterward split by class. The histogram scale is identical for all histograms. The x-axis is clipped to the 0.1%- and 99.9%-quantile of all estimates of a specific estimator.

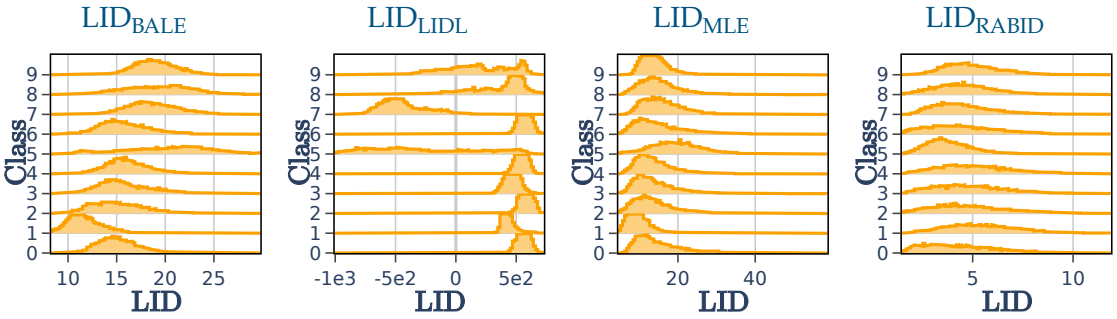


Figure 5.16: Distribution of LID estimates for each class of the Fashion-MNIST dataset using 100 neighbors. The LID estimates were computed on the full dataset and afterward split by class. The histogram scale is identical for all histograms. The histograms of LID_{LIDL} and LID_{MLE} were clipped for better visibility of the other histograms. The x-axis is clipped to the 0.1%- and 99.9%-quantile of all estimates of a specific estimator.

the datasets MNIST and Fashion-MNIST [90] a dataset similar in structure to MNIST but with more complex images of clothing. Both of these datasets were previously used by Tempczyk et al. [82] to evaluate the $\mathrm{LID}_{\mathrm{LIDL}}$ estimator. Of the $\mathrm{LID}_{\mathrm{ALC}}$ family, we only display $\mathrm{LID}_{\mathrm{BALE}}$ in the plots as it performed best on the artificial datasets and since all of the estimators have a pairwise Pearson correlation of at least 0.97 on these datasets. For the $\mathrm{LID}_{\mathrm{LIDL}}$ estimator, we had to reduce the hyperparameters of the density models to fit the available memory on our GPUs of $32GB$ VRAM. The density model requires holding multiple $d \times d$-matrices in memory where d is the data dimensionality, which, even after removing zero-variance dimensions, was too large. As the estimates appear correlated with the estimates of its original publication, we consider the results valid nonetheless. The histograms of the LID estimates for each class of the datasets are displayed in Fig. 5.15 and Fig. 5.16. Overall, the $\mathrm{LID}_{\mathrm{RABID}}$ estimates are the smallest, followed by $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{BALE}}$. This could be explained by $\mathrm{LID}_{\mathrm{RABID}}$ having insufficient neighbors and both $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{BALE}}$ underestimating the LID in line with the artificial sets before. This would imply, that $\mathrm{LID}_{\mathrm{BALE}}$ is likely closer to the true LID than $\mathrm{LID}_{\mathrm{MLE}}$. The relative order of classes of $\mathrm{LID}_{\mathrm{BALE}}$ and $\mathrm{LID}_{\mathrm{MLE}}$ is consistent on both datasets. Meanwhile, the standard deviation of the estimates per class divided by their mean is smaller for $\mathrm{LID}_{\mathrm{BALE}}$ ($\approx$0.23 on both datasets) than for $\mathrm{LID}_{\mathrm{MLE}}$ ($\approx$0.35 on MNIST and $\approx$0.46 on Fashion-MNIST). The $\mathrm{LID}_{\mathrm{RABID}}$ estimates have a too large variance to decide on the relative order of classes. The $\mathrm{LID}_{\mathrm{LIDL}}$ estimates are outside any reasonable range, which is most likely due to an ill choice of radii. Again, the radii "optimized" for the trivial cases appear to fail. Yet, as the different classes occupy visibly different dimensional spaces, according to $\mathrm{LID}_{\mathrm{LIDL}}$, these results are at least interpretable, albeit not in line with the results provided by Tempczyk et al. [82]. In their publication, they claim that almost all $\mathrm{LID}_{\mathrm{LIDL}}$ estimates on MNIST are in the range $[10, 70]$ and on Fashion-MNIST in the range $[100, 600]$. Whilst the MNIST estimates appear to be in the range proposed by other estimators as well, the Fashion-MNIST estimates are far off and likely due to an ill choice of radii as well. The relative order of classes in our experiments, however, mostly agrees with the relative order of classes proposed by $\mathrm{LID}_{\mathrm{BALE}}$ and $\mathrm{LID}_{\mathrm{MLE}}$ on MNIST though diverges on Fashion-MNIST.

5.2.6 Summary

In our evaluation, the estimators of the $\mathrm{LID}_{\mathrm{ALC}}$ family and most specifically $\mathrm{LID}_{\mathrm{BALE}}$ appear to be the most reliable choice for LID estimation in practice for the here pursued interpretation of LID. Whilst it has inferior performance on simple artificial datasets compared to $\mathrm{LID}_{\mathrm{LIDL}}$, it is more robust on more complex datasets. The difficulty of choosing appropriate radii for $\mathrm{LID}_{\mathrm{LIDL}}$ is a major drawback. Although the theory of $\mathrm{LID}_{\mathrm{LIDL}}$ is sound, the practical application is not straightforward. It externalizes the error to the parameterization of a deep neural density model, which is difficult to validate. For an ap-

propriate choice of radii, almost any regime of LID estimates can be achieved, which bears the risk of a circular argument. For any parameterization, the relative order of $\mathrm{LID}_{\mathrm{LIDL}}$ estimates may be usable, but that is unclear as our results are not congruent with the results of Tempczyk et al. [82]. In comparison to the $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$ estimators, $\mathrm{LID}_{\mathrm{ALC}}$ outperforms them in terms of convergence behavior and robustness to skewed neighborhoods. The latter may not be a disadvantage, as different interpretations of the LID may be useful in practice, e. g. to detect outliers. In any case, the $\mathrm{LID}_{\mathrm{ALC}}$ family is an "orthogonal" addition to the field of LID estimation, being differentiated by its performance on skewed and biased neighborhoods.

5.3 Conclusion

In this chapter, we introduced the $\mathrm{LID}_{\mathrm{ALC}}$ family of estimators, which is based on the chordal angles of samples in a neighborhood. The $\mathrm{LID}_{\mathrm{ALC}}$ type 1 and 2 estimators consider neighbors projected on the unit sphere, similar to $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$. These estimators derived from a distributional fit of observed angles have a sound theory and are computationally efficient. The $\mathrm{LID}_{\mathrm{BALL}}$ type 1 and 2 estimators also consider the vector norms of the neighbors. Although the theory behind the $\mathrm{LID}_{\mathrm{BALL}}$ estimators, which are based on an educated guess, is not as sound as for the $\mathrm{LID}_{\mathrm{ALC}}$ estimators, they perform well in practice. The $\mathrm{LID}_{\mathrm{BALE}}$ type 1 and 2 estimators are heuristically improved versions of the $\mathrm{LID}_{\mathrm{BALL}}$ estimators to better fit ideal neighborhoods in small dimensions. These estimators give near identical results compared to the $\mathrm{LID}_{\mathrm{BALL}}$ estimators if the LID is at least 5, but are more robust for lower LID values. They require more complicated mathematical functions (log and Lambert W) but are still computationally efficient. All estimators in the $\mathrm{LID}_{\mathrm{ALC}}$ family make use of a number of angles cubic in the neighborhood size and can thus achieve more reliable results with fewer neighbors than the reference estimators. They better preserve the dimensionality in skewed and biased neighborhoods and are thus more suitable for practical use, given the interpretation of LID pursued here. While the estimators can be "enriched" with mirrored neighbors, this changes the interpretation of the LID and thus achieves worse results in our benchmarks. Especially in skewed and biased neighborhoods, the different behaviors of $\mathrm{LID}_{\mathrm{RABID}}$, $\mathrm{LID}_{\mathrm{ALC}}$, and $\mathrm{LID}_{\mathrm{MLE}}$ may be exploited to derive future ensemble methods. Compared to $\mathrm{LID}_{\mathrm{LIDL}}$, the $\mathrm{LID}_{\mathrm{ALC}}$ family is inferior on high-dimensional simple manifolds, but appears superior in any other case. However, the unclear parameterization and extensive computational requirements of $\mathrm{LID}_{\mathrm{LIDL}}$ make it difficult to compare. In conclusion, we consider $\mathrm{LID}_{\mathrm{ALC}}$ a meaningful addition to the field of LID estimation, especially for practical use and as a basis for ensemble methods in combination with $\mathrm{LID}_{\mathrm{RABID}}$ and $\mathrm{LID}_{\mathrm{MLE}}$ or similar estimators.

Chapter 6

Beyond Euclidean

The theory of most LID estimators is based on Euclidean spaces due to using the Euclidean distance and the corresponding dot product. In general, these theoretical results, therefore, neither transfer to other metrics nor other similarity measures. However, given that some distance function is a metric, i. e. satisfies the triangle inequality, the formulae for most estimators can be practically evaluated on the real numbers without numerical errors such as the division by zero or negative roots. In this chapter, we will discuss some preliminary results of applying our LID estimators to non-Euclidean distances and similarities. We will primarily inspect research questions such as:

How "non-Euclidean" is non-Euclidean space after all?

Is it sufficient, that the formulae are applicable to obtain meaningful results?

Can we get past the limitations to use the estimators on non-Euclidean data?

What results on non-Euclidean distances can be derived from our analysis?

First, we will look at some anecdotal evidence in Section 6.1 to explain why we believe that LID estimation on non-Euclidean distances is a useful method, specifically with the angle-based estimators. We will then take a look at what defines Euclidean space and how to quantify "non-Euclideanness" in Section 6.2. We then inspect the "most non-Euclidean" metric spaces we know of in Section 6.3 and discuss the practical usefulness of very "non-Euclidean" spaces. We will afterward present some preliminary results on how LID estimation on non-Euclidean data can be interpreted in terms of a proxy Euclidean space and what practical implications this has for the analysis of non-Euclidean data in Section 6.4. Section 6.5 concludes this chapter with a summary.

6.1 Some Anecdotal Evidence

In this section, which was in large parts already published in our work on $\mathrm{LID}_{\mathrm{ABID}}$ [85], we will consider anecdotal evidence for the application of LID estimators to non-Euclidean

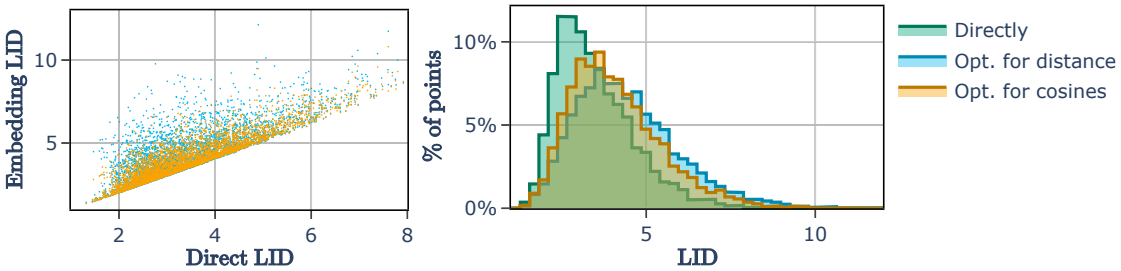


Figure 6.1: $\mathrm{LID}_{\mathrm{ABID}}$ values computed on "cosines" derived from Levenshtein distances on Java 8 class names. LID values are given for "cosines" obtained directly from distances via the law of cosines and those obtained from non-isometric Euclidean embeddings optimized for a minimum squared error on pairwise distances and cosines with respect to the point of interest.

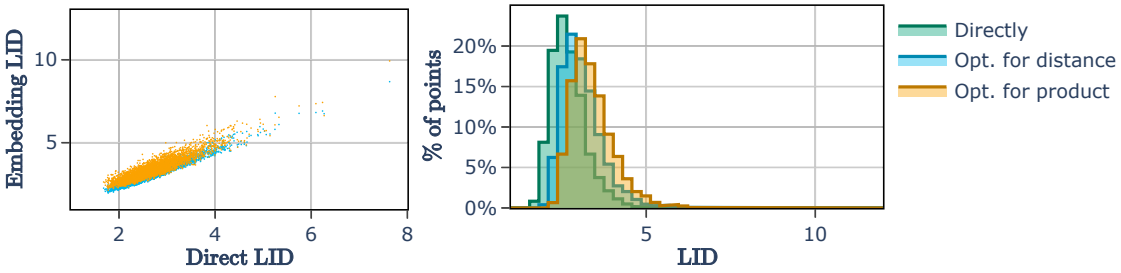


Figure 6.2: $\mathrm{LID}_{\mathrm{ALC}}$ type 1 values computed on "cosines" derived from Levenshtein distances on Java 8 class names. LID values are given for "cosines" obtained directly from distances via the law of cosines and those obtained from non-isometric Euclidean embeddings optimized for a minimum squared error on pairwise distances and inner products with respect to the point of interest.

data. We intend to argue, that $\mathrm{LID}_{\mathrm{ABID}}$, $\mathrm{LID}_{\mathrm{RABID}}$, and $\mathrm{LID}_{\mathrm{ALC}}$ values can be interpreted as geometric complexity measures of implicit isometrically (or approximately) embedded Euclidean spaces for distances (or similarities) obtained in a neighborhood. For the estimators to be applicable in a strictly technical sense, the equations must provide values within the real numbers. Whenever the triangle inequality is satisfied, the law of cosines gives

$$\cos(\angle_{yxz}) = \frac{d(x,y)^2 + d(x,z)^2 - d(y,z)^2}{2d(x,y)d(x,z)} \tag{6.1}$$

$$\cos(\angle_{yxz}) \geq \frac{d(x,y)^2 + d(x,z)^2 - (d(x,y) + d(x,z))^2}{2d(x,y)d(x,z)} = -1 \tag{6.2}$$

$$\cos(\angle_{yxz}) \leq \frac{d(x,y)^2 + d(x,z)^2 - |d(x,y) - d(x,z)|^2}{2d(x,y)d(x,z)} = 1 \tag{6.3}$$

Accordingly, the cosines of the "angles" in metric space are always within the range $[-1, 1]$. Similarly, we have $\cos(\angle_{yxy}) = 1$. As a consequence, $\mathrm{LID}_{\mathrm{ABID}}$ estimates over n neighbors are in the range $[1, n]$. As far as functions on cosines derived from distances are concerned, non-Euclidean metric space so far "looks" Euclidean.

When applying $\mathrm{LID}_{\mathrm{ABID}}$ or $\mathrm{LID}_{\mathrm{ALC}}$ to other metrics than the Euclidean distance by obtaining cosines from distances with the law of cosines, the LID estimates correspond to the intrinsic dimensionality of an implicit Euclidean embedding. In principle, a local Euclidean embeddability is arguably enough to guarantee that the estimators are a measure of local data complexity, which could be checked by investigating each neighborhood separately. The resulting estimates then sufficiently describe the local complexity surrounding a point of interest even if the dataset is not Euclidean embeddable as a whole. An example of such a case is the geodetic distance of a Riemannian manifold, which is locally Euclidean everywhere but in its entirety likely non-Euclidean. For an extended and more detailed discussion on what finite metric spaces are Euclidean embeddable, we recommend the paper by Maehara [58]. For non-isometrically Euclidean embeddable distances, we nevertheless expect the estimators to give an *approximate* estimate of the data complexity. Fig. 6.1 and Fig. 6.2 display one such case, where values for $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{ALC}}$ are computed for all Java 8 class names[1] based on the Levenshtein distance. To evaluate the potential error in computing LID values on a non-isometrically Euclidean embeddable metric, we compare the estimates to those of non-isometric Euclidean embeddings optimized for the least sum of squared errors on both distances and cosines/dot products. The estimates on the approximations appear to be lower bounded by and correlate with the estimates directly computed on the Levenshtein distance itself. Both estimators mostly agree on a LID range of $[2, 4]$, although the difference in estimates hints at errors introduced by the non-Euclideanness of the Levenshtein distance. There, nonetheless, is a clear indication of geometric complexity induced by the LID estimates. Thereby, $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{ALC}}$ values can potentially measure local data complexity even for non-isometrically Euclidean embeddable metrics. For the class names of Java 8, the "independent directions" (i.e. geometric intrinsic dimensions) in each neighborhood are not easily interpretable. The neighborhood of the class `BindingType` ($\mathrm{LID}_{\mathrm{ABID}}$ ≈2.1, $\mathrm{LID}_{\mathrm{ALC}}$ ≈2.3) consists of classes beginning with "Binding" (e.g., `BindingHelper`, `BindingHolder`) or ending with "Type" (e.g., `UnionType`, `WildcardType`). By the Levenshtein distance, `BindingType` lies in between these two groups, i.e. the distances between classes from the "Binding" and the "Type" groups are close to the sum of distances via `BindingType`. The absolute "cosine" for pairs from different groups is very large, whereas the absolute "cosine" for pairs from the same group tends to be small. Contrasting what one would expect from Euclidean distances, pairs from different groups lie in the same/opposite direction, and independent, i.e. orthogonal, transformations are within the same group. The number of groups thus does not correspond to the LID estimate. The directions implied by the intuitive operations "changing the prefix" and "changing the suffix" are not described as easily. For classes with higher LID estimates, describing these directions is even more difficult. Yet, the size of the LID estimates correlates with the diversity of class

[1] https://docs.oracle.com/javase/8/docs/api/allclasses-noframe.html

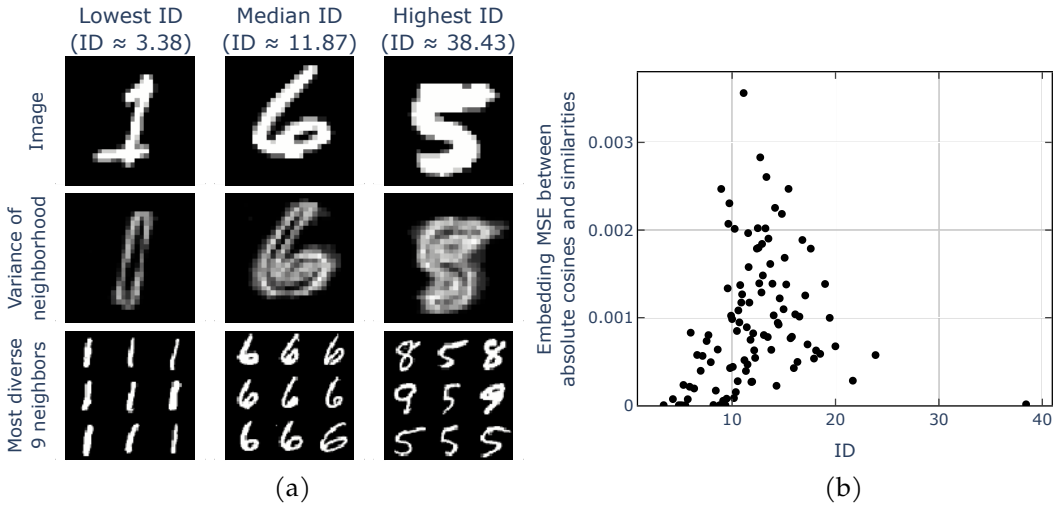


Figure 6.3: (a) Local "diversity" in MNIST images discretized to 8 brightness values over $\mathrm{LID}_{\mathrm{ABID}}$ values computed via normalized conditional mutual information of the brightness vectors. The nearest neighbors were chosen by the largest normalized mutual information. (b) Mean squared error of an approximate Euclidean embedding of the local neighborhoods of MNIST images over their respective $\mathrm{LID}_{\mathrm{ABID}}$ values. The error values are the deviation of normalized conditional mutual information and absolute cosines in the embedding, conveying the same semantic intuition. The selected points correspond to the 1 through 100 percentile of $\mathrm{LID}_{\mathrm{ABID}}$ values.

names in each neighborhood: The neighborhood of `TimeZone` ($\mathrm{LID}_{\mathrm{ABID}} \approx 8.08$, $\mathrm{LID}_{\mathrm{ALC}} \approx 2.8$) encompasses e.g. `TimeUnit`, `MimeType`, `FileStore`, and `TreeModel`. The neighborhood of `ORB` ($\mathrm{LID}_{\mathrm{ABID}} \approx 4.48$, $\mathrm{LID}_{\mathrm{ALC}} \approx 7.63$) encompasses e.g. `Oid`, `URL`, `JMX`, and `CSS`. These are the maximum estimated $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{ALC}}$ values in the dataset. The maximum $\mathrm{LID}_{\mathrm{ABID}}$ is achieved where groups of nearest class names are different, whereas the maximum $\mathrm{LID}_{\mathrm{ALC}}$ is achieved where each triplet of nearest class names is most diverse. The neighborhood of `TimeZone` thus has a lower $\mathrm{LID}_{\mathrm{ALC}}$ estimate, since it contains similar subgroups such as `TimeUnit` and `Timer` whereas the neighborhood of `ORB` gives a lower $\mathrm{LID}_{\mathrm{ABID}}$ estimate as the neighborhood is so diverse, that changing any character of the rather short string is a viable change. Either interpretation can be associated with a "more complex" neighborhood and the difference may in part be due to the rather small neighborhood size of 20.

Further, the $\mathrm{LID}_{\mathrm{ABID}}$ estimator could be applied to similarities giving a LID estimate for a Euclidean embedding with similar local behavior. Opposite to general similarity functions, the cosines involved in our estimators describe the similarity of two points centered on a point of interest, which can best be described as a trivariate similarity function ("similarity of y and z under consideration of x"). If a reasonable definition of such a trivariate similarity function can be given for the objects of interest, the estimators would again imply some embedding and could lead to a measure of local data complexity. One such trivariate similarity function is, e. g. the normalized conditional mutual information,

which is the conditional mutual information [23] divided by the minimum of both conditional entropies as suggested by Kvålseth [53]. We computed neighborhoods of MNIST images (using only a subset, discretized to 8 brightness values) by selecting the 20 images with the highest normalized mutual information to a point of interest as neighbors. Values of the normalized mutual information close to 1 then correspond to very correlated distributions of dark pixels, whereas values close to 0 correspond to independent distributions of dark pixels. The discretization to 8 values effectively transforms most pixels to the darkest and brightest values, while still providing visually interpretable images. By utilizing the normalized conditional mutual information as a similarity measure, we obtained $\mathrm{LID}_{\mathrm{ABID}}$ values for each image that supposedly correlate with the local data complexity, i. e. the diversity of neighbor images. The smallest, median, and largest obtained $\mathrm{LID}_{\mathrm{ABID}}$ values with their corresponding images and the top 9 most diverse neighbors (selection with a minimum of maximum pairwise normalized mutual information) are displayed in Fig. 6.3. The "complexity" of the neighborhood of the digit 5 (highest $\mathrm{LID}_{\mathrm{ABID}}$ value) is larger than that of the digit 1 (lowest $\mathrm{LID}_{\mathrm{ABID}}$ value), indicated by the neighborhood containing fives, eights, and nines rather than just very similar ones. Further, the normalized conditional mutual information values in these small neighborhoods can approximately be embedded into 20-dimensional Euclidean space as cosines as shown in Fig. 6.3. For the most extreme (and arguably most interesting) $\mathrm{LID}_{\mathrm{ABID}}$ values, we could compute Euclidean embeddings with close to no error between similarity values and absolute cosines. Aside from using trivariate similarity functions, we could also just use any bivariate similarity function under the assumption that the data is spherically distributed around a virtual out-of-distribution center. Lastly, similarity functions can be transformed into (not necessarily metric) distance measures, e. g. cosine distances are obtained from cosine similarities, on which the estimators can be evaluated – again without theoretical basis but under the approximative assumptions discussed above.

In conclusion, although we do not provide any theoretical foundation for applying $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{ALC}}$ to non-Euclidean data, preliminary empirical evidence suggests that the measures are correlating to local data complexity for locally (i. e. entire neighborhoods of chosen size) approximately Euclidean embeddable data. The measured LID values correlate with our intuitive understanding of the local complexity of the data. The geometric "directions" implied by these measures, however, might not be easily interpretable. The embedding error of investigated non-Euclidean distances to Euclidean space is surprisingly low, hinting at commonly used distances not striving too far from the Euclidean distance.

6.2 Properties of Euclidean Space and Distance

While the axioms for distance functions and metrics are well-defined, the axioms of the Euclidean distance, or at least a minimum set of such axioms, are not as clear. The axioms we should start with are

Non-Negativity: $d(x,y) \geq 0$

Identity of Indiscernibles: $d(x,y) = 0 \Leftrightarrow x = y$

Symmetry: $d(x,y) = d(y,x)$

Triangle Inequality: $d(x,y) \leq d(x,z) + d(z,y)$

where the first three make out a distance and all four define a metric. As a remark, the Identity of Indiscernibles is often omitted in practical settings to allow for duplicate data points. As for additional axioms to further discriminate non-Euclidean from Euclidean metrics, i. e. metrics that always allow for an isometric embedding into Euclidean space expressed by Cartesian coordinate, potential axioms from the literature are either not sufficient or very "cumbersome" to work with:

Hilbert's Axioms [41]:
A collection of 20 axioms based on the abstract primitives "point", "line", "plane", "Betweenness", "Containment", and "Congruence", e. g. "There exist at least two points on a line. There exist at least three points that do not lie on the same line.", that define the properties of three-dimensional Euclidean space.

Tarski's Axioms [78]:
A collection of 11 axioms similar to Hilbert's Axioms but with significantly fewer axioms and primitives, expandable to d-dimensional space.

Birkhoff's Axioms [12]:
A collection of four axioms that define the properties of two-dimensional Euclidean space, again similar to Hilbert's Axioms but also severly limited in application.

Ptolemy's Inequality:
$d(x,y) \cdot d(z,w) \leq d(x,z) \cdot d(y,w) + d(x,w) \cdot d(y,z)$ (necessary, not sufficient)

Parallelogram Law:
$d(x,y)^2 + d(z,w)^2 = 2 \cdot d(x,z)^2 + 2 \cdot d(y,w)^2$ (necessary, not sufficient)

Inner Product Bounds [87]:
$\langle x,y \rangle \in \left[\sum_{i=1}^{k} \langle x,\hat{r}_i \rangle \langle y,\hat{r}_i \rangle \pm \left(\left(\langle x,x \rangle - \sum_{i=1}^{k} \langle x,\hat{r}_i \rangle^2 \right) \left(\langle y,y \rangle - \sum_{i=1}^{k} \langle y,\hat{r}_i \rangle^2 \right) \right)^{\frac{1}{2}} \right]$
(necessary and sufficient for $k \to \infty$)

where we define $\langle x, y\rangle = d(x,z)^2 + d(y,z)^2 - d(x,y)^2$ for an arbitrary point z. The most "convenient" axioms would be the simple (in)equalities of Ptolemy and the Parallelogram Law, but they are not sufficient to define Euclidean space. The axiom sets of Hilbert, Tarski, and Birkhoff, and other similar axiom sets omitted for brevity, as well as the Inner Product Bounds are likely too complex, at least to automatically derive Euclideanness from a given metric. For axiom sets like those of Hilbert, Tarski, and Birkhoff, that comes to little surprise, as their usefulness is rather in the theoretical analysis of Euclidean space and in expressing the known theory on Euclidean space in a minimal formal way. The Inner Product Bounds, however, are a more practical approach to define Euclidean space but induce a large number of inequalities when the dimensionality of the space is not known. In that case, all infinitely many inequalities would have to be satisfied to ensure Euclideanness, which can only be cut short by proving that the additional terms beyond some k are zero. Given a black-box metric, this is not feasible, as even checking only up to some k induces an exponential number of necessary tests to confirm Euclideanness up to that k. However, in practical settings, we rather encounter datasets in the form of finite samples combined with some distance function. We can thus assert our input to be of the form of a distance matrix $D \in \mathbb{R}^{n\times n}$ where $D_{i,j} = d(x_i, x_j)$ for data points x_i, x_j and a distance function d. We can further define the according Gramian matrix $G^{(k)} \in \mathbb{R}^{(n-1)\times(n-1)}$ as $G^{(k)}_{i,j} = (D^2_{i,k} + D^2_{j,k} - D^2_{i,j})/2$ where we omit the zero row and column k. Using this discrete matrix representation, we obtain a different selection of axioms that are necessary and sufficient for (spherical) Euclideanness:

Gramian Positive Semi-Definiteness [73]:
$G^{(k)}$ is positive semi-definite for all k

Spherical Euclideanness [1, Theorem 4.2]:
$\exists\, \beta > 0 : (\mathbb{1} - \frac{D^{\circ 2}}{\beta})$ is positive semi-definite (special case of the previous axiom)

Spectral Euclideanness [1, Theorem 3.11]:
Exactly one eigenvalue of $D^{\circ 2}$ is positive and $\forall\, w : D^{\circ 2} w = \mathbf{1} \Rightarrow w^T \mathbf{1} \geq 0$

All of these axioms, which individually are necessary and sufficient for (spherical) Euclidean distances, require the distance function to satisfy the three distance axioms and are more practical to work within a finite setting. Solely the triangle inequality is implied by the other properties. Note, that any finite Euclidean sample can be approximated to arbitrary precision by a spherical sample by adding a dimension to the data points such that their norm is some very large number. As that number approaches infinity, the measured spherical distances approach the original non-spherical distances. This relates to the Spherical Euclideanness property by the negative eigenvalues approaching zero to arbitrary precision. When these properties are satisfied, the Hadamard square of the distance Matrix $D^{\circ 2}$ is also – slightly unintuitively – denoted as a Euclidean Distance Matrix

(EDM) in the literature [1]. In this chapter, we will call a distance matrix Euclidean, if a point cloud in some $\mathbb{R}^d$ exists that realizes the distance matrix and we will call it an EDM if we specifically refer to the Hadamard square of the distance matrix. The major hurdle in using these axioms as a test is the evaluation of the spectrum. Typical approaches to decide positive semi-definiteness use Cholesky or Eigendecomposition which both have a complexity of $\mathcal{O}(n^3)$ on $n \times n$ matrices. The third axiom only lends itself to a test if the matrix has full rank, such that the implication can be tested on the preimage of the all-ones vector. In that case, only the two largest eigenvalues need to be computed, which is much faster, but the matrix inversion is still of complexity $\mathcal{O}(n^3)$. A last property, that unites most of the axioms discussed above, is that Euclidean distance has an additive decomposition when squared, i. e. $d(x,y)^2 = d(x,z)^2 + d(z,y)^2$ whenever the directions from z to x and y are independent. Aside from just restating the Pythagorean theorem, we can observe that behavior in the Parallelogram Law, the Inner Product Bounds, the definition of the Gramian, and the Spectral Euclideanness axiom. It appears to be the most fundamental property of Euclidean distance, implying both an additive decomposition of the induced inner product and consequentially a parameterization of the space. It is implied by the axioms, yet most difficult to explicitly state as a sufficient closed-form axiom.

After discussing the properties of Euclidean space, we can now attempt to quantify the "non-Euclideanness" of some distance function. For this, we can consider different transformations of the distance function and use a Euclideanness test as a constraint. Since only the matrix-based tests are feasible in practice, we will focus on transformations of the distance matrix. For one, we can measure the non-Euclideanness of a distance matrix by some matrix distance to the nearest Euclidean distance matrix. Since we expect Euclidean distances to decompose additively when squared, a natural choice for the matrix distance is the Frobenius norm, i. e. the entry-wise Minkowski 2-norm. We could then e. g. use a black-box optimization algorithm to minimize both the Frobenius norm and the sum of squared negative eigenvalues of the Gramian matrix. With proper weighting of the objectives, this would yield a distance matrix that is as close as possible to a Euclidean distance matrix, albeit at questionable computational cost and resulting quality.

Similarly, we can evaluate the matrix distance to the nearest positive semi-definite matrix to the observed Gramian matrix. Extending the previous choice of the matrix distance, a natural choice would then be the entry-wise Minkowski 1-norm. Though, for convenience, the Frobenius norm is a better choice, since theoretical results for the nearest positive semi-definite matrix can be found in the literature, whereas the nearest positive semi-definite matrix under elementwise Minkowski 1-norm appears an unsolved problem. Higham [40] provided a method to obtain the nearest positive semi-definite matrix under the Frobenius norm as

$$G^{(k)*} = \frac{B+H}{2} = \sum_{i=1}^{n-1} \max(0, \lambda_i) v_i v_i^T \tag{6.4}$$

where $B = (G^{(k)} + G^{(k)T})/2$, $B = UH$ is a polar decomposition, and λ_i and v_i are the i-th eigenvalue and eigenvector of $G^{(k)} \in \mathbb{R}^{(n-1)\times(n-1)}$. The second equation only holds for symmetric $G^{(k)}$, which is naturally the case as we require the distance function to be symmetric. The nearest positive semi-definite matrix under the Frobenius norm is thus the same matrix after omitting all negative eigenvalues. The drawback of this approach is that it does not minimize the error on distance values, but rather on derived inner products, which are skewed in magnitude by the choice of k, i. e. the virtual center of the inner product space. Since the set of EDMs is convex [1], i. e. closed under convex combinations, we can use this result to find a relaxed version of the nearest distance matrix under the Frobenius norm, by choosing the nearest convex combination of EDMs corresponding to the closest positive semi-definite matrices of individual $G^{(k)}$. The closest convex combination should then be chosen under element-wise Minkowski 1-norm to compensate for the squaring implied by the EDMs.

Aside from employing matrix distances to evaluate the non-Euclideanness and find approximate Euclidean distance matrices by "locally" changing individual values, we can also use "global" distance transformations. Two useful transformations are the power transform and the t-translation [25]. The power transform is defined as $d_\alpha(x,y) = d(x,y)^\alpha$ for some $\alpha \in \mathbb{R}$. For $\alpha \leq 1$, both metrics and Euclidean distance are closed under the power transform. For sufficiently small α, all distances become more similar and any distance matrix can be transformed into a metric matrix and subsequently to a Euclidean distance matrix. The t-translation is defined as $d_t(x,y) = d(x,y) + t$ if $x \neq y$ and $d_t(x,x) = 0$ for some $t \in \mathbb{R}$. For large enough t, all distances become more similar, again resulting in any distance matrix becoming a metric matrix and subsequently a Euclidean distance matrix. These transformations are particularly useful, due to their simplicity and the guarantee of obtaining a Euclidean distance matrix for an appropriate parameter choice. The "nearest" Euclidean distance matrix under these transformations can then be found by a parameter search, where the smallest viable parameter quantifies some notion of "non-Euclideanness". For already Euclidean distance matrices, the parameters can even be $\alpha > 1$ and $t < 0$. The power transform appears especially meaningful since it induces the additive decomposability of the squared Euclidean distance. It implies that non-Euclidean distances are additively decomposable for a sub-two-power, effectively contracting on some longer distances, similar to shortcuts on the sphere using the big-circle distance.

In summary, we have discussed the properties of Euclidean space and distance, and some methods to quantify the non-Euclideanness of a distance matrix. The properties of Euclidean space are still somewhat elusive on the functional level, as only infinite sets of inequalities are known to define Euclidean distance. For finite datasets, we can employ the distance matrix to test for Euclideanness typically using the derived Gramian matrix, which however requires virtually centering the data at one of its samples. To quantify

the non-Euclideanness we can employ matrix distances to the nearest Euclidean distance matrix, as well as the minimum parameters for global transformations of the distance function to guarantee Euclideanness. None of these methods, however, extend to unseen samples, since we do not analyze the distance function itself, but rather concrete observed distance values.

6.3 Is (Very) Non-Euclidean Useful?

In the previous section, we considered different characterizations of Euclidean space and potential measures of "non-Euclideanness". The motivating anecdotal examples in Section 6.1 hinted at practically relevant metrics to be quite close to the Euclidean distance. In this section, we will investigate the opposite direction. We consider the worst-case scenario for non-Euclidean spaces and discuss whether "really non-Euclidean" metrics are even practically relevant. We will make use of the power transformation of a distance matrix, as the power transformation is linked to the additive decomposition property for squared Euclidean distances. We observed that a block-based distance matrix is especially bad for the power transformation. It has the shape

$$D_n = \begin{pmatrix} 0 & \dots & 1 & \frac{1}{2} & \dots & \frac{1}{2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & \dots & 0 & \frac{1}{2} & \dots & \frac{1}{2} \\ \frac{1}{2} & \dots & \frac{1}{2} & 0 & \dots & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \frac{1}{2} & \dots & \frac{1}{2} & 1 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{n \times n} \tag{6.5}$$

where the first block has length $\lceil n/2 \rceil$ and the second block has length $\lfloor n/2 \rfloor$. All the metric axioms are satisfied and we can rewrite D_n and $D_n^{\circ 2}$ in terms of the identity matrix and outer products:

$$D_n = -\mathbb{I}_n + \tfrac{1}{2}\mathbf{1}_n\mathbf{1}_n^T + \tfrac{1}{2}aa^T + \tfrac{1}{2}bb^T \tag{6.6}$$

$$D_n^{\circ\alpha} = -\mathbb{I}_n + r\mathbf{1}_n\mathbf{1}_n^T + (1-r)aa^T + (1-r)bb^T \tag{6.7}$$

where

$$a = (1, \dots 1, 0 \dots, 0) \text{ with } \sum_{i=1}^{n} a_i = \left\lceil \tfrac{n}{2} \right\rceil \tag{6.8}$$

$$b = (0, \dots 0, 1 \dots, 1) = \mathbf{1}_n - a \tag{6.9}$$

$$r = 2^{-\alpha} \tag{6.10}$$

If n is even and at least 4 and $\alpha > 0$, the positive eigenvalues of that matrix are $((1-r)n-2)/2$ and $((1+r)n-2)/2$ with the eigenvectors $(a-b)/\sqrt{n}$ and $\mathbf{1}_n/\sqrt{n}$. Showing that these eigenvectors and eigenvalues are correct is easily verified by insertion.

$$
\begin{aligned}
D_n^{\circ\alpha}\tfrac{a-b}{\sqrt{n}} &= -\tfrac{a-b}{\sqrt{n}} + \tfrac{r}{\sqrt{n}}(a+b)(a+b)^T(a-b) + \tfrac{1-r}{\sqrt{n}}aa^T(a-b) + \tfrac{1-r}{\sqrt{n}}bb^T(a-b) && (6.11)\\
&= -\tfrac{a-b}{\sqrt{n}} + \tfrac{r}{\sqrt{n}}ab^T(a-b) + \tfrac{r}{\sqrt{n}}ba^T(a-b) + \tfrac{1}{\sqrt{n}}aa^T(a-b) + \tfrac{1}{\sqrt{n}}bb^T(a-b) && (6.12)\\
&= -\tfrac{a-b}{\sqrt{n}} - \tfrac{rn}{2}\cdot\tfrac{a-b}{\sqrt{n}} + \tfrac{n}{2}\cdot\tfrac{a-b}{\sqrt{n}} && (6.13)\\
&= \tfrac{(1-r)n-2}{2}\cdot\tfrac{a-b}{\sqrt{n}} && (6.14)\\
&= \tfrac{3n-8}{8}\cdot\tfrac{a-b}{\sqrt{n}} && (6.15)\\
D_n^{\circ\alpha}\tfrac{a+b}{\sqrt{n}} &= -\tfrac{a+b}{\sqrt{n}} + \tfrac{r}{\sqrt{n}}(a+b)(a+b)^T(a+b) + \tfrac{1-r}{\sqrt{n}}aa^T(a+b) + \tfrac{1-r}{\sqrt{n}}bb^T(a+b) && (6.16)\\
&= -\tfrac{a+b}{\sqrt{n}} + \tfrac{r}{\sqrt{n}}ab^T(a+b) + \tfrac{r}{\sqrt{n}}ba^T(a+b) + \tfrac{1}{\sqrt{n}}aa^T(a+b) + \tfrac{1}{\sqrt{n}}bb^T(a+b) && (6.17)\\
&= -\tfrac{a+b}{\sqrt{n}} + \tfrac{rn}{2}\cdot\tfrac{a+b}{\sqrt{n}} + \tfrac{n}{2}\cdot\tfrac{a+b}{\sqrt{n}} && (6.18)\\
&= \tfrac{(1+r)n-2}{2}\cdot\tfrac{a+b}{\sqrt{n}} = \tfrac{(1+r)n-2}{2}\cdot\tfrac{\mathbf{1}_n}{\sqrt{n}} && (6.19)
\end{aligned}
$$

Proving, that these eigenvalues are the only positive eigenvalues results from the fact that by construction, the eigenvectors corresponding to positive eigenvalues must be positively correlated with either a, b, or $\mathbf{1}_n$. For any third eigenvalue to be larger than zero, the corresponding eigenvector must be positively correlated with a, b, and $\mathbf{1}_n$ but orthogonal to $(a-b)$, $(b-a)$, and $\mathbf{1}_n$. By substituting $a = \mathbf{1}_n - b$ and $b = \mathbf{1}_n - a$, this also implies orthogonality to $2a - \mathbf{1}_n$ and $2b - \mathbf{1}_n$ and consequentially orthogonality to a and b. Thus, such an eigenvector can not exist. As per the Spectral Euclideanness criterion introduced in Section 6.2, the matrix $D_n^{\circ\alpha}$ is an EDM if the second largest eigenvalue is ≤ 0. The second required property is trivially satisfied if the second largest eigenvalue is < 0 since the only positive eigenvalue corresponds to the eigenvector $\mathbf{1}_n/\sqrt{n}$. The critical value of α to have the second largest eigenvalue of the Hadamard square ≤ 0 is

$$\alpha_c = \frac{-\log_2\left(1-\frac{2}{n}\right)}{2}\,. \tag{6.20}$$

Thus, for any $\alpha \leq 2\alpha_c$, the matrix $D_n^{\circ\alpha}$ is an EDM and for $\alpha \leq \alpha_c$, the distance matrix is Euclidean. From that, we can see that the "non-Euclideanness" as quantified by the maximum α to obtain Euclidean distances increases as n increases and we can define a score for the "non-Euclideanness" based on the (empirical) critical α as:

$$\text{non-Euclideanness}(\alpha) = \frac{1}{2^{2\alpha}-1} - \frac{1}{3} \quad \left(= \frac{n}{2} - \frac{4}{3}\right) \tag{6.21}$$

This score increases linearly in n for the discussed hypothetical worst-case distance matrix. It further maps $\alpha = 1$ to a "non-Euclideanness" of 0. For odd n, the eigenvalues and eigenvectors are much less trivial. The critical value for α is empirically lower than the one obtained from (6.20) as can be seen in Fig. 6.4. The empirical α values for odd n

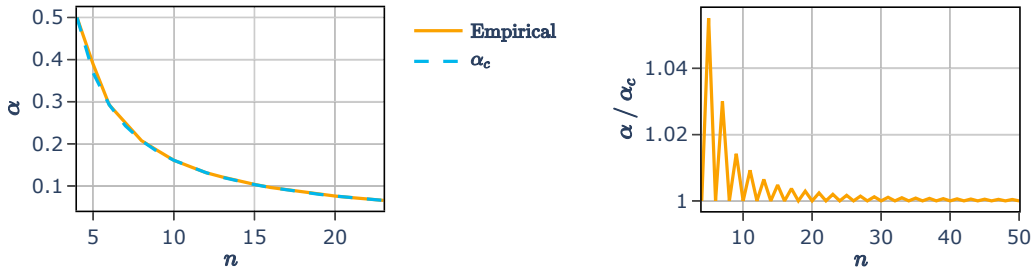


Figure 6.4: Left: The empirically evaluated critical α of D_n and the critical α_c as per (6.20). Right: The ratio of the empirical critical α and α_c as per (6.20). Empirical α values were computed by binary search over power transforms of D_n.

though seemingly converge to the formula for α_c. The α_c formula in (6.20) thus is a lower bound for the critical α of any metric distance matrix if D_n is the worst-case.

Without proof, we hypothesize that D_n is the worst-case scenario for non-Euclidean spaces. No numerical optimization algorithm we tested was able to find a worse distance matrix in terms of the critical α. As per the series expansion of (6.20), the critical α falls off approximately inversely proportional to n. Conversely, the worst non-Euclidean metric for n points is at worst a power transformation of some Euclidean space with a power approximately proportional to n. Applying this result to whole datasets is unfeasible, as the resulting critical power is too high to be of practical relevance. However, applying it to localities of a smaller size might be feasible. Furthermore, the hypothetical worst-case scenario discussed here is a very extreme case, that is unlikely encountered in "natural" metrics. The triangle inequality is tight for all coefficients in the matrix, thus any minor change would make this matrix non-metric. Further, the matrix semantically represents a very specific structure that is unlikely to occur in practice and if it occurred should be considered a mistake in the choice of the metric. The data is partitioned into two sets of points A and B. All points of A are equidistant to any point in B and vice versa. They can be interpreted as lying on a sphere around every point in the other set. Yet, within each set, A and B respectively, each pair of points must lie on opposite ends of each of these spheres. Even in the case of three points equidistant to some fourth "center", having all of these points being pairwise "opposites" of each other, violates the intuition of how we use distances in practice. If a metric were defined to allow for that property, in most cases, it would be deemed ill-fit for the dataset.[2] The worst-case scenario now potentiates that behavior for n virtual centers and $\mathcal{O}(n^2)$ "opposite" pairs for each center. Not only is this worst-case unlikely to be encountered, but it is "not useful" for any practical application and might be considered a contraindication to use the generating metric. The worst-case scenario not only provides a lower bound for the critical α, but it also highlights a key

[2]This is a personal opinion rather than empirical observation, but no such metric with practical use is known to the author.

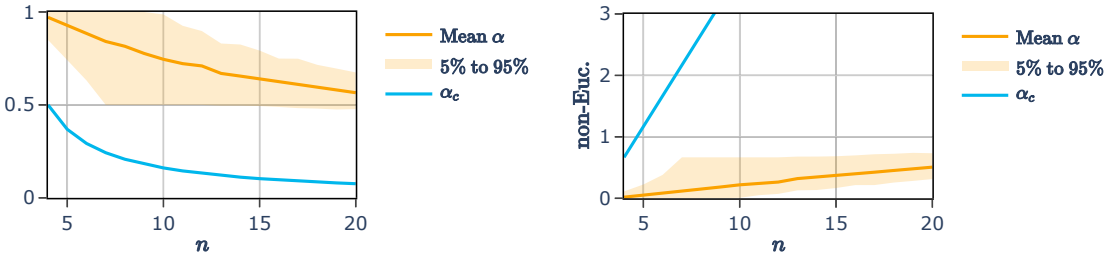


Figure 6.5: The empirical critical α and resulting "non-Euclideanness" score for distance matrices on 500 random $(n-1)$-nearest neighbor neighborhoods as per Levenshtein distance on 4104 Java class names.

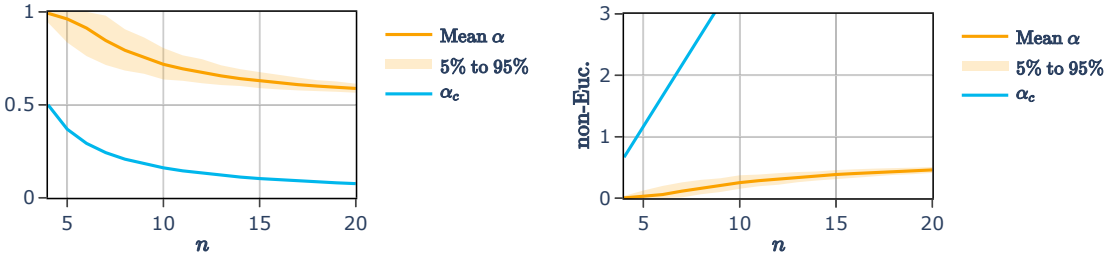


Figure 6.6: The empirical critical α and resulting "non-Euclideanness" score for distance matrices on 500 random $(n-1)$-nearest neighbor neighborhoods as per Manhattan distance on 2000 points sampled from a univariate normal distribution in 5 dimensions.

property of Euclidean spaces (and in relaxation for "useful" metrics): There must not be more than two pairwise opposites on a sphere. This property is similar to the intuition of dimensions as independent directions. For any dimension, we can move along a direction and the opposite direction – the two pairwise opposites on the sphere of normalized directions. Violating the property appears to be the most "non-Euclidean" a metric can be, at least in terms of the power transform. It does not imply that not violating this property is a sufficient condition for a metric to be Euclidean, but it at least allows for it.

To reflect on the anecdotal examples from Section 6.1, the empirical critical α values for the Levenshtein distance are much larger than α_c, and the "non-Euclideanness"

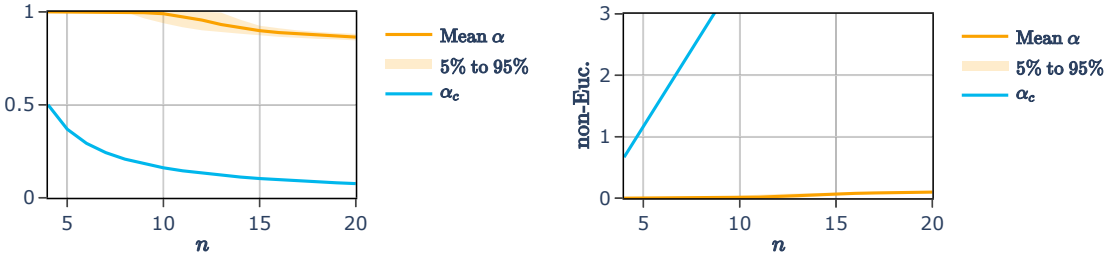


Figure 6.7: The empirical critical α and resulting "non-Euclideanness" score for distance matrices on 500 random $(n-1)$-nearest neighbor neighborhoods as per great-circle distance on 2000 points sampled uniformly from a 15 dimensional unit sphere.

score is very low. Fig. 6.5 displays the empirical critical α values and the resulting "non-Euclideanness" scores. The scores are mostly below the score for D_4, and the required power to make the Levenshtein distance matrices Euclidean is mostly $\geq 1/2$. A critical α of $\approx 1/2$ appears to be a lower bound even as the matrix size increases, suggesting that this property holds even for larger localities. Values for the Manhattan distance and great-circle distance (radians of the angle between two points) – which is not even a metric as the triangle inequality does not hold –, displayed in Fig. 6.6 and Fig. 6.7 respectively, are even closer to Euclidean. The Manhattan distance suggests the same converging behavior, while the critical α values for the great-circle distance are too large to make an equivalent statement. However, these results support the hypothesis that practically useful distances are (at least locally) close to Euclidean.

6.4 LID in non-Euclidean space

After having discussed properties that differentiate Euclidean and non-Euclidean space in Section 6.2 and having made an argument that "useful" distances are probably "close" to Euclidean in Section 6.3, we will now investigate how different LID estimators behave in non-Euclidean spaces. We will base our first part of the analysis on the power transform since it is related to the squared decomposability of the Euclidean distance and provides a measure for "non-Euclideanness" as previously argued. The power transform also allows us to sample non-Euclidean distance matrices since we can compute the interval of powers, for which the distance matrix is non-Euclidean but remains metric. Depending on the initial distance matrix, the range of powers to obtain a metric distance matrix varies and if a metric distance matrix is required, the power has to be chosen accordingly. We also have to keep in mind, that the critical power for previously inspected non-Euclidean distances did not drop much below $1/2$. The second part of the analysis will be focused on an approximate Euclidean distance matrix obtained by the nearest Gramian matrix under the Frobenius norm. The resulting proxy space provides a closer approximation of the distance values and cosines, however, it does not provide a global transformation applicable to unseen data. Yet, the approximate approach rather than a global transformation is more akin to manifold learning techniques like e. g. ISOMAP [83], thereby providing practically relevant insights nonetheless.

The general idea in the remainder of this section is to investigate how LID estimates behave on arbitrary distance matrices compared to power-transformed and approximate Euclidean distance matrices. If the estimates are strongly correlated, we can consider the LID estimator as a geometric complexity measure on an implicit Euclidean space. The resulting questions are to what extent the LID estimates on power-transformed and approximate Euclidean matrices are correlated with their non-Euclidean counterparts and

whether we can make tangible statements based on the numerical values of the LID estimates.

6.4.1 Canonical LID and Canonical Euclidean Space

By definition of the $\mathrm{LID}_{\mathrm{MLE}}$ estimator, estimates based on the power transform scale inversely linear with the applied power, as we can extract the power from the logarithm in its definition. That is

$$\mathrm{LID}_{\mathrm{MLE}}\left(D^{\circ\alpha}\right) = \frac{\mathrm{LID}_{\mathrm{MLE}}\left(D\right)}{\alpha} \tag{6.22}$$

where D are the distances $\mathrm{LID}_{\mathrm{MLE}}$ is computed on. The relative order of LID estimates is preserved, allowing for qualitative comparisons in non-Euclidean spaces in terms of an implicit power-transformed Euclidean space. If we assert an implicit power-transformed Euclidean space, however, the numerical value of the LID estimate is almost arbitrary. Given distance values D, if we constrain $\mathrm{LID}_{\mathrm{MLE}}$ to an implicit Euclidean space, any LID estimate $\geq \mathrm{LID}_{\mathrm{MLE}}\left(D^{\circ\alpha}\right)$, where α is at most the critical power for D, is a valid estimate. That is, even if D describes distances in a δ-dimensional Euclidean space, there exists a power-transformed Euclidean space for any $\delta' \geq \delta$. The exact choice of the implicit space $\mathrm{LID}_{\mathrm{MLE}}$ acts upon is thereby arbitrary and the numerical value of the estimate is not meaningful. To compensate for that issue, we suggest defining a canonical LID: The LID of the Euclidean space obtained from the critical α. We will refer to the corresponding power-transformed space as the *canonical Euclidean space*. As per $\mathrm{LID}_{\mathrm{MLE}}$, the canonical Euclidean space is the lowest intrinsically dimensional Euclidean space that realizes the distances of a power transform.

6.4.2 LID Estimates on Power-Transformed Euclidean Space

To evaluate the LID estimates on power-transformed Euclidean spaces, we need to generate non-Euclidean distance matrices. A rough explanation of the generation process has been given in the introduction of this section. We will now provide a more detailed description of the generation process. We start by sampling 500 of points from a uniform ball distribution in $\mathbb{R}^d$ for a uniform random $d \in [3, 10]$. The origin is added as the first point to the set. We then evaluate the Minkowski p-distance matrix for a random $p \in [10^{-2}, 10^2]$. The p value is chosen with equal probability below or above 2. If p is below 2, it is selected uniform from $[10^{-2}, 2]$. If p is above 2, we choose a uniform number from $[10^{-2}, 1]$, invert it, and add 1. That selection process is supposed to generate a wide range of non-Euclidean distance matrices. Afterward, we reduce the distance matrix to the 51 rows and columns belonging to the origin and its 50 nearest neighbors. We then evaluate the critical α_{Euc} for the reduced matrix, obtaining the canonical Euclidean distance matrix D. We also evaluate the largest α_{met} for which the distance matrix remains

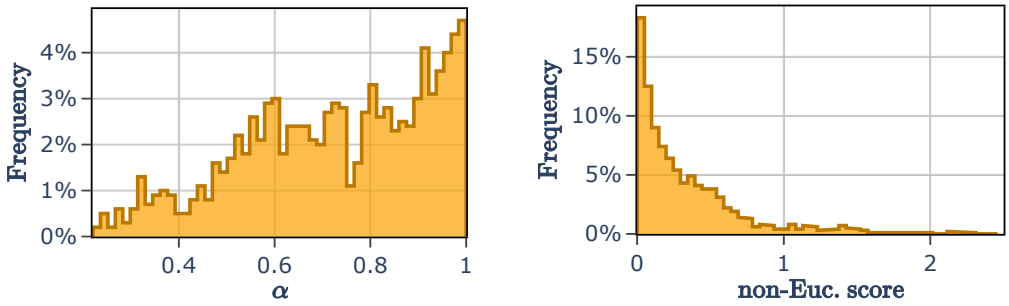


Figure 6.8: Distribution of critical α and corresponding non-Euclideanness score values for the randomly generated non-Euclidean distance matrices.

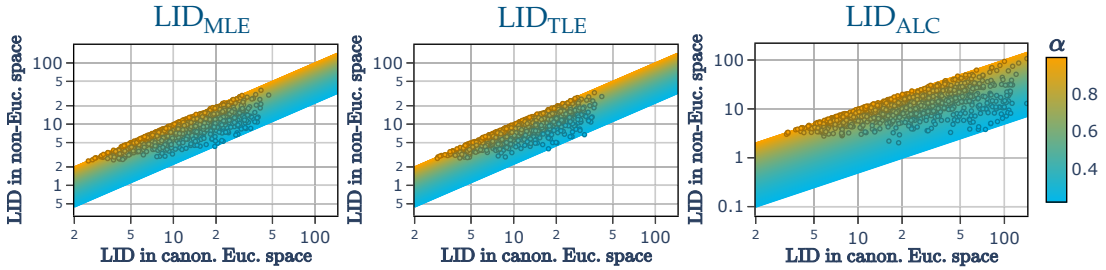


Figure 6.9: Correlation of LID estimates on the canonical Euclidean space compared to the non-Euclidean space. All marks are colored by the effective critical power α. The background is colored such that the coordinate (x, y) has the color of α for which $y = x\alpha$ for LID_{MLE} and LID_{TLE} and $y = x\alpha^2$ for LID_{ALC}.

metric. We then sample a uniformly random value ν from $[\alpha_{Euc}, \alpha_{met}]$ and define the effective critical α as $\alpha = \alpha_{Euc}/\nu \leq 1$. The non-Euclidean distance matrix used is then $D^{\circ 1/\alpha}$. This process is capable of providing a quite diverse set of non-Euclidean distance matrices $D^{\circ 1/\alpha}$ and at the same time the corresponding critical α. We evaluated 1000 of these distance matrices and computed the LID estimates for the canonical Euclidean space and the non-Euclidean space. The distribution of α values is displayed in Fig. 6.8. The distribution covers a wide range of values reaching from $\approx$0.2 to 1. The distribution steeply drops off below 0.5 and is almost uniform above 0.5. These values for the critical power are much broader than the anecdotal values we have seen in Section 6.3 but far from the hypothetical worst-case for $n = 50$ of ≈ 0.029. Similarly, the non-Euclideanness score derived from the critical power is in most cases < 1 and does not exceed 2 by a lot. Even though the generative process is based on non-Euclidean metrics from the Minkowski p-norm family, it appears difficult to sample "very" non-Euclidean distance matrices. Whether this is due to the choice of the generative process or due to the Minkowski p-norm family generally being close to Euclidean is unclear.

In regards to the correlation of LID estimates on non-Euclidean data, we observe a strong correlation between the canonical Euclidean space and the non-Euclidean space for LID_{MLE}, LID_{TLE}, and LID_{ALC} as displayed in Fig. 6.9. For a fixed critical α, the LID

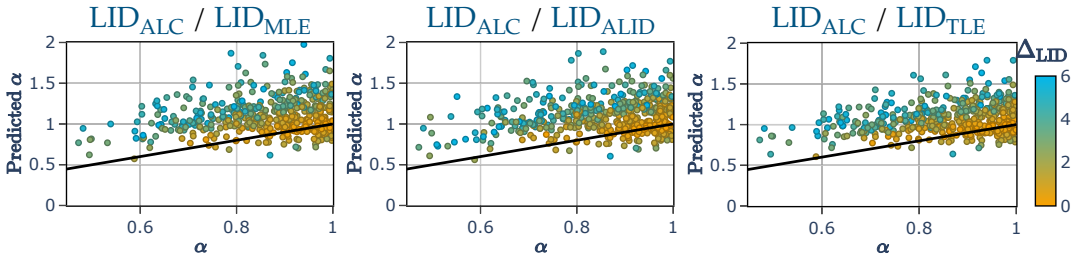


Figure 6.10: Prediction of the critical α based on a quotient of LID estimates on the non-Euclidean space. The coloring is based on the difference of the LID estimates on the canonical Euclidean space and only marks with a maximum difference of 6 are displayed. The line indicates the identity.

estimates on the non-Euclidean space are essentially linear and quadratic functions of α times the LID estimates on the canonical Euclidean space as indicated by the coloring of the marks matching the background coloring. The distance-based estimators follow a linear relationship, while the chordal angles-based $\mathrm{LID}_{\mathrm{ALC}}$ follows a quadratic relationship. Not only does this allow for a relative comparison of LID estimates in non-Euclidean spaces, but it also indicates, that the critical power may be derived from a combination of the LID estimates in the non-Euclidean space. Ideally, the LID estimates in the canonical Euclidean space agree for all estimators. By dividing the $\mathrm{LID}_{\mathrm{ALC}}$ estimates by the distance-based LID estimates, the ideally identical canonical LID estimates cancel out and we obtain a prediction of the critical power. That prediction is displayed in Fig. 6.10. For instances, where the canonical LIDs agree, i. e. on instances where the number of neighbors suffices for a stable estimate with distance-based estimators, the prediction of the critical power is quite accurate. The larger the difference in the LID estimates, the less accurate the prediction becomes. The estimates of the distance-based estimators in the canonical Euclidean space are, however, not congruent enough with the $\mathrm{LID}_{\mathrm{ALC}}$ estimates for a sufficiently exact estimation of the critical power in almost all cases. This is likely due to the insufficiently large neighborhood size of 50 points for the distance-based estimators since most canonical LIDs are ≥ 10, for which distance-based estimators require > 100 neighbors for a stable estimate.

As for $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$, we observe that the canonical LID estimates and the non-Euclidean LID estimates are only correlated for sufficiently large critical powers. The correlation is displayed in Fig. 6.11. When the critical power is large enough, the $\mathrm{LID}_{\mathrm{RABID}}$ estimates on the non-Euclidean distances are nearly identical to the canonical LID estimates. The errors in the non-Euclidean estimates, opposite to the previously discussed estimators, are two-sided, whereby a mean over an entire dataset may cancel the error out. For critical powers around and below 0.8, the correlation is likely not sufficient for practical purposes. Reconsidering the anecdotal examples from Section 6.1, the good performance of $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ on the non-Euclidean examples can not be explained by an implicit power transform. The distance-based estimators and the $\mathrm{LID}_{\mathrm{ALC}}$ estimator

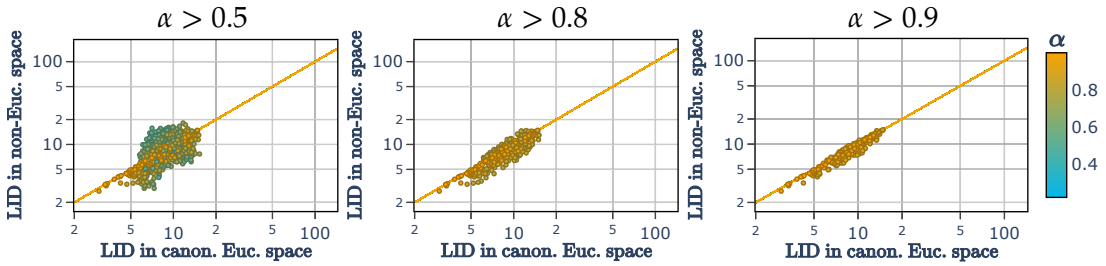


Figure 6.11: Correlation of $\mathrm{LID}_{\mathrm{RABID}}$ estimates on the canonical Euclidean space compared to the non-Euclidean space. All marks are colored by the effective critical power α. The displayed marks are all marks for which the effective critical power is above the respective threshold. The line indicates the identity.

seemingly describe the relative complexity of an implicit power-transformed Euclidean space. The $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ estimators only share that perspective, when the critical α is close to 1. As a next step, we will therefore consider the LID estimates on an approximate Euclidean space.

6.4.3 LID Estimates on Approximate Euclidean Space

To obtain an approximate Euclidean distance matrix, we use the nearest Gramian matrix under the Frobenius norm as described in (6.4). We use the same non-Euclidean distance matrix generation process as described in the previous section and base the Gramian matrix on the added origin of the distribution. This approximates the inner products from the perspective of the point of interest as used in $\mathrm{LID}_{\mathrm{ABID}}$. Accordingly, we would expect the pairwise cosines of the neighborhood to be approximated rather well. As can be seen in Fig. 6.12, the Gramian approximation gives a much closer approximation to the non-Euclidean distances than the power transform. For a fair comparison, the two Euclidean distance matrices were scaled to minimize the Frobenius norm of the difference to the non-Euclidean distance matrix. That is fairer since LID estimators only consider the relative size of distances, and the "best fit" regarding the LID is therefore linear scaling invariant. Subsequently, the approximation error values were divided by the mean Frobenius norm of involved matrices to obtain a relative error. On average, the Gramian approximation is around 2 times closer to the non-Euclidean distances than the power transform. The errors are however correlated and increase with smaller critical α values, as would be expected, again promoting the "non-Euclideanness" score.

The correlation of LID estimates on the Gramian-based approximate Euclidean space compared to the non-Euclidean space is displayed in Fig. 6.13 and Fig. 6.14. Opposite to the power transform, the LID estimates are not linear or quadratic functions of the critical power and the canonical LID estimates. They are rather close to the identity while the critical power causes deviations from the identity. For $\mathrm{LID}_{\mathrm{RABID}}$ and $\mathrm{LID}_{\mathrm{ALC}}$, the er-

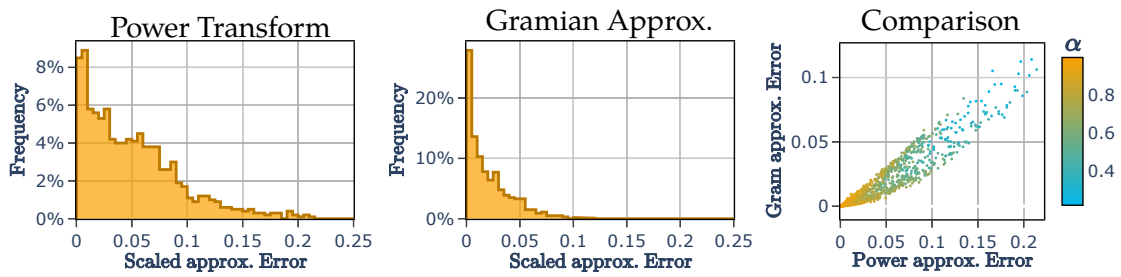


Figure 6.12: Scaled approximation error between the non-Euclidean distance matrix and either the power-transformed or the Gramian-based approximate Euclidean distance matrix. The approximation error is measured as the Frobenius distance between the scaled matrices, where a linear scaling is applied to minimize the error.

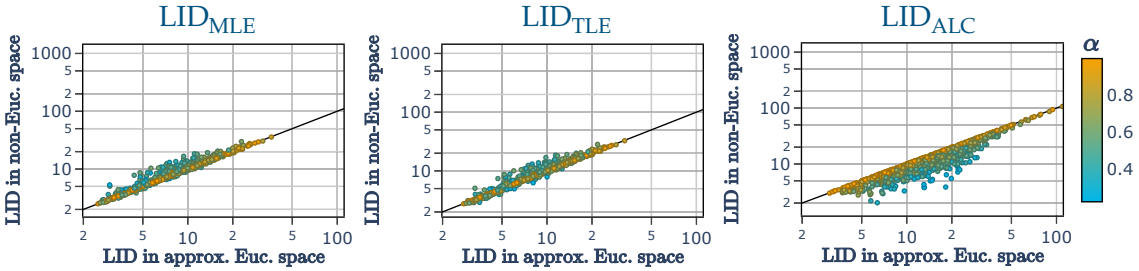


Figure 6.13: Correlation of LID estimates on the Gramian-based approximate Euclidean space compared to the non-Euclidean space. All marks are colored by the effective critical power α. The line indicates the identity.

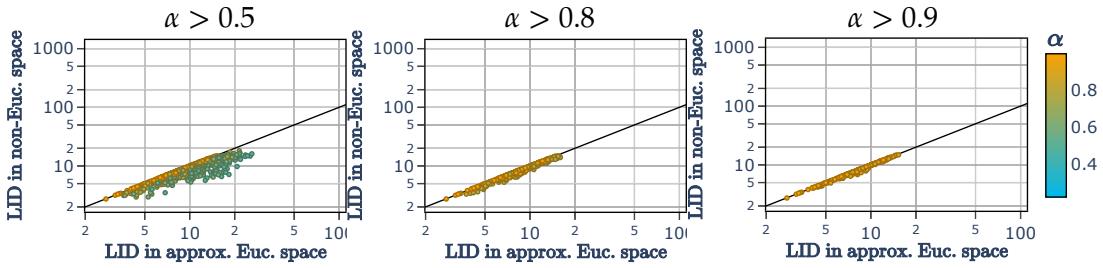


Figure 6.14: Correlation of $\text{LID}_{\text{RABID}}$ estimates on the Gramian-based approximate Euclidean space compared to the non-Euclidean space. All marks are colored by the effective critical power α. The displayed marks are all marks for which the effective critical power is above the respective threshold. The line indicates the identity.

ror appears one-sided, i. e. the non-Euclidean estimate is a lower bound of the approximated Euclidean estimate. That behavior can be observed in the anecdotal example of Levenshtein in Fig. 6.1 and Fig. 6.2. The errors for the distance-based $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{TLE}}$ instead are two-sided and strongly correlated. Considering different thresholds for the critical α, we observe that for any $\alpha > 0.8$, the $\mathrm{LID}_{\mathrm{RABID}}$ estimates are nearly identical to the Gramian approximation LID estimates, while for $\alpha > 0.5$ the correlation is still strong. The expected better approximation of $\mathrm{LID}_{\mathrm{RABID}}$ on Gramian approximated compared to power-transformed Euclidean spaces is thus confirmed. Since all evaluated estimators give correlated LID estimates on the non-Euclidean and Gramian approximated Euclidean distance matrices, we can also consider the LID estimates as a complexity measure on an implicit approximated Euclidean space. That perspective is less useful than the power transformation for $\mathrm{LID}_{\mathrm{ALC}}$ and $\mathrm{LID}_{\mathrm{MLE}}$ and similar estimators since the implicit approximated space is based on localities and does not necessarily give a global comparison. Yet, it better represents the relative sizes of distance values due to the closer approximation of the matrix.

6.4.4 Summary

In this section, we have investigated two interpretations of LID on non-Euclidean distance matrices. First, we considered an implicit power-transformed space, on which distance-based and $\mathrm{LID}_{\mathrm{ALC}}$ estimators give near linear and quadratic LID estimates when controlling for the critical power. The relation is strong enough, that it effectively allows us to estimate the critical power from a quotient of the LID estimates, given that the neighborhood sizes suffice for a congruent estimation in the power-transformed Euclidean space. The $\mathrm{LID}_{\mathrm{RABID}}$ estimator, however, only gives near identical LID estimates for critical powers above 0.8. Second, we considered an approximate Euclidean space based on the nearest Gramian matrix under the Frobenius norm. For these distance matrices, all estimators gave near identical LID estimates compared to the non-Euclidean space, while the "non-Euclideanness" in terms of the critical power increased the error in the LID estimates. From the experiments we derive the claim that LID estimates on non-Euclidean distance matrices can at least in relative magnitude be used as a complexity measure. The estimators are not theoretically sound for these spaces, yet, we can interpret them as approximations in similarly fairing Euclidean proxy spaces. Whether these proxy spaces are approximate or power-transformed does not matter for that purpose. Qualitatively, the $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ are the least correlated estimators for that purpose, unless a low "non-Euclideanness" can be assumed. The investigated practical distance functions exhibit a low enough "non-Euclideanness" to warrant the use of $\mathrm{LID}_{\mathrm{RABID}}$ as a complexity measure. As $\mathrm{LID}_{\mathrm{ALC}}$ estimates are strongly correlated with their counterparts in approximated Euclidean spaces and a linear function of the canonical LID estimates in power-

transformed spaces, it is a good choice for use in non-Euclidean spaces, similar to $\mathrm{LID}_{\mathrm{MLE}}$ and $\mathrm{LID}_{\mathrm{TLE}}$. The almost one-sided error of the different estimators may be exploited to bound the non-Euclidean LID from above with $\mathrm{LID}_{\mathrm{RABID}}$ and $\mathrm{LID}_{\mathrm{ALC}}$ and from below with $\mathrm{LID}_{\mathrm{MLE}}$. Whether this applies well to practical settings remains to be evaluated.

Aside from promoting LID estimators for the use in non-Euclidean space, this section also gives empirical support to the use of the critical power of a non-Euclidean distance matrix as a measure for the "non-Euclideanness" of the space. All other qualitative observations are affected by the critical power. It further implies the "canonical LID" and "canonical Euclidean space", a concept to compensate for the arbitrary numerical values of $\mathrm{LID}_{\mathrm{MLE}}$ estimates on power-transformed spaces and an ideal case for the Euclidean embedding using the power transformation. These terms however are affected by the considered locality size, since the critical power varies with the number of neighbors. The result of the power transform with a global choice of power, thus provides a globally transformed space, that is only locally Euclidean. That concept is familiar to manifold learning techniques, as the geodetic distances on a Riemannian manifold are also only locally Euclidean.

6.5 Conclusion

In this chapter, we took a thorough look at if and how LID estimators can be applied to non-Euclidean spaces and more specifically metric spaces. The theory of the estimators does not allow for the immediate application of the estimators, except for $\mathrm{LID}_{\mathrm{GED}}$ which – with significant analytical work – can be extended to specific non-Euclidean distance functions [47]. However, if we consider an implicit Euclidean space, on which estimates are correlated, then we can interpret the estimates as a complexity measure on these implicit Euclidean spaces. While not necessarily giving exact numbers of coordinates in a Cartesian representation, the relative values can be used to compare the local complexity around different data points. The implicit Euclidean space investigated in this chapter were power transformations and numerical approximations. For both of these implicit Euclidean spaces, the estimates were sufficiently correlated to warrant the use of LID estimators. In the process, we proposed methods to both estimate the maximum power of a power transform and to evaluate the "non-Euclideanness" of observed distance matrices. We gave a specific matrix class that is extremely "non-Euclidean", representing a lower bound for the critical power of a power transform, hypothetically, and argued that practically useful distance functions should be much less "non-Euclidean" than this matrix class. The provided (anecdotal) evidence supports that claim. If a global bound for the critical power can be determined (empirically), local Cartesian representations can even be materialized from the power-transformed distances. While not providing a watertight

theoretical framework, we consider the arguments in this chapter sufficient to warrant the use of LID estimators on non-Euclidean spaces.

Chapter 7

Quo Vadis, LID?

In this chapter, we recapitulate the contents of this thesis and give an outlook on the future of LID estimation. The main focus of the thesis is the introduction of LID estimators with improved properties over previous LID estimators. We introduced the $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ estimators, which work on distributions of pairwise cosines between data points. They expand the $\mathrm{LID}_{\mathrm{MLE}}$ estimator over a bounded geometry and are related to PCA. In empirical studies, $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ showed improved performance over $\mathrm{LID}_{\mathrm{MLE}}$ and familiar estimators. Although it is functionally very similar to the $\mathrm{LID}_{\mathrm{FCI}}$ estimator, it has the benefit of providing a closed form in terms of the cosine matrix, which allows for a thorough analysis of the estimator and derived properties. The $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ estimators are therefore a valuable addition to the toolbox of LID estimators.

To expand on the $\mathrm{LID}_{\mathrm{ABID}}$ and $\mathrm{LID}_{\mathrm{RABID}}$ estimators, we introduced the $\mathrm{LID}_{\mathrm{TRIP}}$ and $\mathrm{LID}_{\mathrm{ALC}}$ estimators, advancing the central concept in $\mathrm{LID}_{\mathrm{RABID}}$ of "angles" to more than two points. The $\mathrm{LID}_{\mathrm{TRIP}}$ estimator extended the pairwise angle to the angle between a point and a linear subspace. The derivation provided not only an approximation for the covariance of normalized samples but also an expected value for the variance under random projections to linear subspaces sampled from a dataset. By parameterizing the estimator with an acceptable loss in accuracy, the $\mathrm{LID}_{\mathrm{TRIP}}$ estimator can be used to estimate the intrinsic dimensionality of a lossy subspace, that retains the most information. We further gave an algorithm to approximate the $\mathrm{LID}_{\mathrm{TRIP}}$ estimator in a computationally efficient manner. Yet, computing the $\mathrm{LID}_{\mathrm{TRIP}}$ estimator exactly is computationally expensive and it is sensitive to residual variance in noisy settings, making it less practical for "all-purpose" LID estimation tasks.

The $\mathrm{LID}_{\mathrm{ALC}}$ estimator instead extends pairwise angles to the angles inside chordal triangles, a measure founded on three rather than two neighbors. Consequentially, it provides a cubic rather than quadratic number of values per estimate, which provides a more exact estimate of the intrinsic dimensionality, especially in very small neighborhoods. The derived variants of the $\mathrm{LID}_{\mathrm{ALC}}$ estimator allow for slightly different settings

and "react" differently to varying scenarios. Aside from being a precise and rather easy-to-implement estimator, this also allows us to employ the estimator variants in ensemble techniques, which could potentially provide even better estimates. We also introduced the empirically derived variants LID_{BALL} and LID_{BALE} to perform the same LID estimation on non-normalized neighborhoods. The LID_{BALE} estimator, albeit lacking a proper theoretical derivation, performed the best in our experiments.

Joining LID estimators in ensemble techniques is a promising approach to improve the accuracy of LID estimates and potentially to detect properties of samples that are not easily detectable by other approaches. LID estimation has, for example, been proposed as a method for Outlier Detection in the past [48]. Whereas a single LID estimator did not suffice for state-of-the-art performance on that task, an ensemble of LID estimators might suffice, since different estimators exhibit vastly different perturbations on outliers and other geometric features. For one, an ensemble of LID estimators could be used to detect the skew of a neighborhood distribution, which is a potential indicator for outliers, that observation could conversely be used to improve the LID estimate quality. With LID_{ABID}, $\text{LID}_{\text{RABID}}$, and LID_{ALC} and its variants, we have introduced a whole collection of estimators that exhibit different perturbations compared to distance-based estimators like LID_{MLE} and to each other. Yet, exploring the ensemble of these estimators is a task for future research.

In Chapter 6, we investigated the applicability of LID estimators to non-Euclidean spaces. We argued that LID estimates on non-Euclidean spaces can be interpreted as a complexity measure on an implicit equivalent or approximate Euclidean space, even if no isometric embedding exists. That circumvents the issue of the estimators' theory not applying to non-Euclidean space. We gave a measure for "non-Euclideanness" and argued that practically useful distance functions are likely not very "non-Euclidean". The empirical evidence provided in this thesis supports this claim. But why was the last chapter of this thesis focused on the extension of LID estimators to non-Euclidean spaces – especially at a time, when latent embeddings of deep neural networks provide Euclidean vectors for almost any input? Why is that extension relevant, and where can the field of LID estimate research go from here?

The investigation is not merely an academic exercise but has practical implications. It effectively allows us to evaluate the complexity of data points given arbitrary similarity or dissimilarity functions following much milder assumptions than the Euclidean case. One such example is the starting point of this thesis: Investigating the "complexity" of words in a corpus, where the similarity function is the cooccurrence of words in documents. Following the interpretation, that the cooccurrence of words hints at semantic relationship [62, Strong Contextual Hypothesis], words that exhibit a large "complexity" should connect multiple semantic subspaces or topics. That indicates stop words, relevant words for multiple topics, or homonyms. While classically not considered homonyms, the first

two cases follow the same concept of homonyms, as they convey different interpretations collected in the same word. The name of a country, e. g., can be used as a geographic descriptor, the collective of a country's citizens, or its government. It is a homonym of these semantic facets. The neural network community attempted to tackle this issue with context rather than word embeddings at the cost of additional computational complexity. Given the assumption that using LID is sound on that similarity function, we should be able to use LID estimates to identify these words. Algorithms could then be developed to better treat these words in natural language processing tasks beyond mere stop word removal.

It further allows the application of LID to control the flow of algorithms on arbitrary data types. An example of such a control mechanism is given by Savic, Kurbalija, and Radovanovic [72]. The authors use a modified $\mathrm{LID}_{\mathrm{MLE}}$ estimator to control random walks on graphs to obtain a better vector representation of graphs by trying not to move past high-LID nodes. The obtained random walks focus on low-LID neighborhoods. Interpreting the nodes as samples on a geometric variety, the high-LID nodes mimic singularities, connecting multiple geodetic localities on an "untangled" manifold. By avoiding these singularities, the random walk can focus on the local structure of the manifold. This concept of "LID-aware" algorithms is founded on the applicability of LID. For spaces that do not admit an isometric Euclidean embedding, the theoretical foundation for the applicability of LID is lackluster. With Chapter 6, we hope to promote the application of LID estimators in LID-aware algorithms beyond the Euclidean case. The most promising application field is perhaps in the control of heuristic algorithms for NP-hard problems. Algorithm schemes like Branch and Bound and Stochastic Local Search use heuristics to guide the search for the optimal solution in super-polynomially hard problems [43] (if P $\neq$ NP). In combinatorial problems, a key problem tackled by heuristics is the selection of the next variable to explore or fix, known as Variable Selection [88, 38]. "Classically" these heuristics are "rules-of-thumb", e. g. considering how many clauses become (un)satisfied by fixing a variable in SAT [88]. More recent approaches employ machine learning to predict the next variable to fix [38]. Both of these approaches give little insight into why a specific heuristic works. We hope, that using LID estimates gives a more principled approach. In a preliminary test on the satisfiability problem on conjunctive normal forms, we created a graph with variables as nodes and edges between any two variables occurring in the same clause. Using the graph distance function, we evaluated $\mathrm{LID}_{\mathrm{ABID}}$ estimates for each variable. First fixing the variables with the highest $\mathrm{LID}_{\mathrm{ABID}}$ value outperformed random choice in terms of visited nodes of the Branch and Bound tree on random formulas. On very hard instances like those of the pigeonhole principle [36], all $\mathrm{LID}_{\mathrm{ABID}}$ estimates were equal. That hints at the estimates being a viable heuristic, as being unable to decide on the next variable to fix is a sign of the problem being very hard. Related to the control of algorithms, there is an additional potential link between LID estimates, i. e.

geometric complexity, and computational complexity. This is due to the observation, that some NP-hard problems are known to be easier to approximate given bounded Euclidean dimensionality [7]. LID may be used to identify more and less complicated parts of the problem space, either to control a solver or purely for analytical purposes.

To summarize, while new estimators that provide better estimates are always welcome, we now have a catalog of estimators that exhibit different perturbations to different geometric features. The next step in improving the LID quality is to combine these estimators in ensemble techniques. Aside from that, we believe that LID estimators should not only be applied to parameterize algorithms but also to control their flow, similar to the advances in LID-aware algorithms [72, 50]. To that end, LID must apply to non-Euclidean spaces, which we have argued in favor of in Chapter 6.

List of Publications

This is a list of all articles written during my doctoral studies. They are listed in chronological order and split into articles that majorly contributed to the thesis, articles that are only cited in the thesis, and papers that did not contribute to the thesis. The authors' contributions to the articles contributing to this thesis are listed under the respective articles.

Articles with major contribution to this thesis

(Best Student Paper)
[84] Erik Thordsen and Erich Schubert. "ABID: Angle Based Intrinsic Dimensionality". In: *Similarity Search and Applications*. Springer International Publishing, 2020, pp. 218–232. DOI: 10.1007/978-3-030-60936-8_17. eprint: 2006.12880
My contribution: Conceptualization, experiments, proofs, writeup. Erich Schubert contributed to the conceptualization and writeup.

(Extended journal version of [84])
[85] Erik Thordsen and Erich Schubert. "ABID: Angle Based Intrinsic Dimensionality - Theory and analysis". In: *Inf. Syst.* 108 (2022), p. 101989. DOI: 10.1016/j.is.2022.101989
My contribution: Conceptualization, experiments, proofs, writeup. Erich Schubert contributed to the conceptualization and writeup.

[87] Erik Thordsen and Erich Schubert. "On Projections to Linear Subspaces". In: *Similarity Search and Applications*. Vol. 13590. Lecture Notes in Computer Science. Springer, 2022, pp. 75–88. DOI: 10.1007/978-3-031-17849-8_7
My contribution: Conceptualization, experiments, proofs, writeup. Erich Schubert contributed to the conceptualization and writeup.

Articles with minor contribution to this thesis

[86] Erik Thordsen and Erich Schubert. "MESS: Manifold Embedding Motivated Super Sampling". In: *Similarity Search and Applications*. Vol. 13058. Lecture Notes in Computer Science. Springer, 2021, pp. 232–246. DOI: 10.1007/978-3-030-89657-7_18
My contribution: Conceptualization, experiments, proofs, writeup. Erich Schubert contributed to the conceptualization and writeup.

Articles with no contribution to this thesis

(Based on my Master's thesis)
Andre Droschinsky, Petra Mutzel, and Erik Thordsen. "Shrinking Trees not Blossoms: A Recursive Maximum Matching Approach". In: *Symposium on Algorithm Engineering and Experiments*. SIAM, 2020, pp. 146–160. DOI: 10.1137/1.9781611976007.12

Erik Thordsen and Erich Schubert. "CANDLE: Classification And Noise Detection With Local Embedding Approximations". In: *Proceedings of the LWDA 2021 Workshops*. Vol. 2993. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 219–231

(Best Paper, Best Student Paper)
Erik Thordsen and Erich Schubert. "An Alternating Optimization Scheme for Binary Sketches for Cosine Similarity Search". In: *Similarity Search and Applications*. Vol. 14289. Lecture Notes in Computer Science. Springer, 2023, pp. 41–55. DOI: 10.1007/978-3-031-46994-7_4

Erik Thordsen and Erich Schubert. "Grouping Sketches to Index High-Dimensional Data in a Resource-Limited Setting". In: *Similarity Search and Applications*. Vol. 15268. Lecture Notes in Computer Science. Springer, 2024, pp. 274–282. DOI: 10.1007/978-3-031-75823-2_23

(Pending submission)
Erik Thordsen and Erich Schubert. "Explicit Formulae to Interchangeably use Hyperplanes and Hyperballs using Inversive Geometry". In: *CoRR* abs/2405.18401 (2024). DOI: 10.48550/ARXIV.2405.18401. arXiv: 2405.18401

(Extended journal version)
Erik Thordsen and Erich Schubert. "An Alternating Optimization Scheme for Binary Sketches". In: *Inf. Syst.* (2025). DOI: 10.1016/j.is.2025.102563

Bibliography

[1] Abdo Y Alfakih et al. *Euclidean distance matrices and their applications in rigidity theory*. Springer, 2018.

[2] Laurent Amsaleg et al. "Estimating Local Intrinsic Dimensionality". In: *International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 29–38. DOI: 10.1145/2783258.2783405.

[3] Laurent Amsaleg et al. "Extreme-value-theoretic estimation of local intrinsic dimensionality". In: *Data Min. Knowl. Discov.* 32.6 (2018), pp. 1768–1805. DOI: 10.1007/s10618-018-0578-6.

[4] Laurent Amsaleg et al. "High Intrinsic Dimensionality Facilitates Adversarial Attack: Theoretical Evidence". In: *IEEE Trans. Inf. Forensics Secur.* 16 (2021), pp. 854–865. DOI: 10.1109/TIFS.2020.3023274.

[5] Laurent Amsaleg et al. "Intrinsic Dimensionality Estimation within Tight Localities". In: *International Conference on Data Mining*. SIAM, 2019, pp. 181–189. DOI: 10.1137/1.9781611975673.21.

[6] Mihael Ankerst et al. "OPTICS: Ordering Points To Identify the Clustering Structure". In: *SIGMOD International Conference on Management of Data*. ACM Press, 1999, pp. 49–60. DOI: 10.1145/304182.304187.

[7] Sanjeev Arora. "Polynomial Time Approximation Schemes for Euclidean Traveling Salesman and other Geometric Problems". In: *J. ACM* 45.5 (1998), pp. 753–782. DOI: 10.1145/290179.290180.

[8] Yong Bao and Raymond Kan. "On the moments of ratios of quadratic forms in normal random variables". In: *J. Multivar. Anal.* 117 (2013), pp. 229–245. DOI: 10.1016/j.jmva.2013.03.002.

[9] Sukarna Barua et al. "Quality Evaluation of GANs Using Cross Local Intrinsic Dimensionality". In: *CoRR* abs/1905.00643 (2019). arXiv: 1905.00643.

[10] Ruben Becker et al. "Subspace Determination Through Local Intrinsic Dimensional Decomposition". In: *Similarity Search and Applications*. Vol. 11807. Lecture Notes in Computer Science. Springer, 2019, pp. 281–289. DOI: 10.1007/978-3-030-32047-8_25.

[11] Patrick Billingsley. *Measure and Probability*. John Wiley & Sons: New York, 1995.

[12] George D. Birkhoff. "A Set of Postulates for Plane Geometry, Based on Scale and Protractor". In: *Annals of Mathematics* 33.2 (1932), pp. 329–345.

[13] Christopher M. Bishop. "Bayesian PCA". In: *Advances in Neural Information Processing Systems*. The MIT Press, 1998, pp. 382–388.

[14] T. Tony Cai, Jianqing Fan, and Tiefeng Jiang. "Distributions of angles in random packing on spheres". In: *J. Mach. Learn. Res.* 14.1 (2013), pp. 1837–1864. DOI: 10.5555/2567709.2567722.

[15] T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. "Optimal rates of convergence for covariance matrix estimation". In: *The Annals of Statistics* 38.4 (2010), pp. 2118–2144. DOI: 10.1214/09-AOS752.

[16] Kevin M. Carter, Raviv Raich, and Alfred O. Hero III. "On Local Intrinsic Dimension Estimation and Its Applications". In: *IEEE Transactions on Signal Processing* 58.2 (2010), pp. 650–663. DOI: 10.1109/TSP.2009.2031722.

[17] Claudio Ceruti et al. "DANCo: An intrinsic dimensionality estimator exploiting angle and norm concentration". In: *Pattern Recognit.* 47.8 (2014), pp. 2569–2581. DOI: 10.1016/j.patcog.2014.02.013.

[18] Edgar Chávez et al. "Searching in metric spaces". In: *ACM Comput. Surv.* 33.3 (2001), pp. 273–321. DOI: 10.1145/502807.502808.

[19] Oussama Chelly, M. E. Houle, and K. Kawarabayashi. "Enhanced Estimation of Local Intrinsic Dimensionality Using Auxiliary Distances". In: *National Institute of Informatic* (2016).

[20] Ji Young Choi and Jonathan D. H. Smith. "On the Unimodality and Combinatorics of Bessel Numbers". In: *Discret. Math.* 264 (2003), pp. 45–54.

[21] Richard Cole and Lee-Ad Gottlieb. "Searching dynamic point sets in spaces with bounded doubling dimension". In: *Symposium on Theory of Computing*. ACM, 2006, pp. 574–583. DOI: 10.1145/1132516.1132599.

[22] J.A. Costa and A.O. Hero. "Geodesic entropic graphs for dimension and entropy estimation in manifold learning". In: *IEEE Trans. Signal Process.* 52.8 (2004), pp. 2210–2221. DOI: 10.1109/TSP.2004.831130.

[23] T. Cover and Joy A. Thomas. *Elements of Information Theory, Second Edition*. John Wiley & Sons, Ltd, 2005. DOI: 10.1002/047174882X.

[24] Shay Deutsch and G. Medioni. "Learning the Geometric Structure of Manifolds with Singularities Using the Tensor Voting Graph". In: *Journal of Mathematical Imaging and Vision* (2017). DOI: 10.1007/s10851-016-0684-2.

[25] Michel Marie Deza and Elena Deza. *Encyclopedia of distances*. Berlin: Springer, 2009. DOI: 10.1007/978-3-642-00234-2.

[26] Mateo Díaz, Adolfo J. Quiroz, and Mauricio Velasco. "Local angles and dimension estimation from data on manifolds". In: *J. Multivar. Anal.* 173 (2019), pp. 229–247. DOI: 10.1016/j.jmva.2019.02.014.

[27] J. Eckmann and D. Ruelle. "Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems". In: *Physica D: Nonlinear Phenomena* (1992). DOI: 10.1016/0167-2789(92)90023-G.

[28] Vittorio Erba, Marco Gherardi, and Pietro Rotondo. "Intrinsic dimension estimation for locally undersampled data". In: *Scientific Reports volume 9, Article number: 17133 (2019)* (June 2019). DOI: 10.1038/s41598-019-53549-9. arXiv: 1906.07670 [cs.LG].

[29] Richard M. Everson and Stephen J. Roberts. "Inferring the eigenvalues of covariance matrices from limited, noisy data". In: *IEEE Trans. Signal Process.* 48.7 (2000), pp. 2083–2091. DOI: 10.1109/78.847792.

[30] Mingyu Fan et al. "Intrinsic dimension estimation of data by principal component analysis". In: (2010). DOI: 10.48550/arxiv.1002.2050. arXiv: 1002.2050.

[31] Keinosuke Fukunaga and David R. Olsen. "An Algorithm for Finding Intrinsic Dimensionality of Data". In: *IEEE Trans. Computers* 20.2 (1971), pp. 176–183. DOI: 10.1109/T-C.1971.223208.

[32] Walter Gautschi. "Some Elementary Inequalities Relating to the Gamma and Incomplete Gamma Function". In: *Journal of Mathematics and Physics* 38 (1959), pp. 77–81.

[33] Peter Grassberger and Itamar Procaccia. "Characterization of Strange Attractors". In: *Phys. Rev. Lett.* 50 (5 1983), pp. 346–349. DOI: 10.1103/PhysRevLett.50.346.

[34] Dominik Gruntz. *On computing limits in a symbolic manipulation system*. ETH Zürich, 1996.

[35] Anupam Gupta, Robert Krauthgamer, and James R. Lee. "Bounded Geometries, Fractals, and Low-Distortion Embeddings". In: *Foundations of Computer Science*. IEEE Computer Society, 2003, pp. 534–543. DOI: 10.1109/SFCS.2003.1238226.

[36] Armin Haken. "The intractability of resolution". In: *Theoretical Computer Science* 39 (1985). Third Conference on Foundations of Software Technology and Theoretical Computer Science, pp. 297–308. DOI: 10.1016/0304-3975(85)90144-6.

[37] Felix Hausdorff. "Dimension und äußeres Maß". In: *Mathematische Annalen* 79.1-2 (1918), pp. 157–179. DOI: 10.1007/BF01457179.

[38] He He, Hal Daumé III, and Jason Eisner. "Learning to Search in Branch and Bound Algorithms". In: *Advances in Neural Information Processing Systems*. 2014, pp. 3293–3301.

[39] Matthias Hein and Jean-Yves Audibert. "Intrinsic dimensionality estimation of submanifolds in R^d". In: *International Conference on Machine Learning*. Vol. 119. ACM International Conference Proceeding Series. ACM, 2005, pp. 289–296. DOI: 10.1145/1102351.1102388.

[40] Nicholas J. Higham. "Computing a nearest symmetric positive semidefinite matrix". In: *Linear Algebra and its Applications* 103 (1988), pp. 103–118. DOI: 10.1016/0024-3795(88)90223-6.

[41] David Hilbert. *Grundlagen der Geometrie*. Springer, 1922.

[42] Bruce M. Hill. "A simple general approach to inference about the tail of a distribution". In: *The Annals of Statistics* 3.5 (1975), pp. 1163–1174.

[43] Holger H. Hoos and Thomas Stützle. "Stochastic Local Search". In: *Handbook of Approximation Algorithms and Metaheuristics*. Chapman and Hall/CRC, 2007. DOI: 10.1201/9781420010749.CH19.

[44] Michael E. Houle. "Local Intrinsic Dimensionality I: An Extreme-Value-Theoretic Foundation for Similarity Applications". In: *Similarity Search and Applications*. Vol. 10609. Lecture Notes in Computer Science. Springer, 2017, pp. 64–79. DOI: 10.1007/978-3-319-68474-1_5.

[45] Michael E. Houle. "Local Intrinsic Dimensionality II: Multivariate Analysis and Distributional Support". In: *Similarity Search and Applications*. Vol. 10609. Lecture Notes in Computer Science. Springer, 2017, pp. 80–95. DOI: 10.1007/978-3-319-68474-1_6.

[46] Michael E. Houle. "Local Intrinsic Dimensionality III: Density and Similarity". In: *Similarity Search and Applications*. Vol. 12440. Lecture Notes in Computer Science. Springer, 2020, pp. 248–260. DOI: 10.1007/978-3-030-60936-8_19.

[47] Michael E. Houle, Hisashi Kashima, and Michael Nett. "Generalized Expansion Dimension". In: *International Conference on Data Mining Workshops*. IEEE Computer Society, 2012, pp. 587–594. DOI: 10.1109/ICDMW.2012.94.

[48] Michael E. Houle, Erich Schubert, and Arthur Zimek. "On the Correlation Between Local Intrinsic Dimensionality and Outlierness". In: *Similarity Search and Applications*. Vol. 11223. Lecture Notes in Computer Science. Springer, 2018, pp. 177–191. DOI: 10.1007/978-3-030-02224-2_14.

[49] David R. Karger and Matthias Ruhl. "Finding nearest neighbors in growth-restricted metrics". In: *Symposium on Theory of Computing*. ACM, 2002, pp. 741–750. DOI: 10.1145/509907.510013.

[50] Dušica Knežević et al. "Evaluation of LID-Aware Graph Embedding Methods for Node Clustering". In: *Similarity Search and Applications*. Springer International Publishing, 2022, pp. 222–233. DOI: 10.1007/978-3-031-17849-8_18.

[51] Mark A. Kramer. "Nonlinear principal component analysis using autoassociative neural networks". In: *AIChE Journal* 37.2 (1991), pp. 233–243. DOI: 10.1002/AIC.690370209.

[52] Hans-Peter Kriegel et al. "A General Framework for Increasing the Robustness of PCA-Based Correlation Clustering Algorithms". In: *Scientific and Statistical Database Management*. Vol. 5069. Lecture Notes in Computer Science. Springer, 2008, pp. 418–435. DOI: 10.1007/978-3-540-69497-7_27.

[53] Tarald O. Kvålseth. "Entropy and Correlation: Some Comments". In: *IEEE Trans. Syst. Man Cybern.* 17.3 (1987), pp. 517–519. DOI: 10.1109/TSMC.1987.4309069.

[54] Elizaveta Levina and Peter J. Bickel. "Maximum Likelihood Estimation of Intrinsic Dimension". In: *Advances in Neural Information Processing Systems*. 2004, pp. 777–784.

[55] E. Lorenz. "Deterministic nonperiodic flow". In: *Journal of the Atmospheric Sciences* (1963). DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

[56] Xingjun Ma et al. "Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality". In: *International Conference on Learning Representations*. 2018.

[57] L. V. D. Maaten and Geoffrey E. Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* (2008).

[58] Hiroshi Maehara. "Euclidean embeddings of finite metric spaces". In: *Discret. Math.* 313.23 (2013), pp. 2848–2856. DOI: 10.1016/j.disc.2013.08.029.

[59] Justin Matejka and George W. Fitzmaurice. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing". In: *CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 1290–1294. DOI: 10.1145/3025453.3025912.

[60] Aaron Meurer et al. "SymPy: symbolic computing in Python". In: *PeerJ Computer Science* 3 (Jan. 2017), e103. DOI: 10.7717/peerj-cs.103.

[61] Luisa Micó, José Oncina, and Enrique Vidal. "A new version of the nearest-neighbour approximating and eliminating search algorithm (AESA) with linear preprocessing time and memory requirements". In: *Pattern Recognit. Lett.* 15.1 (1994), pp. 9–17. DOI: 10.1016/0167-8655(94)90095-7.

[62] G. Miller and W. Charles. "Contextual correlates of semantic similarity". In: *Language and Cognitive Processes* 6.1 (1991), pp. 1–28. DOI: 10.1080/01690969108406936.

[63] T. Minka. "Automatic Choice of Dimensionality for PCA". In: *Advances in Neural Information Processing Systems*. Vol. 13. MIT Press, 2000.

[64] Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics. Wiley, 1982. DOI: 10.1002/9780470316559.

[65] Stephen M. Omohundro. *Five balltree construction algorithms*. TR / International Computer Science Institute 89,63. Berkeley, Calif.: International Computer Science Inst., 1989. 22 pp.

[66] Karl Pearson. "On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11 (1901), pp. 559–572.

[67] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *J. Mach. Learn. Res.* 12 (2011), pp. 2825–2830. DOI: 10.5555/1953048.2078195.

[68] Julius von Rohrscheidt and Bastian Rieck. "Topological Singularity Detection at Multiple Scales". In: *International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 35175–35197.

[69] Simone Romano et al. "Measuring dependency via intrinsic dimensionality". In: *International Conference on Pattern Recognition*. IEEE, 2016, pp. 1207–1212. DOI: 10.1109/ICPR.2016.7899801.

[70] Sam T. Roweis and Lawrence K. Saul. "Nonlinear Dimensionality Reduction by Locally Linear Embedding". In: *Science* 290.5500 (2000), pp. 2323–2326. DOI: 10.1126/SCIENCE.290.5500.2323.

[71] Alessandro Rozza et al. "Novel high intrinsic dimensionality estimators". In: *Mach. Learn.* 89.1-2 (2012), pp. 37–65. DOI: 10.1007/s10994-012-5294-7.

[72] Milos Savic, Vladimir Kurbalija, and Milos Radovanovic. "Local Intrinsic Dimensionality and Graphs: Towards LID-aware Graph Embedding Algorithms". In: *Similarity Search and Applications*. Vol. 13058. Lecture Notes in Computer Science. Springer, 2021, pp. 159–172. DOI: 10.1007/978-3-030-89657-7_13.

[73] I. J. Schoenberg. "Remarks to Maurice Frechet's Article "Sur La Definition Axiomatique D'Une Classe D'Espace Distances Vectoriellement Applicable Sur L'Espace De Hilbert". In: *Annals of Mathematics* (1935). DOI: 10.2307/1968654.

[74] Erich Schubert. "A Triangle Inequality for Cosine Similarity". In: *Similarity Search and Applications*. Vol. 13058. Lecture Notes in Computer Science. Springer, 2021, pp. 32–44. DOI: 10.1007/978-3-030-89657-7_3.

[75] Erich Schubert, Andreas Lang, and Gloria Feher. "Accelerating Spherical k-Means". In: *Similarity Search and Applications*. Vol. 13058. Lecture Notes in Computer Science. Springer, 2021, pp. 217–231. DOI: 10.1007/978-3-030-89657-7_17.

[76] Erich Schubert and Arthur Zimek. "ELKI: A large open-source library for data analysis - ELKI Release 0.7.5 "Heidelberg"". In: *CoRR* abs/1902.03616 (2019). arXiv: 1902.03616.

[77] Erich Schubert and Arthur Zimek. *ELKI Multi-View Clustering Data Sets Based on the Amsterdam Library of Object Images (ALOI)*. Zenodo. 2010. DOI: 10.5281/zenodo.6355684.

[78] Wolfram Schwabhäuser, Wanda Szmielew, and Alfred Tarski. *Metamathematische Methoden in der Geometrie*. Springer Berlin Heidelberg, 1983.

[79] Panagiotis Sidiropoulos. "N-sphere chord length distribution". In: *CoRR* (2014). arXiv: 1411.5639 [math.PR].

[80] F Takens. "On the numerical determination of the dimension of an attractor". In: *Dynamical Systems and Bifurcations*. Springer Berlin Heidelberg, 1985, pp. 99–106. DOI: 10.1007/BFB0075637.

[81] Yee Whye Teh, David Newman, and Max Welling. "A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation". In: *Advances in Neural Information Processing Systems*. MIT Press, 2006, pp. 1353–1360.

[82] Piotr Tempczyk et al. "LIDL: Local Intrinsic Dimension Estimation Using Approximate Likelihood". In: *International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 21205–21231.

[83] J. Tenenbaum, V. Silva, and J. Langford. "A global geometric framework for nonlinear dimensionality reduction." In: *Science* 290.5500 (2000), pp. 2319–2323. DOI: 10.1126/SCIENCE.290.5500.2319.

[84] Erik Thordsen and Erich Schubert. "ABID: Angle Based Intrinsic Dimensionality". In: *Similarity Search and Applications*. Springer International Publishing, 2020, pp. 218–232. DOI: 10.1007/978-3-030-60936-8_17. eprint: 2006.12880.

[85] Erik Thordsen and Erich Schubert. "ABID: Angle Based Intrinsic Dimensionality - Theory and analysis". In: *Inf. Syst.* 108 (2022), p. 101989. DOI: 10.1016/j.is.2022.101989.

[86] Erik Thordsen and Erich Schubert. "MESS: Manifold Embedding Motivated Super Sampling". In: *Similarity Search and Applications*. Vol. 13058. Lecture Notes in Computer Science. Springer, 2021, pp. 232–246. DOI: 10.1007/978-3-030-89657-7_18.

[87] Erik Thordsen and Erich Schubert. "On Projections to Linear Subspaces". In: *Similarity Search and Applications*. Vol. 13590. Lecture Notes in Computer Science. Springer, 2022, pp. 75–88. DOI: 10.1007/978-3-031-17849-8_7.

[88] Dave A. D. Tompkins, Adrian Balint, and Holger H. Hoos. "Captain Jack: New Variable Selection Heuristics in Local Search for SAT". In: *Theory and Applications of Satisfiability Testing - SAT*. Vol. 6695. Lecture Notes in Computer Science. Springer, 2011, pp. 302–316. DOI: 10.1007/978-3-642-21581-0_24.

[89] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature Methods* 17 (Feb. 2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2. arXiv: 1907.10121 [cs.MS].

[90] Han Xiao, Kashif Rasul, and Roland Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. Aug. 28, 2017. arXiv: cs.LG/1708.07747 [cs.LG].