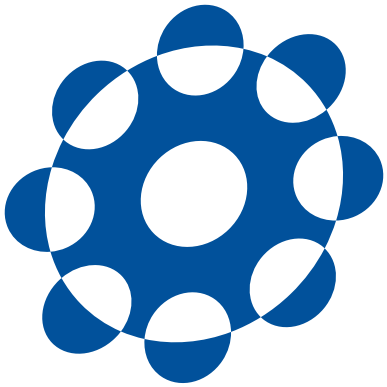
Intrinsic Dimensionality: from Estimation to Applications

Oussama Chelly

Doctor of Philosophy

Department of Informatics
School of Multidisciplinary Sciences
SOKENDAI (The Graduate University for Advanced Studies)



Intrinsic Dimensionality: from Estimation to Applications

本質次元： 推定からアプリケーションへ

Doctoral Dissertation

Oussama Chelly

(ウセーマ・シェリー)

The Graduate University for Advanced Studies

December 2016

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Abstract

The 21st century has witnessed the advent of the Big Data era. Given the important data-related challenges, interest in analyzing and processing large datasets became more prominent than ever.

Whenever performing a task on a large data set fails, it is very hard to assess with certainty whether the failure is related to the difficulty of the task, to the choice of the algorithm and its settings, or to the data itself. In fact, data can sometimes be intrinsically hard.

The number of instances or the number of features alone do not fully describe the degree of difficulty in the data. Therefore many metrics have been designed in order to quantify data complexity. Many metrics such as entropy attempt to describe the data complexity from an information-theoretic point of view. Other metrics such as Intrinsic Dimensionality (ID) describe the geometric complexity. Intrinsic Dimensionality can be defined as the minimum number of attributes required to describe the data without information loss, or as the space-filling capacity of the data.

The accuracy and efficiency of many algorithms in the areas of Artificial Intelligence, Data Mining, Machine Learning, Pattern Recognition and Similarity Search depend on the quality of ID measures. Therefore, many algorithms for estimating ID have been proposed.

The main focus in this thesis is the estimation of Intrinsic Dimensionality.

Statement of Originality

In this statement, I list the contributions reported in this dissertation. Some of these contributions were obtained as part of collaborations with my co-advisors and fellow researchers. This statement aims at discerning my contributions from the work of my collaborators for the purposes of evaluation of this dissertation. My collaborators deserve all the credit for the guidance, the constructive comments, and the advice they provided.

- In Chapter 2 I surveyed the state of the art of Intrinsic Dimensionality estimation.
- In Chapter 3 I surveyed the elements of Extreme Value Theory which are necessary to the understanding of this work, as well as the main methods for estimating the tail index of probability distributions.
- In Chapter 4 I summarized the main results on Local Intrinsic Dimensionality obtained by Houle [Hou15] upon which parts of this work are based.
- Chapter 5 presents research work done in collaboration with Laurent Amsaleg, Teddy Furon, Stéphane Girard, Michael E. Houle, Ken-ichi Kawarabayashi, and Michael Nett. An initial version of this work appeared in the paper entitled "Estimating Local Intrinsic Dimensionality" [ACF+15] presented in the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2015). A more complete version entitled "Extreme-Value-Theoretic Estimation of Local Intrinsic Dimensionality" has been accepted for publication in the Journal of Data Mining and Knowledge Discovery (DAMI). Concretely, the MLE, MoM, and PWM estimators were obtained by Michael Nett and myself under the guidance of Michael E. Houle, and verified by the other coauthors. The work on RV estimators was carried out by Stéphane Girard and myself. Specifically, I proved that both Hill's and Karger & Ruhl's estimators are special cases within the RV family, and proved the main theoretical result on RV estimators in Lemma 1 which shows how to obtain an estimator with minimum variance. The experiments on artificial distance distributions, on artificial manifolds, and the entire framework involving approximate nearest neighbor distances were done by myself, while Laurent Amsaleg,

Teddy Furon, Michael Nett, and I worked together on experimentation with real world data sets. I took the lead in writing the research papers that summarize the outcomes of this research. These papers were polished by Michael E. Houle, and had inputs and comments from all coauthors.

- Chapter 6 presents research I carried out under the co-supervision of Michael E. Houle and Ken-ichi Kawarabayashi. The theoretical result on the cumulated volume of internally tangent balls was obtained by Michael E. Houle and myself. The rest of the results as well as the experimental study are my own work.

- Chapter 7 is unpublished research I performed under the co-supervision of Michael E. Houle and Ken-ichi Kawarabayashi.

Every contribution that is not listed above is my own work.

— Oussama Chelly, December 2016

Acknowledgements

I am obliged to my co-advisors Michael E. Houle and Ken-ichi Kawarabayashi not only for their academic guidance but also for their personal assistance and financial support. Throughout the course of my Ph.D., they have always set the highest standards, have sustained an endless encouragement and help, and have offered the most valuable advice.

I would like to acknowledge the financial support received in the form of a Research Assistantship from the Japan Science and Technology Agency (JST) Large Graph Project, and from the Japan Society for the Promotion of Science (JSPS) Kakenhi Kiban B Project "Practical and Effective Data Mining Via Local Intrinsic Dimensional Modeling".

I am furthermore honored to have collaborated with Laurent Amsaleg, James Bailey, Sarah Erfani, Teddy Furon, Stéphane Girard, Kohei Hayashi, Hervé Jégou, Peer Kröger, Michael Nett, Vinh Nguyen, Vincent Oria, Miloš Radovanović, Mahito Sugiyama, Erich Schubert, and Arthur Zimek. Our challenging discussions enriched my research experience. I am particularly grateful for Erich Schubert and Arthur Zimek for hosting my visit to Ludwig Maximilian University of Munich.

My thanks go to all of my labmates: Faizy Ahsan, Dominique Barbe, Nawal Benabbou, William Blacoe, Brankica Bratić, Guillaume Casanova, Jean Coquet, Elias Englmeier, Xueyao Jiang, Marie Kiermeier, Kamil Krynicki, Xiguo Ma, Tobias Moritz, Kha Nguyen, Simone Romano, Jichao Sun, Natalie Thurlby, Weeris Treeratanajaru, Jorge Valhondo, Jonathan von Brünken, Arwa Wali, Xiaoting Wang, Simon Wollwage, Hanting Xie, and Shaonan Zhang. My thanks also go to all of my university colleagues: Nestor Alvaro, Vanessa Bracamonte, Rathachai Chawuthai, Aniek Dharmayanthi, Viktors Garkavijs, Kent Kawashima, Lika Okamoto, Rungsiman Nararatwong, Nhi Quach, Arunima Sikdar, Xuan Thien, Ario Widyoutomo, and Lihua Zhao.

My thanks go to Henri Angelino, Yu Horishita, Tsukushi Inagaki, Mineyo Iwase, Mio Kobayashi, Kyoko Nakajima, Yoko Nakayama, Motoko Okumoto, Mio Takahashi, Hiroko Tokuda, Maiko Tsuruoka, Satoko Tsushima, and Mayuko Yamaguchi for their assistance with the administrative matters, and the activities organized in the university. I sincerely salute all those who work behind the scenes to make The National Institute of Informatics and The Graduate University for Advanced Studies an enjoyable place to study and work.

I am indebted to all of my teachers and professors from L'École Primaire Farhat Hached - La Marsa, Le Collège Taieb M'hiri - La Marsa, Le Lycée Bourguiba Pilote de Tunis, L'Institut Préparatoire aux Études Scientifiques et Techniques (IPEST), L'École Nationale d'Informatique et de Mathématiques Appliquées de Grenoble (Ensimag), L'Institut National Politechnique de Grenoble (Grenoble INP), La Universidad Politécnica de Madrid (UPM), The National Institute of Informatics (NII), and The Graduate University for Advanced Studies (Sokendai). I could not go through the Ph.D. experience without the knowledge I learned from them.

My deepest gratitude goes to Chiheb, Douaa, Lassad, Mariame, Meriem, Mohamed, Mourad, Nawel, Olfa, Omar, Rami, Sahar, Sonia, Souha, Yamen, and to all of my friends at the Tokyo International Exchange Center (TIEC) for being the family I relied upon during the hard times.

Last but not least, I want to thank my family Jalel, Essia, Syrine and Yasmine and dedicate this thesis to show my gratitude for their endless love, care, and support.

— Oussama Chelly, January 2017

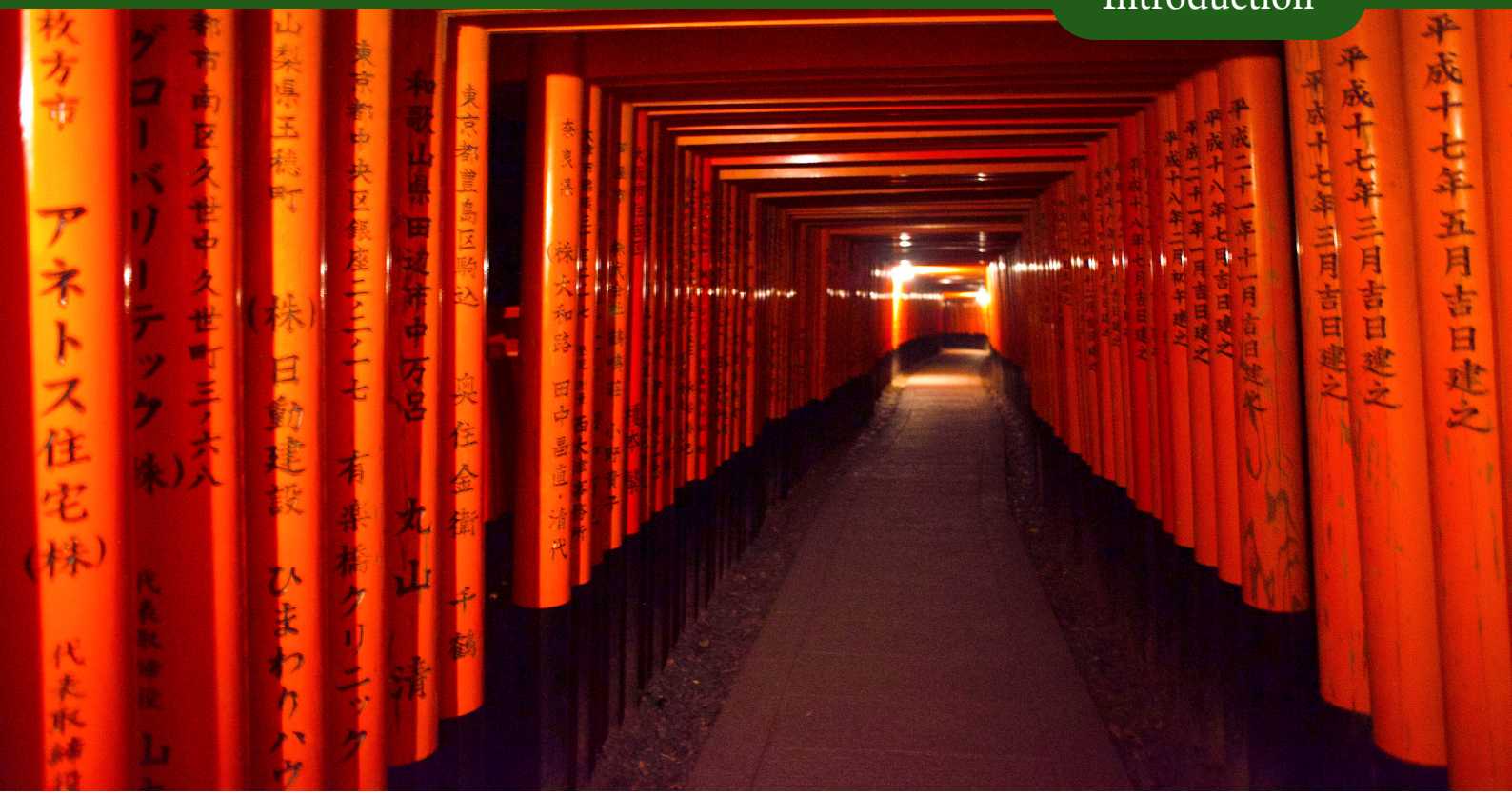
Contents

Part I

Introduction

Chapter 1

Introduction



The 21st century has witnessed the advent of the Big Data era. Given the important data-related challenges, interest in analyzing and processing large datasets became more prominent than ever.

Whenever performing a task on a large data set fails, it is very hard to assess with certainty whether the failure is related to the difficulty of the task, to the choice of the algorithm and its settings, or to the data itself. In fact, data can sometimes be intrinsically hard. The number of instances or the number of features alone do not fully describe the degree of difficulty in the data.

Both the efficiency and efficacy of fundamental operations in areas such as search and retrieval, data mining, and machine learning commonly depend on the interplay between measures of data similarity and the choice of features by which objects are represented. In settings where the number of features (the so-called representational dimension) is high, similarity values tend to concentrate strongly about their respective means, a phenomenon widely referred to as 'the curse of dimensionality'. Consequently, as the dimensionality increases, the discriminative ability of similarity measures diminishes to a point where methods that depend on them lose their

effectiveness [WSB98, BGRS99, Pes00].

Many metrics such as entropy attempt to describe the data complexity from an information-theoretic point of view. Other metrics such as Intrinsic Dimensionality (ID) describe the geometric complexity. Intrinsic Dimensionality can be defined as the minimum number of attributes required to describe the data without information loss, or as the space-filling capacity of the data.

The representational dimension alone cannot explain the curse of dimensionality. This can be seen from the fact that the number of degrees of freedom within a subspace or manifold is independent of the dimension of the space in which it is embedded. This number is often described as the 'intrinsic dimensionality' of the manifold or subspace.

In an attempt to improve the discriminability of similarity measures, and the scalability of methods that depend on them, much attention has been given in the areas of machine learning, databases, and data mining to the development of dimensional reduction techniques. Linear techniques for dimensionality reduction include Principal Component Analysis (PCA) and its variants [Jol86, BCG11, GR70]. Nonlinear dimensionality reduction methods (also known as manifold learning methods) include Isometric Mapping [TDSL00], Multi-Dimensional Scaling [TDSL00, VK06], Locally Linear Embedding and its variants [RS00], Hessian Eigenmapping Spectral Embedding [DG03], Local Tangent Space Alignment [ZZ04], and Non-Linear Component Analysis [SSM98]. Most dimensional reduction techniques require that a target dimension be provided by the user, although some attempt to determine an appropriate dimension automatically. Ideally, the supplied dimension should depend on the intrinsic dimensionality (ID) of the data. This has served as a prime motivation for the development of models of ID, as well as accurate estimators.

Over the past few decades, many practical models of the intrinsic dimensionality of datasets have been proposed. Examples include the previously mentioned Principal Component Analysis and its variants [Jol86, BCG11], as well as several manifold learning techniques [SSM98, RS00, VK06, KJ94]. Topological approaches to ID estimate the basis dimension of the tangent space of the data manifold from local samples [FO71, BS98, PBJD79, VD95]. Fractal measures such as the Correlation Dimension (CD) estimate ID from the space-filling capacity of the data [FK94, CV02, GKL03]. Graph-based methods use the k-nearest neighbor graph along with density in order to measure ID [CHI04]. Parametric modeling and estimation of distribu-

tion often allow for estimators of intrinsic dimension to be derived [LL02, LB04].

The aforementioned ID measures can be described as 'global', in that they consider the dimensionality of a given set as a whole, without any individual object being given a special role. In contrast, 'local' intrinsic dimensionality models can provide different ID measurements depending on the location within the same dataset. These local approaches can be very useful when the data underlying model consists of several manifolds of heterogeneous dimensionality. Several local intrinsic dimensionality models have been proposed recently, such as the expansion dimension (ED) [KR02], the generalized expansion dimension (GED) [HKN12], the minimum neighbor distance (MiND) [RLC+12], and local continuous intrinsic dimension (LID) [Hou13]. These models quantify ID in terms of the rate at which the number of encountered objects grows as the considered range of distances expands from a reference location.

In general, machine learning techniques that rely too strongly on local information can be accused of overfitting the data. This has motivated the development of global techniques for manifold learning such as Local Tangent Space Alignment, which first identifies manifolds restricted to neighborhoods of selected points, and then optimizes the alignment of these local structures in order to produce a more complex description of the data [ZZ04]. The alignment process often involves an explicit penalty for overfitting. In general, local learning can compensate for overfitting by accounting for it in the final optimization process for the alignment of the local manifolds.

In addition to applications in manifold learning, measures of local ID have been used in the context of similarity search, where they are used to assess the complexity of a search query [KR02], or to control the early termination of search [HMNO12, HMOS14]. They have also found applications in outlier detection, in the analysis of a projection-based heuristic [dVCH12], and in the estimation of local density [vBHZ15]. The efficiency and effectiveness of the algorithmic applications of local intrinsic dimensional estimation (such as [HMNO12, HMOS14]) depends heavily on the quality of of the estimators employed.

In contexts where the end application uses local ID measurements only without any need for locally modeling data in terms of manifolds [KR02, HMNO12, HMOS14, dVCH12, vBHZ15], distance-based ID models are often very useful. Indeed, these local ID models such as the ED [KR02] and the LID [Hou13] rely on the distances to the nearest neighbors in order to assess dimensionality. This type

of local estimators is well-suited for many of the applications in question since the nearest neighbor distances are precomputed and available for the estimators.

When data points are viewed as a sample drawn from an underlying continuous distribution, distances from a fixed query location to the data points can be seen as realizations of a continuous positive random distance variable. In this case, the smallest distances (i.e. distances to the nearest neighbors) encountered would be 'extreme events' associated with the lower tail of the distance distribution.

In Extreme Value Theory (EVT), a discipline of statistics concerned with the study of tails of continuous probability distributions, the random variable associated with nearest neighbor distances can be assumed to follow a power-law distribution, where the exponent can be viewed as a form of dimension [CBTD01]. Specifically, continuous lower-bounded random variables are known to asymptotically converge to the Weibull distribution as the sample size grows, regardless of the original distance measure and its distribution. In an equivalent formulation of EVT due to Karamata, the cumulative distribution function of a tail distribution can be represented in terms of a regularly-varying (RV) function whose dominant factor is a polynomial in the distance [CBTD01, Hou15]; the degree (or 'index') of this polynomial factor determines the shape parameter of the associated Weibull distribution, or equivalently the exponent of the associated power law. The index has been interpreted as a form of intrinsic dimension [CBTD01]. Maximum likelihood estimation of the index leads to the well-known Hill estimator for power-law distributions [H$^+$75].

While EVT provides an asymptotic description of tail distributions, in the case of continuous distance distributions, the distribution can be exactly characterized in terms of LID [Hou15]. The LID model introduces a function that assesses the discriminative power of the distribution at any given distance value [Hou13,Hou15]. A distance measure is described as 'discriminative' when an expansion in the distance results in a relatively small increase in the number of observations. This function is shown to fully characterize the cumulative distribution function without the explicit involvement of the probability density [Hou15]. The limit of this function yields the skewness of the Weibull distribution (or equivalently, the Karamata representation index, or power law exponent) associated with the lower tail.

Within the LID model, intrinsic dimensionality can be interpreted interms of the indiscriminability of the distance measure. In fact, when the cumulative distribution function is differentiable, intrinsic dimensionality is the inverse of an expansion-

based measure of indiscriminability. Hence, in addition to the more traditional applications stated earlier, LID has the potential for wide application in many machine learning and data mining contexts, as it makes no assumptions on the nature of the data distribution other than continuity. Moreover, the interpretation of LID in terms of the the indiscriminability of the distance measure naturally lends itself to the design of outlier detection techniques [vBHZ15], and in the understanding of density-related phenomena such as the hubness of data [RNI10a, RNI10b, Hou15].

The main focus of this thesis is the estimation of Local Intrinsic Dimensionality. The central chapters of the dissertation propose an elaborate way of estimating LID. The proposed estimators have better theoretical foundations than state-of-the-art distance-based estimators since they make weaker assumptions. Precisely, the proposed estimators assume a continuous distance with a differentiable cumulative distribution function unlike their counterparts where assumptions on the nature of the underlying manifolds are necessary [LB04, RLC+12]. In practice, the theoretical foundations of LID allow the use of any distance sample not necessarily radial (i.e. from a reference point) provided that the associated distance variable is well defined and follows the previously mentioned assumptions. In addition, our best estimators are more practical in terms of computational time complexity than many of the state-of-the-art estimators and have the advantage of converging faster than other distance-based estimators.

Part I of this thesis contains the introductory material. The next Chapter is dedicated to surveying the state of the art of ID estimation. We first describe the attributes desired from an ID estimator, Then we survey both global and local ID estimation methods. In Chapter 3 we introduce the notions of Extreme Value Theory that are related to this work. We then summarize the different approaches used in EVT to model the tails of probability distributions, and we survey some of the estimators of the index of EVT distributions. In the last introductory chapter, we introduce Local Intrinsic Dimensionality (LID), a recent model of ID upon which our estimators are built.

Part II is the core of this thesis. Chapters 5 and 6 establish connections between the estimation of tail indices in EVT on the one hand, and the estimation of LID in the theory of machine learning and data mining on the other hand. In Chapter 5, we use statistical methods to estimate LID. The main theoretical contributions made in this chapter include a framework for the estimation of LID based on commonly-used statistical estimation techniques such as Maximum Likelihood Esti-

mation, the Method of Moments, and the method of Probability-weighted moments; a new family of estimators based on the Regularly-Varying functions model of EVT where several existing models are shown to be special cases; as well as confidence intervals for the proposed estimators. The main experimental contributions include an empirical study using both artificial and real data to show the advantages and the limitations of our methods when compared with the state of the art global and local estimators; experiments on artificial distance distributions confirming the theoretical convergence of our estimators; experiments validating the advantage of using local ID estimators over their global counterparts in the case of non-linear manifolds; profiles of well-known real-world datasets in terms of LID highlighting the variability of data complexity from region to region within one same dataset; and an experimental study showing the robustness of our estimators when approximate nearest-neighbor distances are used instead of exact distances.

In Chapter 5, the proposed estimators use direct distances from a reference point to its neighbors. In order to enhance the distance sample size, it is possible to use 'auxiliary' distances which are the distances between pairs of neighbors without increasing the neighborhood size. In Chapter 6, we develop a new family of ID estimators that improve the convergence of our previously proposed estimators by using these auxiliary distances. Using all pairwise distances within a neighborhood leads to biased estimators. Our main contribution in this chapter is choice of distances that do not introduce bias, which are those between a neighbor and its neighbors contained in a ball internally tangent to the original locality. Other contributions made in this Chapter include a new local ID estimator that uses auxiliary distances, a theoretical analysis on the expected number of distance samples available to the estimator; an experimental comparison of our estimators with state-of-the-art estimators that extends the experiments of Chapter 5, and which shows the convergence and bias of our estimators on both synthetic and real data.

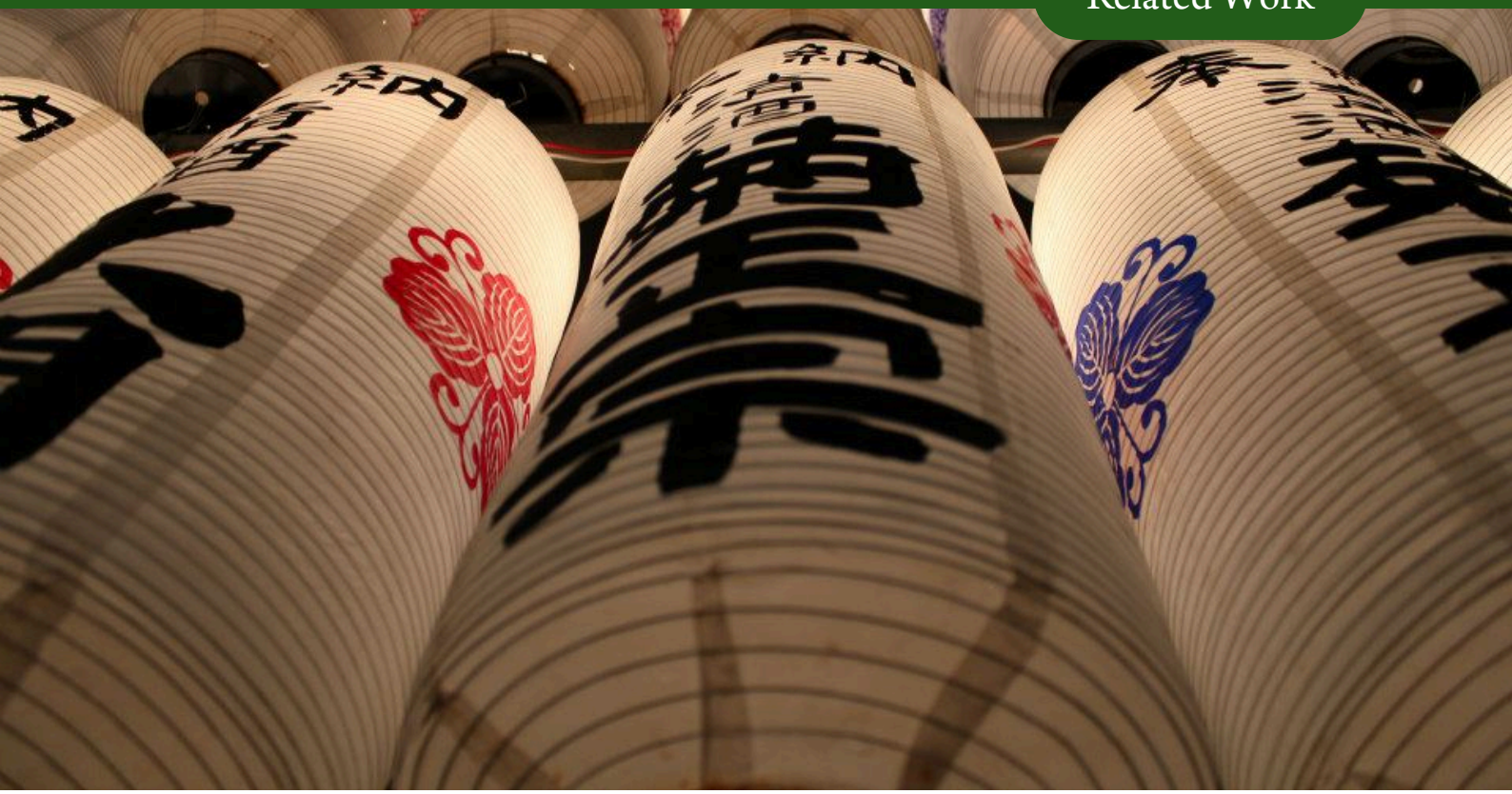
Part III of this thesis provides an example of how ID estimates can be used as indiscriminability measurements in feature selection. Unlike other feature selection algorithms where ID is only used to estimate the number of features to be conserved, we propose in Chapter 7 new feature selection algorithms where ID is used to guide the selection process. Two algorithms are proposed, one being univariate and the other being multivariate. They are a use case for LID as an indiscriminability measure. The proposed algorithms perform better than the state of the art on high-dimensional large data. Their failure on small data sets highlights the limits of

the LID model. Indeed, when the available data sample is small, the number of nearest neighbor distances used in the estimation of LID can no longer be small enough (as compared with the size of the full data sample) to remain within the EVT assumptions. This limitation applies to other local ID estimators but the LID model has the advantage of providing theoretical explanations. In addition to the univariate and the multivariate feature selection algorithms that we propose, the contributions include a theoretical analysis based on the notion of submodularity for the multivariate algorithm and an experimental framework for feature selection setting random as a baseline. Experiments show the advantage of using our methods in high-dimensional settings, and the limitations of our methods on small datasets.

Part IV is the conclusion where we summarize the main research findings, and where we propose some directions for future research in this area. We finally discuss the implications of this research for machine learning and data mining.

Chapter 2

Related Work



Intrinsic Dimensionality (ID) is not a new topic in machine learning. However, it gained a lot of attention over the past few years due to its theoretical and practical implications on several machine learning and data mining algorithms.

Dimensionality does not have a unique definition. Different mathematical disciplines use different definitions of the notion of dimensionality. Various definitions have been proposed such as the Hausdorff Dimension [Hau18] and the Information Dimension [Rén59, Ish93]. We can intuitively define ID as the space-filling capacity of data, or as the minimum number of attributes required to represent given data. From a topological point of view, ID has to be an integer. However, from an information theory point of view ID can be any real number.

It is possible to classify ID estimation approaches according to many criteria. In this work, we adopt the locality criterion. Hence, an estimator is said to be 'global' when dimensionality is assumed to be constant across all data. By contrast, a model is described as 'local' when it assumes that dimensionality can vary from one location to another within the same data set.

In this chapter, we first define the attributes desired from an 'ideal' ID estima-

tor. Then, we survey the state of the art for both local and global approached for measuring intrinsic dimensionality.

2.1 Properties desired in an Intrinsic Dimensionality estimator

Intrinsic Dimensionality estimation has a wide range of applications in machine learning and data mining. Camastra's [CS16] extends builds on Pestov's [Pes08] work in an attempt to enumerate the attributes of an 'ideal' ID estimator. However with each end application having its own limitations and requirements, it is difficult to define such universally 'perfect' ID estimator that can be used across all tasks. Among the qualities that are usually desired from an ID estimator independently of the end application, we cite the following properties:

1. consistency, i.e. convergence to the true ID being estimated as the data sample size increases,
2. accuracy independently of the true ID being estimated,
3. having a 'reasonable' computational complexity,
4. independence of the dimensionality of the representational space.

The first property is an essential attribute of any ID estimator. As the size of the sample used in the estimation increases, the ID measurement provided by the estimator must approach the true ID value of the data generation model. Having an estimator where adding samples does not guarantee a better measurement is in an absolute sense not a desirable effect. The second property means that the true ID of data should not influence the estimation process. In practice, the true ID does have an impact on the convergence and on the bias of the estimation. Many if not most estimators tend to underestimate high values of ID. Moreover, the convergence tends to be slower for these values. The third attribute often depends on the end application. Ideally, the computation of ID should not slow down the algorithm where ID estimates are being used. In practice, there are cases where a higher computational cost is traded for better learning results. Some estimators do not satisfy the fourth and final property since the dimensionality of the space is sometimes directly or inherently an input of the estimator.

It is hard to conceive an ID estimator that satisfies all of these requirements simultaneously. In fact, these properties often conflict one another. As the true ID increases, many estimators tend to underestimate dimensionality thereby violating the consistency property. Moreover, the huge sample size required for convergence especially for high ID values may sometimes render the estimator computationally impractical. Additionally, the dependence on the dimensionality of the representational space sometimes introduces a bias that cannot be resolved by increasing the sample size. In some contexts obtaining a solution which is at the same time optimal and computationally scalable is unrealistic. A concrete example is that of ID models where the estimation requires the solution of an NP-hard problem [FQZ09]. On the one hand, attempting to find an optimal solution to the problem leads to a combinatorial explosion. On the other hand, using heuristics to approximate a solution for the NP-hard step only comes at the cost of optimality. Finally, a higher representational dimension leads to a higher complexity as accessing information from the sample would require more computation.

Besides the conflicts between the mentioned attributes, different applications of ID estimation may have different requirements. In fact, an estimator that is well-suited for a particular task may well be unsuited for a another task. For example, while many feature engineering algorithms require a single ID measurement that describes the entire data set, subspace clustering algorithms require different local estimates of ID for various data subsets.

The task dependency of the qualities desired from an ID estimator can be highlighted by the estimators' sensitivity to data scale and to noise. In fact, the sensitivity of ID estimators to data scale is required in subspace clustering, while independence of data scale is often a desired attribute when ID estimates are used in feature engineering. The sensitivity to noise is also related to data scale. Indeed it can be claimed that noise has its own underlying ID, and thereby it can be argued that the presence of noise modifies the dimensional properties of data, in particular at a high scale. Thereupon, it is nearly impossible to engineer an ID estimator that can be used across all learning tasks.

The relative importance of the mentioned attributes depends on the task where ID estimation is needed. In many point-wise algorithms for classification and for outlier detection methods [vBHZ15], the ID contrast between different data subset or point locations is far more important than the actual ID values. In this context, accuracy is not as necessary as scalability in these algorithms. Hence, priority

is given to reducing the computational complexity and reducing the variance in ID estimation at the cost of accuracy.

An 'ideal' ID estimator is unrealistic. In fact, as of the time of writing this work no known estimator satisfies all mentioned requirements. Various approaches have been proposed and satisfy some of the aforementioned requirements to various degrees.

2.2 Global models of Intrinsic Dimensionality

Global ID estimation algorithms are those that assume data to have a uniform dimensionality across all data points. Consequently, global estimators provide a single dimensionality measurement on a given data set.

Historically, global estimators appeared before local estimators since data sets were relatively small and the need to describe data fragments locally was limited. When data originates from multiple hidden models with heterogeneous dimensionalities most global estimators tend to detect the highest ID amongst the dimensionalities of the data fragments.

Based on the assumptions they require global estimators of intrinsic dimensionality can be separated into five groups: topological methods, multidimensional scaling, fractal models, geodesic models, and statistical models.

2.2.1 Topological Models

Topological methods, also known as projection methods, assume data to be distributed on a single manifold. They measure intrinsic dimensionality as the dimension of the hyperplane where data can be projected while preserving variance. Due to noise, the projection is usually done at the cost of a given error threshold θ.

2.2.1.1 Principal Component Analysis

Principal Component Analysis (PCA) [Pea01, Jol86] is an orthogonal linear transformation where data is projected into a new vector base. In the new coordinate system, the vectors of the base are ordered in decreasing order of the data variance on each vector's orientation. The vectors in this new base are called principal components.

PCA algorithm can be summarized as follows:

Algorithm 1 PCA given a dataset X and a threshold θ

1. Compute the covariance matrix of X.
2. Permutate the eigenvectors of the covariance matrix such that the corresponding eigenvalues $(\lambda_i)_{i\in[1,D]}$ are in decreasing order, i.e. $\lambda_i > \lambda_j \, \forall i < j$.
3. Return the eigenvectors i such that $\lambda_i/\lambda_1 < \theta \, \forall i < J$.

Given a projection error threshold θ, the first J components such that the corresponding normalized variance $\lambda_i/\lambda_1 < \theta \, \forall i < J$ are called 'principal components'. While PCA is not an ID estimator per se, the number of principal components is viewed as an ID estimate since this number indicates the dimension of the hyperplane required to host the data at the cost of a projection error θ.

PCA has several drawbacks:

- PCA is extremely sensitive to noise. In fact, a single outlier to the point set can change the orientation of the principal components leading to a reassessment of the eigenvalues. The presence of noise and outliers often makes PCA overestimate the ID.
- By assuming data to be laying on a hyperplane PCA overestimates the dimensionality of non-linear manifolds. Consider a cloud of points forming a circle in a 2-dimensional space. ID as estimated by PCA will always take a value of 2, even though it can be claimed that a circle is a one-dimensional geometric setting.
- The choice of the threshold θ has an immediate impact on the ID estimate. Since the choice of the threshold is heuristic, PCA as an estimator of ID is not trustworthy.
- A large data sample is required for obtaining the principal components [Ale76, BK81, GV88, CL92]. More specifically, the sample size required by PCA is exponential on the dimensionality [BY95, HATW98, OC09, MS10].
- PCA is subject to overfitting, particularly when the sample size across the principal components is very small [BY95, HATW98, OC09], or very heterogeneous [MWZH99].

2.2.1.2 Nonlinear Principal Component Analysis

PCA assumes the underlying manifold that contains data to be a linear hyperplane. As a result, PCA has tends to overestimate the ID of non-linear manifolds. A 5-layer Auto-associative Neural Networks (ANN) can be used to avoid PCA's assumption on the linearity of manifolds [KJ94]. The ANN model has a bottleneck structure with 5 layers that are called: input, mapping, bottleneck, demapping, and output. The first and fifth layers have the same number of neurons, so do the second and fourth. The ID estimate in this method is the number of neurons in the bottleneck layer.

Even though the ANN-based approach is better than the original PCA on non-linear manifolds, ANN projections are sometimes sub-optimal [Mal98] and therefore they can lead to inaccurate ID estimates.

2.2.1.3 Bayesian Principal Component Analysis

PCA and its nonlinear variants are deterministic models approaches. They do not associate data with a probabilistic model. Moreover, they lack a method for selecting the number of PCs to be retained. The Probabilistic Principal Component Analysis [TB99] views observed data as the realization of a latent multidimensional variable. Assuming a d-dimensional latent variable U, the prior is distributed as a zero-mean normal distribution with the d-dimensional identity matrix I_d as a covariance matrix. Observed data is a D-dimensional matrix X that relates to the latent model through the equation

$$T = W \cdot U + \mu + \epsilon,$$

where W is a projection matrix, μ is a D-dimensional vector. Noise in the D-dimensional representational space is modeled by ϵ which is a zero-mean normal distribution with variance $\sigma^2 I_D$, with σ being a constant.

Bayesian Principal Component Analysis (BPCA) [TB99] is a Maximum Likelihood Estimation of d. Despite its theoretical foundations, BPCA has strong assumptions on the nature of noise leading it to fail in situations with uniformly distributed noise.

2.2.2 Multidimensional Scaling

While topological models attempt to find a variance-preserving projection, Multidimensional Scaling (MDS) [CC00] models attempt to find a distance-preserving projection. The distortion of the distance measures after the projection is called 'stress' of the projection. The projection with the lowest stress is evaluated for each choice of the dimensionality of the target space. There obviously is no stress when the input space and the output space have the same dimensionality. Then as the dimension of the target space decreases, the stress increases. ID is estimated as the lowest dimensionality for which the gain in terms of stress reduction becomes negligible. In practice, the minimum stress is plotted as a function of the dimensionality, and the estimated ID is the point where the curve starts to flatten.

2.2.2.1 Sammon's Mapping

Sammon's Mapping [Sam69] is a multidimensional scaling method where the stress ϵ is defined as:

$$\epsilon_{\text{Sammon}} = \left[\sum_{x_i, x_j \in X} \delta(x_i, x_j) \right]^{-1} \sum_{x_i, x_j \in X} \frac{[\delta(x_i, x_j) - \delta(h(x_i), h(x_j))]^2}{\delta(x_i, x_j)},$$

where h is the projection and $\delta(x_i, x_j)$ is the distance between x_i and x_j. In order to minimize the stress, Sammon's Mapping uses the gradient-descent algorithm.

2.2.2.2 MDSCAL

MDSCAL [Kru64,RSN72a,RSN72b] was the first method to introduce multidimensional scaling. The stress in this method proposed by Kruskal is similar to Sammon's Mapping, except for using ranks of distances instead of the distances themselves. The stress ϵ can be expressed as:

$$\epsilon_{\text{Kruskal}} = \left[\frac{\sum_{x_i, x_j \in X} [\text{rank}(\delta(x_i, x_j)) - \text{rank}(\delta(h(x_i), h(x_j)))]^2}{\sum_{x_i, x_j \in X} \text{rank}(\delta(x_i, x_j))^2} \right]^{\frac{1}{2}},$$

where h is the projection, $\delta(x_i, x_j)$ is the distance between x_i and x_j, and the rank of the distance amongst all distances between pairs of original points is indicated by $\text{rank}(\delta(x_i, x_j))$.

Sammon's Mapping being both simpler and more accurate than MDSCAL, the latter is no longer of interest for ID estimation.

2.2.2.3 Bennett's algorithm

In 1969, Bennett [Ben69] proposed the first algorithm designed specifically for measuring ID. The algorithm assumes data to be uniformly distributed in a d-dimensional space Under such assumptions, the variance of distances within such sphere is inversely proportional to the dimension d.

The algorithm operates iteratively. Every iteration the algorithm iteration two steps. In the first step, points are relocated in the representational space so as to maximize the variance of pairwise distances. In the second step, the position of the points is adjusted in a way that allows the ranking of every pairwise distant to remain unchanged within its locality. These two steps are repeated until the variance of pairwise distances converges. Then as in PCA, the covariance matrix is computed and the number of 'important' eigenvalues is the dimensionality estimation. Bennett's algorithm combines the defects of both PCA and multiscaling method, such as the need for a threshold to decide the number of prominent eigenvalues. Therefore, it only has historical value [CS16].

2.2.2.4 Chen & Andrews' algorithm

Chen & Andrews [CA74] proposed an improvement to the second step of the original algorithm's iteration. Indeed, they propose a new function for evaluating the quality of local rankings of distances between pairs of points The improvement is minor and does not overcome the limitations encountered in the original algorithm.

2.2.2.5 Curvilinear Component Analysis

Curvilinear Component Analysis (CCA) [DH97] uses a self-organizing neural networks. Self-organizing maps (SOM) [Koh95] are first used for vector quantization. Then vectors are projected non-linearly into a space of a lower dimension. The dimension of the new space is CCA's estimation of the ID.

2.2.3 Fractal Models

Unlike topological models where dimensionality is restricted to being an integer, fractal models estimate ID as a real number. While a non-integer dimensionality

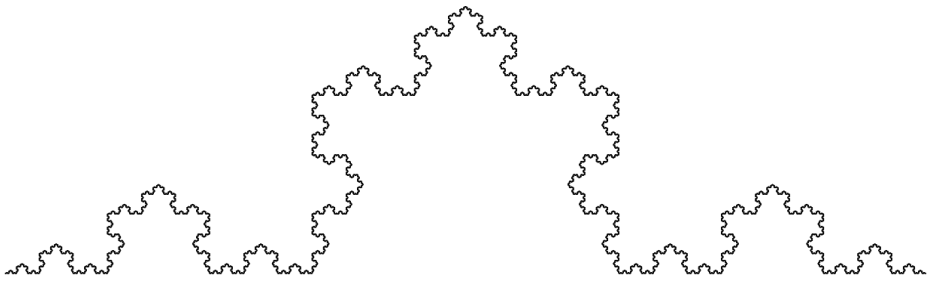


Figure 2.1: Koch Snowflake

may seem incorrect from a topological point of view, the notion of fractal dimensionality is well-founded in fractal geometry since it was first introduced by Mandelbrot [Man67]. Examples of point sets with a fractal dimension include Cantor set [SSS78] with a dimensionality of $\log_3(2)$ [Mar87], Koch curve also known as Koch Snowflake (c.f. Figure 2.1) with a dimensionality of $\log_3(4)$ [Mar87], and Ikeda map [Ike79] (c.f. Figure 2.2) with a dimensionality of 1.7.

Fractal dimensionality measures the space-filling capacity of a point set. Measuring this capacity requires huge point samples. In fact, is the set has an intrinsic dimensionality of d no less than $10^{d/2}$ samples are required to obtain an accurate estimation [ER92, Smi88]. This is the major drawback of fractal models.

The Koch curve for example cannot be projected on a one-dimensional line without loss of information. Nonetheless, Koch curve cannot populate uniformly a two-dimensional space. Hence, the ID of Koch snowflake is between 1 and 2. It was theoretically proved that the Koch Snowflake has an ID equal to $\log_3(4)$, which is often correctly estimated by fractal dimensionality estimators. However, for sets with a higher dimensionality the estimation has an important negative bias.

2.2.3.1 Hausdorff Dimension

Fractal models are heuristics used to estimate the Hausdorff Dimension [Hau18], or more precisely its upper bound the Box-Counting Dimension [Ott02]. In order to define the Hausdorff Dimension we need to introduce the quantity

$$\Gamma^d(r) = \inf_{s_i} \sum_i r_i^m, \tag{2.1}$$

where the set of balls s_i of diameter $r_i < r$ cover the data set.

Figure 2.2: Ikeda map attractor

Definition 1 Hausdorff Dimension The d-dimensional Hausdorff measure is defined as:

$$\Gamma^d = \lim_{r \to 0} \Gamma^d(r). \tag{2.2}$$

The Hausdorff Dimension is the critical value d^* such that $\Gamma^d = +\infty$ if $d > d^*$, and $\Gamma^d = 0$ if $d < d^*$.

Since the Hausdorff Dimension is hard to estimate, fractal models attempt to estimate an upper bound called the Box-Counting Dimension, also known as the Kolmogorov Capacity.

Definition 2 Box-Counting Dimension Let $\nu(r)$ be the number of hypercubes of size r required to cover the point set X. The Box-Counting Dimension is

$$\mathrm{ID}_{\mathrm{BC}} = -\lim_{r \to 0} \frac{\ln \nu(r)}{\ln r}$$

whenever the limit exists.

This definition is motivated by the observation that $\nu(r)$ is proportional to $(1/r)^{\mathrm{ID}_{\mathrm{HD}}}$, where $\mathrm{ID}_{\mathrm{HD}}$ is the Hausdorff Dimension of the set X.

2.2.3.2 Kégl's algorithm

Kégl's Algorithm [Kég02] is an approximation of the Box-Counting Dimension. The algorithm is based on the equivalence between the box cardinality $\nu(r)$ and the cardinality of the maximum independent vertex set of the graph G_r, where G_r is the graph with vertex set X and where the edges are pairs of points such that the pairwise distance is smaller than r.

A greedy heuristic can be used to estimate the maximum independent vertex set of G_r (and therefore the value of $\nu(r)$). Starting from an empty set S, we iterate over X adding to S points that are at distance at least r from all points already in S. The cardinality of S at the end of the iteration is an approximation of the cardinality of the maximum independent vertex set of G_r, and therefore an approximation of $\nu(r)$.

Hence, for a choice of a pair of distances r_1 and r_2 Kégl's estimator of ID is:

$$\widehat{\mathrm{ID}}_{\text{Kégl}} = -\frac{\ln \hat{\nu}(r_2) - \ln \hat{\nu}(r_1)}{\ln(r_2) - \ln(r_1)},$$

where $\hat{\nu}$ denotes the aforementioned approximation of $\nu(r)$ by the maximum independent vertex set of G_r.

The heuristic choice of the distance pair (r_1, r_2) is a drawback of Kégl's method since this choice has a huge impact on the final estimation. Furthermore, Kégl's algorithm is not robust.

2.2.3.3 Quantization-based estimation of intrinsic dimension

Raginsky & Lazebnik's method [RL05] is an improvement to Kégl's estimator where vector quantizers are used instead of box counting in order to assess ID.

It is claimed that Kégl's approach leads to a negative bias due to overfitting, a problem that vector quantizers supposedly address. Nonetheless, the method is computationally expensive and hence is not applicable on large datasets.

2.2.3.4 Grassberger & Procaccia's algorithm

Grassberger & Procaccia's algorithm [GP04] estimates the Correlation Dimension which is a lower bound of the Box-Counting Dimension, and which can be defined as follows.

Definition 3 Correlation Dimension

$$\mathrm{CD} = \lim_{r \to 0} \frac{\ln C(r)}{\ln r},$$

where the correlation integral $C(r)$ is defined as

$$C(r) = \lim_{n \to \infty} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}_{\delta(x_i, x_j) < r}.$$

$\delta(x_i, x_j)$ denotes the Euclidean distance between x_i and x_j. $\mathbb{1}$ denotes the indicator function such that $\mathbb{1}_{d(x_i,x_j)<r} = 1$ if $d(x_i, x_j) < r$ and 0 otherwise.

In Grassberger & Procaccia's algorithm, $C(r)$ is plotted as a function of $\ln(r)$ for different choices of distance r, then the Correlation Dimension is estimated as the slope of the curve for lower values of r for which the curve is usually linear.

2.2.3.5 Fan & al's algorithm

While the estimator originally proposed by Grassberger and Procaccia provides real-valued estimates of the Correlation Dimension, Fan & al [FQZ09] propose to fit the curve point cloud formed by the pairs $(\ln(r), C(r))$ using a polynomial function. The degree of the polynomial that best fits the point cloud $(\ln(r), C(r))$ provides an integer-valued estimation of the dimension.

2.2.3.6 Hein & Audibert's algorithm

The algorithm proposed by Hein and Audibert [HA05] generalizes the Correlation Dimension to any kernel. In other words, the algorithm is a generalization of In Grassberger & Procaccia's approach.

2.2.3.7 Takens' algorithm

Takens' method [Tak85] is the Maximum Likelihood Estimation of the Correlation Dimension and can be obtained as:

$$\widehat{\mathrm{CD}}_{\mathrm{Takens}} = -\frac{1}{|Q(r)|} \sum_{i \in Q(r)} i,$$

where $Q(r) = \{\delta(x,y) | x \in X, y \in X, \delta(x,y) < r\}$ is the set of pairwise distances in the set X which are lesser than a given value r, and $|Q(r)|$ denotes the cardinality of the set $Q(r)$.

In Takens' approach, the choice of r is set by heuristics [The90] which have limited theoretical foundations. Moreover, the optimality of the method is not guaranteed unless the correlation integral is of the form $C(r) = ar^{\mathrm{CD}}(1 + br^2 + o(r^2))$ with $a, b \in \mathbb{R}$ [The88].

2.2.3.8 Camastra & Vinciarelli's algorithm

Fractal-based methods require an unrealistically huge number of samples to obtain an accurate ID estimation [ER92, Smi88]. In real situations where such number of points is unavailable, fractal-based methods tend to underestimate ID with a negative bias. Camastra & Vinciarelli's Algorithm [CV02] is an attempt to correct the negative bias in Grassberger & Procaccia's algorithm. The method is —according to its own authors— not theoretically well-founded, and thus can lead to wrong ID estimates.

2.2.4 Other Global Models

2.2.4.1 Costa & Hero's k-NNG

The method is based on properties of the k Nearest Neighbor Graph [CHI04]. Assuming data to be distributed on a d-dimensional manifold, the method starts by first constructing the k Nearest Neighbor Graph, then computing the Minimum Spanning Tree [Kru56, Pri57]. ID is then estimated from the Geodetic Minimum Spanning Tree Length.

2.2.4.2 Riemannian Manifold Learning

The method [LZ08], as suggested by its name, assumes data to lie on a Riemannian manifold. Then the manifold is reconstructed using simplices. The ID is then estimated as that of the simplex with highest dimensionality. Simplicial reconstruction from data samples is an open problem, and although methods have been proposed [Fre02] their reliability in ID estimation is questionable [CV02].

2.2.4.3 IDEA

The Intrinsic Dimension Estimation Algorithm (IDEA) [RLR+11,RLC+12] is based on the following consideration. Uniformly sampling a d-dimensional point x from a uniformly dense d-dimensional hypersphere is equivalent to generating a point y from the multidimensional Gaussian $\mathcal{N}(0, \mathbb{I}_d)$ then scaling its norm accordingly:

$$x = \frac{u^{\frac{1}{d}}}{\|y\|} \cdot y,$$

where u is uniformly sampled from $[0, 1[$.

Noting that $E[1 - \|x\|] = 1/(1 + d)$, the method estimates the expectation of distances from each point j in the set X:

$$\rho_j = \sum_{i=1}^{k} \frac{x_i}{x_k},$$

with x_i being the distance from the point j to its i-th nearest neighbor. IDEA estimates dimensionality from the different expectation of distances in the data set X of size n and representational dimension D as:

$$\widehat{\mathrm{ID}}_{\mathrm{IDEA}} = \frac{\frac{1}{nD}\sum_{j \in X} \rho_j}{1 - \frac{1}{nD}\sum_{j \in X} \rho_j}.$$

It must be noted that IDEA estimate of ID does depend on the representational dimension D. This dependency is a major drawback of the method. Moreover, the number of samples required for the convergence of IDEA is not known.

2.2.4.4 DANCo

DANCo is an algorithm that infers dimensionality from not only norm concentration but also angle concentration effects [CBR+14].

2.2.4.5 Wang & Marron's algorithm

Wang & Marron's algorithm [WM+08] is an algorithm that produces an ID function based on an input scaling parameter. The method being valid only on data of small intrinsic dimensionality, the change in the scale input parameter does not drastically affect the outcome. In practice, the method has little importance since other

estimators are more robust and provide a single estimate when the true ID of the data in hands is low.

2.3 Local models of Intrinsic Dimensionality

More often than not, data does not come from a single underlying model but from multiple models. Consequently, a unique ID measurement does not fully describe the complexity and the disparity across data. It is more adequate to have ID estimates that are different from one locality to another within the same data space.

Under such view, an ID measurement provided by a global estimator does not fully describe the dimensional properties of the data. Thereupon, it is necessary to have an ID estimate for each group or cluster of points that are similar and are very likely to be on the same local manifold. This motivates the use of 'local ID estimators' that account for the disparity between local manifolds in terms of their respective intrinsic dimensionalities.

Local ID estimators use information contained in small groups of points in order to provide an ID measurement for each group. Many local ID estimators use a clustering method in order to separate the groups of points where ID is to measured. Some local ID estimators operate in the neighborhood of a given reference point. They are referred to as 'pointwise' estimators since they provide an estimation that is specific to the reference point.

2.3.1 Topological models

Topological models for local ID estimation are similar to the topological approach in global ID estimation. Indeed, they assume data to be locally embedded in a manifold. Then, ID is estimated as the dimension of the manifold.

2.3.1.1 Fukunaga-Olsen's algorithm

Fukunaga-Olsen's algorithm [FO71] can be viewed as a local approach for the original PCA algorithm [Jol86]. Assuming data to be locally embedded in a linear subspace, the ID would be equal to the number of nonzero eigenvalues of the covariance matrix. In real-world data, it is nearly impossible to obtain eigenvalues equal to zero in the covariance matrix. In practice, eigenvalues are normalized by the highest eigenvalue and the normalized eigenvalues that are lesser than a threshold θ are as-

sumed to be zero. Accordingly, ID is estimated as the number of normalized eigenvalues that are higher than the threshold θ given as a parameter of the algorithm.

Fukunaga-Olsen's algorithm can be summarized as follows:

Algorithm 2 Fukunaga-Olsen's algorithm given a dataset X and a threshold θ:

1. split the dataset X into different clusters using the K-means algorithm.
2. obtain a Voronoi tesselation [AK00, DFG99] of the space based on the clustering.
3. for each Voronoi set, compute the eigenvalues of the covariance matrix.
4. for each covariance matrix, normalize the eigenvalues by the highest eigenvalue.
5. for each Voronoi set, ID is estimated as the number of normalized eigenvalues that are higher than a predefined threshold θ.

This approach has some limitations. First, not only is the choice of the threshold θ heuristic and not universal to all data, but this choice should not be uniformly used across the different localities of the same data. Second, relying on clustering algorithms decreases the quality of the estimates [VD95]. Knowing that clustering is an unsupervised task and that many clustering algorithms are not deterministic, it is hard to obtain reliable ID estimates.

Various methods have been proposed based on Fukunaga-Olsen's approach. In these methods, data is fragmented into various subsets using alternatives to the K-means clustering.

2.3.1.2 Bruske & Sommer's algorithm

Bruske & Sommer's algorithm [BS98] is a variation of Fukunaga-Olsen's algorithm where instead of clustering a Topology Representing Network (TRN) [MS94] is used to obtain a Voronoi tesselation of the representational space.

2.3.1.3 Fan & al's algorithm

The approach of Fan & al [FQZ09] is also similar Fukunaga-Olsen's algorithm except for the methods used to fragment the data set. Instead of the K-means clustering used in Fukunaga-Olsen's algorithm or the Topology Representing Network

used in Bruske-Sommer's algorithm, Fan & al opt for minimal cover sets approximation. Because of the NP-hardness of computing set covers, this algorithm has an impractical computational complexity especially when applied to voluminous data sets.

2.3.1.4 Multiscale SVD

Multiscale Singular Value Decomposition [LJM09,LMR16] estimates ID as the number of significant singular values. Even though the method accounts for data granularity (i.e. scale), it is strictly similar to PCA-based approaches, since the singular values are the squares of the eigenvalues of the covariance matrix.

2.3.1.5 ID estimation from Expected Simplex Skewness

Johnson & al's algorithm [JSF15] estimates ID from the skewness of simplices. The vertices of each simplex consist of a subset of data points in addition to the data centroid. The 'skewness' of each simplex is then estimated as the volume if the vertices incident to the centroid were orthogonal. ID is derived from each simplex skewness based on the phenomenon of concentration of measure [Pes00].

2.3.2 Local Multidimensional Scaling

Multidimensional Scaling (MDS) approaches attempt to find a subspace with the lowest dimensionality such that distance between pairs of points in the data set are preserved. Similarly local MDS methods operate on subsets of points that are located

2.3.2.1 Isometric Feature Mapping

Isometric Feature Mapping (ISOMAP) [TDSL00] consists of finding a continuous bijection (i.e. a homeomorphism) between a given location in the data representational space and a d-dimensional hyperplane. ISOMAP assumes that an isometric homeomorphism h exists. We recall that an isometric transformation is a transformation that preserves geodesic distances between all pairs of points.

Since the local underlying manifold is not known, the ISOMAP algorithm computes the neighborhood graph, and the shortest paths in the neighborhood graph. Then ISOMAP uses an MDS algorithm [CC00], to approximate the best-fitting d-dimensional manifold that conserves the neighborhood distances.

2.3.2.2 Locally Linear Embedding

Locally Linear Embedding (LLE) [RS00] is a multidimensional scaling method that avoids the preservation of pairwise distances between pairs of points that are very distant from each other. LLE uses locally linear manifolds in order to reconstruct the global nonlinear structure.

LLE has several applications, for example in text mining and image processing. However since the global ID is an input of the LLE algorithm [CS16], it relies on other estimation methods. Hence, this algorithm has limited interest in the context of ID estimation

2.3.2.3 Laplacian Eigenmaps

Laplacian Eigenmaps [BN03] is similar to LLE in that it requires the global ID as an input of the algorithm [CS16]. In the context of ID estimation, the method cannot provide an ID estimate without an input provided by the practitioner or obtained through a different estimator.

2.3.3 Expansion Dimension

The dimensionality m in an Euclidean vector space can be determined using the ratios of two volume and two distance measurements. Consider the balls $B(x_1, r_1)$ and $B(x_2, r_2)$ centered respectively at x_1 and x_2 of radii $0 < r_1 < r_2$. Let λ be a (Lebesgue) volume measure. The ratio of volumes leads to a simple closed-form expression for the dimensionality m of Euclidean spaces:

$$\frac{\lambda(B(x_2, r_2))}{\lambda(B(x_1, r_1))} = \left(\frac{r_2}{r_1}\right)^m$$
$$\implies m = \frac{\ln \lambda(B(x_2, r_2)) - \ln \lambda(B(x_1, r_1))}{\ln r_2 - \ln r_1}.$$

An oracle that accepts two concentric balls as a query, and returns the dimension of the space, is not immediately useful. However, if volume is estimated by the number of points at distance r from a reference point x, then the returned value can serve as a measure of intrinsic dimensionality, the local Expansion Dimension (ED).

2.3.3.1 Karger & Ruhl's Expansion Dimension

In Karger & Ruhl's method [KR02] the two concentric balls have a doubling radius, and the volume is measured in terms of number of points contained in each ball. Explicitly, Karger & Ruhl's Expansion Dimension is:

$$\widehat{\mathrm{ED}} = \frac{\ln |B(x, 2r)| - \ln |B(x, r)|}{\ln 2} .$$

In this method, different choices of the radius r may lead to different estimates. This is known as the scaling problem. Moreover, a value of r for a given locality may well be inappropriate for a different locality. Furthermore, using two balls with a doubling radius has no theoretical foundations.

2.3.3.2 Generalized Expansion Dimension

Generalized Expansion Dimension (GED) [HKN12] is a heuristic allowing an adaptive choice for the radius and for the relative size of the two concentric balls. GED attempts to find the most robust choice of radii by trying all possible pairs of balls with radii corresponding to the nearest neighbor distances up to a predefined rank k.

Explicitly, the GED is calculated as follows:

Algorithm 3 GED estimate at query location $q \in X$

1. Let $K = [k-, k+]$ be the range of considered neighborhood sizes, where $0 < k- < k+$. For any choice of $k \in K$, let $A_k = \{(k, i) : i \in Kandi = k\}$.
2. Let $Q = (\delta_{q,k-}, ..., \delta_{q,k+})$ be the ordered list of distances to the query point q, for those neighbors of q with ranks in K.
3. For any $k \in K$ let $\hat{d}_k$ be the median of the individual ED estimates tests involving the sphere containing k neighbors.
4. Report the median.

2.3.3.3 He & al algorithm

He & al' s algorithm [HDJ+14] assumes that for a dense sample the density can be estimated by the ratio of the number of points by the volume. The algorithm

uses the graph distance as an approximation of the geodesic distance then uses an ED approach to estimate the ID locally.

2.3.4 Distance-based Estimation

Data points can be viewed as a sample drawn from an underlying continuous distribution often modeled in terms of manifolds. Under such modeling considerations, distances from a fixed query location to the data points can be seen as realizations of a continuous positive random distance variable. This observation leads to many distance-based ID estimators

2.3.4.1 Levina & Bickel's algorithm

Levina & Bickel's algorithm is a maximum likelihood estimation of ID [LB04] provided that neighbors of a given point are viewed as realizations of a Poisson process. More precisely, the 'observed' data is assumed to be a mapping of a d-dimensional sample to a D-dimensional representation. The sample is viewed as realizations of a d-dimensional random variable with a smooth probability density. The smoothness of the probability density guarantees that neighbors in the latent space are mapped into neighbors in the representational space.

Assuming the probability density f to be constant within a ball $B(q, w)$ with radius w and centered at q, points in the neighborhood of a fixed point q are the realizations of a Poisson process $P(r)$ with $r < w$. $P(r)$ indicates the number of points within distance r of q. The rate of this process is

$$\lambda(r) = f \cdot S(q, r) = f \cdot \frac{\pi^{d/2} d r^{d-1}}{\Gamma(1 + d/2)}$$

where $S(q, t)$ is the surface area of the sphere of radius t centered at q. The corresponding log-likelihood function is

$$\mathcal{L}(d, \log f) = \int_0^w \log \lambda(r) dP(r) - \int_0^w \log \lambda(r) dr.$$

Maximizing the log-likelihood estimation leads to the following local estimator. For a point $q \in X$:

$$\hat{d}_k(q) = -\left[\frac{1}{k} \sum_{i=1}^{k} \ln \left(\frac{\delta_i(q)}{\delta_k(q)}\right)\right]^{-1},$$

where δ_i indicates the distance from the point q to its i-th nearest neighbor.

In order to have a global estimation from the individual local estimates, Levina and Bickel take the average of the point-wise estimates over the entire data set:

$$\hat{d}_k = \frac{1}{n} \sum_{q \in X} \hat{d}_k(q).$$

This approach was later criticized by Mackay and Ghamarani who explain in an unpublished note [MG05] that taking the average is not theoretically well founded. Assuming the different points of the data set to be independent, maximizing the likelihood function of the dimensionality for all points simultaneously yields that the global ID estimate should be the harmonic mean of the point-wise estimates. In addition to the theoretical foundations, using the harmonic mean instead of the average leads to a smaller bias for smaller values of the neighborhood size k.

In order to adjust for scaling, Levina and Bickel measure the global ID for various choices of neighborhood size k then report the average:

$$\widehat{\mathrm{ID}}_{\mathrm{L.\&B.}} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \hat{d}_k.$$

2.3.4.2 MiND

Minimum Neighbor Distance (MiND) is a framework for ID estimation [RLC+12]. The framework is based on the observation that a point x sampled from a uniformly dense d-dimensional hypersphere can be obtained with the same probability distribution by generating a point y from the multidimensional Gaussian $\mathcal{N}(0, \mathbb{I}_d)$ then scaling its norm accordingly:

$$x = \frac{u^{\frac{1}{d}}}{\|y\|} \cdot y,$$

where u is uniformly sampled from $[0, 1[$.

The volume of the d-dimensional hypersphere is:

$$V_r = \frac{\pi^{d/2}}{\Gamma(1 + d/2)} \cdot r^d,$$

with Γ denoting the Gamma function. Assuming this volume to be proportional to the probability of having a neighbor at distance r, the log-likelihood as a function

of the dimensionality d over a sample X of n points is:

$$\mathcal{L}(d) = n\log(k) + n\log(d) + (d-1)\sum_{x_i \in X} \log(\rho(x_i)) + (k-1)\sum_{x_i \in X} \log(1-\rho^d(x_i)),$$

where $\rho(x_i)$ indicates the ratio of the distance to the nearest neighbor of x_i to the distance to the k-th nearest neighbor of x_i.

Then maximizing the likelihood yields different estimators. The authors propose an integer-valued estimation, where ID is the integer in $[1, D]$ that maximizes the likelihood function. They also propose to simplify the problem by taking a value of $k = 2$ which leads to the same form (c.f. Equation 2.3.4.1) seen in Levina & Bickel estimator [LB04]. For a general choice of k it is possible to use numerical optimization [CL96]. Alternatively, Kullback-Leibler divergence estimation [WKV06] can lead to an estimate of ID.

Even though the original algorithm was proposed as a global estimator, the global estimate can be viewed as an aggregation of point-wise ID estimates. Hence, this approach falls in the local estimators' category.

2.3.4.3 Manifold-Adaptive Dimension

The Manifold-Adaptive Dimension [FSA07] is a distance-based estimator based on the following assumptions. Given a query point q, the probability of encountering a neighbor x_i at distance δ_i from q lesser than r is expressed as:

$$\Pr[\delta_i \leq r] = \eta(q, r)r^d,$$

or equivalently

$$\ln \Pr[\delta_i \leq r] = \ln \eta(q, r) + d \ln r,$$

where d indicates the ID and $\eta(q, r)$ indicates the density inside the ball $B(q, r)$.

Noticing that the last equation is linear in d, and assuming that $\eta(q, r)$ as a function of the distance r is constant for small values of r, the ID can be estimated as:

$$\hat{d}_q = \frac{\ln 2}{\ln \delta_k - \ln \delta_{\lceil k/2 \rceil}}.$$

This assumes that δ_k is small enough to not reach the curvature of the manifold.

A global estimate is obtained by taking the average of all local estimates across

the entire dataset:

$$\widehat{\mathrm{ID}}_{\mathrm{MAD}} = \frac{1}{n} \sum_{q \in X} \max(\hat{d}_q, D)$$

with X being the dataset, n its cardinality, and D the representational dimension.

2.3.5 Other methods

2.3.5.1 Mordohai & Medioni's algorithm

Mordohai & Medioni' s algorithm [MM10] estimates dimensionality based on tensor voting. The local ID for each point is provided by the largest gap between the eigenvalues of the tensor. A global estimate can be obtained by averaging the local estimates. The computational complexity of the tensor voting is prohibitive in contexts where the representational dimension of the data is high. Among other computational limitations, the memory complexity is exponential in the representational dimension D.

2.3.5.2 Brand's Charting

Brand's charting algorithm [Bra02] maps the original representation space of dimension D to a Euclidean vector space of dimension d such that curves that are locally parallel in the original space do not intersect in the new space. The estimates ID as the smallest value of d for which such mapping is possible. Brand's method is not robust to noise, is not applicable to nonlinear manifolds.

Chapter 3

Extreme Value Theory

Extreme Value theory (EVT) is a discipline in statistics concerned with the modeling of what can be regarded as the extreme behavior of stochastic processes. This discipline emerged in the late 19th century with the work of Vilfredo Pareto (1848-1923). Then it was developed over the course of the 20th century and was applied to various fields.

EVT has seen applications in areas such as civil engineering [Har01], operations research [DM01] [TC00, MSR+00, DM01], risk assessment [LC00], material sciences [Gri93], bioinformatics [Rob00], geophysics [LC00], and multimedia [FJ13].

In this work, distances to neighbors are considered to be the sample drawn from a random distance variable. Hence distances to the nearest neighbors can be viewed as 'extreme events'. Under these modeling considerations EVT results can be applied to the lower tail of the distance distribution.

In this Chapter, we summarize the main results of EVT, then we survey the main state-of-the-art estimators for the Extreme Value Index (EVI).

3.1 Main results in Extreme Value Theory

Extreme Value Theory is concerned with the study of occurrences of a random variable that can be described as 'extremely' distant from the variable's mean. These 'extreme events' are associated with the upper or lower tails of probability density functions, as opposed to 'usual events' which are associated with the modes.

Various approaches have been used to model these extreme events, and to study the asymptotic behavior of the tail of probability distributions. Three equivalent approaches have been developed which are: the Block Maxima Method, the Peak Over Threshold method, and the Regularly-Varying functions representation.

3.1.1 Distribution of maxima

The first approach to modeling extreme values is the Block Maxima Method historically known as 'Annual Maxima Series' (AMS). Given a data sample from an set of random variables assumed to be independent and identically distributed, the approach relies on segmenting the sample into blocks and viewing the maximum (or minimum) of each block as an extreme event. The best known theorem obtained under these modeling constraints, attributed in parts to Fisher and Tippett [FT28], and Gnedenko [Gne43], states that the maximum of n independent and identically-distributed random variables (after proper renormalization) converges in distribution to a Generalized Extreme Value (GEV) distribution as n goes to infinity.

Definition 4 Let $\mu, \xi \in \mathbb{R}$ and $\sigma > 0$. The family of generalized extreme value distributions $\mathcal{F}_{\text{GEV}}$ covers distributions whose cumulative distribution function can be expressed in the form

$$\mathcal{F}_{\text{GEV}} = \left\{ \exp\left(- \left[1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right) \right\}.$$

A distribution $G \in \mathcal{F}_{\text{GEV}}$ has support $\text{supp}(G) = [\mu - \frac{\sigma}{\xi}, \infty)$ whenever $\xi > 0$ and $\text{supp}(G) = (-\infty, \mu - \frac{\sigma}{\xi}]$ when $\xi < 0$. If $\xi = 0$, the support covers the complete real line. The parameters μ, σ and $\gamma = -1/\xi$ are respectively called the location, the spread, and the index (or shape).

Theorem 1 (Fisher-Tippet-Gnedenko) Let $(\mathrm{X}_i)_{i\in\mathbb{N}}$ be a sequence of independent identically-distributed random variables and let $\mathrm{M}_n = \max_{1\leq i\leq n} \mathrm{X}_i$. If there exist a sequence of positive constants $(a_i)_{i\in\mathbb{N}}$, and a sequence of constants $(b_i)_{i\in\mathbb{N}}$, such that

$$\lim_{n\to\infty} \Pr\left[\frac{\mathrm{M}_n - b_n}{a_n} \leq x\right] = F(x),$$

for any $x \in [0,1]$, where F is a non-degenerate distribution function, then $F \in \mathcal{F}_{\text{GEV}}$.

3.1.2 Conditional excess distribution

Extreme value theory mainly draws its power from two major results due to Fisher, Tippett and Gnedenko, as well as Balkema, de Haan and Pickands [FT28, Gne43, BDH74, PI75]. The second approach is known as the Conditional Excess Method, or 'Peak Over Threshold' (POT) method. Balkema-de Haan-Pickands theorem is useful when observing the occurrences of a continuous random variable that exceed a given high threshold (or fall below a low threshold). The theorem states that excesses over a fixed threshold converge in distribution to the Generalized Pareto Distribution (GPD).

Consider the following two definitions.

Definition 5 Let $\xi \in \mathbb{R}$ and $\sigma > 0$. The family of generalized Pareto distributions $\mathcal{F}_{\text{GPD}}$ is defined by its cumulative distribution function

$$\mathcal{F}_{\text{GPD}} = \left\{1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-\frac{1}{\xi}}\right\}.$$

Every distribution $G \in \mathcal{F}_{\text{GPD}}$ has support $\text{supp}(G) = (\max\{0, -\frac{\sigma}{\xi}\}, \infty)$. The parameters μ, σ and $\gamma = -1/\xi$ are respectively called the location, the spread, and the index (or shape).

Definition 6 Let X be a random variable whose distribution F_{X} has the upper endpoint $x^+ \in \mathbb{R} \cup \{\infty\}$. Given $w < x^+$, the conditional excess distribution $F_{\mathrm{X,w}}$ of X is the distribution of $\mathrm{X} - w$ conditioned on the event $\mathrm{X} > w$:

$$F_{\mathrm{X,w}}(x) = \frac{F_{\mathrm{X}}(w+x) - F_{\mathrm{X}}(w)}{1 - F_{\mathrm{X}}(w)}.$$

We are now in a position to introduce a powerful theorem due to Pickands, Balkema and de Haan [BDH74, PI75], which can be regarded as the counterpart to the central limit theorem for extremal statistics.

Theorem 2 (Pickands-Balkema-de Haan) Let $(\mathrm{X}_i)_{i\in\mathbb{N}}$ be a sequence of independent random variables with identical distribution function F_{X} satisfying the conditions of the Fisher-Tippett-Gnedenko Theorem. As $w \to x^+$, the conditional excess distribution $F_{\mathrm{X,w}}(x)$ converges to a distribution in $\mathcal{F}_{\text{GPD}}$.

3.1.3 Regularly-Varying functions representation

The Fisher-Tippett-Gnedenko Theorem and the Pickands-Balkema-de Haan Theorem have been shown to be equivalent to a third characterization of the tail behavior in terms of regularly-varying (RV) functions, known as Karamata Representation. The asymptotic cumulative distribution of X in the tail $[0, w)$ can be expressed as $F_{\mathrm{X}}(x) = x^{\gamma}\ell_{\mathrm{X}}(1/x)$, where ℓ_{X} is differentiable and slowly varying; that is, for all $c > 0$, ℓ_{X} satisfies

$$\lim_{t\to\infty} \frac{\ell_{\mathrm{X}}(ct)}{\ell_{\mathrm{X}}(t)} = 1.$$

F_{X} restricted to $[0, w)$ is itself said to be regularly varying with index γ. In particular, a cumulative distribution $F \in \mathcal{F}_{\text{GEV}}$ has $\xi < 0$ if and only if F is RV and has a finite endpoint. Note that the slowly-varying component $\ell_{\mathrm{X}}(1/x)$ of F_{X} is not necessarily constant as x tends to zero. For a detailed account of RV functions, we refer the reader to [BGT89].

3.2 Extreme Value Index estimators

We are only interested in estimating the index of Extreme Value Distributions. Parametric, semi-parametric, and non-parametric approaches were proposed to estimate the Extreme Value Index. In this section, we will survey the main semi-parametric approaches. In the rest of this section, we assume the sample $x_1, x_2, ..., x_n$ to be drawn from independent and identically distributed random variables, and ordered in increasing order.

3.2.1 Hill Estimator

There does not exist a Maximum Likelihood Estimator that estimates the index $\gamma \in \mathbb{R}$ without any restriction on the range of γ [DHF07]. Hill's estimator [H$^+$75, Wei78] is the Maximum Likelihood Estimator for the Extreme Value Index whenever γ is restricted to $\gamma > 0$. The properties of Hill's estimator such as its almost sure convergence [DHM88] and asymptotic normality [HT85, dHR98] have been thoroughly studied [Hal82, Mas82, DR84, CM85, GS87, dHP98] which reflects its importance in the EVT literature.

Hill's estimator can be expressed as follows:

$$\hat{\gamma}_H = \frac{1}{k} \sum_{i=1}^{k} \ln x_{n-i} - \ln x_{n-k}.$$

3.2.2 Generalized Hill Estimator

Studies made by Beirlant & al [BVT$^+$96a, BDG$^+$05] used the Hill estimator in order to propose an estimator that is valid for $\gamma \in \mathbb{R}$ (unlike the original Hill estimator which is valid only for $\gamma > 0$). However, the proposed estimator was not obtained using the standard MLE approach. The generalized Hill estimator can be expressed as follows:

$$\hat{\gamma}_{\text{GH}} = \hat{\gamma}_{H,k} + \sum_{i=k}^{n} [\ln \hat{\gamma}_{H,i} - \ln \hat{\gamma}_{H,k}],$$

where $\hat{\gamma}_{H,i}$ is the Hill estimator using the first i observations.

3.2.3 Pickands' Estimator

Pickand's estimator [PI75] applies the percentile method to the top observations in the sample. The method consists of approximating the quantiles of probability density by the quantiles observed in the sample, then solving for the EVI in the equations obtained from the approximation. Pickand's estimator has the form:

$$\hat{\gamma}_P = \frac{1}{\ln 2} \ln \frac{x_{n-\lfloor k/4 \rfloor+1} - x_{n-\lfloor k/2 \rfloor+1}}{x_{n-\lfloor k/2 \rfloor+1} - x_{n-k+1}}.$$

3.2.4 The Moments' Estimator

The Method of Moments leads to the following estimator [DEdH89]:

$$\hat{\gamma}_M = \hat{m}_1 + \frac{1}{2}\left[1 - \frac{\hat{m}_1^2 - 1}{\hat{m}_2}\right],$$

where the j-th moment of the log-excess is estimated as:

$$\hat{m}_j = \frac{1}{k}\sum_{i=1}^{k}\left[\ln x_{n-i+1} - \ln x_{n-k}\right]^j.$$

Note that the Hill estimators corresponds to the first moment of the log-excess ($\hat{\gamma}_H = \hat{m}_1$).

3.2.5 The Peak Over Threshold Maximum Likelihood Estimator

The following estimator based on equations similar to the Maximum Likelihood equations [Dav84] is widely regarded as the MLE for the index [Smi87]:

$$\hat{\gamma}_{\text{ML}} = \frac{1}{k}\sum_{i=1}^{k}\ln[1 + \hat{\sigma}(x_{n-i+1} - x_{n-k})],$$

where $\hat{\sigma}$ is the Maximum Likelihood Estimation of the scale parameter.

3.2.6 Kernel Estimators

For a positive Extreme Value Index, a family of estimators known as 'kernel estimators' is given by [CDM85, GLDW$^+$03]:

$$\hat{\gamma}_{\mathcal{K}} = \frac{\sum_{i=1}^{k}\mathcal{K}(i/k)[\ln x_{n-i+1} - \ln x_{n-k}]}{\sum_{i=1}^{k}\mathcal{K}(i/k)},$$

where $\mathcal{K}$ is a non-negative non-increasing kernel defined on $\mathbb{R}+$ and summing to 1.

Note that Hill's estimator is the special case where the kernel $\mathcal{K}$ is identically equal to 1. This family of estimators covers a large set of known estimators of EVI known as QQ-estimators [BVT96b, KR96, CV98, OGFA06].

3.3 Second order Extreme Value Index

In statistics in general and in EVT in particular, statisticians have an interest in assessing the speed of convergence of estimators. In the case of the first order Extreme Value Index (EVI) which is the growth rate of the probability density function, the convergence of estimators is ruled by a second order growth rate, which is the growth rate of the EVI itself. This second order growth rate is a statistic commonly known in EVT as the 'second order EVI' and often denoted ρ. Higher values of $|\rho|$ indicate a higher convergence rate of EVI estimators, and lower values of $|\rho|$ indicate that estimators of the EVI require larger sample sizes.

If F is the cumulative distribution function of a heavy-tailed Extreme Value Distribution ($\gamma > 0$), and if U denotes the quantile function ($U(t) = F^{\leftarrow}(1 - \frac{1}{t})$, where $F^{\leftarrow}(s) \triangleq \inf\{y : F(y) \geq s\}$), then, and the parameter ρ satisfies

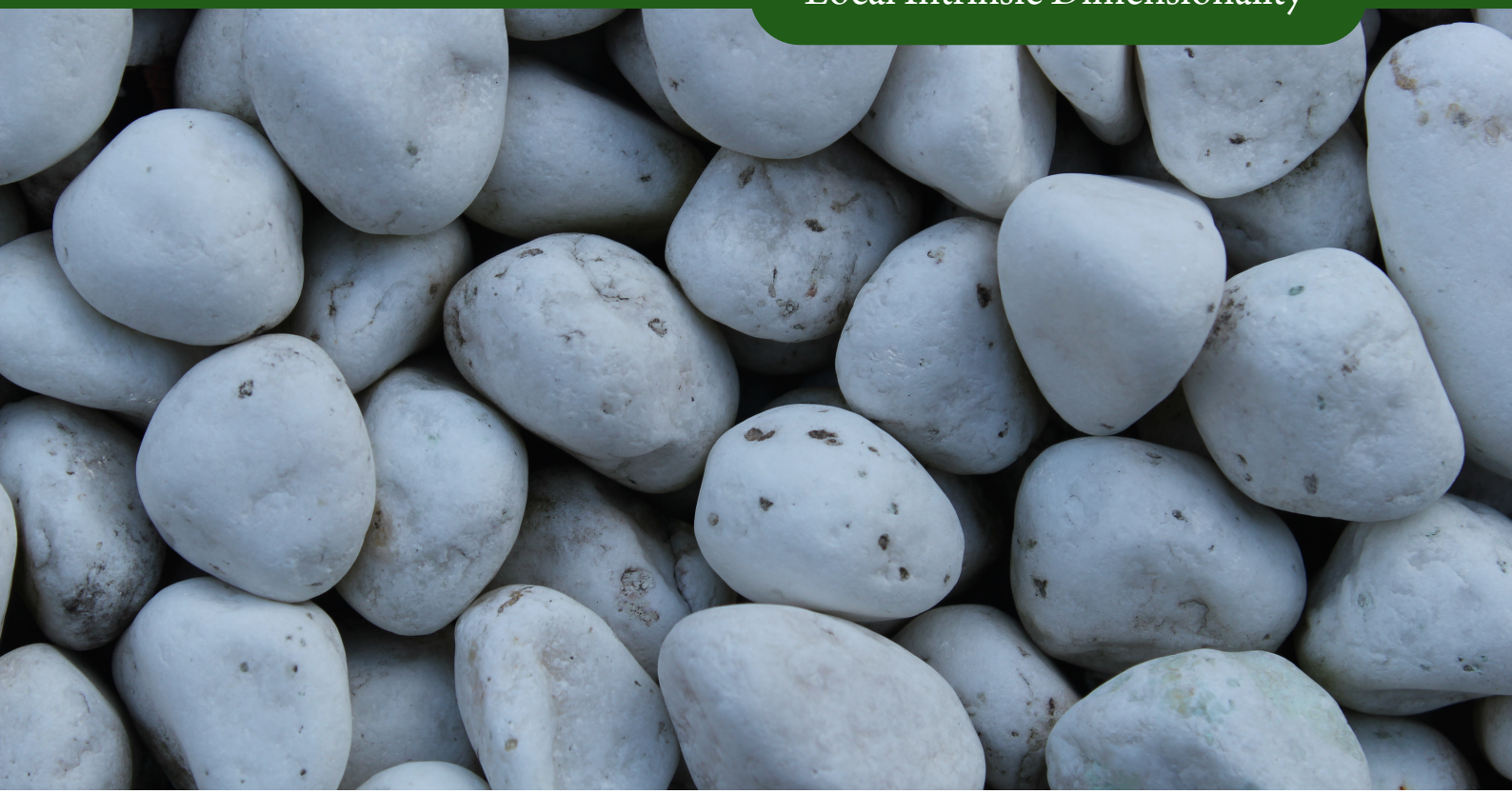
$$\lim_{t\to\infty} \frac{\ln U(tx) - \ln U(t) - \gamma \ln(x)}{A(t)} = \frac{x^{\rho} - 1}{\rho},$$

where $A(t) = \gamma\theta t^{\rho}$ is regularly varying with index $\rho < 0$, with θ being a constant [HW^{+}85, GMN07]. This condition which ensures the convergence of the first order EVI is referred to as the 'second order condition'. The convergence of the estimators of the EVI γ for heavy-tailed distributions ($\gamma > 0$) is ruled by the index ρ referred to as the 'second order EVI'.

The EVT community has developed estimators for the second order EVI over the past few decades. Many estimators have been proposed [HW^{+}85, DK98, Pen98, GDHP02]. The best estimators so far rely on the Method of Moments [AGdH03], and assume the second order EVI to be negative.

Chapter 4

Local Intrinsic Dimensionality



Expansion Models of Intrinsic Dimensionality can be described as heuristics, since they are based on observations and are not supported by an underlying theory. Local Intrinsic Dimensionality (LID) is a theoretical framework that models the Expansion Dimension in terms of random distance variables.

This model of dimensionality is shown to be equivalent to a measure of indiscriminability makes it very interesting from a Machine Learning point of view. The model being general and its assumptions being limited to the continuity of the distance variable, LID is suited for applications in similarity search and unsupervised learning among other applications in Machine Learning and Data Mining.

In this Chapter, we summarize the theory and modeling of LID, we show how LID is equal to a measure of indiscriminability, then we show the connection between LID and EVT.

4.1 Local Intrinsic Dimension as an expansion model

Local Intrinsic Dimensionality (LID) is an extension of a well-studied model of intrinsic dimensionality to continuous distributions of distances proposed in [Hou13]. LID aims to quantify the local ID of a feature space exclusively in terms of the distribution of inter-point distances. Formally, let $(\mathbb{R}^m, \mathrm{d})$ be a domain equipped with a non-negative distance function d. Let us consider the distribution of distances within the domain with respect to some fixed point of reference. We model this distribution in terms of a random variable X with support $[0, \infty)$. X is said to have probability density f_{X}, where f_{X} is a non-negative Lebesgue-integrable function, if and only if

$$\mathsf{Pr}[a \leq \mathrm{X} \leq b] = \int_a^b f_{\mathrm{X}}(x)\,\mathrm{d}x,$$

for any $a, b \in [0, \infty)$ such that $a \leq b$. The corresponding cumulative density function F_{X} is canonically defined as

$$F_{\mathrm{X}}(x) = \mathsf{Pr}[\mathrm{X} \leq x] = \int_0^x f_{\mathrm{X}}(u)\,\mathrm{d}u.$$

Accordingly, whenever X is absolutely continuous at x, F_{X} is differentiable at x and its first-order derivative is $f_{\mathrm{X}}(x)$. For such settings, the local intrinsic dimension is defined as follows.

> Definition 7 ([Hou13]) Given an absolutely continuous random distance variable X, for any distance threshold x such that $F_{\mathrm{X}}(x) > 0$, the local continuous intrinsic dimension of X at distance x is given by
>
> $$\mathrm{ID}_{\mathrm{X}}(x) \triangleq \lim_{\epsilon \to 0^+} \frac{\ln F_{\mathrm{X}}\left((1+\epsilon)x\right) - \ln F_{\mathrm{X}}(x)}{\ln(1+\epsilon)},$$
>
> wherever the limit exists.

With respect to the generalized expansion dimension [HKN12], a precursor of LID, the above definition of $\mathrm{ID}_{\mathrm{X}}(x)$ is the outcome of a dimensional test of neighborhoods of radii x and $(1+\epsilon)x$ in which the neighborhood cardinalities are replaced by the expected number of neighbors. LID also turns out to be equivalent to a formulation of the (lack of) discriminative power of a distance measure, as both formulations have the same closed form:

Theorem 3 ([Hou13]) Let X be an absolutely continuous random distance variable. If F_X is both positive and differentiable at x, then

$$\mathrm{ID}_X(x) = \frac{x f_X(x)}{F_X(x)}.$$

Local ID has potential for wide application thanks to its very general treatment of distances as continuous random variable. Direct estimation of ID_X, however, requires the knowledge of the distribution of X. Extreme value theory, which we survey in the following section, allows the estimation of the limit of $\mathrm{ID}_X(x)$ as x tends to 0 without any explicit assumptions of the data distribution other than continuity.

4.2 Indiscriminability

A distance measure is 'discriminative' when an expansion in distance results in a relatively small increase in the number of observations. Discriminability of points improves the efficiency of data mining and machine learning tasks, minimizes error and ensures robustness. Capturing a huge number of points by a small increase in distance measure greatly increases the computational cost of applications, and has a negative impact on effectiveness. Moreover, the relative ranking of points with reference to an indiscriminative distance measure can easily be affected by noise. When selecting features, it is strongly desirable that they be chosen so as to ensure better discriminability.

Formally, let x be a reference point and let R be an absolutely continuous random distance variable with respect to that reference point, as defined in Section 4.1. For any distance r from x such that $F_R(r) > 0$, the indiscriminability of R at r is given by the following limit wherever it exists [Hou13]:

$$\begin{aligned}\mathrm{InDiscr}_{F_R}(r) &\triangleq \lim_{\epsilon \to 0^+} \left(\frac{F_R((1+\epsilon)r) - F_R(r)}{\epsilon \cdot F_R(r)} \right) \\ &= \frac{r \cdot \phi_R(r)}{F_R(r)} = \mathrm{ID}_{F_R}(r).\end{aligned}$$

Note that this definition of indiscriminability is unitless, does not require knowledge of the statistical parameters of the underlying distance distribution, and coincides precisely with local ID.

4.3 Connection between LID and EVT

In the following we demonstrate a direct relation between LID and extreme value theory, which arises as an implication of Theorem 2. Note that any choice of distance threshold w corresponds to a neighborhood of radius w based at the reference point, or equivalently, to the tail of the distribution of distances on $[0, w)$. As discussed in [CBTD01], Theorem 2 also applies to lower tails: one can reason about minima using the transformation $\mathrm{Y} = -\mathrm{X}$. The distribution of the excess $\mathrm{Y} - (-w)$ (conditioned on $\mathrm{Y} > -w$) then tends to a distribution in $\mathcal{F}_{\text{GPD}}$, as w tends to the lower endpoint of F_{X} located at zero [Net14]. Accordingly, as w tends to zero, the distribution in the tail $[0, w)$ can be restated as follows [CBTD01].

> Lemma 1 Let X be an absolutely continuous random distance variable with support $[0, \infty)$ and cumulative distribution function F_{X} such that $F_{\mathrm{X}}(x) > 0$ if $x > 0$. Let $c \in (0, 1)$ be an arbitrary constant. Let $w > 0$ be a distance threshold, and consider x restricted to the range $[cw, w)$. As w tends to zero, the distribution of X restricted to the tail $[cw, w)$ satisfies, for some fixed $\xi < 0$,
>
> $$\frac{(x/w)^{-\frac{1}{\xi}}}{F_{\mathrm{X,w}}(x)} \to 1.$$

Note that the distribution of excess distance $w - \mathrm{X}$ is bounded from above by w which, according to [CBTD01], enforces that $\xi < 0$.

Proof Consider the distribution of threshold excess $w - \mathrm{X}$ with X being restricted to $[cw, w)$. According to Theorem 2, $w - \mathrm{X}$ asymptotically follows a generalized Pareto distribution:

$$\mathsf{Pr}[w - \mathrm{X} \leq y] \to 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}},$$

with $\sigma > 0$ and $\xi < 0$, so that

$$\mathsf{Pr}[\mathrm{X} \leq w - y] \to \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}}.$$

Since a distance x corresponds to a threshold excess of $w - y$,

$$F_{\mathrm{X,w}}(x) = \mathsf{Pr}[\mathrm{X} \leq x] \to \left(1 + \frac{\xi(w-x)}{\sigma}\right)^{-\frac{1}{\xi}}.$$

We see that $F_{\mathrm{X,w}}(0) = 0$ holds if and only if

$$\left(1 + \frac{\xi(w-x)}{\sigma}\right)^{-\frac{1}{\xi}} = 0,$$

implying that $\sigma = -\xi w$. With this additional constraint, the distribution of distances in the tail $[cw, w)$ simplifies to

$$\frac{(x/w)^{-\frac{1}{\xi}}}{F_{\mathrm{X,w}}(x)} \to 1.$$

■

To summarize, whenever Theorem 2 applies to a distance variable X, the cumulative distribution of distances within a radius-w neighborhood is asymptotically determined by a single parameter $\xi < 0$. We can prove the following statement concerning LID.

Theorem 4 Let X be an absolutely continuous random distance variable with support $[0, \infty)$, satisfying the conditions of Theorem 2, and $w > 0$ be a distance threshold. Then, as w tends to zero,

$$\mathrm{ID}_{\mathrm{X}}(w) \to -\frac{1}{\xi} =: \mathrm{ID}_{\mathrm{X}}.$$

Proof (Sketch only. For a more detailed and rigorous treatment, see [Hou15].) Lemma 1 states that under the conditions of Theorem 4, the cumulative excess distribution $F_{\mathrm{X,w}}$ follows

$$\frac{(x/w)^{-\frac{1}{\xi}}}{F_{\mathrm{X,w}}(x)} \to 1$$

as the threshold w approaches zero. The probability density $f_{\mathrm{X,w}}$ in the tail of the distribution is obtained by taking the derivative with respect to x:

$$f_{\mathrm{X,w}}(x) \approx \frac{\partial}{\partial x}\left(\frac{x}{w}\right)^{-\frac{1}{\xi}} = -\frac{1}{\xi w}\left(\frac{x}{w}\right)^{\frac{1}{\xi}-1}.$$

Applying Theorem 3 gives

$$\mathrm{ID}_{\mathrm{X}}(x) \approx \frac{x \cdot f_{\mathrm{X,w}}(x)}{F_{\mathrm{X,w}}(x)} \to -\frac{1}{\xi}.$$

■

Note that together Lemma 1 and Theorem 4 allow us to restate the asymptotic cumulative distribution of distances in the tail $[cw, w)$ as

$$\frac{(x/w)^{\mathrm{ID}_{\mathrm{X}}}}{F_{\mathrm{X,w}}(x)} \to 1.$$

4.4 Second order LID

In this section, we are interested in understanding the relation between the change in LID in a small neighbor of a given reference point and the properties of the distance distribution. A complete description can be found in this work [Hou15].

We generalize the ID_F notation for any continuous function F defined and differentiable on $[0, +\infty)$. If f denotes the derivative of F, we define:

$$\mathrm{ID}_F(x) = \frac{x \cdot f(x)}{F(x)}.$$

In the case where F is the cumulative distribution function of a continuous distance variable, we refer to ID_F as the LID function. The limit $\lim_{x \to 0} \mathrm{ID}_F(x)$ is the LID.

When expanding the local neighborhood from a reference point outwards (i.e. the distance x increases), an increase in discriminability (i.e. $\mathrm{ID}_F(x) < \mathrm{ID}_F(0)$ with $0 < x < w$) indicates that the the discriminability of neighbors at distance x is higher than the discriminability at distance 0. In other words, this increase indicates that the growth rate in probability measure is lower than the growth rate that would be encountered in a hypothetical locally-uniform distribution of points within a manifold of dimension $\mathrm{ID}_F(0)$. The interpretation of this increase in discriminability is that the reference point from which distances are being measured is an inlier with reference to its nearest neighbors.

The limit effect as the distance x approaches 0 of the condition $\mathrm{ID}_F(x) < \mathrm{ID}_F(0)$ is equivalent to the condition $\mathrm{ID}'_F(x) < 0$. After normalization by the LID so as to make the comparison possible across manifolds of different dimensions, the

condition can be equivalently expressed as $\mathrm{ID}_{\mathrm{ID}_{F_X}}(x) = \frac{x \cdot \mathrm{ID}'_F(x)}{\mathrm{ID}_F(x)} < 0$. Hence the function $\mathrm{ID}_{\mathrm{ID}_{F_X}}$, referred to as the second order LID function, indicates the strength of inlierness of the reference point when negative and the strength of outlierness of the reference point when negative [Hou15]. Since $\lim_{x \to 0} \mathrm{ID}_{\mathrm{ID}_{F_X}}(x) = 0$, the inlierness (or outlierness) at the reference point (i.e. when the distance x approaches 0) is indicated by the limit $\lim_{x \to 0} \mathrm{ID}_{|\mathrm{ID}_{\mathrm{ID}_{F_X}}|}(x)$. We refer to the last quantity by the second order LID.

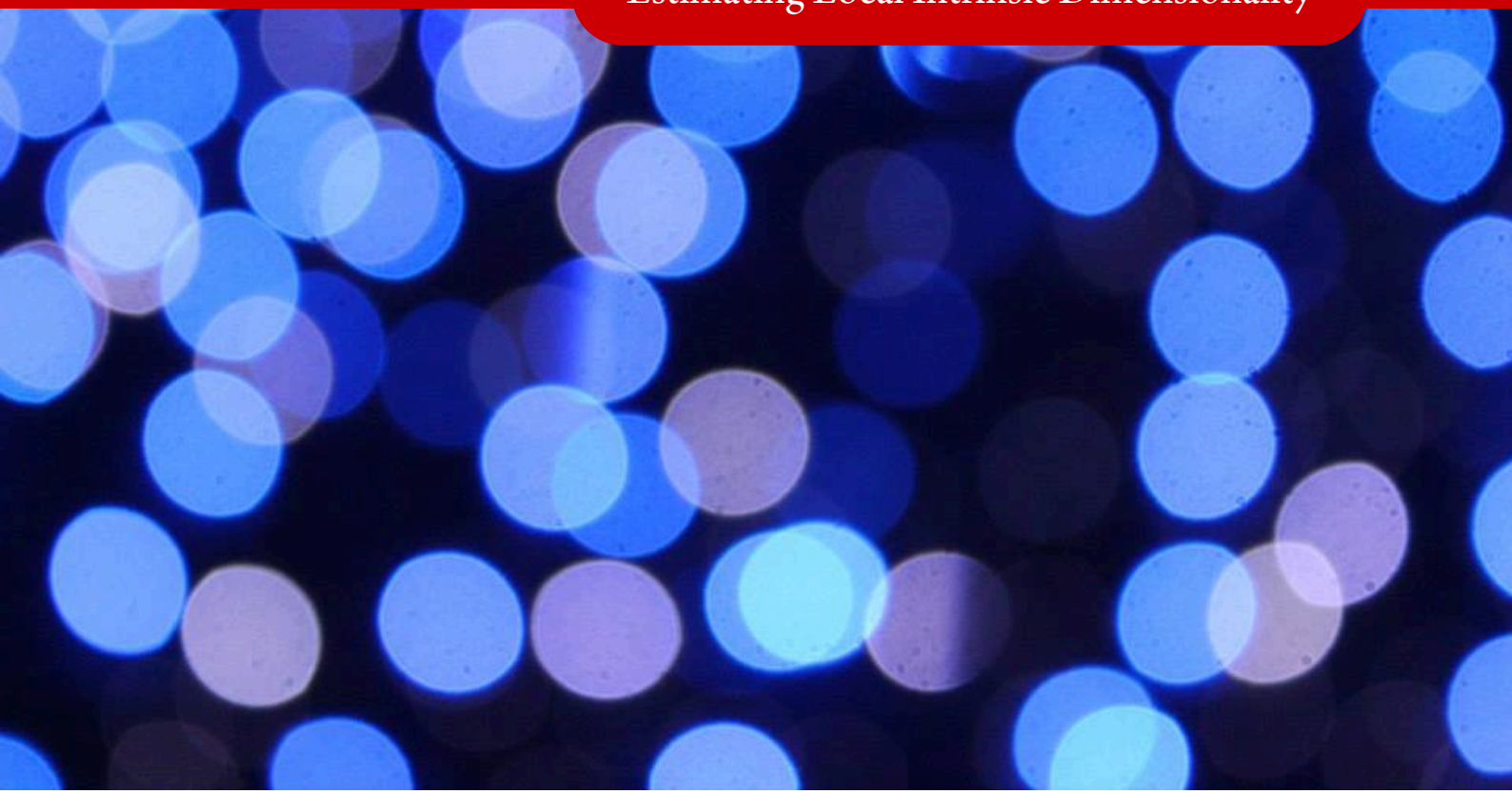
A measurement of the second order LID can indicate inlierness, which can be of interest to many Data Mining and Machine Learning applications. An estimate of inlierness can have applications in clustering, classification and outlier detection. Note that estimating second order LID is outside the scope of this thesis.

Part II

Estimation

Chapter 5

Estimating Local Intrinsic Dimensionality



This chapter is concerned with the estimation of a local measure of intrinsic dimensionality (ID) recently proposed by Houle. The local model can be regarded as an extension of Karger and Ruhl's expansion dimension to a statistical setting in which the distribution of distances to a query point is modeled in terms of a continuous random variable. This form of intrinsic dimensionality can be particularly useful in search, classification, outlier detection, and other contexts in machine learning, databases, and data mining, as it has been shown to be equivalent to a measure of the discriminative power of similarity functions. Several estimators of local ID are proposed and analyzed based on extreme value theory, using maximum likelihood estimation, the method of moments, probability weighted moments, and regularly varying functions. An experimental evaluation is also provided, using both real and artificial data.

The original contributions of this work can be summarized as:

- A framework for the estimation of local continuous intrinsic dimension (LID) using well-established techniques: the maximum likelihood estimation (MLE), the method of moments (MoM), and the method of probability-weighted mo-

ments (PWM). In particular, we verify that applying MLE to LID leads to the well-known Hill estimator [H+75].

- A new family of estimators based on the extreme-value-theoretic notion of regularly varying functions. Several existing dimensionality models (ED, GED, and MiND) are shown to be special cases of this family,

- confidence intervals for the variance and convergence of the estimators we propose.

- An experimental study using artificial data and synthetic distance distributions, in which we compare our estimators with state-of-the-art global and local estimators. We also show that the empirical variance and convergence rates of the MLE (Hill) and MoM estimators are superior to those of the other local estimators studied.

- Experiments showing that local estimators are more robust than global ones in the presence of noise in nonlinear manifolds. Our experiments show that our approaches are very competitive in this regard with other methods, both local and global.

- An experimental study showing the effectiveness of LID estimation when using approximate nearest neighbors.

- Profiles of several real-world datasets in terms of LID, illustrating the degree of variability of complexity from region to region within a dataset. The profiles demonstrate that a single 'global' ID value is in general not sufficient to fully characterize the complexity of real-world data.

The remainder of the Chapter is structured as follows. In Section 5.1 we propose and analyze several estimators of continuous ID, using maximum likelihood estimation (MLE, which yields the Hill estimator), the method of moments (MoM), probability weighted moments (PWM), and regularly varying functions (RV). In Section 5.2 we present our experimental study, and discuss the practical performance of our proposed estimators. We conclude the Chapter in Section 5.4 with a discussion of potential future applications.

5.1 Estimation

This section is concerned with practical methods for the estimation of the local intrinsic dimension of a random distance variable X. In particular, we adapt known methods for GPD parameter estimation such as the Maximum Likelihood Estimation (in Section 5.1.1) and Moment-based estimation (in Sections 5.1.2 and 5.1.3), and propose a new family of estimators based on Regularly Varying functions (in Section 5.1.4).

For the remainder of this discussion we assume that we are given a distance threshold $w > 0$ and a sequence $x_1, \ldots, x_n$ of observations of a random distance variable X with support $[0, w)$. Without loss of generality, we assume that the observations are given in ascending order — that is, $x_1 \leq x_2 \leq \cdots \leq x_n$.

5.1.1 Maximum Likelihood Estimation

Maximization of the likelihood function is one of the most widely used parameter estimation techniques in statistics. The Maximum Likelihood Estimator (MLE) has no optimality guarantees for finite samples, but has the advantage of being asymptotically consistent, optimal, and efficient (in that it achieves the Cramer-Rao bound).

Definition 8 Given a random variable X with parameter θ, the likelihood of θ as a function of observations x_1, x_2, ..., x_n is defined as

$$L(\theta|\, x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\, \theta)\,.$$

Note that θ can be multivariate. In the case of our study, we are interested in a single parameter of the distribution, namely the shape parameter ξ.

Maximizing the likelihood function is mathematically equivalent to maximizing its logarithm. It is often more convenient to work with the 'log-likelihood' function defined as follows:

Definition 9 Given a random variable X with parameter θ, the log-likelihood of θ as a function of observations x_1, x_2, ..., x_n is defined as

$$\mathcal{L}(\theta|\, x_1, \ldots, x_n) = \ln L(\theta|\, x_1, \ldots, x_n)\,.$$

Definition 10 Given a random variable X with parameter θ, and a set of observations x_1, x_2, ..., x_n, the Maximum Likelihood Estimator (MLE) of θ is the value for which $L(\theta|\, x_1, \ldots, x_n)$ is maximized:

$$\hat{\theta} \;=\; \operatorname*{argmax}_{\theta} L(\theta|\, x_1, \ldots, x_n) \;=\; \operatorname*{argmax}_{\theta} \mathcal{L}(\theta|\, x_1, \ldots, x_n).$$

For convenience, when the sample x_1, x_2, ..., x_n is understood, we will denote the likelihood and log-likelihood of θ by $L(\theta)$ and $\mathcal{L}(\theta)$, respectively.

Using the asymptotic expression of the distance distribution given in Lemma 1, for a given sample of neighborhood distances x_1, x_2, ..., x_n, we see that the log-likelihood of $\mathrm{ID_X}$ is given by

$$\begin{aligned}
\mathcal{L}(\mathrm{ID_X}) &= \ln\left[\prod_{i=1}^{n} f_{\mathrm{X,w}}(x_i)\right] \\
&= \ln\left[\prod_{i=1}^{n} \frac{F_{\mathrm{X,w}}(w)\,\mathrm{ID_X}}{w}\left(\frac{x_i}{w}\right)^{\mathrm{ID_X}-1}\right] \\
&= n\ln\frac{F_{\mathrm{X,w}}(w)}{w} + n\ln\mathrm{ID_X} + (\mathrm{ID_X}-1)\sum_{i=1}^{n}\ln\frac{x_i}{w}.
\end{aligned}$$

The first- and second-order derivatives of the log-likelihood function are respectively

$$\mathcal{L}'(\mathrm{ID_X}) \;=\; -\frac{n}{\mathrm{ID_X}} - \sum\nolimits_{i=1}^{n}\ln\frac{x_i}{w} \quad\text{and}\quad \mathcal{L}''(\mathrm{ID_X}) \;=\; \frac{n}{\mathrm{ID_X^2}}\,.$$

Accordingly, the maximum-likelihood estimate $\widehat{\mathrm{ID}}_{\mathrm{X}}$ is

$$\widehat{\mathrm{ID}}_{\mathrm{X}} = -\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}\ln\frac{x_i}{w}\right)^{-1},$$

which follows the form of the well-known Hill estimator for the scaling exponent of a power-law tail distribution [H$^+$75].

The variance is asymptotically given by the inverse of the Fisher information, defined as

$$I = \mathsf{E}\left[-\frac{\partial^2\mathcal{L}(\mathrm{ID_X})}{\partial\,\mathrm{ID_X^2}}\right] = \frac{n}{\mathrm{ID_X^2}},$$

where $\mathsf{E}[\cdot]$ denotes the expectation. Therefore, if the number of samples n is sufficiently large, we have $\widehat{\mathrm{ID}}_{\mathrm{X}} \sim \mathcal{N}(\mathrm{ID_X}, \mathrm{ID_X^2}\,/\,n)$. Accordingly, with probability $1-\beta$,

a sample of n distances in $[0, w)$ provides an estimate $\widehat{\mathrm{ID}}_{\mathrm{X}}$ lying within

$$\mathrm{ID}_{\mathrm{X}} \pm \frac{\mathrm{ID}_{\mathrm{X}}}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\beta}{2}\right).$$

In other words, the $1 - \beta$ confidence interval is

$$\left[\frac{\widehat{\mathrm{ID}}_{\mathrm{X}}}{1 + n^{-1/2}\Phi^{-1}(1 - \beta/2)}, \frac{\widehat{\mathrm{ID}}_{\mathrm{X}}}{1 - n^{-1/2}\Phi^{-1}(1 - \beta/2)}\right].$$

Using different assumptions, Levina and Bickel obtained the same ID estimator [LB04]. However, their estimator was obtained following several restrictive assumptions. In fact, they assume that a mapping exists between the data and a lower-dimensional embedding. The also assume the manifold to be locally smooth and the density to be locally constant. The Levina & Bickel estimator is viewed by many as a global estimator because their original approach used the local estimates as a step to obtain a single global estimate.

5.1.2 Method of Moments

For any choice of $k \in \mathbb{N}$, the k-th order non-central moment μ_k of the random distance X is

$$\mu_k = \mathsf{E}\left[\mathrm{X}^k\right] = \int_{x=0}^{w} x^k f_{\mathrm{X}}(x)\, \mathrm{d}x = w^k \frac{\mathrm{ID}_{\mathrm{X}}}{\mathrm{ID}_{\mathrm{X}} + k}.$$

Solving for the intrinsic dimension gives

$$\mathrm{ID}_{\mathrm{X}} = -k\frac{\mu_k}{\mu_k - w^k} = g\left(\frac{\mu_k}{w^k}\right),$$

with $g(x) = k\frac{x}{1-x}$. When estimating the order-k moment by its empirical counterpart $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$, we see that $\mathsf{E}[\hat{\mu}_k] = \mu_k$ and $\mathsf{E}[\hat{\mu}_r^2] = (n\mu_{2k} + n(n-1)\mu_k^2)n^{-2}$, so that

$$\mathsf{Var}[\hat{\mu}_k^2] = \frac{\mu_{2k} - \mu_k^2}{n} = \frac{w^{2k}\mathrm{ID}_{\mathrm{X}}k^2}{n(\mathrm{ID}_{\mathrm{X}} + 2k)(\mathrm{ID}_{\mathrm{X}} + k)^2}.$$

Assuming the convergence of the empirical moments, the distribution of $\frac{\hat{\mu}_k}{w^k}$ is therefore asymptotically normal, with

$$\frac{\hat{\mu}_k}{w^k} \sim \mathcal{N}\left(\frac{\mathrm{ID}_\mathrm{X}}{\mathrm{ID}_\mathrm{X}+k}; \frac{\mathrm{ID}_\mathrm{X} k^2}{n(\mathrm{ID}_\mathrm{X}+2k)(\mathrm{ID}_\mathrm{X}+k)^2}\right).$$

According to [Rao09, Th. 6a2.9], if $x \sim \mathcal{N}(\mu; \sigma^2 n^{-1})$ asymptotically, then $g(x) \sim \mathcal{N}(g(\mu); \sigma^2 n^{-1} g'(\mu)^2)$, where g' is the first-order derivative of g. Therefore, asymptotically

$$\widehat{\mathrm{ID}}_\mathrm{X} \sim \mathcal{N}\left(\mathrm{ID}_\mathrm{X}; \frac{\mathrm{ID}_\mathrm{X}^2}{n}\left(1 + \frac{(k/\mathrm{ID}_\mathrm{X})^2}{\mathrm{ID}_\mathrm{X}^2(1+2k/\mathrm{ID}_\mathrm{X})}\right)\right).$$

This variance is monotonically increasing in k/ID_X, which indicates that we should use moments of small order k. When k/ID_X tends to zero, the variance converges to $\mathrm{ID}_\mathrm{X}^2/n$, the variance of the maximum-likelihood estimator (see Section 5.1.1). Note that an upper bound on ID_X implies that the variance is bounded. In this case we can derive confidence intervals similar to Section 5.1.1.

5.1.3 Probability-Weighted Moments

General probability-weighted moments are defined as

$$m_{k,l,m} = \mathsf{E}\left[F_\mathrm{X}(\mathrm{X})^k (1 - F_\mathrm{X}(\mathrm{X}))^l \mathrm{X}^m\right].$$

We restrict here our attention to a subfamily: for any choice of $k \in \mathbb{N}$, ν_k is defined as

$$\nu_k \triangleq \mathsf{E}\left[F_\mathrm{X}(\mathrm{X})^k \mathrm{X}\right] = \int_{x=0}^{w} F_\mathrm{X}(x)^k x f_\mathrm{X}(x)\,\mathrm{d}x = \frac{\mathrm{ID}_\mathrm{X}\, w}{\mathrm{ID}_\mathrm{X}\, k + \mathrm{ID}_\mathrm{X} + 1};$$

solving for the intrinsic dimension yields

$$\mathrm{ID}_\mathrm{X} = \frac{\nu_k}{w - \nu_k(k+1)} = h\left(\frac{\nu_k}{w}\right), \quad \text{where} \quad h(x) = \frac{x}{1-(k+1)x}.$$

According to [HW87] and [LMW79], a commonly-used estimator of the k-th probability-weighted moment of this form is

$$\hat{\nu}_k = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{i-0.35}{n}\right)^k x_i.$$

Analogously to the previous section, we can show that this estimator has variance

$$\mathsf{Var}[\hat{\nu}_k] = \frac{\mathrm{ID_X}\, w^2}{(\mathrm{ID_X}\, k + \mathrm{ID_X} + 1)(2\,\mathrm{ID_X}\, k + \mathrm{ID_X} + 2)}.$$

Similarly, we find that asymptotically

$$\widehat{\mathrm{ID}}_\mathrm{X} \sim \mathcal{N}\left(\mathrm{ID_X}; \frac{\mathrm{ID_X^2}}{n}\left(1 + \frac{(\mathrm{ID_X}k + 1)^2}{\mathrm{ID_X}(2\mathrm{ID_X}k + \mathrm{ID_X} + 2)}\right)\right).$$

For $k = 0$, the variance is equivalent to that of the moment-based estimator with $k = 1$ (see Section 5.1.2). Since the variance increases monotonically with k for any fixed $\mathrm{ID_X}$, the use of lower-order probability-weighted moments is advisable.

5.1.4 Estimation Using Regularly Varying Functions

In this section we introduce an ad hoc estimator for the intrinsic dimensionality based on the characterization of distribution tails as regularly varying functions. Consider the empirical distribution function $\hat{F}_\mathrm{X}$, defined as

$$\hat{F}_\mathrm{X}(x) = \frac{1}{n}\sum\nolimits_{j=1}^{n} [\![x_j < x]\!],$$

where $[\![\varphi]\!]$ refers to the Iverson bracket, which evaluates to 1 if φ is true, and 0 otherwise. We propose the following estimator for the index κ of F_X.

Definition 11 Let X be an absolutely continuous random distance variable restricted to $[0, w)$. The local intrinsic dimension $\mathrm{ID_X}$ can be estimated as

$$\widehat{\mathrm{ID}}_\mathrm{X} = \hat{\kappa} = \frac{\sum_{j=1}^{J} \alpha_j \ln\left[\hat{F}_\mathrm{X}((1+\tau_j\delta_n)x_n)/\hat{F}_\mathrm{X}(x_n)\right]}{\sum_{j=1}^{J} \alpha_j \ln(1+\tau_j\delta_n)},$$

under the assumption that $\delta_n \to 0$ as $n \to \infty$, where $(\alpha_j)_{1\le j\le J}$ and $(\tau_j)_{1\le j\le J}$ are sequences.

We will refer to this family of estimators as RV, for 'regularly varying'. Note that since RV estimators involve only the products $\tau_j\delta_n$ for $1 \le j \le J$, we may assume without loss of generality that $\tau_1 + \cdots + \tau_J = 1$. The estimators are based on the observation that, for all $1 \le j \le J$,

$$\begin{aligned}
&\ln\left[F_{\mathrm{X}}((1+\tau_j\delta_n)x_n)/F_{\mathrm{X}}(x_n)\right] \\
&= \kappa\ln(1+\tau_j\delta_n) + \ln\left[\ell_{\mathrm{X}}((1+\tau_j\delta_n)x_n)/\ell_{\mathrm{X}}(x_n)\right] \\
&\simeq \kappa\ln(1+\tau_j\delta_n).
\end{aligned}$$

The RV family covers several of the known local estimators of intrinsic dimensionality. For the parameter choices $J = 1$ and $\epsilon = \tau\delta_n$, the RV estimator reduces to the GED formulation proposed in [HKN12]:

$$\widehat{\mathrm{ID}}_{\mathrm{X}} = \frac{\ln\left[\hat{F}_{\mathrm{X}}((1+\epsilon)x_n)/\hat{F}_{\mathrm{X}}(x_n)\right]}{\ln(1+\epsilon)}.$$

By setting $\epsilon = 1$, Karger & Ruhl's expansion dimension is obtained, while by setting x_n as the distance to the k-nearest neighbor and ϵ such as $(1+\epsilon)x_n$ as the distance to the nearest neighbor, we find a special case of the MiND family (precisely MiND_{ml1}) [RLC+12].

Alternatively, by setting $J = n$, $\alpha_i = 1$ for all $i \in [1..n]$, and choosing the vector τ such that $1 + \tau_i\delta_n = \frac{x_i}{x_n}$, the RV estimator becomes

$$\widehat{\mathrm{ID}}_{\mathrm{X}} = \frac{\sum_{j=1}^{n}\ln[j/n]}{\sum_{j=1}^{n}\ln[x_j/x_n]} \approx \frac{\ln\sqrt{2\pi n} - n}{\sum_{j=1}^{n}\ln[x_j/x_n]}.$$

As $n \to \infty$, this converges to the MLE (Hill) estimator presented in Section 5.1.1, with $w = x_n$.

We now turn our attention to an analysis of the variation of RV estimators. First, we introduce an auxiliary function which drives the speed of convergence of the estimator proposed in Definition 11. For $x \in \mathbb{R}$ let $\varepsilon_{\mathrm{X}}(x)$ be defined as

$$\varepsilon_{\mathrm{X}}(x) \triangleq \frac{x\ell'_{\mathrm{X}}(x)}{\ell_{\mathrm{X}}(x)}.$$

In [AGd03, AdL03], the auxiliary function is assumed to be regularly varying, and the estimation of the corresponding regular variation index is addressed. Within this article, so as to prove the following results, we limit ourselves to the assumption that ε_{X} is ultimately non-increasing.

Theorem 5 Let X be a random distance variable over $[0, w)$ with distribution function $F_{\mathrm{X}}(x) = x^{\kappa}\ell_{\mathrm{X}}(1/x)$, and let $\tau_{\max} \triangleq \max_{1\le j\le J}\tau_j$. Furthermore, let

$\delta_n, x_n \to 0$ so that $n\, F_{\mathrm{X}}(x_n)\delta_n \to \infty$ and $\sqrt{nF_{\mathrm{X}}(x_n)\delta_n}\varepsilon_{\mathrm{X}}(1/[(1+\tau_{\max}\delta_n)x_n]) \to 0$ as n approaches infinity. If the auxiliary function ε_{X} is ultimately non-increasing, then $\sqrt{nF_{\mathrm{X}}(x_n)\delta_n}\cdot[\mathrm{ID}_{\mathrm{X}} - \widehat{\mathrm{ID}}_{\mathrm{X}}]$ converges to a centered Gaussian with variance

$$\mathrm{ID}_{\mathrm{X}} V_{\alpha,\tau} = \mathrm{ID}_{\mathrm{X}} \frac{\alpha^\top S \alpha}{(\alpha^\top \tau)^2},$$

where $S_{a,b} = (|\tau_a| \wedge |\tau_b|)[\![\tau_a\tau_b > 0]\!]$ for $(a,b) \in \{1,\ldots,J\}^2$. ($A \wedge B$ denotes the minimum of A and B.)

Note that the requirement $nF_{\mathrm{X}}(x_n)\delta_n \to \infty$ can be interpreted as a necessary and sufficient condition for the almost sure presence of at least one distance sample in the interval $[x_n, (1+\tau_j\delta_n)x_n)]$. In addition, the condition

$$\sqrt{nF_{\mathrm{X}}(x_n)\delta_n}\varepsilon_{\mathrm{X}}(1/[r_n(1+\tau_{\max}\delta_n)]) \to 0$$

enforces that the approximation bias $\varepsilon_{\mathrm{X}}(1/[(1+\delta_n)x_n])$ is negligible compared to the standard deviation of the estimate, $1/\sqrt{nF_{\mathrm{X}}(x_n)\delta_n}$. We continue the analysis by proposing choices of α that minimize the variance in Theorem 5.

Lemma 2 The weight vector $\alpha = (\alpha_1,\ldots,\alpha_J)^\top$ minimizing $V_{\alpha,\tau}$ is proportional to $\alpha_0 = S^{-1}\tau = (1,0,\ldots,0)^\top$, and the associated optimal variance is given by $V_0(\tau) = \left(\tau^\top S^{-1}\tau\right)^{-1}$.

Proof The maximum of the Rayleigh functional $\alpha^\top\tau\tau^\top\alpha\left(\alpha^\top S\alpha\right)^{-1}$ is known to be attained when α is proportional to the eigenvector associated with the largest eigenvalue of $S^{-1}\tau\tau^\top$. Since $S^{-1}\tau\tau^\top$ is a rank-one matrix, the eigenvector corresponding to the unique non-zero eigenvalue is $S^{-1}\tau$. Without any loss of generality, we permute the entries of the vector τ such that $\tau_a < \tau_b$ for all $a < b$. Asymptotically, we have $0 < \tau_1 < \cdots < \tau_J$. Noting that the first column of the matrix S is $(\tau_1, \tau_2, ..., \tau_J)^\top$, we can infer that the vector $(1,0,\ldots,0)^\top$ is a solution of the equation $S.\alpha_0 = \tau$. Since S is invertible, the solution α_0 must be unique. We therefore conclude that $\alpha_0 = (1,0,\ldots,0)^\top$. ■

For the case $J = 1$, we see that $\tau = (1)^\top$ and $V_0(1) = 1$. This indicates that the GED minimizes the variance of estimation. However, different choices can be made regarding the weight vector τ and regarding the criterion to use in order to optimize the choice of α. Minimizing variance is one choice explored in this paper,

but other criteria can be used. In general, however, the following confidence interval holds for RV estimators:

Lemma 3 Let $\beta \in (0,1)$, and assume that the assumptions of Theorem 5 hold with $\alpha = S^{-1}\tau$. Let $u_\beta = \Phi^{-1}((1+\beta)/2)$, where Φ is the cumulative distribution function of the standard Gaussian distribution. Then

$$\mathrm{ID}_{\mathrm{X}} \pm u_\beta \left(n\delta_n V_0(\tau) \widehat{\mathrm{ID}}_{\mathrm{X}} \hat{F}_{\mathrm{X}}(x_n) \right)^{-1/2}$$

are the boundaries of the asymptotic confidence interval of level β for $\widehat{\mathrm{ID}}_{\mathrm{X}}$.

Proof Lemma 3 is a direct consequence of the asymptotic distribution established in Theorem 5 and the convergence of $\hat{F}_{\mathrm{X}}(x_n)$ to $F_{\mathrm{X}}(x_n)$ as $n \to \infty$. ■

5.2 Experimental Framework

As part of our evaluation of our estimators of local intrinsic dimension, we investigate their performance (as well as those of competing estimators) on a series of data distributions, both real and artificially generated. While trials involving real application data are primarily of practical interest, the study of artificial data allows to systematically assess the ability of the individual methods to identify data dimensionality.

5.2.1 Methods

The methods used in this study include MLE, MoM, PWM, and RV. For all estimators, the neighborhood size is set to $k = 100$. The RV estimators are evaluated for the choices $J = 1$ and $J = 2$, as follows:

$$\widehat{\mathrm{ID}}_{\mathrm{RV}} = \begin{cases} \frac{\ln n - \ln(n/2)}{\ln x_n - \ln x_{\lfloor n/2 \rfloor}}, & \text{if} \quad J = 1 \\ \frac{\ln(n/j) - (p-1)\ln(i/j)}{\ln x_n/x_j + (p-1)\ln x_i/x_j}, & \text{if} \quad J = 2, \end{cases}$$

where $p = (x_i - 2x_j + x_n)/(x_n - x_j)$, $i = \lfloor n/2 \rfloor$, and $j = \lfloor 3n/4 \rfloor$. Note that the estimator RV for $J = 1$ is a form of generalized expansion dimension (GED) [HKN12]. For every dataset, we report the average of ID estimates across all the points in the dataset. All estimators in our study can be computed in time linear in the number of sample points.

Method	Parameters
PCA	threshold $= 0.025$
kNNG_1	$k = 100$, $\gamma = 1$, $M = 1$, $N = 10$
kNNG_2	$k = 100$, $\gamma = 1$, $M = 10$, $N = 1$
MiND_{ml1}	None
MiND_{mli}	$k = 100$

Table 5.1: Parameter choices used in the experiments.

Our experimental framework includes several state-of-the-art estimators of intrinsic dimensionality, both local and global. The global estimators consist of a projection method (PCA), fractal methods (CD [CV02], Hein [HA05], Takens [Tak85]), and graph-based methods (kNNG_1, kNNG_2 [CHI04]). The local distance-based estimators are MiND_{ml1} and MiND_{mli} [RLC+12]. Table 5.1 summarizes the parameter choices for every method, except for the fractal methods, which do not involve any parameter.

The MiND variants makes more restrictive assumptions than our methods: they assume the data to be uniformly distributed on a hypersphere, with a locally isometric smooth map between the hypersphere and the representational space. MiND uses only the two extreme samples (smallest and largest), and requires knowledge of the dimension of the space (D). In contrast, our approach assumes only that the nearest neighbor distances are in the lower tail of the distance distribution, where EVT estimation can be performed.

5.2.2 Artificial Distance Distributions

In the following we propose a set of experiments concerning artificial data, and describe the method employed for the generation of test data.

First, consider a reference point P drawn uniformly at random from within the m-dimensional unit sphere, for some choice of $m \in \mathbb{N}$. According to the method of normal variates, we define $\mathrm{P} = \mathrm{Z}^{1/m}\mathrm{Y}\|\mathrm{Y}\|^{-1}$, where Z is uniformly distributed on $[0, 1]$, and Y is a random vector in $\mathbb{R}^m$ whose coefficients follow the standard normal distribution. The distance of P, with respect to our choice of reference point at location $0 \in \mathbb{R}^m$, is distributed as follows:

$$\mathrm{X} = \frac{\|\mathrm{Z}^{1/m}\mathrm{Y}\|}{\|\mathrm{Y}\|} = \mathrm{Z}^{1/m}.$$

Note that, by measuring LID purely based on distance values with respect to a reference point, the model does not require that the data have an underlying spatial representation. As such, non-integer values of $m \in \mathbb{R}$ can be selected for the generation of distances, if desired.

For choices of $m \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, we draw 100 independent sequences of sample distance values from the distribution described above, and record the estimates produced by each of our methods for sample sizes n between 10 and 10^4.

5.2.3 Artificial Data

The datasets used in our experiments have been proposed in [RLC+12]. They consist of 15 manifolds of various stuctures and intrinsic dimensionalities (d) represented in spaces of different dimensions (D). They are summarized in Table 5.2.

Manifold	d	D	Description
1	10	11	Uniformly sampled sphere.
2	3	5	Affine space.
3	4	6	Concentrated figure confusable with a 3d one.
4	4	8	Non-linear manifold.
5	2	3	2-d Helix
6	6	36	Non-linear manifold.
7	2	3	Swiss-Roll.
8	12	72	Non-linear manifold.
9	20	20	Affine space.
10a	10	11	Uniformly sampled hypercube.
10b	17	18	Uniformly sampled hypercube.
10c	24	25	Uniformly sampled hypercube.
11	2	3	Möbius band 10-times twisted.
12	20	20	Isotropic multivariate Gaussian.
13	1	13	Curve.

Table 5.2: Artificial datasets used in the experiments.

These datasets were generated in different sizes (10^3, 10^4, and 10^5 points) in order to evaluate the effect of the number of points on the quality of the different estimators. For each dataset and for each of the three sizes, we average the estimates over 20 instances.

In order to evaluate the robustness of the estimators, we also prepared versions of these datasets with noise added. For each attribute f, we added normally distributed noise with mean equal to zero and standard deviation $\sigma_n = p \cdot \sigma_f$ where σ_f is the standard deviation of the attribute itself, and $p \in \{0.01, 0.04, 0.16, 0.64\}$. For attributes with $\sigma_f = 0$, the noise was generated with standard deviation $\sigma_n = p \cdot \sigma_f^*$ where σ_f^* is the minimum of the nonzero standard deviations over all attributes.

5.2.4 Real Data

Not only can a reliable estimation of ID greatly benefit the practical performance of many applications [KR02, BKL06, HMNO12], it also serves as a characterization of high-dimensional datasets and the potential problems associated with their use in practice. To this end, we investigate the distribution of LID estimates on the following datasets, each taken from a real-world application scenario.

- The ALOI (Amsterdam Library of Object Images) data set contains a total of 110250 color photos of 1000 different objects taken from varying viewpoints under various illumination conditions. Each image is described by a 641-dimensional vector of color and texture features [BFF+01].

- The ANN_SIFT1B dataset consists of one billion 128-dimensional SIFT descriptors randomly selected from the dataset ANN_SIFT, consisting of $2.8 \cdot 10^{10}$ SIFT descriptors extracted from $3 \cdot 10^7$ images. These sets have been created for the evaluation of nearest-neighbor search strategies at very large scales [JTDA11].

- BCI5 [Mil04] is a brain-computer interface dataset in which the classes correspond to brainwave readings taken while the subject contemplated one of three different actions (movement of the right hand, movement of the left hand, and the subvocalization of words beginning with the same letter).

- Gisette [GGBHD04] is a subset of the MNIST [LBBH98] handwritten digit image dataset, consisting of 50-by-50-pixel images of the highly confusable digits '4' and '9'. 2500 random features were artificially generated and added to the original 2500 features, so as to embed the data into a higher dimensional feature space. As the dataset was created for the NIPS 2003 feature

selection challenge, the precise generation mechanism of the random features was not made public.

- Isolet [CF90] is a set of 7797 human voice recordings in which 150 subjects recite each of the 26 letters of the alphabet twice. Each entry consists of 617 features representing selected utterances of the recording.

- The MNIST database [LBBH98] contains of 70000 recordings of handwritten digits. The images have been normalized and discretized to a 28×28-pixel grid. The gray-scale values of the resulting 784 pixels are used to form the feature vectors.

5.2.5 Approximate Nearest Neighbors

For many datasets, various approximate nearest neighbor (ANN) methods can generate neighborhood sets much faster than would be possible using an exact indexing method. As a rule, with ANN indexing methods it is possible to influence the trade-off between accuracy and time complexity by means of parameter choices at query time, design choices at construction time, or both. However, the use of approximate neighborhood information can lead to a degradation in the quality of data statistics that rely on it. In particular, the question arises as to how the quality of LID estimators are affected when applied to distance samples generated from approximate neighborhoods of diminishing accuracy. In this part of the experimental study, we investigate the relationship between the accuracy of neighborhood sets and the accuracy of LID estimates. Here, accuracy is measured as the proportion of distance samples in the exact neighborhood that also appear in the approximate neighborhood under consideration. Under the assumption that the exact and approximate neighborhoods all have the same size k, this notion of accuracy coincides with those of both recall and precision.

For any given dataset, we can generate approximate k-NN sets with carefully controlled levels of accuracy, through the sparsification of exact neighbor sets of size greater than k. The sparsification is done in two steps. In the first step, we randomly select a proportion of the exact k nearest neighbors at the desired level of accuracy. In the second step, we complete the new approximate list with nearest neighbors drawn from outside the exact k-NN list, in a way that the selection rate matches the accuracy. More precisely, let r be the target level of accuracy, expressed as a

proportion between 0 and 1. Initially, the approximate neighborhood distance sample is constructed by randomly selecting $\lfloor rk \rfloor$ elements of the approximate neighborhood (without replacement) from among the first k elements in the exact k-NN set. Next, an additional $k - \lfloor rk \rfloor$ elements are randomly selected from among those ranked between $k+1$ and $K = \lceil k/r \rceil$ in the exact K-NN set, and add their distances to the sample. With this choice of K, the accuracy of the approximate k-NN query result is almost identical to that of the K-NN query result:

- for neighbors ranked between 1 and $\lfloor rk \rfloor$, the accuracy is $\lfloor rk \rfloor / k$, where

$$r \cdot \left(1 - \frac{1}{k}\right) < \frac{\lfloor rk \rfloor}{k} \leq r \,,$$

- for neighbors ranked between 1 and $\lceil k/r \rceil$, the accuracy is $k/\lceil k/r \rceil$, where

$$r \cdot \left(1 - \frac{1}{k+1}\right) < \frac{k}{\lceil k/r \rceil} \leq r \,.$$

As k increases, these upper and lower bounds converge to r.

In our experiments, to observe the effect of using ANN on LID values, we use MLE estimation with $k = 100$. The accuracy r is chosen from the range 0.5 to 1.0, since for these values, the maximum size of the exact neighborhoods required for the experimentation is a manageable $2k = 200$.

5.2.6 Nearest Neighbor Descent

The computational and storage costs associated with the construction of an exact k-nearest neighbor graph (similarity graph) is a limitation in many machine learning algorithms. Particularly in high-dimensional settings, the cost of generating all exact k-nearest neighbor lists can be quadratic in the number of data objects, which for large datasets can be prohibitively high. Many approximation methods exist for the construction of nearest neighbor (ANN) with computation costs much less than those of exact methods, though at the expense of accuracy.

Algorithm 4 NN-descent [DML11] given a dataset D, a distance function d, and a neighborhood size k:

1. For each data point $q \in D$:

- Initialize G by randomly generating a tentative k-NN list for q with an assigned distance of $+\infty$;
- Compute the RNN (reverse nearest neighbor) lists for q.

2. Repeat
   - For each data point $q \in D$:
     - Check different pairs of q's neighbors (u, v) in q's k-NN and RNN lists, and compute $\mathrm{d}(u, v)$;
     - Use $\langle u, \mathrm{d}(u, v)\rangle$ to update v's k-NN list, and use $\langle v, \mathrm{d}(u, v)\rangle$ to update u's k-NN list;
3. until the k-NN graph G converges.
4. Return the k-NN graph G.

We conducted an experiment to show that the process of obtaining the neighborhoods necessary for LID estimation can be considerably accelerated using a state-of-the-art ANN method, with little or no effect on LID estimates. From among the many ANN algorithms available, we chose the state-of-the-art Nearest Neighbor Descent (NN-Descent) [DML11] algorithm for our experimentation. The NN-Descent algorithm is based on the assumption of transitivity of the similarity measure — in other words, that two neighbors of a given data object are also likely to be neighbors of one another. As shown in the pseudo-code description of Algorithm 4, all points are initially associated with randomly built 'k-NN lists' which are then iteratively updated. At every iteration, a pivot element q is selected, and each possible pair (u, v) of q's neighbors is considered for mutual updates. If the distance $\mathrm{d}(u, v)$ is smaller than the distance to the last element in u's k-NN list, then the list is updated by inserting v in the appropriate location. The same test is applied to the k-NN list of v. In addition, similar tests are applied to the reverse (inverted) k-NN list of q. The algorithm converges when a pivot selection round completes without updates are made to the k-NN lists. As recommended in the original paper [DML11], we modified the convergence condition so as to terminate after a maximum of 7 rounds of the loop in lines 2-3.

Dataset	d	D	ID_{MLE}	ID_{MoM}	ID_{PWM}	ID_{GED}	ID_{RVE}	$MiND_{ml1}$	$MiND_{mli}$	CD	Hein	Takens	kNNG$_1$	kNNG$_2$	PCA
m1	10	11	8.07	8.08	8.14	7.91	7.79	9.50	8.95	9.24	5.35	9.44	7.96	7.02	11.00
m2	3	5	2.67	2.67	2.68	2.65	2.60	2.94	3.00	2.87	2.75	2.91	2.53	2.52	3.00
m3	4	6	3.56	3.56	3.59	3.55	3.49	3.88	4.00	3.63	3.70	3.66	4.00	2.88	5.30
m4	4	8	4.76	4.93	5.18	5.16	5.06	3.90	4.00	3.93	5.00	3.78	6.04	3.38	8.00
m5	2	3	1.98	2.03	2.07	2.03	2.00	1.97	2.00	1.95	2.30	1.98	2.27	1.99	3.00
m6	6	36	7.08	7.18	7.39	7.24	7.13	6.00	7.00	5.73	2.85	5.73	9.43	8.30	12.00
m7	2	3	2.49	2.80	3.04	3.22	3.12	2.00	2.00	1.95	1.90	1.95	3.10	2.86	3.00
m8	12	72	12.29	12.33	12.51	11.97	11.79	13.49	13.00	11.00	3.60	11.85	14.28	12.56	24.00
m9	20	20	12.39	12.40	12.50	11.96	11.79	15.03	13.50	12.84	4.30	14.68	19.68	10.84	20.00
m10a	10	11	7.39	7.40	7.47	7.28	7.16	8.50	8.00	8.42	8.15	8.45	10.69	6.65	10.00
m10b	17	18	11.06	11.07	11.15	10.73	10.56	13.40	12.00	9.35	7.05	13.16	12.42	14.45	17.00
m10c	24	25	14.05	14.07	14.22	13.52	13.32	17.69	15.35	16.82	6.05	16.90	17.31	29.77	24.00
m11	2	3	2.49	2.74	2.94	3.05	2.97	2.01	2.00	1.99	2.70	2.00	2.83	2.59	3.00
m12	20	20	12.48	12.46	12.43	11.85	11.67	16.79	14.00	13.69	3.70	13.64	11.71	5.13	20.00
m13	1	13	1.35	1.75	2.11	2.22	2.08	1.01	1.00	1.01	1.15	1.01	1.46	1.36	7.90

Table 5.3: Average ID estimates for 1000-point-manifolds using 100 nearest neighbors.

5.3 Experimental Results

5.3.1 Artificial Distance Distributions

We begin our experimental study with an assessment — in terms of bias, variance, and convergence — of the ability of each estimator to identify the ID of a sample of distance values generated according to different choices of target ID. Note that for these trials, the distributional model asserted in Lemma 1 holds everywhere on the range $[0, w)$ by construction (with $w = 1$).

Figures 5.1 and 5.2 show the behavior of MLE, MoM, and RV (for choices of $J = 1$ and $J = 2$). The convergence to the target ID value observed in every case empirically confirms the consistency of these estimators. Likewise, PWM is consistent however, one should beware of PWM's susceptibility to the effects of numerical instability.

We also note that the RV estimator with $J = 1$ (GED) — which asymptotically minimizes variance according to Lemma 2 — is not the choice that minimizes variance when the number of samples is limited. Faster initial convergence favors the choice of MLE and MoM for applications where the number of available query-to-neighbor distances is limited, or where time complexity is an issue.

5.3.2 Artificial Data

In Tables 5.3 and 5.4, for each of the estimators considered in this study, we present ID estimates for the artifical datasets, averaged over 20 runs each. It should be noted that as PCA and $MiND_{mli}$ estimates are restricted to integer values, their bias is

Dataset	d	D	ID_{MLE}	ID_{MoM}	ID_{PWM}	ID_{GED}	ID_{RVE}	$MiND_{ml1}$	$MiND_{mli}$	CD	Hein	Takens	$kNNG_1$	$kNNG_2$	PCA
m1	10	11	9.04	9.10	9.32	9.06	8.92	9.61	9.00	9.56	8.95	9.59	9.20	9.87	11.00
m2	3	5	2.88	2.90	2.94	2.90	2.85	2.96	3.00	3.08	3.55	2.98	2.77	2.44	3.00
m3	4	6	3.86	3.90	3.97	3.92	3.85	3.92	4.00	3.75	3.90	3.76	3.94	3.94	5.05
m4	4	8	4.06	4.14	4.27	4.23	4.15	3.91	4.00	3.83	4.65	3.84	3.84	3.84	8.00
m5	2	3	1.98	2.01	2.04	2.01	1.98	1.90	1.95	2.05	2.20	2.00	2.02	2.02	3.00
m6	6	36	6.64	6.78	7.11	7.01	6.89	5.85	6.00	5.05	4.30	5.66	3.34	3.34	12.00
m7	2	3	1.96	1.99	2.02	1.99	1.95	1.99	2.00	1.97	1.95	1.98	1.83	1.83	3.00
m8	12	72	13.72	13.86	14.50	13.91	13.69	12.91	14.00	11.95	8.10	11.92	14.08	14.08	24.00
m9	20	20	14.47	14.56	15.08	14.41	14.18	15.95	15.00	15.69	2.65	15.74	10.11	10.11	20.00
m10a	10	11	8.20	8.25	8.43	8.21	8.08	8.86	8.00	8.87	9.10	8.92	6.55	6.55	10.00
m10b	17	18	12.72	12.80	13.21	12.69	12.49	13.95	13.00	13.82	6.70	13.85	19.52	19.52	17.00
m10c	24	25	16.66	16.77	17.45	16.54	16.28	18.50	17.00	18.08	10.90	18.13	15.00	15.00	24.00
m11	2	3	1.99	2.03	2.06	2.04	2.00	1.99	2.00	1.99	2.00	2.00	1.84	1.84	3.00
m12	20	20	15.46	15.54	16.03	15.23	15.00	17.74	16.00	15.04	3.70	15.00	37.63	37.63	20.00
m13	1	13	1.01	1.04	1.06	1.03	1.01	0.00	1.00	1.00	1.00	1.00	0.85	0.85	8.00

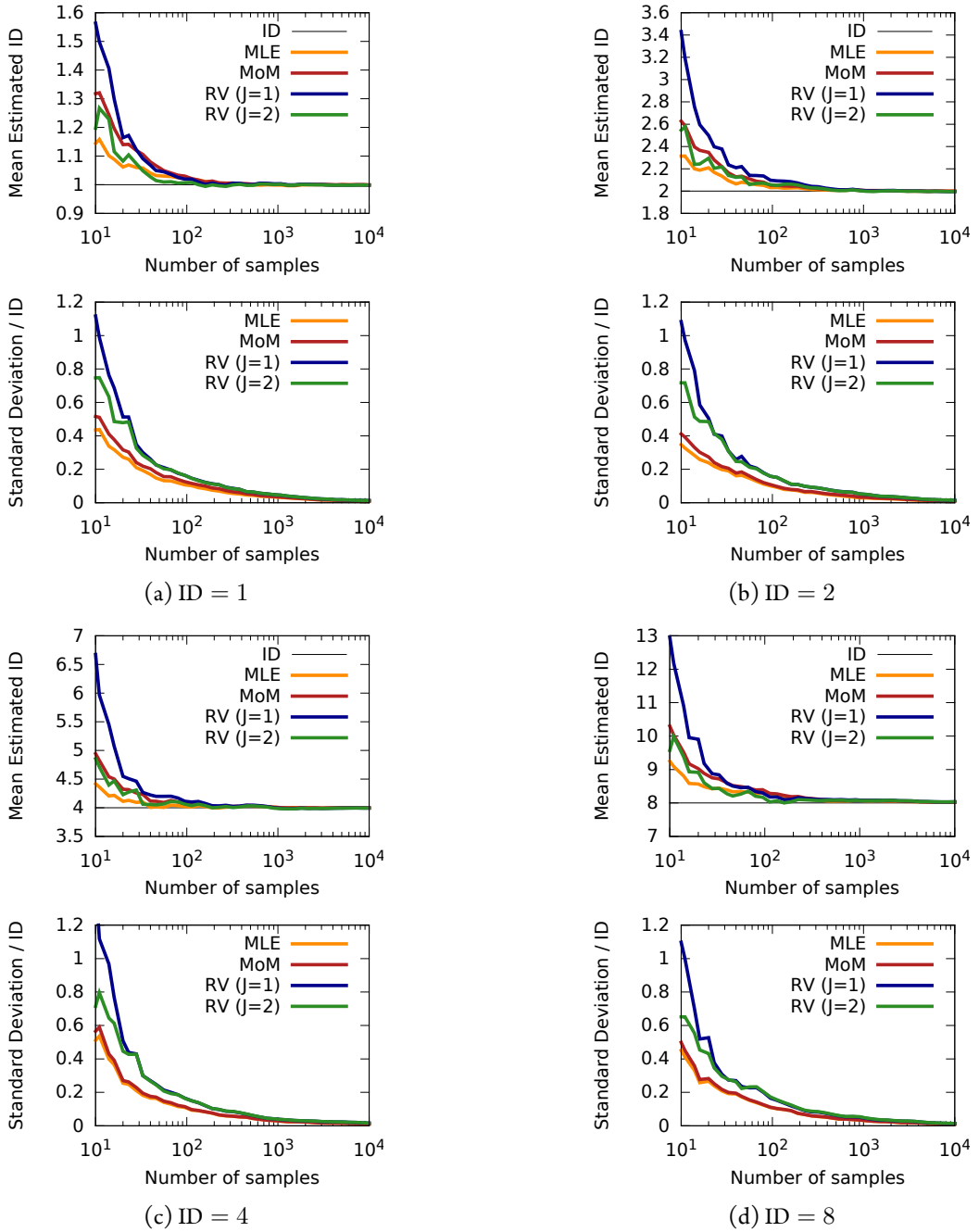
Table 5.4: Average ID estimates for 10000-point-manifolds using 100 nearest neighbors.

Dataset	d	D	ID_{MLE}	ID_{MoM}	ID_{PWM}	ID_{GED}	ID_{RVE}	$MiND_{ml1}$	$MiND_{mli}$	CD	Hein	Takens	$kNNG_1$	$kNNG_2$	PCA
m1	10	11	-10.07	-10.55	-11.80	-11.81	-11.88	-1.56	-2.78	-11.82	-38.55	-12.10	-62.17	-64.74	-22.73
m2	3	5	2.43	-1.03	-3.06	-3.10	-3.51	36.49	0.00	12.01	-18.31	23.15	16.97	32.79	-33.33
m3	4	6	-30.83	-32.05	-33.25	-33.16	-33.25	-23.47	-25.00	-22.13	-41.03	-22.34	-35.79	-35.79	-60.40
m4	4	8	65.02	62.32	59.25	57.21	57.59	88.75	70.00	78.85	-6.45	78.12	21.61	21.61	-18.75
m5	2	3	-48.48	-48.26	-48.04	-48.26	-48.48	-37.37	-48.72	-71.71	-54.55	-49.50	-25.74	-25.74	-66.67
m6	6	36	166.11	161.65	157.10	144.94	145.43	282.22	220.83	261.58	-30.23	221.02	219.76	219.76	131.25
m7	2	3	-8.67	-14.57	-16.34	-15.58	-15.90	34.17	0.00	14.21	10.26	9.60	-44.81	-44.81	-66.67
m8	12	72	44.17	43.00	39.79	35.44	35.65	115.49	60.71	86.53	25.93	85.99	93.68	93.68	95.21
m9	20	20	-21.77	-22.25	-24.34	-24.01	-23.98	-9.97	-17.00	-22.12	167.92	-22.62	157.17	157.17	-31.75
m10a	10	11	21.46	21.45	21.59	20.83	20.92	22.12	25.00	9.02	-64.29	8.07	338.17	338.17	10.00
m10b	17	18	12.89	12.73	12.49	11.66	11.69	17.35	15.38	1.37	-29.85	0.65	-36.37	-36.37	5.88
m10c	24	25	7.98	7.87	7.45	6.83	6.88	14.76	11.76	-2.99	-74.31	-3.75	-177.73	-177.73	4.17
m11	2	3	32.16	29.56	28.64	28.43	28.50	47.74	0.00	41.21	10.00	40.50	195.65	195.65	-35.00
m12	20	20	-22.83	-23.10	-24.52	-23.90	-23.93	-16.52	-19.69	-16.22	13.51	-16.27	-84.45	-84.45	-26.00
m13	1	13	376.24	353.85	337.74	339.81	341.58	inf	500.00	524.00	305.00	527.00	363.53	363.53	-75.00

Table 5.5: Deviation of ID estimates for 10000-point-manifolds with added noise (p=0.01) using 100 nearest neighbors.

Dataset	d	D	ID_{MLE}	ID_{MoM}	ID_{PWM}	ID_{GED}	ID_{RVE}	$MiND_{ml1}$	$MiND_{mli}$	CD	Hein	Takens	$kNNG_1$	$kNNG_2$	PCA
m1	10	11	-10.18	-10.66	-11.91	-11.92	-12.00	-1.87	-2.78	-17.05	-63.69	-12.20	-341.09	-324.72	-23.18
m2	3	5	2.43	-1.03	-3.06	-3.45	-3.51	37.16	0.00	18.83	-9.86	22.82	-7.94	4.51	-33.33
m3	4	6	-30.57	-32.05	-33.25	-33.16	-33.25	-23.47	-25.00	-26.40	-42.31	-22.07	-31.47	-31.47	-60.40
m4	4	8	65.02	62.32	59.25	56.97	57.35	89.00	71.25	55.61	-17.20	77.34	131.25	131.25	-20.00
m5	2	3	-48.48	-48.26	-48.04	-48.26	-48.48	-37.37	-48.72	-69.76	-54.55	-49.50	-50.99	-50.99	-66.67
m6	6	36	165.96	161.50	156.82	144.79	145.28	281.20	220.83	260.79	-2.33	220.49	714.07	714.07	130.83
m7	2	3	-8.67	-14.57	-16.83	-15.58	-15.90	34.17	0.00	15.74	7.69	11.62	-38.25	-38.25	-66.67
m8	12	72	44.24	43.07	39.86	35.59	35.72	116.42	60.71	86.69	46.30	86.16	9.52	9.52	95.21
m9	20	20	-21.77	-22.25	-24.27	-24.01	-23.91	-10.22	-17.33	-22.31	132.08	-22.74	15.73	15.73	-31.75
m10a	10	11	21.46	21.45	21.59	20.83	20.79	21.78	25.00	3.04	-48.35	7.96	25.65	25.65	10.00
m10b	17	18	12.89	12.73	12.49	11.66	11.69	17.20	15.38	-3.84	-36.57	0.72	-38.17	-38.17	5.88
m10c	24	25	8.04	7.87	7.51	6.83	6.94	14.49	11.76	-7.85	-59.17	-3.53	-18.80	-18.80	4.17
m11	2	3	32.16	29.56	28.64	28.43	28.50	46.73	0.00	40.20	37.50	39.00	255.43	255.43	-35.00
m12	20	20	-22.83	-23.10	-24.52	-23.90	-23.93	-16.18	-19.37	-16.16	33.78	-16.33	-174.25	-174.25	-26.00
m13	1	13	376.24	353.85	337.74	339.81	341.58	inf	500.00	525.00	220.00	528.00	327.06	327.06	-75.00

Table 5.6: Deviation of ID estimates for 10000-point-manifolds with added noise (p=0.04) using 100 nearest neighbors.



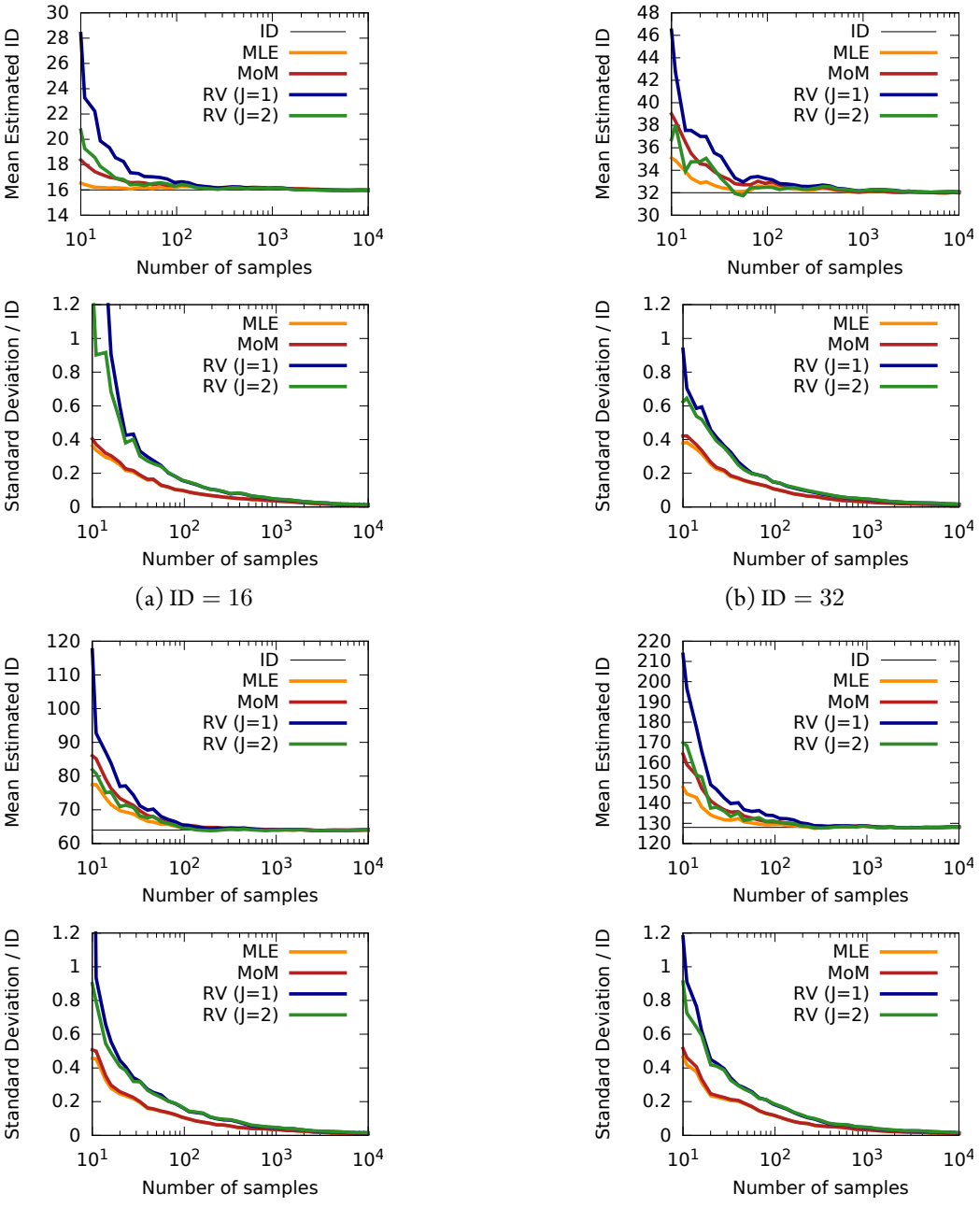
(a) ID = 1

(b) ID = 2

(c) ID = 4

(d) ID = 8

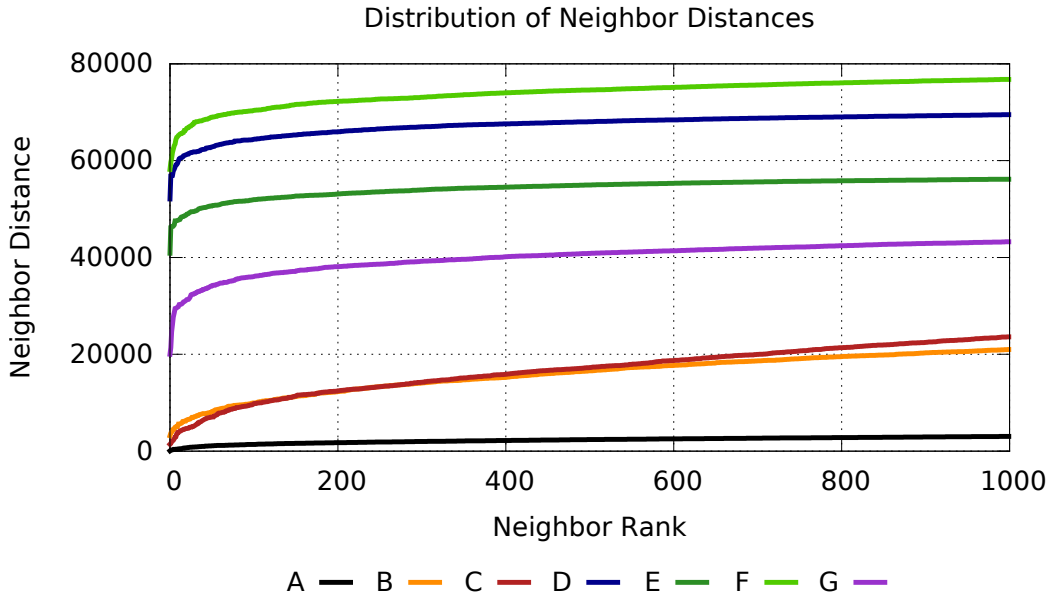
Figure 5.1: Comparison of the mean and standard deviation of LID estimates provided by MLE, MoM and RV (for $J = 1$ and $J = 2$) on increasingly large samples drawn from artificially-generated distance distributions. The results cover target dimensionality values between 1 and 8. The values are marked in the corresponding plots.



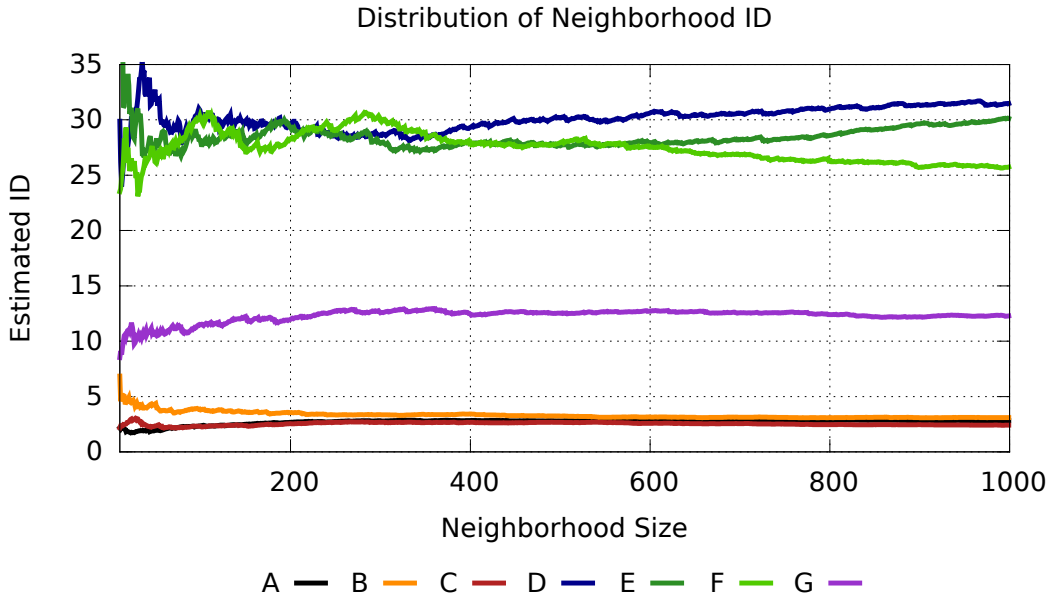
(a) ID = 16 (b) ID = 32

(c) ID = 64 (d) ID = 128

Figure 5.2: Comparison of the mean and standard deviation of LID estimates provided by MLE, MoM and RV (for $J = 1$ and $J = 2$) on increasingly large samples drawn from artificially-generated distance distributions. The results cover target dimensionality values between 16 and 128. The values are marked in the corresponding plots.



(a) Illustration of the distribution of k-nearest neighbor distances for $k \in [1, 1000]$ with respect to 7 points of interest.



(b) Distribution of LID estimates based on k-nearest neighbor sets for $k \in [10, 1000]$ with respect to 7 points of interest.

Figure 5.3: Distribution of $\mathrm{ID}_{\mathrm{MLE}}$ estimates and distance values across neighborhoods around the points of interest.

Dataset	d	D	ID_{MLE}	ID_{MoM}	ID_{PWM}	ID_{GED}	ID_{RVE}	$MiND_{mli}$	$MiND_{mli}$	CD	Hein	Takens	kNNG$_1$	kNNG$_2$	PCA
m1	10	11	-10.18	-10.66	-11.80	-11.81	-11.88	-1.77	-2.78	-16.95	-35.75	-12.10	-37.61	-41.84	-22.73
m2	3	5	2.43	-1.03	-3.06	-3.10	-3.51	37.16	0.00	19.48	-18.31	23.49	-24.19	-13.93	-33.33
m3	4	6	-30.83	-32.05	-33.25	-33.42	-33.25	-22.96	-25.00	-31.20	-35.90	-22.34	-35.03	-35.03	-60.40
m4	4	8	65.02	62.08	59.02	56.97	57.35	88.75	70.00	67.10	8.60	77.86	86.98	86.98	-20.00
m5	2	3	-48.48	-48.26	-48.04	-48.26	-48.48	-37.37	-48.72	-73.66	-54.55	-49.50	-48.02	-48.02	-66.67
m6	6	36	166.11	161.65	157.10	144.94	145.57	281.71	220.83	261.19	32.56	220.67	424.55	424.55	130.83
m7	2	3	-8.67	-14.57	-16.34	-15.58	-15.90	34.17	0.00	19.29	18.46	15.15	4.37	4.37	-66.67
m8	12	72	44.17	43.00	39.79	35.44	35.57	115.72	59.64	85.94	-11.11	85.65	-11.93	-11.93	95.21
m9	20	20	-21.77	-22.25	-24.27	-24.01	-23.98	-9.66	-17.00	-22.12	100.00	-22.68	-907.22	-907.22	-31.75
m10a	10	11	21.46	21.45	21.59	20.83	20.79	21.22	25.00	9.02	-39.56	8.18	34.35	34.35	10.00
m10b	17	18	12.89	12.73	12.49	11.66	11.69	17.28	15.38	1.16	-58.21	0.51	-38.78	-38.78	5.88
m10c	24	25	8.04	7.93	7.51	6.89	6.88	14.43	11.76	-2.71	-30.73	-3.42	-610.20	-610.20	4.17
m11	2	3	31.66	29.06	28.64	28.43	28.50	46.73	0.00	27.64	10.00	39.00	3811.41	3811.41	-35.00
m12	20	20	-22.83	-23.17	-24.52	-23.90	-23.93	-16.52	-19.69	-16.16	6.76	-16.27	-835.80	-835.80	-26.00
m13	1	13	376.24	352.88	336.79	339.81	340.59	inf	500.00	491.00	270.00	528.00	387.06	387.06	-75.00

Table 5.7: Deviation of ID estimates for 10000-point-manifolds with added noise (p=0.16) using 100 nearest neighbors.

lower for examples having integer ground-truth intrinsic dimension, especially when this dimensionality is small. Also, unlike the other estimators tested, MiND estimators also require that an upper bound on the ID be supplied (set to D in these experiments). PCA requires a threshold parameter to be supplied, the value of which can greatly influence the estimation.

The experimental results indicate that local estimators tend to over-estimate dimensionality in the case of non-linear manifolds (sets m3, m4, m5, m6, m7, m8, m11 and m13) and to under-estimate it in the case of linear manifolds (sets m1, m2, m9, m10a, m10b, m10c and m12). The experimental results with higher sampling rates confirm the reduction in bias that would be expected with smaller k-nearest-neighbor distances, as the local manifold structure more closely approximates the tangent space.

For highly non-linear manifolds, such as the Swiss Roll (m7) or the Möbius band (m11), global estimators have difficulty in identifying the intrinsic dimension. As one might expect, the local estimators ID and MiND are more accurate for such cases. Although high local curvature is reflected in the distance distribution, and consequently the local dimensional estimates as well, the effect is much smaller than for global estimators. With a higher sampling rate, k-nearest neighbor distances are diminished, and the curvature becomes locally less significant. The local manifold structure tends to that of its tangent space, reducing the bias of local estimation. We also note that the bias is proportional to the intrinsic dimensionality of the manifold. As dimensionality increases, a higher sampling rate is required in order to reduce the bias.

To show the effects of noise on the estimators, we display in Tables 5.5, 5.6 and 5.7 for each method the deviation of every estimate in the presence of noise as a proportion of the estimate obtained in the absence of noise. In other words we divide the difference between the estimate on the noisy manifold and the estimate on the original manifold by the latter of the two. On the one hand, we note that global methods, k-NNG in particular, are significantly affected by noise: their estimates diverge very quickly as noise is being introduced. It is not necessarily a disadvantage since the structure of the manifold has been drastically changed. On the other hand, the local estimators display more resistance to noise in the case of non-linear manifolds; among the local estimators, our EVT estimators tend to outperform the MiND variants.

We note that the additive noise considered in this experiment does not drastically impact the intrinsic dimensionality in the case of hypercubes. (sets m10a, m10b and m10c). That explains why PCA appears resistant to noise for the sets m10a, m10b and m10c. However, noise in these manifolds may drive points far from their original positions, which may explain the relatively high estimates obtained by local intrinsic dimensionality estimators on these sets.

The robustness of local estimation is of great importance for many applications such as search and outlier detection. The resistance to noise seems to be generally higher in the case of manifolds of higher intrinsic dimensionality. It is important that our estimates can be trusted on these complex manifolds where the concentration effect is more important. In datasets of smaller intrinsic dimensionality, our noise model raises the dimensionality agressively which does not happen very often in real world situations.

5.3.3 Real Data

Based on our experiments on synthetic data, we expect the performance of our proposed estimators to be largely in agreement with one another. Accordingly, for clarity of presentation, for the experimentation on real data, we show results only for the MLE estimator.

For each of the datasets considered in this study, Figures 5.4 and 5.5 illustrates the distribution of LID estimates based at reference points drawn from the data. Due to its large size, for the ANN_SIFT1B dataset, the reference set was generated by selecting 10^4 items uniformly at random. For the other datasets, the entire

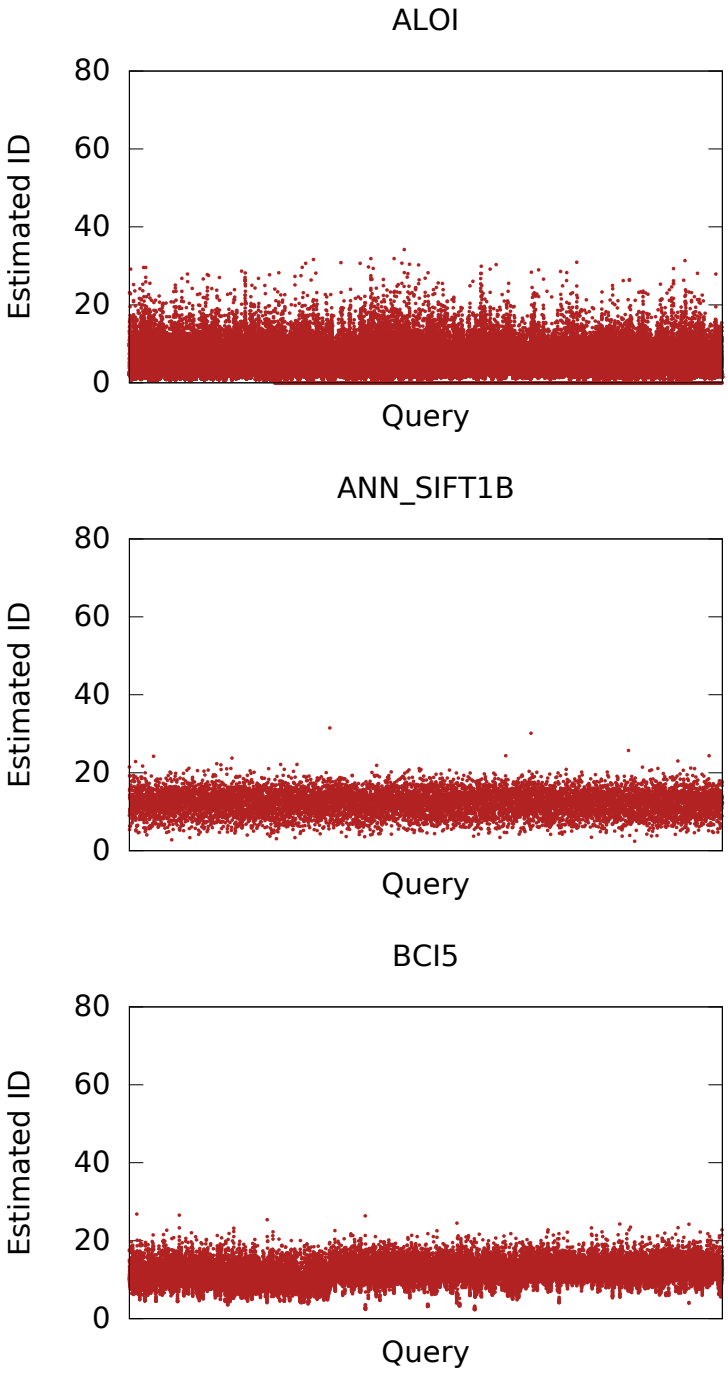


Figure 5.4: Plots of the distribution of LID values across each dataset. The LID values were obtained using the MLE estimator on the size-100 neighborhoods of the individual reference points.

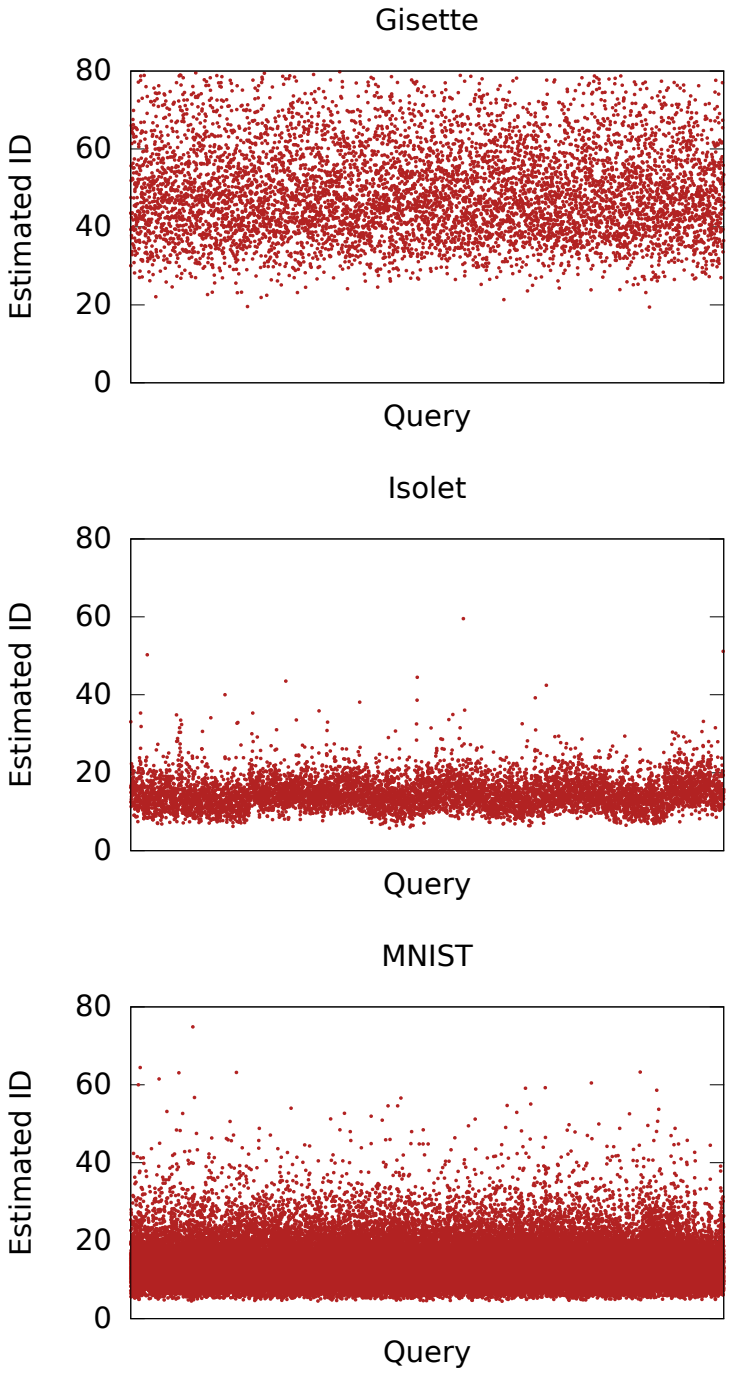


Figure 5.5: Plots of the distribution of LID values across each dataset. The LID values were obtained using the MLE estimator on the size-100 neighborhoods of the individual reference points.

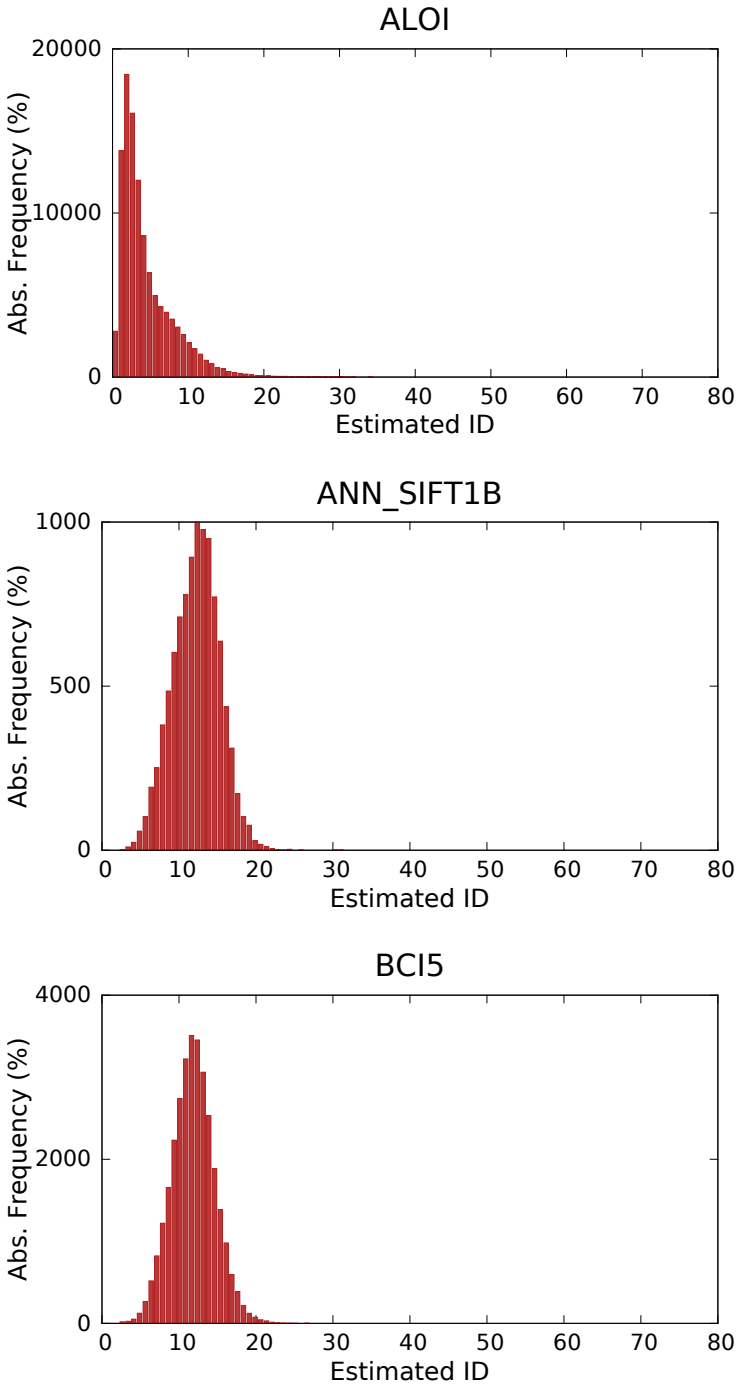


Figure 5.6: Histograms of LID values across each dataset, obtained using the MLE estimator on the size-100 neighborhoods of the individual reference points.

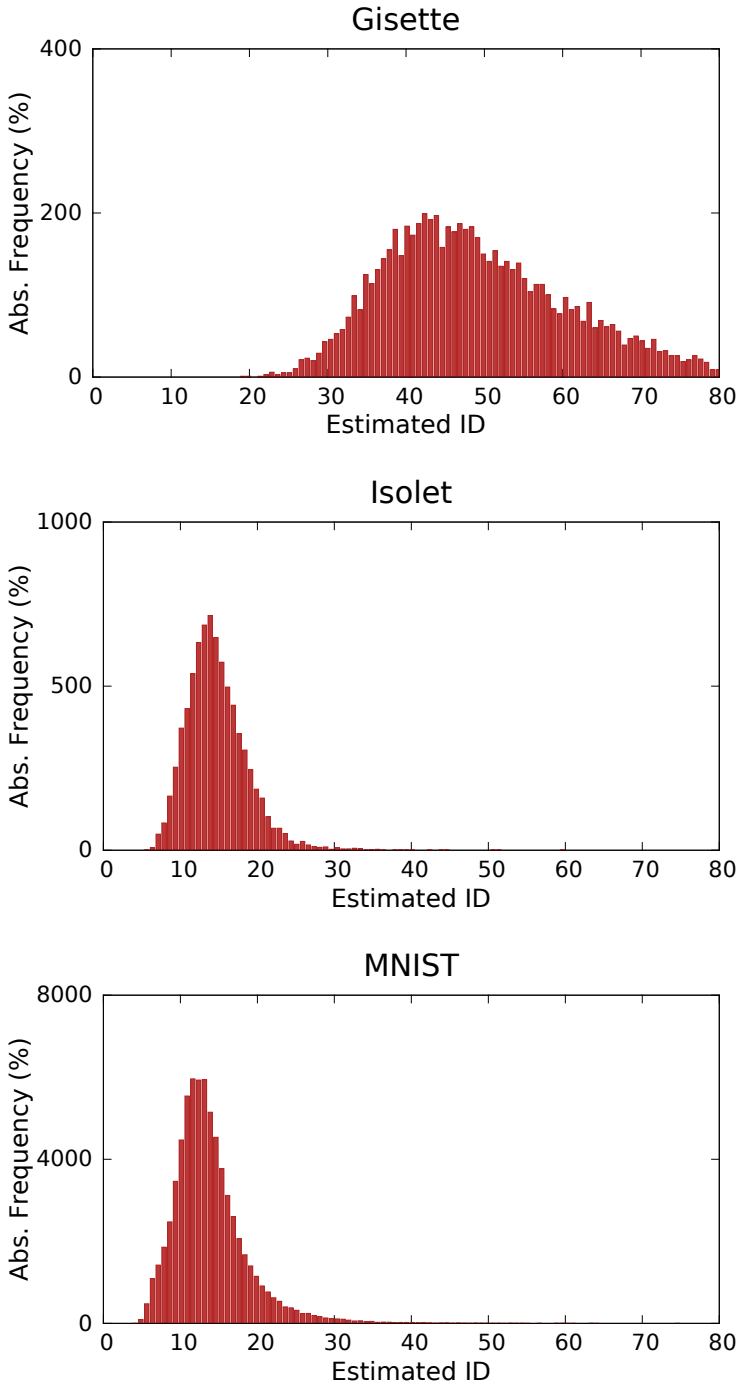


Figure 5.7: Histograms of LID values across each dataset, obtained using the MLE estimator on the size-100 neighborhoods of the individual reference points.

dataset was used as the reference set. We observe clear differences in the distribution of LID values among the datasets; for example, the center and spread of the LID estimates for ALOI are considerably lower than those obtained for the other datasets, whereas the LID estimates for Gisette are clearly higher. More precisely, we observe mean values of $\mu_{\text{ALOI}} = 4.4$, $\mu_{\text{ANN\_SIFT1B}} = 12.3$, and $\mu_{\text{Gisette}} = 49.4$. with the corresponding standard deviations of $\sigma_{\text{ALOI}} = 3.5$, $\sigma_{\text{ANN\_SIFT1B}} = 3.0$, and $\sigma_{\text{Gisette}} = 12.4$. It should be noted that the measured ID within the neighborhoods that were tested is far smaller than the dimension of the full feature spaces. By plotting the same data as histograms in Figure 6.7, we can furthermore see that the individual distributions of LID values differ in kurtosis and skewness as well.

Figure 6.7 shows that the LID estimates for the Gisette dataset are very high compared to those of the other 5 sets. In particular, they are much higher than the LID values for MNIST, the original data set from which Gisette was constructed. It is clear from the LID histograms that the addition of artificial noise features in Gisette drastically inflates the LID values in the dataset, revealing that the generation mechanism underlying these noise features is very different from that of real-world datasets. Although this generation mechanism was not revealed by the creators of Gisette, local intrinsic dimension — as a measure of the subspace-filling capacity of the data — is capable of differentiating between artificial noise and natural noise.

For the ANN_SIFT1B dataset, from among the points of interest highlighted in the scatter plot in Figure 5.4 and 5.5, A, B and C correspond to the objects for which the three lowest LID values have been estimated ($\text{ID}_A \approx 2.8$, $\text{ID}_B \approx 3.1$, and $\text{ID}_C \approx 2.4$). Likewise, the objects corresponding to D, E and F achieved the three greatest ID values at $\text{ID}_D \approx 31.5$, $\text{ID}_E \approx 30.1$, and $\text{ID}_F \approx 25.7$. The object G has been chosen as its associated dimensionality estimate ($\text{ID}_G \approx 12.3$) is closest to the mean. We subsequently investigated the distribution of distances in the neighborhoods of these points so as to gain a better understanding of why the corresponding dimensionality estimates take such low, high, or average values.

The most striking difference between the individual points of interest are the distances to their respective k-nearest neighbors. Figure 5.3a displays for each point of interest the specific distribution of neighbor-distances for all values of k between 1 and 1000. Interestingly, the ID measured at the points of interest appears to be associated with other properties of the respective objects. For example, distribution of neighbor-distances for objects with high corresponding dimensionality (D, E and F) indicate that these points are in some sense outliers. On the other hand, despite

Dataset	d	D	r=.5	r=.6	r=.7	r=.8	r=.9
m1	10	11	7.33	7.49	7.59	7.67	7.75
m2	3	5	2.56	2.59	2.62	2.64	2.66
m3	4	6	3.35	3.42	3.47	3.50	3.53
m4	4	8	4.87	4.86	4.85	4.83	4.79
m5	2	3	2.10	2.04	2.01	2.00	1.99
m6	6	36	6.84	6.93	7.00	7.02	7.06
m7	2	3	2.61	2.58	2.56	2.55	2.52
m8	12	72	11.28	11.60	11.85	12.01	12.16
m9	20	20	11.29	11.61	11.87	12.08	12.22
m10a	10	11	6.98	7.10	7.20	7.28	7.34
m10b	17	18	10.15	10.43	10.64	10.80	10.95
m10c	24	25	12.73	13.12	13.44	13.65	13.88
m11	2	3	2.14	2.33	2.44	2.49	2.50
m12	20	20	11.10	11.52	11.84	12.08	12.31
m13	1	13	1.92	1.74	1.62	1.51	1.42

Table 5.8: Average ID (MLE) estimates for 1000-point manifolds using 100 approximate nearest neighbors with controlled recall.

their distance distributions being quite dissimilar, the LID values measured at A, B, and C are nearly identical.

5.3.4 Approximate Nearest Neighbors

This set of experiments shows that using approximate neighbors reduces the overall computation time of LID at the cost of an increase in bias. In an approximate k-NN query result, only a certain proportion of the observed distance values (equal to the accuracy of the result) correspond to distances values associated with members of an exact k-NN result. The distances associated with the approximate result can be regarded as having been generated by first sampling the dataset, and taking the distance values associated with the exact k-NN set with respect to the sample. The bias of the LID estimates for the approximate neighborhood can therefore be regarded as the result of a sparsification of the available distance information.

The results presented in tables 5.8 and 5.9 show that using distances drawn from approximate neighborhoods does not lead to significant changes in estimated LID values, provided that the accuracy of the neighborhoods is reasonably high. In fact, for the datasets studied, the change in estimated LID values did not exceed 18% of

Dataset	d	D	r=.5	r=.6	r=.7	r=.8	r=.9
m1	10	11	8.81	8.90	8.97	8.97	9.02
m2	3	5	2.82	2.85	2.87	2.86	2.87
m3	4	6	3.79	3.83	3.85	3.84	3.86
m4	4	8	4.19	4.16	4.14	4.09	4.08
m5	2	3	1.97	1.98	1.98	1.98	1.98
m6	6	36	6.85	6.82	6.78	6.71	6.68
m7	2	3	1.95	1.96	1.96	1.96	1.96
m8	12	72	13.47	13.60	13.68	13.66	13.71
m9	20	20	13.97	14.16	14.30	14.32	14.41
m10a	10	11	8.00	8.08	8.15	8.14	8.18
m10b	17	18	12.32	12.47	12.59	12.60	12.67
m10c	24	25	16.02	16.26	16.43	16.47	16.59
m11	2	3	2.02	2.01	2.01	2.00	2.00
m12	20	20	14.72	14.98	15.18	15.24	15.37
m13	1	13	1.03	1.02	1.02	1.01	1.01

Table 5.9: Average ID (MLE) estimates for 10000-point manifolds using 100 approximate nearest neighbors with controlled recall.

the ground truth intrinsic dimension in the worst case, even with a neighborhood accuracy of 50%.

We observe that for each of the datasets, the observed bias is inversely proportional to the neighborhood accuracy: a higher accuracy always corresponds to a lower bias, although the relationship is not linear. We also observe that the sign of the bias depends on the curvature of the underlying manifolds within which the datasets are distributed. This trend is clear even when only 1000 points were generated within the manifolds (see Table 5.8). The bias is positive for the non-convex sets (m4, m5 , m7, and m13). For these sets of high curvature, distance sparsification has a proportionally greater effect on the smaller distances, as compared to when the manifolds are linear. When the loss of instances of smaller distance values is higher than for larger distance values, the estimates of LID would be expected to rise.

It is important to note that estimation over neighborhoods of size 100 within a dataset of size 1000 is not in line with the asymptotic assumptions of EVT, since the neighborhood here can hardly be viewed as being derived from an extreme lower tail of the underlying distribution. However, estimation over neighborhoods of size 100

Dataset	Exact NN	NN-Descent		
	Time (s)	Time (s)	Time prop.	Accuracy
ALOI	85168	2558	0.030	0.999968
ANN_SIFT1B	2020520	13305	0.007	0.945113
BCI5	1466	209	0.143	0.999995
Isolet	2523	590	0.234	1.000000
Gisette	230	111	0.486	0.999999
MNIST	50211	2943	0.059	0.999960

Table 5.10: Effect of using NN-Descent

within a dataset of size 10,000 would be expected to lead to more stable results, due to the much smaller ratio of the neighborhood set size to the full dataset size. This is borne out by the experimental results shown in Table 5.8, where it can be seen that the approximation of neighborhood distance values has very little effect on the quality of ID estimation.

For the artificial datasets, as a representative ANN method, NN-Descent achieves extremely high accuracies while achieving useful speedups over sequential search (especially for the larger datasets). As seen in Figure 5.8 and 5.9, average accuracies range between 99.9982% and 100%, while average execution costs range between 3 and 8 times faster than exact k-NN computation time for sets of 10000 points, and between 15 and 41 times faster than exact k-NN computation time for sets of 100000 points. Under these conditions, the LID estimates for all artificial datasets included in this experiment remain unchanged. For the datasets of size 1000 or less, the execution cost of NN-Descent is dominated by the overheads associated with the underlying data structures. However, as shown in Figure 5.9 for datasets of 100000 points, the benefit of estimating LID with approximate neighborhoods quickly becomes apparent as the dataset size rises.

For the real-world datasets, NN-Descent achieves very high accuracies as well, while achieving important speedups over exact nearest neighbor computation. Average accuracies in all cases were at least 94.5%, as can be seen from Table 5.10. On the small datasets, NN-Descent accelerates the computation of nearest neighbors by no more than a factor of 2. For these small datasets, the time gain is limited by the overheads in maintaining the data structures required for NN-Descent. On the large datasets of this study, approximate nearest neighbors are obtained in up to 151 times faster than exact nearest neighbors. Due to the high accuracy of

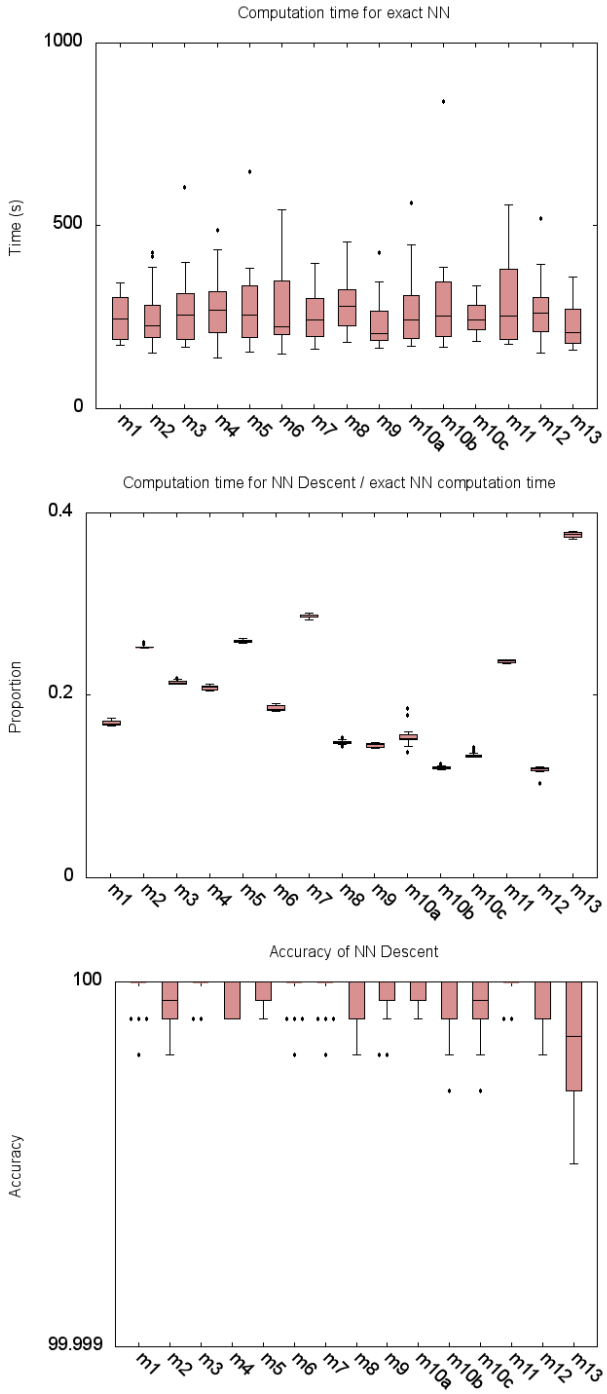


Figure 5.8: Execution time and accuracy of NN-Descent compared with exact nearest neighbors' computation for 20 runs on the 10000-point datasets.

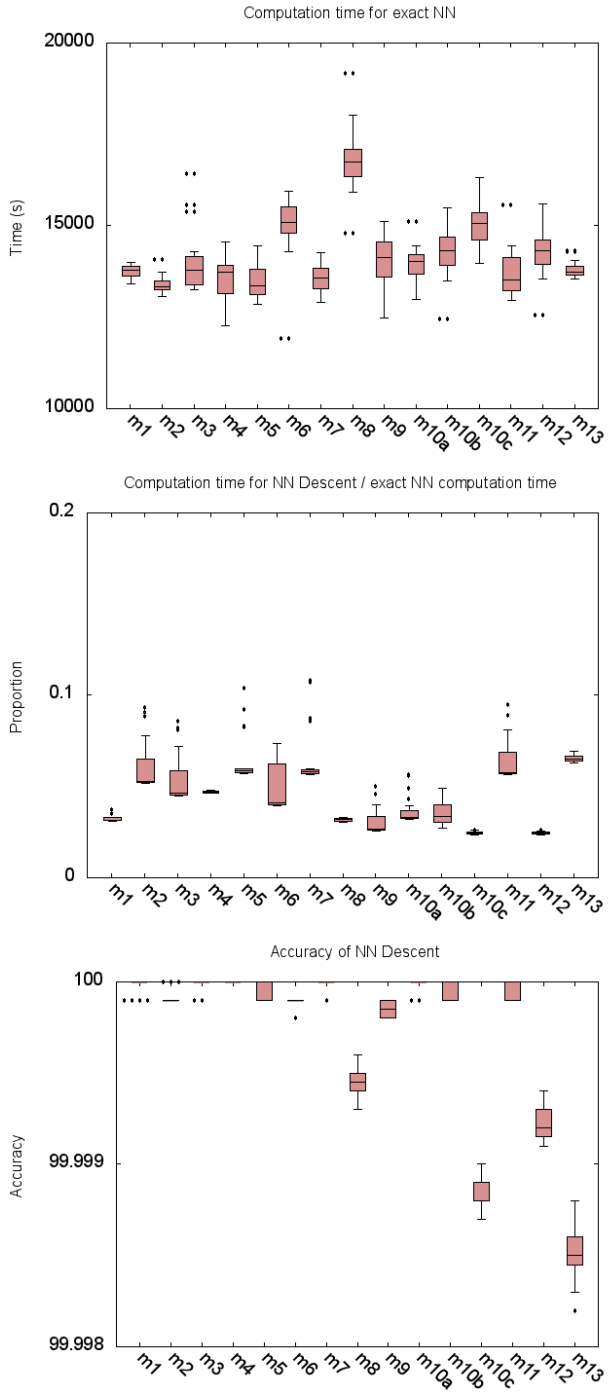


Figure 5.9: Execution time and accuracy of NN-Descent compared with exact nearest neighbors' computation for 20 runs on the 100000-point datasets.

neighborhoods, LID estimates remain essentially unchanged for all datasets except for ANN_SIFT1B, where they deviate by only -1.82% from their original values. For most machine learning applications, such small changes in LID values would likely have little or no impact on the usefulness of the estimates.

Through these experiments, we can conclude that the use of approximate nearest neighbor computation allows LID estimation to be effectively applied at large scales. LID estimation can therefore be a viable option even for those machine learning and data mining applications where scalability is an important issue.

5.4 Discussion

Our experimental results on synthetic data show that for all of the estimators of LID that we propose, the estimation stabilizes for sample sizes on the order of 100. However, for Theorem 2 to be applicable, one must set a sufficiently small threshold on the lower tail of the distribution, which may severely limit the number of samples falling within the tail. Although there is a conflict between the accuracy of the estimator and the validity of the model, this conflict is resolved as the size of the dataset scales upward; it is in precisely such situations where the applications of ID have the most impact.

For situations where exact neighborhood information is impractical to compute, our experimental results show that LID estimation is effective even when only approximate neighborhood information is available. Consequently, learning machines that exploit LID values need not suffer from the high computational cost associated with the computation of exact neighborhoods.

Estimates of local ID constitute a measure of the complexity of data. Along with other indicators such as contrast [SR06], LID could give researchers and practitioners more insight into the nature of their data, and therefore help them improve the efficiency and efficacy of their applications. As a tool for guiding learning processes, the proposed estimators could serve in many ways. Data collected during the retrieval processes could be automatically filtered out as noise, whenever they are associated with an unusually high ID value. In this way, the quality of query results may be enhanced as well.

The performance of content-based retrieval systems is usually assessed in terms of the precision and recall of queries on a ground truth dataset. However, in high-dimensional settings it is often the case that some points are much less likely to ap-

pear in a query result than others. Unlike LID, conventional measures of complexity or performance do not account for this difficulty. LID has therefore the potential to aid in the design of fair benchmarks that truly reflect the power of retrieval systems, according to a sound, mathematically-grounded procedure.

Chapter 6

Estimating LID Using Auxiliary Distances

A global estimator can be adapted for local estimation of ID simply by applying it to the subset of the data lying within some region surrounding a point of interest. Global methods typically make use of many (if not most or all) of the pairwise relationships within the data; however, 'clipping' of the data set to a region, by discounting some of these relationships while preserving others, may lead to estimation bias whenever the boundary shape is not properly accounted for in the ID model or estimation strategy. On the other hand, implicit in their design, local estimators of ID avoid the negative affect of clipping, by considering only the direct relationships between a reference point and its nearest neighbors. The sample boundary is usually set to the distance from the reference point to the farthest object in the neighborhood. With this distinction in mind, application of global estimators within the neighborhood of a given reference point should not be regarded as truly 'local'.

Local estimators of ID can potentially have significant impact when used in subspace outlier detection, subspace clustering, or other applications in which the intrinsic dimensionality is assumed to vary from location to location. However, in practical settings, the natural groups within the data are often too small to provide the

number of samples necessary for accurate estimation of ID — in the LID framework, for example, approximately one hundred distance values are usually required for convergence [ACF+15]. Simply choosing a number of samples sufficient for the convergence of the estimator can lead to a violation of the locality constraint, as the sample could consist of points from several different natural groups, each with their own intrinsic dimensionalities. When the cluster memberships and size are not known in advance, in order to ensure that the majority of the points are drawn from the same group, it is necessary to use estimators that can cope with the smallest possible sample sizes [ACF+15, RLC+12]. Thus, the development of local ID estimators with faster convergence properties is essential for the effectiveness and the efficiency of subspace-based applications.

One possible strategy for improving the convergence properties of estimation without violating locality is to draw more measurements from smaller data samples — however, for the case of distance-based local estimation from neighborhood samples, this would require the use of distances between pairs of neighbors, and not merely the distances from the reference point to its neighbors. Indeed, the global distance-based correlation dimension (CD) [Tak85], if restricted to a neighborhood, would use all pairwise distances within the neighborhood to achieve its estimate. Although for a given neighborhood size this local use of CD would be expected to converge much faster than true local ID estimators, the result would be biased due to the clipping.

In this Chapter, we show that the convergence properties of LID estimation can be improved by augmenting it with distance measurements from members of the neighbor set to their own nearest neighbors. The sizes of these 'auxiliary' neighborhoods is restricted so that they are completely contained within the original, 'primary' neighborhood, thus preserving the locality of the estimation. Within a given primary neighborhood of k elements, the number of distance measurements thus could range between a minimum of k and a maximum of $k(k+1)/2$. We show that under certain assumptions, the number of measurements available depends on the local ID itself, with the greatest number of auxiliary distance measurements being available when the ID is small. The main contributions of this Chapter include:

- the augmented local ID estimator, ALID;
- for the case of uniform data distributions in Euclidean space, a theoretical analysis of the expected number of auxiliary distances available in terms of

ID;

- an experimental comparison of the bias, variation, and convergence properties of ALID with LID and other local and global estimators of ID, on both synthetic and real data sets.

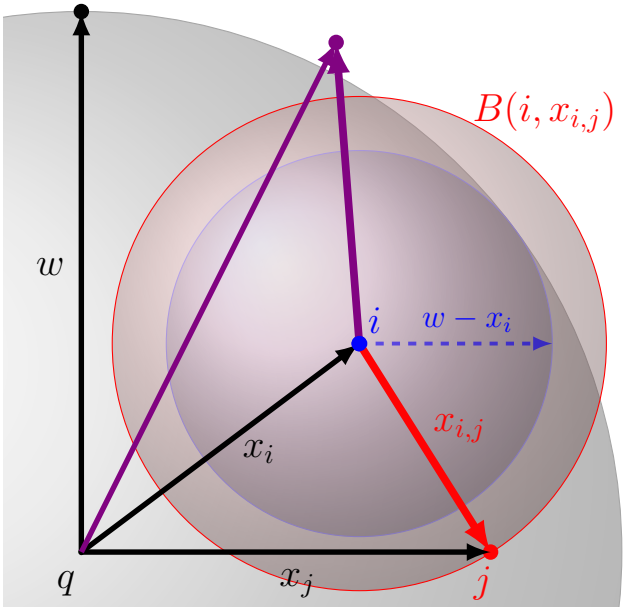
The remainder of the Chapter is structured as follows. In Section 6.1 we introduce our proposed estimator, and present a theoretical analysis relating the expected number of auxiliary distances with the intrinsic dimensionality. In Section 7.3 our experimental framework is described in detail, and in Section 7.4 we present our experimental comparison of ALID with existing local and global ID estimators. In this latter section we also validate our theoretical analysis empirically, by showing the number of auxiliary measurements available and comparing them to the numbers predicted by the theory. We conclude the with a short discussion.

6.1 Augmented Local ID Estimation

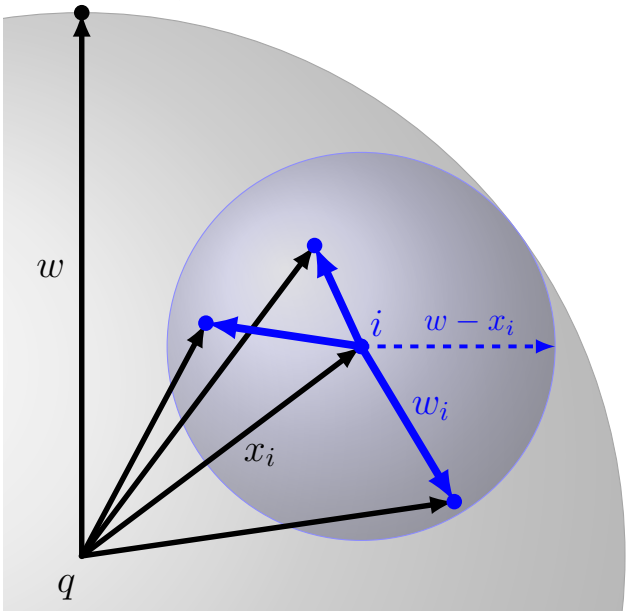
6.1.1 MLE estimation for ALID.

Global estimators based on correlation dimension use the smallest pairwise distances within the data in order to measure the global ID. In particular, the Takens estimator [Tak85] uses all pairwise distances within balls of a fixed radius and evaluates ID using the same Hill estimator. With this approach, intracluster distances are likely to dominate intercluster distances that may occur whenever the radius is too high.

Restricting the computation of correlation dimension to a neighborhood is not a satisfactory estimation strategy for local ID. Consider a query point q and its neighborhood $B(q, w)$ of radius w, and a neighbor i at distance $x_i \leq w$ from q. If the distance $x_{i,j}$ from i to a neighbor j is such that $x_i + x_{i,j} > w$, then the ball $B(i, x_{i,j})$ centered at i with radius $x_{i,j}$ (shown in red in Figure 6.1a) would not be completely contained within $B(q, w)$. Restricting (or 'clipping') the estimation to points located both in $B(i, x_{i,j})$ and $B(q, w)$ would result in an estimation error: points located inside $B(i, x_{i,j})$ but outside the original neighborhood would not be accounted for by the dimensionality estimator, and would thereby induce a bias. Alternatively, estimating over all points of $B(i, x_{i,j})$ would require points from outside the neighborhood, which would violate the locality assumption.



(a) Neither auxiliary nor direct distances (in purple) to neighbors that are outside the locality can be used. Moreover, auxiliary distances where the corresponding ball (in red) crosses over the original locality can not be used for the estimation.



(b) Pairwise distances that remain within internally tangent balls can be used in the ID estimation without introducing distortions. In this figure we consider only one nearest neighbor (i) and the corresponding usable auxiliary distances (in blue).

Figure 6.1: State-of-the-art local ID estimators use only direct distances (in black). The proposed estimator $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ uses additional distances between pairs of neighbors. Some of these distances (in blue) can be used, while others (in purple and red) cannot.

To avoid the negative effects of clipping, ALID makes use of an auxiliary distance measurement $x_{i,j}$ from i only if the ball of radius $x_{i,j}$ centered at i is entirely contained within the original neighborhood (as shown in Figure 6.1b). This condition can be stated as $x_{i,j} \leq w - x_i$.

The proposed auxiliary-distance estimator ($\widehat{\mathrm{ID}}_{\mathrm{ALID}}$) can be regarded as an aggregation of Hill estimates ($\widehat{\mathrm{ID}}_{\mathrm{MLE}}$) calculated at q as well as its neighbors located within distance w. We implicitly assume that the $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ estimates at these neighbors converge to the ID of q, as w tends to zero.

Let X_i be the random distance variable from the neighbor i in the range $[0, w - x_i)$, and let $f_{X_i,w-x_i}$ and $F_{X_i,w-x_i}$ be respectively the pdf and cdf associated with X_i. To simplify the notation, we assign the rank $i = 0$ to the test point. The log-likelihood function is:

$$\begin{aligned}
\mathcal{L}(\mathrm{ID_X}) &= \ln\left[\prod_{\substack{x_{i,j}+x_i<w \\ i,j\in[0,k]}} f_{X_i,w-x_i}(x_{i,j})\right] \\
&= \ln\left[\prod_{\substack{x_{i,j}+x_i<w \\ i,j\in[0,k]}} \mathrm{ID_X}\frac{F_{X_i,w-x_i}(w-x_i)}{w-x_i}\right. \\
&\qquad \left.\cdot\left(\frac{x_{i,j}}{w-x_i}\right)^{\mathrm{ID_X}-1}\right] \\
&= (k+\rho(w))\cdot \mathrm{ID_X} \\
&\quad + (\mathrm{ID_X}-1)\sum_{\substack{x_{i,j}+x_i<w \\ i,j\in[0,k]}} \ln\left[\frac{x_{i,j}}{w-x_i}\right] \\
&\quad + \sum_{\substack{x_{i,j}+x_i<w \\ i,j\in[0,k]}} \ln\left[\frac{F_{X_i,w-x_i}(w-x_i)}{w-x_i}\right],
\end{aligned}$$

where $\rho(w) = \sum_{i,j\in[1,k]} \mathbb{1}[x_{i,j} + x_i < w]$ denotes the number of auxiliary distances used in the estimation. Accordingly, our auxiliary-distance MLE estimator is

$$\widehat{\mathrm{ID}}_{\mathrm{ALID}} = -\left(\frac{1}{k+\rho(w)}\sum_{\substack{x_{i,j}<w-x_i \\ i,j\in[0,k]}} \ln\left[\frac{x_{i,j}}{w-x_i}\right]\right)^{-1}.$$

A confidence interval can be obtained for $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ using a derivation similar to that of $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ in [ACF+15], with the number of distance samples being $\rho(w)$:

$$\left[\frac{\widehat{\mathrm{ID}}_{\mathrm{ALID}}}{1+\rho(w)^{-1/2}\Phi^{-1}(1-\frac{\beta}{2})}, \frac{\widehat{\mathrm{ID}}_{\mathrm{ALID}}}{1-\rho(w)^{-1/2}\Phi^{-1}(1-\frac{\beta}{2})}\right].$$

Here, Φ denotes the quantile function of the normal distribution.

The number of available auxiliary distance measurements $\rho(w)$ varies from data set to data set, and even from one locality within the set to another. However, under certain simplifying assumptions, it is possible to show that this quantity depends on the local intrinsic dimensionality. If the data distribution is locally uniform in the vicinity of the test point, the expected number of points within a volume would be proportional to the volume itself. Accordingly, the following theorem determines the cumulative volume of all maximal ball placements centered at locations within a neighborhood ball — or in other words, the cumulative volume of all internally tangent balls.

Theorem 6 In a Euclidean manifold of dimensionality α, let us consider a ball of radius w, and volume $V_\alpha(w)$. The total volume of all internally tangent balls is:

$$\rho_\alpha(w) = \frac{V_\alpha(w)^2}{2} \cdot \frac{\Gamma(\alpha)\Gamma(\alpha+1)}{\Gamma(2\alpha)}.$$

Proof (Sketch only) In order to measure the total volume of all internally tangent balls, it is possible to integrate the volumes of all balls of volume $V_\alpha(w-r)$ with centers located on the surface of a sphere of radius r, over values of $r \in [0, w]$. The total volume is given by

$$\rho_\alpha(w) = \int_0^w A_\alpha(r) \cdot V_\alpha(w-r)\, dr \tag{6.1}$$

$$= \int_0^w \frac{2\pi^{\alpha/2}}{\Gamma(\frac{\alpha}{2})} r^{\alpha-1} \cdot \frac{\pi^{\alpha/2}}{\Gamma(\frac{\alpha}{2}+1)} (w-r)^\alpha\, dr, \tag{6.2}$$

where $A_\alpha(r)$ is the surface area of a sphere of radius r in a manifold of intrinsic dimensionality α.

Using the variable changes $r = u + \frac{w}{2}$ and $u = \frac{w}{2}\sin\theta$, we show that:

$$\int_0^w r^{\alpha-1} \cdot (w-r)^{\alpha}\, dr = \frac{w^{2\alpha}}{2}\frac{\Gamma(\alpha)^2}{\Gamma(2\alpha)}.$$

Substituting into Equation 6.2, it follows that:

$$\rho_\alpha(w) = \frac{V_\alpha(w)^2}{2} \cdot \frac{\Gamma(\alpha)\Gamma(\alpha+1)}{\Gamma(2\alpha)}.$$

6.1.2 Complexity of the ALID estimator.

Under the assumption that the expected number of points in a volume is proportional to the volume itself, Theorem 6 implies that the expected number of distances $\rho(w)$ in a neighborhood of radius $w = x_k$ is $\frac{k^2}{2}\frac{\Gamma(\mathrm{ID})\Gamma(\mathrm{ID}+1)}{\Gamma(2\mathrm{ID})}$.

Unlike most local estimators of ID, the complexity of the auxiliary-distance estimator depends on the ID itself. When the assumptions of Theorem 6 apply, we can infer that $C_{\widehat{\mathrm{ID}}_{\mathrm{ALID}}} = O(k \cdot (1 + k\frac{\Gamma(\mathrm{ID})\Gamma(\mathrm{ID}+1)}{\Gamma(2\mathrm{ID})}))$. Thus, the complexity is linear when the estimated ID is high, matching the complexity of $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ and $\widehat{\mathrm{ID}}_{\mathrm{MoM}}$. When the estimated ID is low, $C_{\widehat{\mathrm{ID}}_{\mathrm{ALID}}}$ becomes quadratic in the number of neighbors, like $\widehat{\mathrm{ID}}_{\mathrm{GED}}$ or the Levina & Bickel estimator.

6.2 Experimental framework

Method	Parameters
$\widehat{\mathrm{ID}}_{\mathrm{ALID}}$	$k = 100$
$\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ [ACF+15]	$k = 100$
$\widehat{\mathrm{ID}}_{\mathrm{MoM}}$ [ACF+15]	$k = 100$
kNNG [CHI04]	$k = 100$, $\gamma = 1$, $M = 1$, $N = 10$
l-PCA [Jol86]	$k = 100$, $\theta = 0.025$
MiND_{ml1} [RLC+12]	None
MiND_{mli} [RLC+12]	$k = 100$
PCA [Jol86]	$\theta = 0.025$

Table 6.1: Parameter choices for the methods used in the experiments.

Manifold	d	D	Description
h-d	d	d	Uniformly sampled hypercube.
m1	10	11	Uniformly sampled sphere.
m2	3	5	Affine space.
m3	4	6	Concentrated figure confusable with a 3d one.
m4	4	8	Non-linear manifold.
m5	2	3	2-d Helix
m6	6	36	Non-linear manifold.
m7	2	3	Swiss-Roll.
m8	12	72	Non-linear manifold.
m9	20	20	Affine space.
m10a	10	11	Uniformly sampled hypercube.
m10b	17	18	Uniformly sampled hypercube.
m10c	24	25	Uniformly sampled hypercube.
m11	2	3	Möbius band 10-times twisted.
m12	20	20	Isotropic multivariate Gaussian.
m13	1	13	Curve.

Table 6.2: Artificial datasets used in the experiments.

6.2.1 Competing estimation methods.

In this framework, to show the advantages and limitations of ALID, we compared our proposed estimator $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ with other popular estimators, both local and global. The fractal methods used in our experiments (Grassberger-Procaccia's Correlation Dimension (CD) [GP04], Hein [HA05], and Takens [Tak85]) do not require any parameters to be set, while the parameter choices for the remaining methods are summarized in Table 6.1. We denote by l-PCA the estimator obtained by applying PCA on the respective neighborhoods of size $k = 100$.

It must be noted that PCA variants and methods from the MiND family must be provided with knowledge of the representational dimension, which may give them an advantage in head-to-head comparison with other methods. Moreover, when applied to synthetic data sets, PCA variants and $\mathrm{MiND}_{\mathrm{mli}}$ can often return the exact dimension, since they can return only integer-valued estimates. While it may be claimed that the intrinsic dimension should ideally be an integer, for real data this is not always the case. For example, LID has been shown to be equivalent to a measure of the indiscriminability of the distance measure, which is in general not an

Dataset	Instances	Dim.	Classes
ALOI [BFF+01]	110250	641	1000
ANN_SIFT1B [JTDA11]	10^9	128	$3 \cdot 10^7$
BCI5 [Mil04]	31216	96	3
CoverType [BD99]	581012	54	7
Gisette [GGBHD04]	7000	5000	2
Isolet [CF90]	7797	617	26
MNIST [LBBH98]	70000	784	10
MSD [BMEWL11]	515345	90	90

Table 6.3: Real datasets used in the experiments.

integer [ACF+15]. Furthermore, non-integer values of ID can indicate non-linear properties of an underlying manifold, such as convexity.

6.2.2 Synthetic data.

Our study includes two families of synthetic datasets. For each manifold we generated 20 sets of 10^3 and 10^4 points, and in each experiment we report the average ID measures over the 20 sets. The first family (h) is a set of hypercubes meant to evaluate the convergence of local ID estimators. The second (m) is a benchmark of various types of manifolds [RLC+12, ACF+15].

6.2.3 Real data.

The use of real-world datasets lacks the ground truth available for synthetic data. Therefore, to evaluate our proposed estimator on such sets, we must compare the convergence, bias, and variance characteristics directly against competing methods. In particular, we test the consistency of $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ for the same suite of experiments provided for $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ in [ACF+15], using the 8 real datasets listed in Table 6.3.

- The ALOI (Amsterdam Library of Object Images) data consists of 110250 color photos of 1000 different objects. Photos are taken from varying angles under various illumination conditions. Each image is described by a 641-dimensional vector of color and texture features [BFF+01].

- The ANN_SIFT1B dataset consists of 10^9 128-dimensional SIFT descriptors randomly selected from the dataset ANN_SIFT which contains $2.8{\cdot}10^{10}$ SIFT

descriptors extracted from $3 \cdot 10^7$ images. These sets have been created for the evaluation of nearest-neighbor search strategies at very large scales [JTDA11].

- BCI5 [Mil04] is a brain-computer interface dataset in which the classes correspond to brain signal recordings taken while the subject contemplated one of three different actions (movement of the right hand, movement of the left hand, and the utterance of words beginning with the same letter).
- CoverType [BD99] consists of 581012 geographical locations (a surface of 30 by 30 meters) described by 54 attributes. each location is majorly covered by one of seven tree species.
- Gisette [GGBHD04] is a subset of the MNIST [LBBH98] handwritten digit image dataset, consisting of 50-by-50-pixel images of the highly confusable digits '4' and '9'. 2500 random features were artificially generated and added to the original 2500 features, so as to embed the data into a higher-dimensional feature space.
- Isolet [CF90] is a set of 7797 human voice recordings in which 150 subjects read each of the 26 letters of the alphabet twice. Each entry consists of 617 features representing utterances of the recording.
- The MNIST database [LBBH98] contains of 70000 recordings of handwritten digits. The images have been normalized and discretized to a 28×28-pixel grid. The gray-scale values of the resulting 784 pixels are used to form the feature vectors.
- MSD [BMEWL11] is a subset of the 'Million Song Database' which is a set of radio recordings (from the years 1922 to 2011) described by 12 timbre averages and 78 timbre covariances.

6.3 Results

6.3.1 Experiments with synthetic data.

We first examined the effect of clipping (discussed in Section 6.1.1) by introducing a variant of $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ that makes use of all distance pairs within the neighborhood of the query q. Over all sythetic datasets tested, estimates for the all-pairs variant

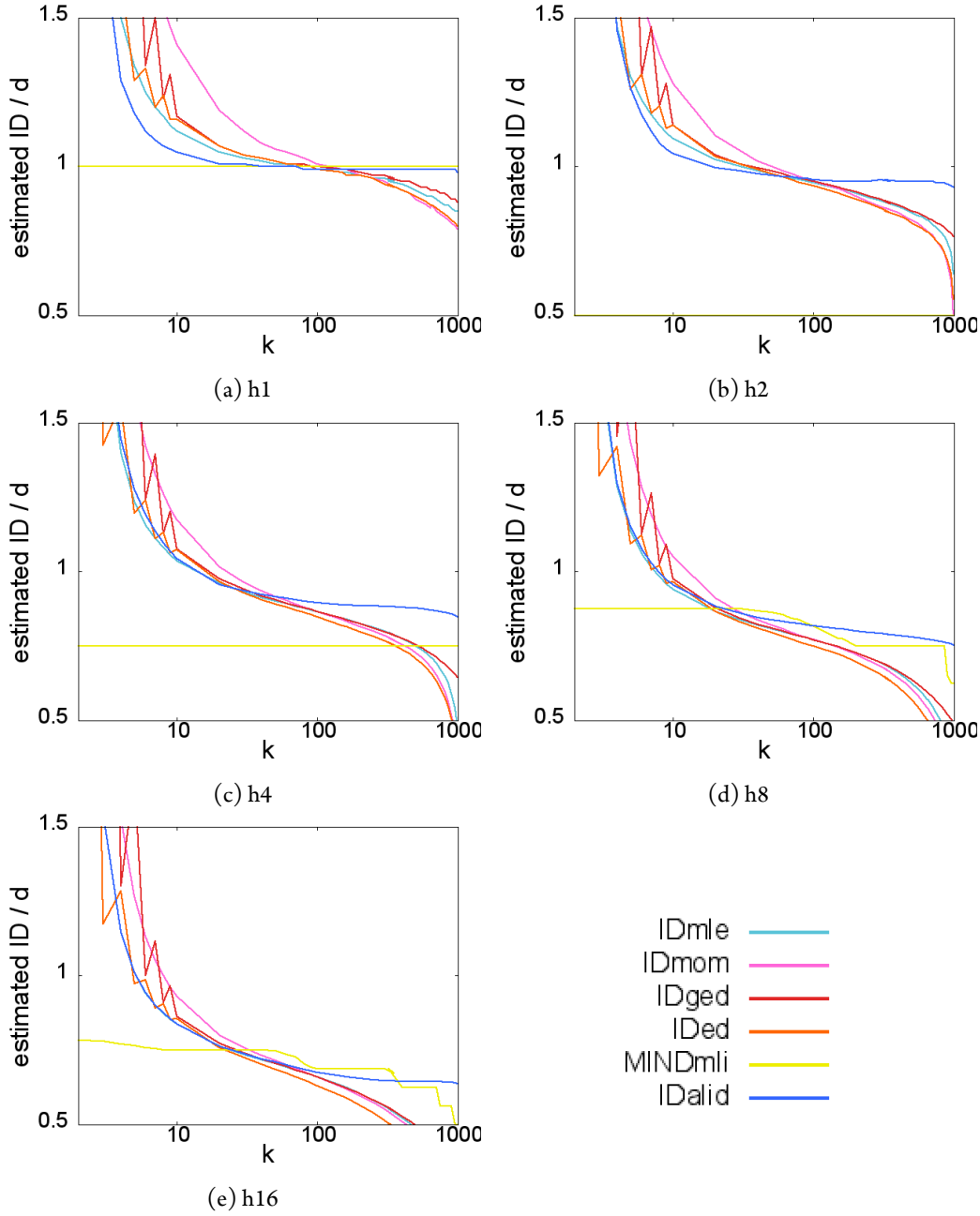
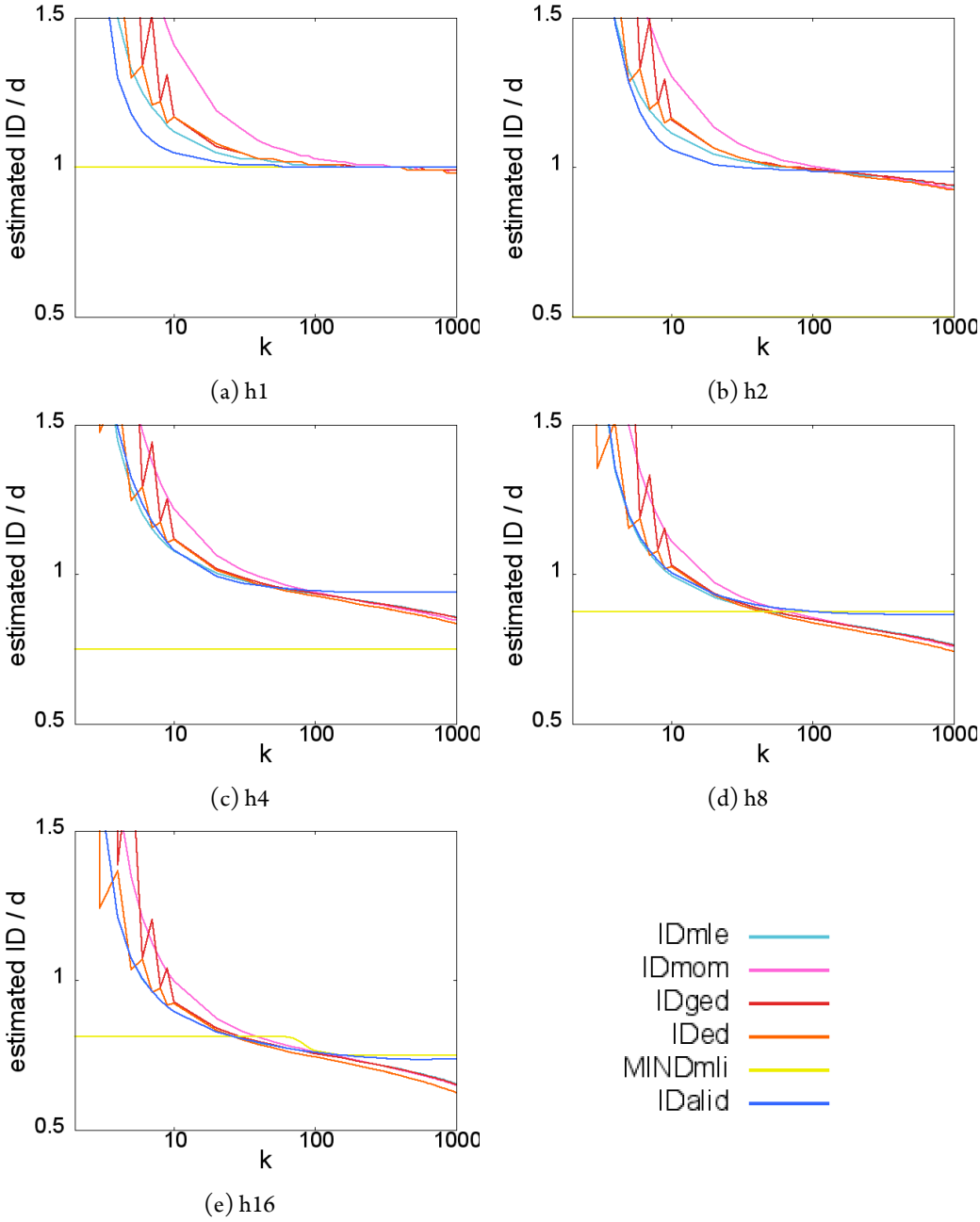


Figure 6.2: Convergence of local ID estimators in 1000-point-sets uniformly sampled from d-dimensional hypercubes.

$\widehat{\text{ID}}_{\text{all-pairs}}$ were averaged over 1000 queries using $k = 100$; overall, the ground truth dimension was underestimated by over 58%. As an example, on dataset m1 (where $d = 10$), the average $\widehat{\text{ID}}_{\text{all-pairs}}$ value was 3.25, while the average for $\widehat{\text{ID}}_{\text{ALID}}$ was 8.42. Therefore, for reliable LID estimation, the clipping effect must be accounted



(a) h1 (b) h2

(c) h4 (d) h8

(e) h16

Figure 6.3: Convergence of local ID estimators in 10000-point-sets uniformly sampled from d-dimensional hypercubes.

for. Hereafter, we report results only for the proposed ALID estimation.

In Figures 6.2 and 6.3, we show the convergence properties of the local ID estimators on two artificial data sets. As the neighborhood size k increases, $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ is the first estimator to stabilize. For the lower-dimensional manifolds, $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ requires

in the order of 100 neighbors to converge [ACF+15], whereas the many auxiliary distance measurements allow $\widehat{\text{ID}}_{\text{ALID}}$ to converge much faster — it requires fewer than 10 neighbors to draw within 10% of the true dimensionality. Meanwhile, as predicted by Theorem 6, as the dimensionality increases, the performance of $\widehat{\text{ID}}_{\text{ALID}}$ tends to that of $\widehat{\text{ID}}_{\text{MLE}}$.

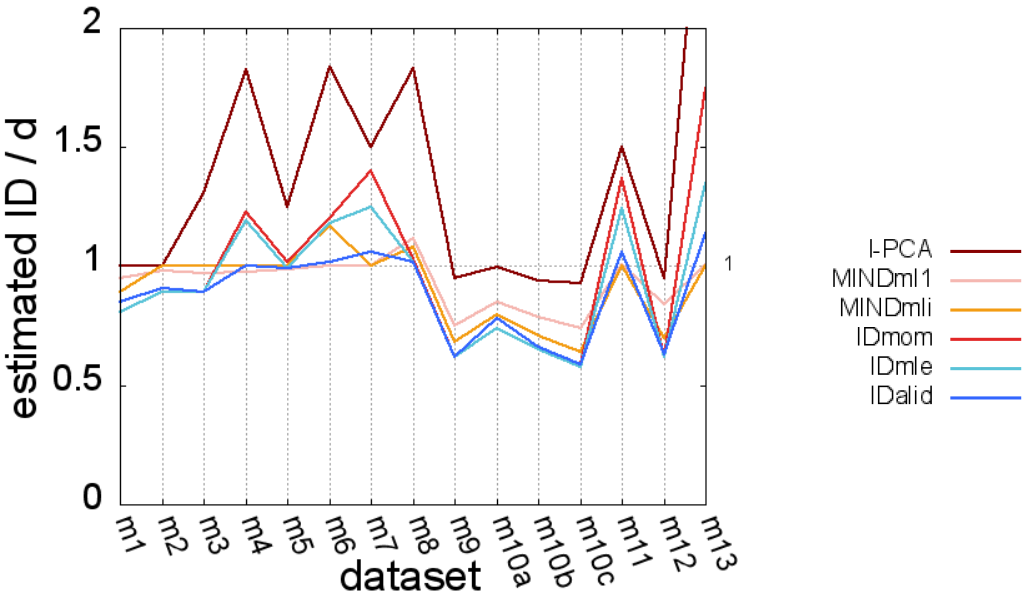
For the experiment shown in Figure 6.3, we evaluated the cumulative absolute error $e = \int_{k=2}^{k=1000} (\widehat{\text{ID}}/d)\, d \log k$ (the normalized difference between the estimate and the true ID value). For data set h1, $\widehat{\text{ID}}_{\text{ALID}}$ has the smallest error (8.78), with $\widehat{\text{ID}}_{\text{MLE}}$ coming in second (9.28). As the dimensionality increases, $\widehat{\text{ID}}_{\text{ALID}}$ converges to $\widehat{\text{ID}}_{\text{MLE}}$, since the proportion of auxiliary distances used tends to zero. This is reflected in the respective errors achieved for h4 $(7.50$ and $7.54)$ and h16 $(7.46$ and $7.54)$.

Overall, the results lead us to two conclusions: (i) our estimator converges faster than its competitors, and (ii) is among the least affected when the neighborhood size k is large.

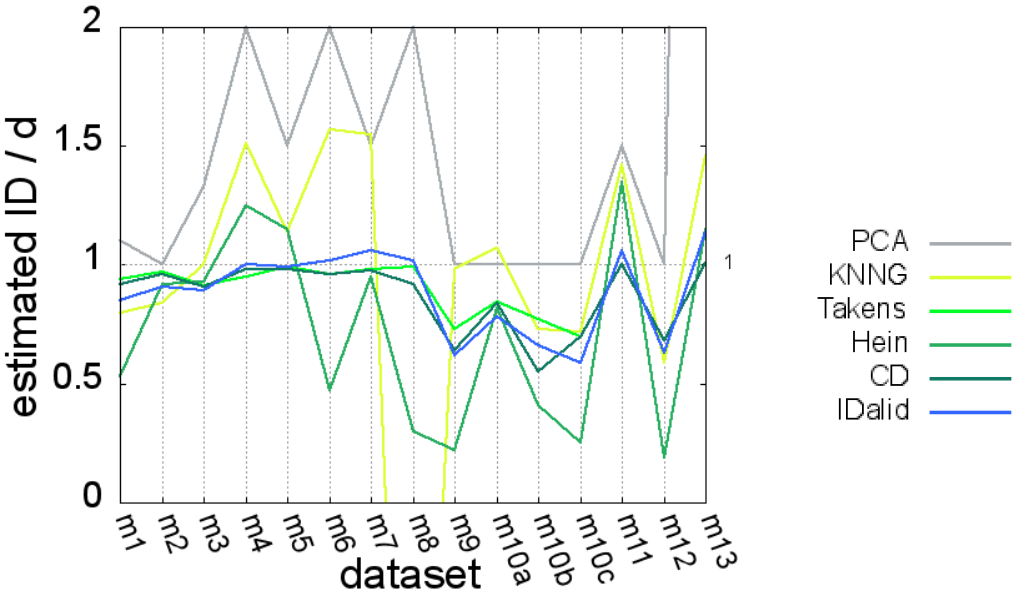
In the second experiment, we estimated the ID on various types of manifolds, with different dimensionalities as summarized in Table 6.3. Local estimators consistently underestimate the dimensionality on linear manifolds (m1, m2, m9, m10, and m12), due to clipping bias. However, local estimators tend to overestimate the dimensionality of nonconvex manifolds (m7, m11, and m13). In both cases, this bias is reduced as the sample sizes increase (Figure 6.5a). As shown in Figure 6.6, on nonlinear and nonconvex manifolds, $\widehat{\text{ID}}_{\text{ALID}}$ has the smallest bias and variance, with the exception of MiND_{mli} on linear manifolds, due to its advantage in having been provided the representational dimension.

In convex and linear manifolds, l-PCA appears to provide consistently accurate estimates with the least bias and variance. However, in real data where the manifolds are not convex, probabilistic local ID methods provide the best trade-off (c.f. Figure 6.6). When PCA is used locally, the variance along a given component coincides with the global variance when the manifold is linear and homogeneous. Whenever the manifold is nonlinear or nonconvex, the local components are very likely to be different from the global components.

Global estimators can be split into two groups based on the experimental results shown in figure 6.5b. Topological estimators (PCA) return the exact dimensionality only when the manifold is linear. However they tend to overestimate the ID on nonlinear manifolds, and perform poorly when the manifold is nonconvex. The



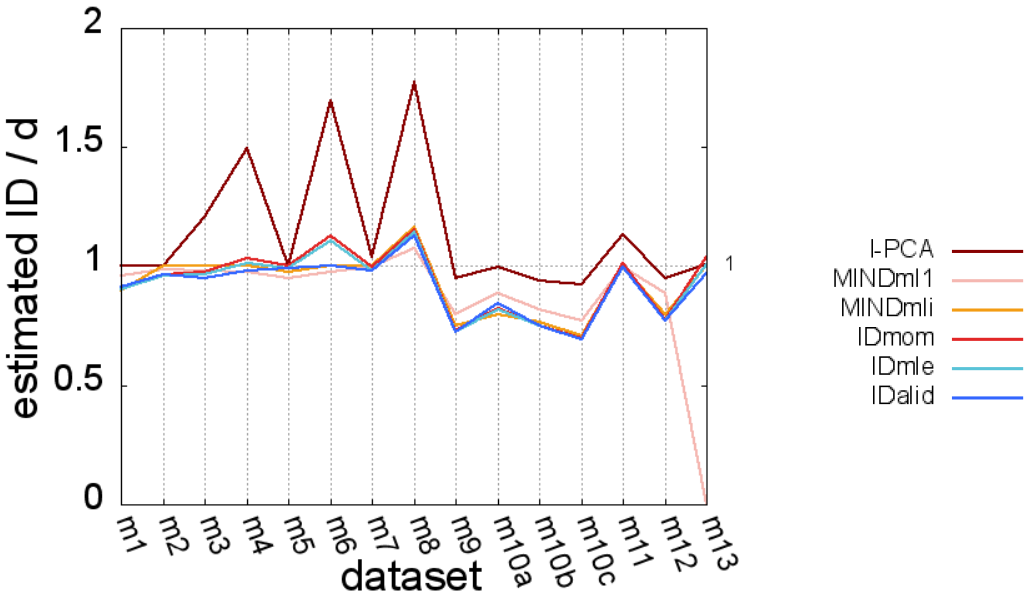
(a) Local ID estimators



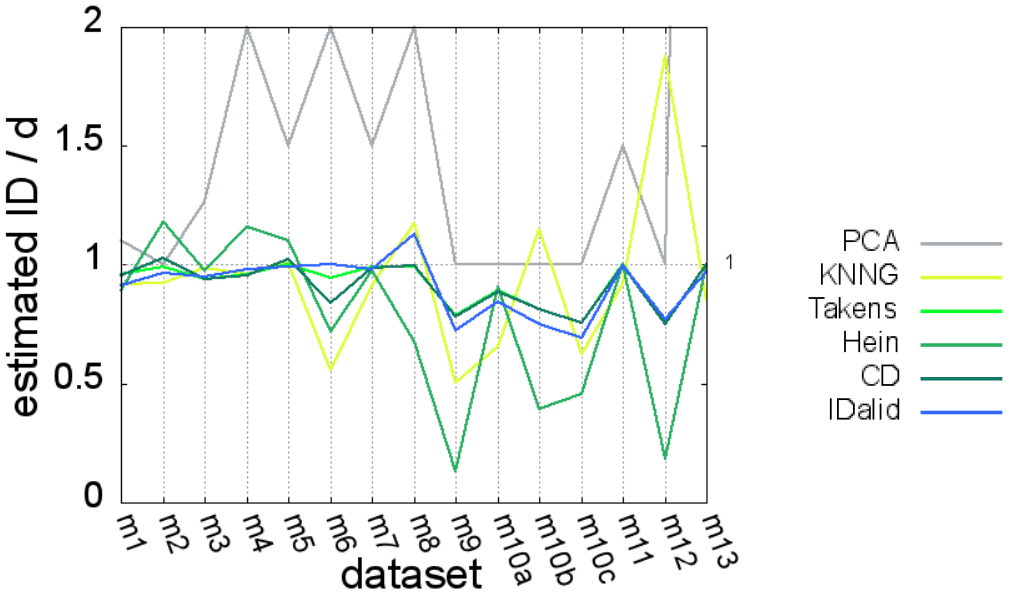
(b) Global ID estimators

Figure 6.4: Comparison of $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ with state-of-the-art ID estimators on 1000-point manifolds of various dimensionalities.

remaining global estimators tend to behave similarly to local estimators in their dependency on linearity and convexity, and on sample size.



(a) Local ID estimators



(b) Global ID estimators

Figure 6.5: Comparison of $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ with state-of-the-art ID estimators on 10000-point manifolds of various dimensionalities.

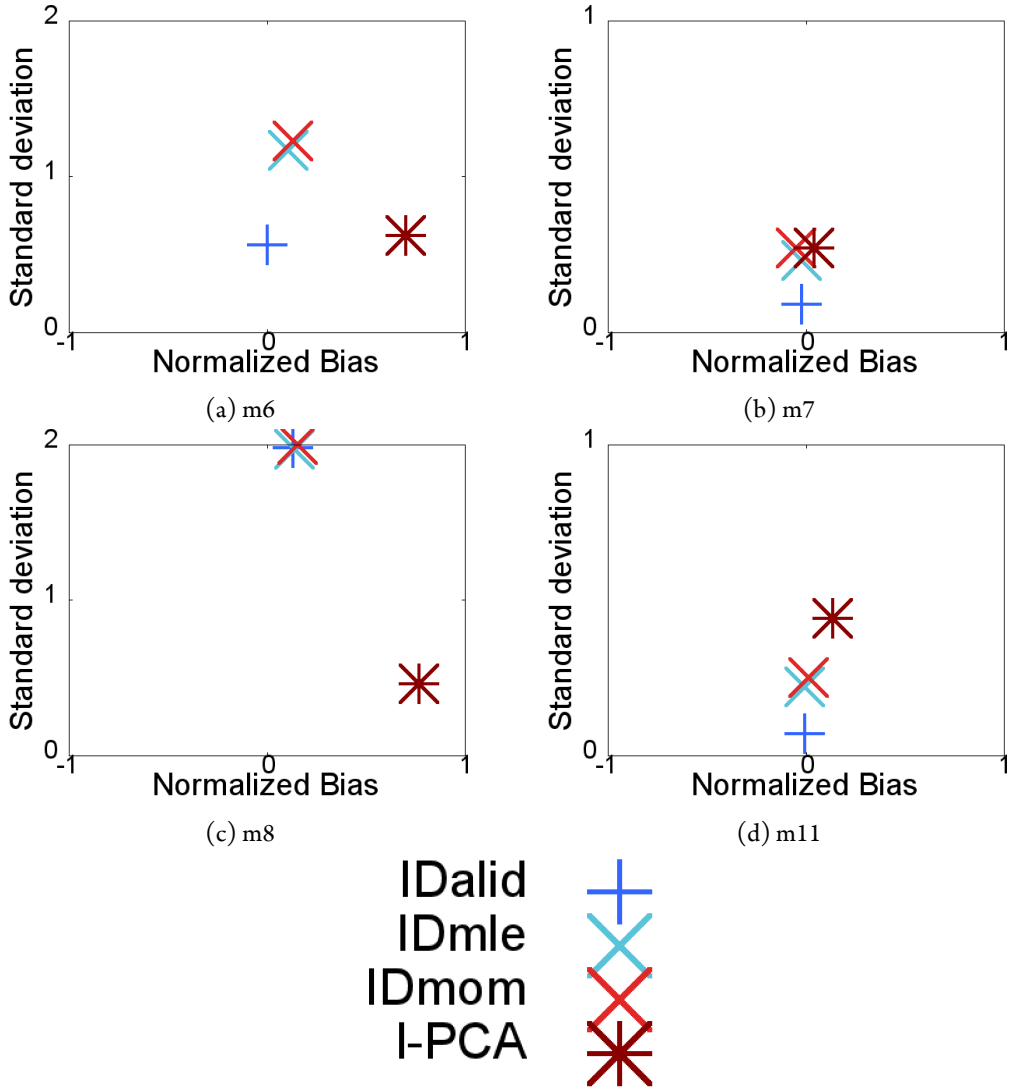
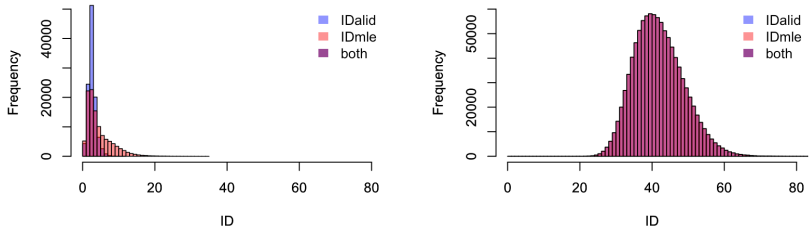


Figure 6.6: Bias and standard deviation of local ID estimators on nonconvex and nonlinear manifolds.

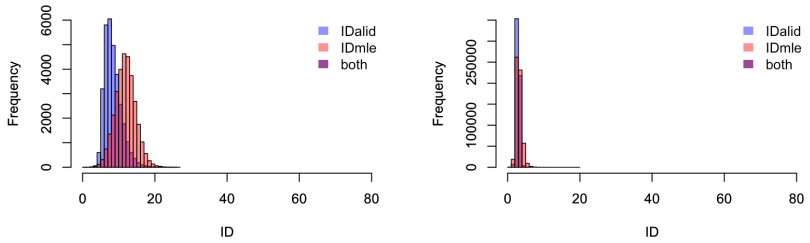
6.3.2 Experiments with real-world data.

As a first step, we evaluated ID on 8 publicly available datasets using $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ and $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ (see Figure 6.7). In all of the real-world datasets, the results are consistent with the theory, in that the estimates of $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ are much sharper than those of $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ when the ID is small, but tend to those of $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ as ID increases.

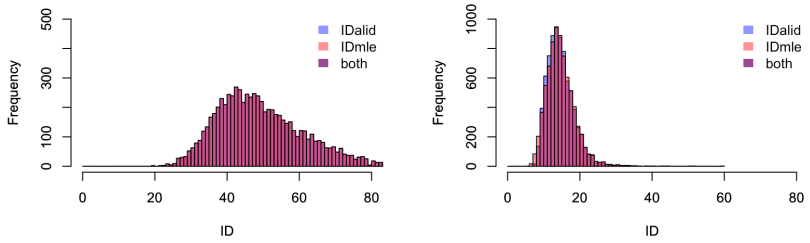
In a second experiment, we show the stability and robustness of $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ across various values of k, as compared to $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$. Figure 6.8 shows the ID estimates on



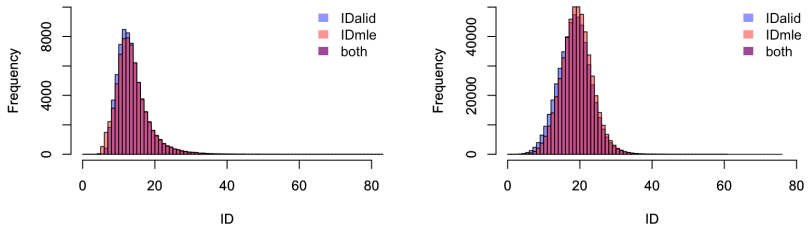
(a) ALOI (b) ANN_SIFT



(c) BCI5 (d) CoverType

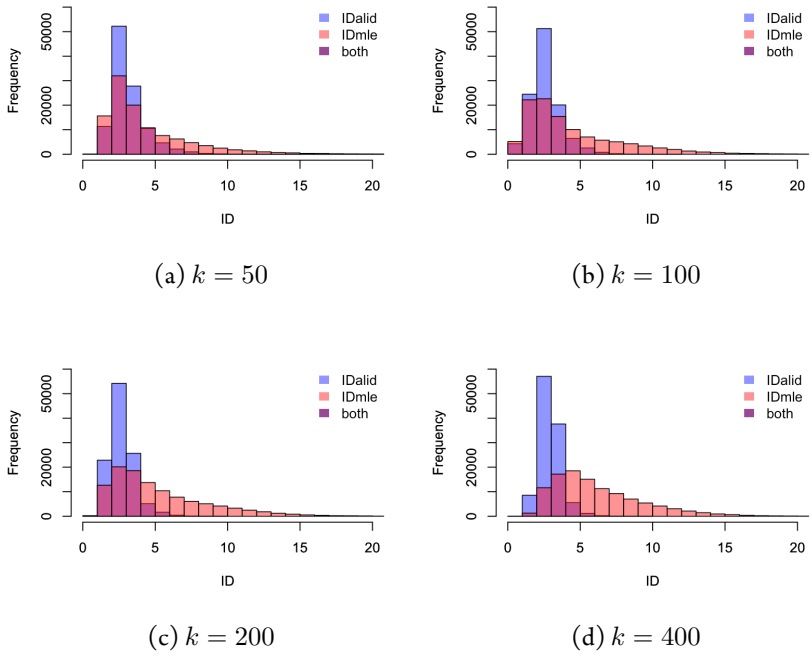


(e) Gisette (f) Isolet



(g) MNIST (h) MSD

Figure 6.7: Histograms of LID values across each dataset, obtained using the $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ and $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ estimators on the size-100 neighborhoods of the individual reference points.



(a) $k = 50$ (b) $k = 100$

(c) $k = 200$ (d) $k = 400$

Figure 6.8: Histograms of LID values across ALOI dataset, obtained using the $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ and $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ estimators on the size-100 neighborhoods of the individual reference points.

the ALOI data set, which consists of 1000 image classes of size approximately 110. The proportion of $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ estimates smaller than 4 consistently increases with k from 83% when $k = 50$ to 94% when $k = 400$. Meanwhile, $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ estimates in the range $[0, 4]$ decrease from 61% when $k = 50$ down to to 27% when $k = 400$. While 50 neighbors are probably not sufficient for the convergence of $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$, using more than 110 neighbors results in using points from outside the cluster. For example with 400 neighbors, distances to neighbors from at least 4 different clusters are used in the estimation process. $\widehat{\mathrm{ID}}_{\mathrm{MLE}}$ estimates use only direct distances that reflect the inter-cluster dimensional properties of the data, whereas $\widehat{\mathrm{ID}}_{\mathrm{ALID}}$ uses auxiliary distances as well which predominate at low ID to enhance the detection of the local dimensional properties of the data.

6.4 Discussion

In models such as the Correlation Dimension, pairwise distance measurements have been successfully used in order to estimate global intrinsic dimensionality. How-

ever, to the best of our knowledge, none of the existing models of local intrinsic dimensionality take advantage of distances other than those from a test point to the members of its neighborhood. Our proposed estimation strategy, ALID, makes use of a subset of the available intra-neighborhood distances to achieve faster convergence with fewer samples, and can thus be used on applications in which the data consists of many natural groups of small size. Moreover, it has a smaller bias and variance than state-of-the-art estimators, especially on nonlinear subspaces. Consequently, this estimator can achieve more accurate ID estimates within a smaller locality than the traditional estimators. This has the potential to improve the quality of algorithms where locality is an important factor, such as subspace clustering and subspace outlier detection.

Possible directions for future work include the development of a Method of Moments' estimator using auxiliary distances. Also, for cases where the neighborhood is very small, estimation can potentially be improved by using a limited number of points from outside the locality, at the cost of a slight increase in bias.

Part III

Applications

Chapter 7

Feature Selection Using Local ID

As the dimensionality of data increases, the efficiency and effectiveness of various learning algorithms tends to degrade. In this chapter, we propose new filter approaches for unsupervised feature selection whose selection criteria assess the ability of features to discriminate within the neighborhoods of data points, according to a recent model of the local intrinsic dimensionality of continuous distance distributions. By ranking and selecting those features which are most discriminative under the model, our method seeks to improve the overall local discriminability of the distance measure.

Advances in computer technology are opening the way for the handling of increasingly complex data. The scale of data can be measured with respect to volume (the number of data instances) and dimensionality (the number of features that describe these instances). When dealing with colossal data volume on high-capacity computing platforms, volume reduction techniques such as sampling [FL12, ZHMY13, KGKB03] and parallelization [PLC13, ZCC+08, CKO+06] can allow the exploration and analysis of data with millions or even billions of data entries.

Feature reduction, the dimensional analogue of volume reduction, has several

motivations [GE03]: reducing the costs of data collection, storage and processing; improving the model generalizability and interpretability; and improving the discriminative power of feature-based distance measures [BGRS99]. In general, high dimensionality is associated with a degradation in the efficiency and effectiveness of fundamental data mining and machine learning tasks such as clustering and anomaly detection [IM98] — an effect often referred to as the 'curse of dimensionality'. As the dimensionality rises, distance values between points tend to concentrate around their mean values [Pes00] due to the effects of noise and other sources of error, which drastically alters the discriminative power of the distance measure [Hug68, BGRS99].

Feature reduction methods fall into two main categories: extraction and selection. Feature selection, as the name suggests, retains a subset of the original feature values, chosen with the aim of improving the performance and efficiency of data mining and machine learning tasks. In feature extraction, the attributes are transformed to a smaller set within an artificial feature space, with each new feature value depending on the values of many (or even all) of the original features. Many feature extraction algorithms are based on Principal Component Analysis (PCA) [Cam03, XXZC08] or Linear Discriminant Analysis [NO09].

Within the context of feature extraction, feature selection can be regarded as a special case in which transformation is restricted to axis-aligned projections. For applications in which extraction methods are appropriate, extracted features are very likely to outperform any equal number of selected features, due to the greater flexibility in determining a data transformation that maximizes the quality criteria used to guide the feature generation process. However, in some contexts where feature semantics are of major importance and where artificial features are meaningless, feature extraction cannot be used. Such applications are very common in bioinformatics [SIL07] (gene annotation, microarray analysis) and chemistry [SIL07, HGV11] (mass spectra analysis, interpretability of molecular signatures), among others. In such contexts, however, feature selection can still be applied to reduce the dimension of the data.

Even when feature extraction methods are applicable, they are generally not suitable for very large datasets, as the transformation of the full dataset (as well as every element to be added subsequently) to a new full-dimensional basis can incur a prohibitively high computational cost. This computational complexity, together with a lack of robustness in the presence of noise [FXY12, TP07, FJC06] make feature ex-

traction less useful when the instances outnumber the features. Furthermore, feature extraction has limited use when the goal is to remove irrelevant features, since all of the original features must be retained in order to compute values within the transformed feature space.

In this chapter, we consider the case of unsupervised feature selection, where class label information is unavailable. Unsupervised feature selection is a more difficult problem than supervised feature selection, due to the difficulty in evaluating the quality of features. In the absence of ground truth information, there are two main types of quality criteria that are commonly used to guide the selection process: redundancy elimination, where an attribute is discarded whenever it can be entirely or partially inferred from other attributes, and similarity conservation, in which a feature is selected according to the degree to which it helps to conserve the similarity between data points. Of the two, similarity conservation has received relatively less attention in the research literature.

Depending on their evaluation strategy, feature selection frameworks can be categorized as either filters or wrappers (or embedded). Wrapper methods execute a predictive learning task over the data for many choices of feature subsets, and retain candidate features for which the predictive model achieves high performances [GE03, WKN13]. In practical contexts involving data of even moderate volume or dimension, the cost of wrapper methods quickly becomes prohibitively high. A more affordable alternative is that of filter methods, which evaluate feature quality without resorting to an external learning machine. Although filters generally cannot match the quality of wrappers on small-scale data sets, on data sets of larger scale filters generally require far less computation time, and are less susceptible to overfitting.

Many unsupervised feature selection algorithms (such as LapAOFS and LapDOFS [HJZB11]), require that a target number of features be supplied as an input parameter, while others (such as GLFS [WG14]) automatically determine the number of selected features. With algorithms of the former type, the user is free to increase the number of features if needed, although the entire feature selection task usually must be reiterated. On the other hand, with most algorithms of the latter type, the user has no control over the number of features selected. When dealing with large datasets, a third and more convenient alternative is to produce a ranked list of all features, so as to allow any downstream processes to select the number of features that are most appropriate to the task at hand, given the amount of computational resources available. The most widely known unsupervised feature rank-

ing framework is that of spectral feature selection [ZL07], which includes Laplacian Score for Feature Selection (LS) [HCN05]. LS attempts to produce a reduced feature set that preserves the original neighborhood information to the greatest possible extent. More recent work on filters for unsupervised feature ranking includes Multi-Cluster Feature Selection (MCFS) [CZH10], which attempts to preserve the cluster structure of the data.

Almost all existing filter methods for feature selection are suitable only for data of relatively small volume and dimensionality, such as those found in the context of bioinformatics. In this work, we target the problem of unsupervised feature ranking at higher scales, where the data volume and data dimension together preclude the use of wrapper methods and computationally expensive filters. Such feature selection could be of particular benefit for such database and data mining operations as similarity search, neighborhood-based classification, and clustering.

As a criterion for guiding the selection process, we propose that features be assessed according to their ability to discriminate between the distances encountered in the neighborhoods of data points, in an attempt to alleviate the effects of the curse of dimensionality. Accordingly, we develop feature ranking methods that assess the discriminative power of individual features according to a recently-proposed model of the local intrinsic dimensionality (LID) of continuous distance distributions [ACF+15,Hou13]. While most feature selection methods employ indirect strategies in order to (implicitly) improve the discriminability of distance measures, the use of LID as guiding criterion allows for a more explicit approach, due to a formal equivalence established in [Hou13] between the discriminability of distance measures and low intrinsic dimensionality. Taking into account the performance issues due to larger data scales (in terms of both volume and dimensionality), we develop forward filter algorithms generally applicable to continuously-valued numerical data.

The specific contributions of this Chapter include:

- two greedy forward filter feature ranking algorithms, one univariate (in that it scores features independently) and the other multivariate (in that it scores feature subsets of size greater than 1);
- for the greedy multivariate method, a theoretical analysis based on submodularity;
- an experimental framework setting random feature selection as a baseline for

feature selection, and showing the advantage of using our methods with dense high-dimensional data.

The chapter is organized as follows. In the next section we present a survey of existing methods for feature selection. In Section 7.2, we present the details of two feature ranking methods, the first using direct ranking based on LID, and the second using a greedy feature ranking framework. The experimental framework is presented in Section 7.3, in which our methods are compared against state-of-the-art unsupervised filter-based feature selection algorithms. The experimental results are discussed in Section 7.4. Finally, in the last section we conclude by summarizing the advantages and the limitations of our feature selection strategy, and propose some possible extensions of our algorithm as future work.

7.1 Feature selection

Feature selection consists of choosing a subset of the original features so as to best satisfy a quality criterion. The ultimate goal is to improve model generalizability and interpretability, as well as to lower the computational costs of learning algorithms that use the data.

Formally, let us assume that we are given a set of n points in a space of dimension m represented by the matrix $X \in M_{n,m}(\mathbb{R})$

$$X = (x_1, x_2, ..., x_n) = (f_1, f_2, ..., f_m)^\top,$$

where $x_i \in \mathbb{R}^m$ are points and $f_i \in \mathbb{R}^n$ are features. Feature selection consists of finding $\Omega^* = \{f_1^*, ..., f_d^*\}$, a subset of $\Omega = \{f_1, ..., f_m\}$ that best satisfies an optimality condition.

Feature selection algorithms can be differentiated in several ways. They can be

- wrappers or filters, according to whether or not they employ a learning task as part of its feature evaluation strategy (with embedded methods combining the characteristics of both);

- forward or backward depending on their search direction — whether they build a feature set incrementally starting from an empty set, or cull features incrementally starting from the full set.

- univariate or multivariate, according to whether they assess features individually or in groups.

Wrapper methods apply a predictive learning machine to the data using the candidate feature subsets; the objective function is defined by the performance of this predictive model on a task for which the feature selection is targeted [WKN13]. Wrappers can achieve high performances with the learning machine used in the selection process. However, they are model-specific [GE03], as the quality of the features selected using a given learning model are often inadequate for others. Wrapper methods are also computationally expensive [GE03]: in a wrapper selection process, assessing the quality of each candidate subset requires one run of the learning machine, and one measurement of the quality of its output. Moreover, wrappers present a risk of overfitting not only to the predictive model, but also to the data.

In contrast, filter methods use properties of the data to compare the candidate subsets: they evaluate candidate feature subsets in terms of some quality criterion, often information-theoretic (Minimum-Redundancy Maximum-Relevance [PLD05], Information Gain [CT91], Gini Index [Gin12]), correlation-based (Correlation-based Feature Selection [HS99]), statistical (t-test and chi-squared [LS95]), or based on interclass distances (Fisher Score [DHS01], supervised Spectral Feature Selection framework [ZL07,HCN05]). Although they cannot match the performance of wrappers on the learning tasks used to guide the feature selection process, filters have the advantage of not being specific to any particular learning model. Furthermore, filters generally require far less computation time, and are less susceptible to overfitting. In the context of complex data, the advantages of filters are crucial.

Feature selection algorithms can also be categorized according to the order in which candidate feature subsets are considered. With backward elimination, the first feature subset candidate assessed is the full dataset F; new candidates subsets are constructed by incrementally removing features that underperform with respect to some criterion. The opposite approach, forward selection, consists of starting with an empty set, and building up candidate feature subsets through the incremental introduction of individual features. When the number of features is very high, examining candidate subsets of smaller cardinality is more affordable. In the context of high dimensional data, this argument is largely in favor of forward selection.

Feature selection methods also differ in the way they group features for assessment: univariate methods assess features invidually and independently, while multi-

variate methods consider sets of features during the selection process. Since combinations of features are not considered, the univariate approach cannot account for redundancy of features. However, univariate methods have the general advantage of being faster — since they consider attributes separately, their complexity is usually linear in the number of features. Although the multivariate approach is usually more computationally expensive, in the context of high-dimensional data, multivariate methods that strictly limit the numbers of candidates assessed can sometimes be more affordable than their univariate competitors.

In some cases, a vector $L = (l_1, l_2, ..., l_n)$ of class labels associated with X may also be available. Feature selection methods that make use of such label information as a ground truth are said to be 'supervised'; if label information is not used, the method is said to be 'unsupervised'. In general, wrapper methods tend to be supervised (due to the need to assess the result on a specific learning task).

In the context of big data, where ground truth label information is rarely available, the complexity argument favors unsupervised feature selection filters which operate in a forward direction. Among all existing candidates that fit these criteria, scalability concerns prevent the use of computationally expensive algorithms such as LapDOFS and LapAOFS [HJZB11], all of which require time more than quadratic in the number of points and number of features. Among all algorithms with reasonable scalability characteristics, the unsupervised general-purpose forward-filter Laplacian Score (LS) feature selection method [HCN05] is perhaps the most popular. It belongs to the general framework of spectral feature selection [ZL07], and has a computational complexity of $C_{\text{LS}} = mn^2$. The spectral framework offers three different approaches to scoring features, with the objective always being to preserve the graph structure of the dataset described by a spectral matrix S. Of the three scoring functions available, the second is used by LS:

$$\begin{aligned} \text{Score}_{\text{LS}}(f_t) &= \sum_{x_i, x_j \in X} \frac{(x_{i,t} - x_{j,t})^2}{\text{var}(f_t)} s_{i,j}, \\ \text{where} \quad s_{i,j} &= \delta_{i,j} e^{-\frac{d_{i,j}^2}{\theta}}; \end{aligned}$$

$\text{var}(f_t)$ is the variance of the feature f_t; $d_{i,j}$ denotes the Euclidean distance between x_i and x_j; $\delta_{i,j}$ is 1 when x_i or x_j is among the k-nearest neighbors of the other, and 0 otherwise; and θ represents the variance within a set of distance values, but

is not explicitly defined in the original work. Features are ranked in ascending order of Score_{LS}.

Feature selection can also be guided according to how well the new features preserve the cluster structure of the data. Given a similarity graph represented as a matrix W with $(W)_{i,j} = \delta_{i,j}$, and a target number of clusters c, Multi-Cluster Feature Selection (MCFS) [CZH10] computes the top c solutions $(y_1, ..., y_c)$ for the generalized eigenproblem $Ly = \lambda Dy$, where D is the diagonal matrix $(D)_{i,i} = \sum_{x_i \in X} w_{i,j}$, and $L = W - D$. Least Angle Regression (LAR) is then used to solve the r regression problems

$$\min_{a_i} ||y_i - X^T a_i^2||^2, \quad \text{where} \quad |a_i| < \gamma.$$

Finally, features are scored by

$$\text{Score}_{\text{MCFS}}(f_t) = \max_i |a_{i,t}|.$$

Features are ranked in descending order of $\text{Score}_{\text{MCFS}}$.

MCFS has limited use in practice due to its high computational complexity:

$$C_{\text{MCFS}} = mn^2 + cm^3 + cnm^2.$$

7.2 Method description

7.2.1 LID-based quality scores

For a given reference point, the LID measure indicates the difficulty in discriminating among neighbors. In principle, the easier it is to discriminate among the points in a local neighborhood, the better the performance of distance-based learning tasks.

The selection strategy employed by our methods assesses the (in)discriminability of individual features, in terms of the 1-dimensional distance values encountered in the neighborhoods of the objects of the dataset. The overall quality of the feature is evaluated by aggregating the associated LID estimates for each neighborhood. For computational reasons, it may not be necessary to compute LID estimates at all the data points. Instead, as a heuristic, we may choose to estimate LID for a sample of points $X_* \subseteq X$. One rationale for this is that objects from the same cluster are likely to have similar LID values; if the sample X_* is sufficiently dense, most of the

clusters will have an influence on the the estimation process. It should be noted that although only subsets of the neighborhoods are considered, all of the k-nearest neighbors are used in the estimation (not just those appearing in X_*).

The discriminative power of a feature depends on the location in the dataset from which distances are measured: while the feature may be discriminative among the points of a certain neighborhood (or cluster), it may be indiscriminative for other neighborhoods. It is therefore appropriate to allow points to 'vote' for the most discriminative features in their localities, and then to aggregate these votes.

7.2.2 Proposed algorithms

We propose two different methods of scoring features, one univariate (denoted IDFS) and the other multivariate (referred to as IDFS-m). In IDFS, the score of a feature f is the indiscriminability threshold $s(f)$ for which a proportion q of the points of X_* $(0 < q \leq 1)$ achieve LID values less than or equal to $s(f)$:

$$s(f) = \{\mathrm{ID}_f(x), x \in X_*\}_{(q)}.$$

For a given value of q, a low value of $s(f)$ indicates that f performs well over this proportion of the data. Parameter q must be set high, so that features may be assessed over most of the data regions sampled by X_*. However, setting q too close to 1 would allow the threshold to be determined by 'noise' or 'outlier' elements whose associated LID estimates are very high. For this reason, if α is the anticipated proportion of outliers in the dataset X, then we should ensure that $q < 1 - \alpha$. The proportion of outliers being difficult to predict, it is important to note that it is more harmful to involve an outlier in the feature selection process than to exclude an inlier.

Algorithm 5 ID-based selection of d feature given a dataset $X = (x_1, x_2, ..., x_n) = (f_1, f_2, ..., f_m)^\top$, a neighborhood range k, and a quantile $q \in [0, 1]$

1. Calculate dimensionality estimates $\mathrm{ID}_f(x)$ for each point $x \in X_*$ and for each feature: $\mathrm{ID}_f(x), x \in X, f \in \{f_1, f_2, ..., f_m\}$.

2. Score each feature by the q-quantile of the dimensionality estimates over the subset X_*: $s(f) = \{\mathrm{ID}_f(x), x \in X\}_{(q)}$.

3. Rank the features individually based on the scores they obtain: $\Omega^* = (f_1^*, f_2^*, ..., f_m^*)$, where $s(f_i^*) \leq s(f_j^*), \forall i < j$.

4. Return the d top ranked features in Ω^*: $\Omega_d^* = \{f_1^*, ..., f_d^*\}$.

Intuitively speaking, with IDFS-m, the impact of adding a new feature to an existing subset diminishes as the size of the subset increases. This property, called submodularity [KG12], is consistent with the fact that the first few highest-ranked features should have more impact on discriminability than the same number of later-ranked features.

Definition 12 (Discrete derivative) For a set function $\Psi : 2^V \to \mathbb{R}$, $S \subseteq V$, and $u \in V$,
$\Delta_\Psi(e|S) = \Psi(S \cup \{u\}) - \Psi(S)$ is the discrete derivative of Ψ at S with respect to u.

Definition 13 (Submodular function) A set function $\Psi : 2^V \to \mathbb{R}$ is submodular if
$\forall A \subseteq B \subseteq V, \forall u \in V \setminus B, \Delta_\Psi(u|A) \geq \Delta_\Psi(u|B)$.

For IDFS-m, we evaluate a feature subset $A \in \Omega$ according to the following score:

$$\Psi(A) = \sum_{x \in X} \psi(A, x),$$
$$\text{where} \quad \psi(A, x) = \sum_{a \in A} \alpha(\rho(a, A, x)) \cdot \beta(\mathrm{ID}_a(x)).$$

Here, for an object x, $\rho(a, A, x)$ is the rank of the feature a in the set A with respect to an increasing ordering of $\mathrm{ID}_a(x)$, and $\alpha : \mathbb{N} \to \mathbb{R}$ and $\beta : \mathbb{R} \to \mathbb{R}$ are weighting functions that apply respectively to the feature ranks of a and to the ID associated with the feature.

ID can be viewed as a measure of indiscriminability [ACF+15]. In practice, β can be chosen as the reciprocal function, so that $\beta(\mathrm{ID}_a(x))$ would measure the discriminability of the feature a. For the purposes of the analysis, β can be any monotonically decaying function. The function α determines the relative weight associated with $\beta(\mathrm{ID}_a(x))$. In order to favor the selection of top ranked features, the weighting function α must be monotonically decaying.

Theorem 7 If α is monotonically decreasing and convex, then Ψ is submodular.

Proof Let A and B be two sets of features such that $A \subset B \subset \Omega$. First, we show —by induction on $|B \setminus A|$— that for any $x \in X$, ψ is submodular. The base case ($|A \setminus A| = 0$) is trivial (since $\Delta_\psi(u|A) \geq \Delta_\psi(u|A)$).

Let A' be a set such that $A \subset A' \subset B$ and $B \setminus A' = \{t\}$. Supposing that $\forall x \in X, \forall u \in \Omega \setminus A', \Delta_\psi(u|A) \geq \Delta_\psi(u|A')$, by aligning terms in the summation and taking differences, we can show that $\forall x \in X, \forall u \in \Omega \setminus B, \Delta_\psi(u|A) \geq \Delta_\psi(u|B)$.

Let a_i be the i-th element of A' based on an increasing order of $\mathrm{ID}_a(x)$ (i.e. $i = \rho(a_i, A', x)$). For any $u \in \Omega \setminus B$,

$$\begin{aligned}\Delta(u|A') &\triangleq \psi(A' \cup \{u\}, x) - \psi(A', x) \\ &= \alpha(\epsilon) \cdot \beta(\mathrm{ID}_e(x)) \\ &\quad + \sum_{i=\epsilon}^{|A'|} [\alpha(i+1) - \alpha(i)] \cdot \beta(\mathrm{ID}_{a_i}(x)),\end{aligned}$$

where $\epsilon = \rho(u, A' \cup \{u\}, x)$.

If $\mathrm{ID}_u(x) > \mathrm{ID}_t(x)$, then

$$\begin{aligned}&\Delta(u|A') - \Delta(u|B) \\ &\quad = [\alpha(\epsilon) - \alpha(\epsilon+1)] \cdot \beta(\mathrm{ID}_u(x)) \\ &\qquad + \sum_{i=\epsilon}^{|A'|} [(\alpha(i+1) - \alpha(i)) - (\alpha(i+2) - \alpha(i+1))] \\ &\qquad \cdot \beta(\mathrm{ID}_{a_i}(x)).\end{aligned}$$

A similar argument applies when $\mathrm{ID}_u(x) \leq \mathrm{ID}_t(x)$.

$$\begin{aligned}\Delta(u|B) &= \psi(B \cup \{u\}, x) - \psi(B, x) \\ &= \alpha(\epsilon) \cdot \beta(\mathrm{ID}_u(x)) \\ &\quad + \sum_{i=\epsilon}^{\tau-1} [\alpha(i+1) - \alpha(i)] \cdot \beta(\mathrm{ID}_{a_i}(x)) \\ &\quad + [\alpha(\tau+1) - \alpha(\tau)] \cdot \beta(\mathrm{ID}_t(x)) \\ &\quad + \sum_{i=\tau}^{|A'|} [\alpha(i+2) - \alpha(i+1)] \cdot \beta(\mathrm{ID}_{a_i}(x)),\end{aligned}$$

and thus

$$\begin{aligned}&\Delta(u|A) - \Delta(u|B) \\ &\quad = [\alpha(\tau) - \alpha(\tau+1)] \cdot \beta(\mathrm{ID}_t(x)) \\ &\qquad + \sum_{i=\tau}^{|A'|} [(\alpha(i+1) - \alpha(i)) - (\alpha(i+2) - \alpha(i+1))] \\ &\qquad \cdot \beta(\mathrm{ID}_{a_i}(x)).\end{aligned}$$

If α is monotonically decreasing and convex, then $\Delta(u|A') \geq \Delta(u|B)$. From the assumption that $\Delta(u|A) \geq \Delta(u|A')$, we have that $\Delta(u|A) \geq \Delta(u|B)$, which implies that ψ is submodular. Since Ψ is a finite sum of submodular functions, we conclude that Ψ is submodular as well.

Heuristically maximizing the submodular function Ψ by means of greedy selection leads to an optimality guarantee of $(1 - 1/e)$ [KG12]. Greedy selection also ensures that the features are ranked in decreasing order of quality.

Algorithm 6 Selection of d features given a dataset $X = (x_1, x_2, ..., x_n) = (f_1, f_2, ..., f_m)^\top$ using a greedy feature ranking framework.

1. for each point $x \in X_*$ and each feature $f \in \Omega$ estimate $\mathrm{ID}_f(x)$.
2. $A := \{\}$
3. for each point $x \in X_*$:

(a) rank features $f \in \Omega \setminus A$ by increasing $\mathrm{ID}_f(x)$.

(b) for each feature $f \in \Omega \setminus A$, evaluate $\psi(A \cup \{f\})$.

4. for each feature $f \in \Omega \setminus A$, evaluate $\Psi(A \cup \{f\})$:

5. $f^* := argmax_{f \in \Omega \setminus A} \Psi(A \cup \{f\}) \quad ; \quad A := A \cup \{f^*\}$.

6. repeat 3-5 until $|A| = d$.

7. return A.

7.2.3 Complexity of the proposed algorithms

The asymptotic complexity is assessed in terms of the number of instances ($n = |X|$), the number of features ($m = |\Omega|$), the target number of features ($d < m$), the size of the subset of points where LID is estimated ($n_* = |X_*| < n$), the size of the neighborhoods ($k < n$), and the complexity (denoted $\gamma_n(m, n_*, k)$) of computing an index for a given subset of points.

Theorem 8 Algorithm 5 is of complexity

$$C_{\mathrm{ID}_1}(n, m, d, n_*, k) = O(\gamma_n(m, n_*, k) + mn_*k + m \log m).$$

If the nearest neighbors are computed using brute force, then $\gamma_n(m, n_*, k) = O(mnn_*)$, in which case

$$C_{\mathrm{ID}_1}(n, m, d, n_*, k) = O(mnn_*).$$

Proof The first step of the algorithm consists of computing the nearest neighbors for a subset of n_* points. LID estimates are then evaluated for each feature in Ω, and for each point in X_*. Once the required nearest neighbor distances have been computed, all LID estimates can be generated in $O(mn_*k)$ additional time. The determination of the q-quantile for a given feature can be performed in $O(m)$ time, and thus the calculation of scores in Step 2 requires $O(mn_*)$ time. Finally, sorting the features requires $O(m \log m)$ operations. In total,

$$C_{\mathrm{ID}_1}(n, m, d, n_*, k, q) = O(\gamma_n(m, n_*, k) + mn_*k + m \log m).$$

IDFS is therefore linear in the number of features, and subquadratic in the number of instances. In practice, it is reasonable to assume that $\log m \ll n_* k$. If so, then

$$C_{\text{ID}_1}(n, m, d, n_*, k, q) = O(\gamma_n(m, n_*, k) + mn_*k).$$

We note that under these assumptions, the state-of-the-art Laplacian Score method is of complexity $C_{\text{LS}}(n, m, d) = O(mn^2)$, which indicates that Algorithm 5 is more asymptotically scalable.

Theorem 9 Algorithm 6 is of complexity

$$C_{\text{ID}_2}(n, m, d, n_*, k) = O(\gamma_n(m, n_*, k) + mn_* d \log d).$$

If the nearest neighbors are computed using brute force, then $\gamma_n(m, n_*, k) = O(mnn_*)$, in which case

$$C_{\text{ID}_2}(n, m, d, n_*, k) = O(mn_*(n + d \log d)).$$

Proof As was the case with Algorithm 5, the first step of Algorithm 6 consists of computing the nearest neighbors for a subset of n_* points. LID estimates are then evaluated for each feature in Ω, and for each point in X_*. Once the required nearest neighbor distances have been computed, all LID estimates can be generated in $O(mn_*k)$ additional time. Next, over all points in $|X_*|$, ranking the features in terms of their LID scores requires a total of $O(n_* m \log m)$ time. The main loop of the algorithm is iterated d times. At the i-th iteration, the size of set A is simply $|A| = i$. For each selected point in X_*, and for each $f \in \Omega \setminus A$ of the $(m - i)$ candidate features, $O(\log i)$ operations are needed to rank the new candidate feature in $A \cup f$. This implies that the complexity of the i-th iteration is of the order of $O(n_*(m - i)(\log i))$. All iterations together require $O(n_* md \log d)$ operations. Finally, the full algorithm has a time complexity of

$$\begin{aligned} & C_{\text{ID}_2}(n, m, d, n_*, k) \\ & \quad = \quad O(\gamma_n(m, n_*, k) + mn_* d \log d). \end{aligned}$$

Consequently, IDFS-m remains as scalable as IDFS in terms of volume, but is less scalable in terms of the number of features. Assuming that $\log m \ll d \log d$ and

that $k \ll d \log d$, the complexity bound simplifies to

$$C_{\mathrm{ID}_2}(n, m, d, n_*, k) = O(\gamma_n(m, n_*, k) + mn_* d \log d).$$

In practice, Algorithm 6 can be terminated early if the submodular function Ψ converges to a constant value. This can be judged by setting a small threshold on the minimum required change in Ψ; if the amount of change falls below the threshold, the algorithm is terminated. However, in our experiments, no early termination was performed.

7.3 Experimental framework

In the proposed framework, we compare our methods with the state-of-the-art feature selection algorithms LS and MCFS. As a baseline for comparison, we also report results for a random feature selection strategy. We include PCA in our study in order to emphasize the gap in performance between feature extraction and feature selection methods in those settings where extraction is feasible.

Each method was used to produce a ranking of the features, from which a certain top-ranked proportion were used for a follow-on task — indexing, K-means clustering, or k-NN classification. The proportions of features considered ranged from 2% to 100%.

7.3.1 Methods

For our solutions, we set the neighborhood range of interest at $k = 100$, a value for which the ID estimators studied in [ACF+15] typically converge. The sample size is chosen to be an order of magnitude smaller than the dataset size ($|X_*| = \frac{n}{10}$).

In addition, for the univariate algorithm IDFS, we set $q = 0.95$, and for IDFS-m, we set $\alpha : x \to \frac{1}{x}$ and $\beta : x \to \frac{1}{x}$. For LS, we set the parameter θ to 0.01, 0.1 and 1 so as to cover the range of squared distance values for the datasets studied. The range of neighbors to preserve is set to $k = 100$ (as recommended in [HCN05]), so as to favor features that preserve the 100-nearest-neighbor graph. With MCFS, we set the number of eigenproblems to be the number of classes in the dataset, and the range of neighbors to $k = 5$ (as suggested by the authors of [CZH10]). Note that supplying MCFS with the number of classes gives it an advantage over the other methods in the study.

As the simplest feature selection strategy, uniform random selection should be a baseline for any experimental comparison. However, unlike for many other machine learning tasks that implicitly normalize against the expected result of randomness, random baselines are not commonly encountered in the feature selection literature. To the best of our knowledge, only one such paper exists: a bioinformatics paper that used paired ANOVA tests to conclude that of the 32 methods studied, "no method was significantly better than the random selection strategy" [HGV11]. As a baseline for comparison (which we refer to as 'Random'), we generate 10 random feature rankings, and report the average performance of our learning tasks over these feature sets.

For all methods, the execution time was limited to 10 days (on an a 48-core Intel® Xeon® CPU E5-2670).

7.3.2 Tasks

Although the features selected by our algorithms can be used in unsupervised learning, it is difficult to assess the quality of features based solely on unsupervised learning tasks. For this reason, in our experimentation, we choose to evaluate feature selection methods according to their performance on unsupervised indexing and K-means clustering tasks, and supervised k-nearest neighbor (k-NN) classification tasks. Among these tasks, indexing is the most fundamental, as features that produce inaccurate neighbor sets are very likely to lead to an incorrect classification. Similarly, classification is a less demanding task than clustering: a feature space where the classification performance is poor is unlikely to permit the identification of the component clusters from which the class membership is constituted.

7.3.2.1 Indexing

The membership of the full k-NN query result set can be used as a form of unsupervised 'ground truth'. For a given feature set, we measure its accuracy for indexing as follows:

$$\text{Accuracy} = \frac{1}{n}\sum_{x \in X}\frac{1}{k}|\delta_{k,x} \cap \delta'_{k,x}|,$$

where $\delta_{k,x}$ and $\delta'_{k,x}$ are the k-NN sets of the point x before and after feature selection, respectively.

In our framework, we report the average accuracy over all k-NN lists for the entire dataset, with $k = 100$.

7.3.2.2 k-NN classification

A k-NN classification was performed on various datasets using the features selected by each algorithm. Given labeled (training) and unlabeled (test) points, the k-NN classification task is straightforward: the class of each unlabeled point is set to that of the most frequent class label from among the k nearest labeled points. For $k = 10$ and $k = 100$, and using Euclidean distances, we performed a 10-fold cross validation: the dataset was randomly partitioned into 10 slices of equal size, and the task was executed 10 times, each time with a different slice as the test data, and the remainder as training data.

The most commonly-used measure of classification quality is the accuracy with which classification result C' relates to the original class labels C. If C_x and C'_x denote the labels of the point x in C and C' respectively, and δ is the function that evaluates to 1 if its two members coincide (and 0 otherwise), then the accuracy is given by:

$$\text{Accuracy}(C, C') = \frac{1}{n} \sum_{x \in X} \delta(C_x, C'_x).$$

Normalized Mutual Information (NMI) is a commonly-used measure of dependence between two data groupings. Here, we use it to measure the dependence of the predicted labeling C' on the original class labels in C. It has values ranging from 0 when the clusterings are completely uncorrelated, to 1 if $C = C'$. For a formal definition of the NMI measure, see [SHH99].

7.3.2.3 K-means clustering

Clustering is an unsupervised task where data is split into K groups of instances with similar characteristics. K-means is a clustering algorithm where K centroids points are initially chosen at random. Then, in each iteration every point of the dataset is labeled with its closest centroid, and the centroid is updated to the center of mass of the points assigned to it. This process is repeated until convergence. It is common practice to stop the computation and return the current cluster assignment if no convergence is reached within a fixed number of iterations (100 in our implementation).

Dataset	Instances	Attributes	Classes
ALOI [GBS05, BFF+01]	110250	641	1000
BCI5 [Mil04]	31216	96	3
Gisette [GGBHD04]	7000	5000	2
Isolet [CF90]	7797	617	26

Table 7.1: Characteristics of the datasets used in the experimental framework.

In addition to NMI, we assess the quality of clustering results using a standard measure, the Adjusted Rand Index [Ran71] (ARI). ARI is an adjustment for chance of the Rand Index (RI), which focuses on the numbers of pairs of points whose cluster set membership relationships are preserved across two data groupings. When the degree of preservation is greater than what would be expected by chance, the ARI is positive. When the set membership is perfectly preserved, the ARI score is 1. For a formal definition of the ARI measure, see [Ran71].

7.3.3 Datasets used for the experiments

For our experimental study, we used 4 publicly available datasets covering a range of cardinalities, dimensionalities, domains and types of instances (cf. Table 7.1). Only two of the datasets, ALOI and BCI5, are of size sufficiently large so as to be considered consistent with our extreme-value-theoretic ID modeling assumptions.

7.4 Results

The performance of our proposed methods, together with those of LS, MCFS, Random, and PCA, are shown in Figures 7.1, 7.2, 7.3, and 7.4. The performances plotted for small numbers of selected features show the capacity of the various algorithms to identify good features early in the ranking, while the performances plotted for large numbers of selected features show the capacity of the algorithms to discard undesirable attributes. The left-hand portions of the performance curves are much more significant than the right-hand portions, where the curves all converge to the performance of the full feature set.

As expected, it is clear from the results of the different tasks that features extracted by PCA are consistently of better quality than those produced by any of the selection methods considered in this study. PCA is included in this study not as a

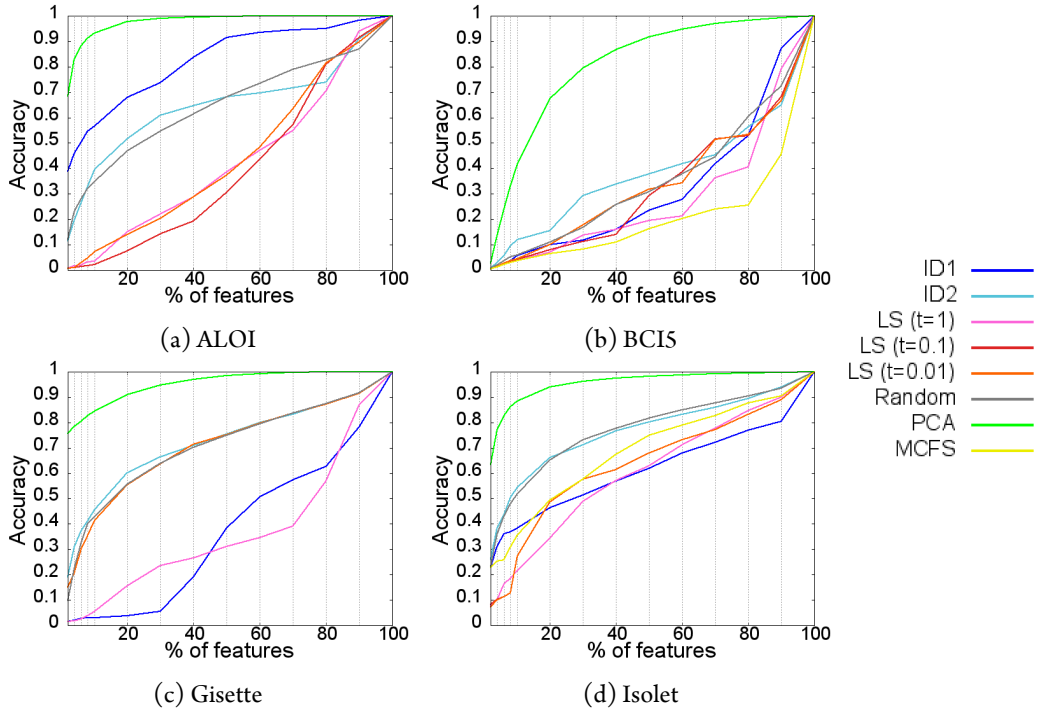
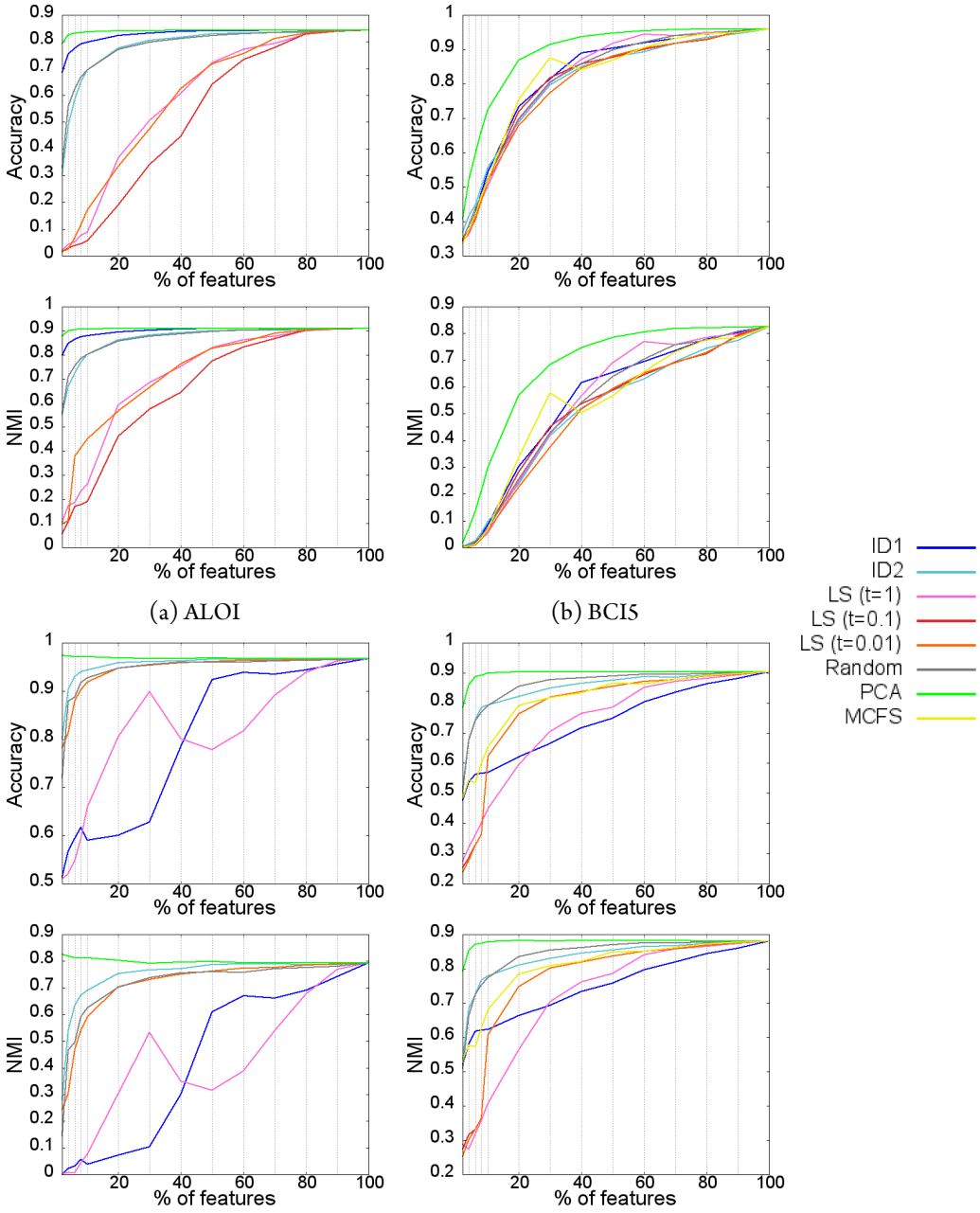


Figure 7.1: Average accuracy of indexing 100-NN using features as ranked by the selection algorithms.

baseline for selection, but only to illustrate the performance gap between selection and extraction for datasets that support both.

For the BCI5 dataset, as shown in Figure 7.4a, the results for K-means clusterings are of very poor quality (NMI below 5% and ARI less than 2.5% even when using the entire feature set) despite the features being of good quality for the classification task, as shown in Figure 7.2b. Like many clustering algorithms, K-means requires a random initialization resulting in a different result each time. Such variability makes it even more difficult to assess the difference in quality between features. In the case where clustering performances are of acceptable quality (Figure 7.4b), they do not contradict the quality of indexing and classification results (as shown in Figures 7.1d, 7.2d, and 7.3d). Thus, our experimentation confirms that clustering tasks may not always be appropriate for the guidance and assessment of feature selection methods.

For the Aloi and Gisette datasets, MCFS failed to converge within the 10-day limit set on execution time — particularly surprising given the comparatively small size of Gisette. Unlike LS and our methods, MCFS matrix operations cannot be ex-

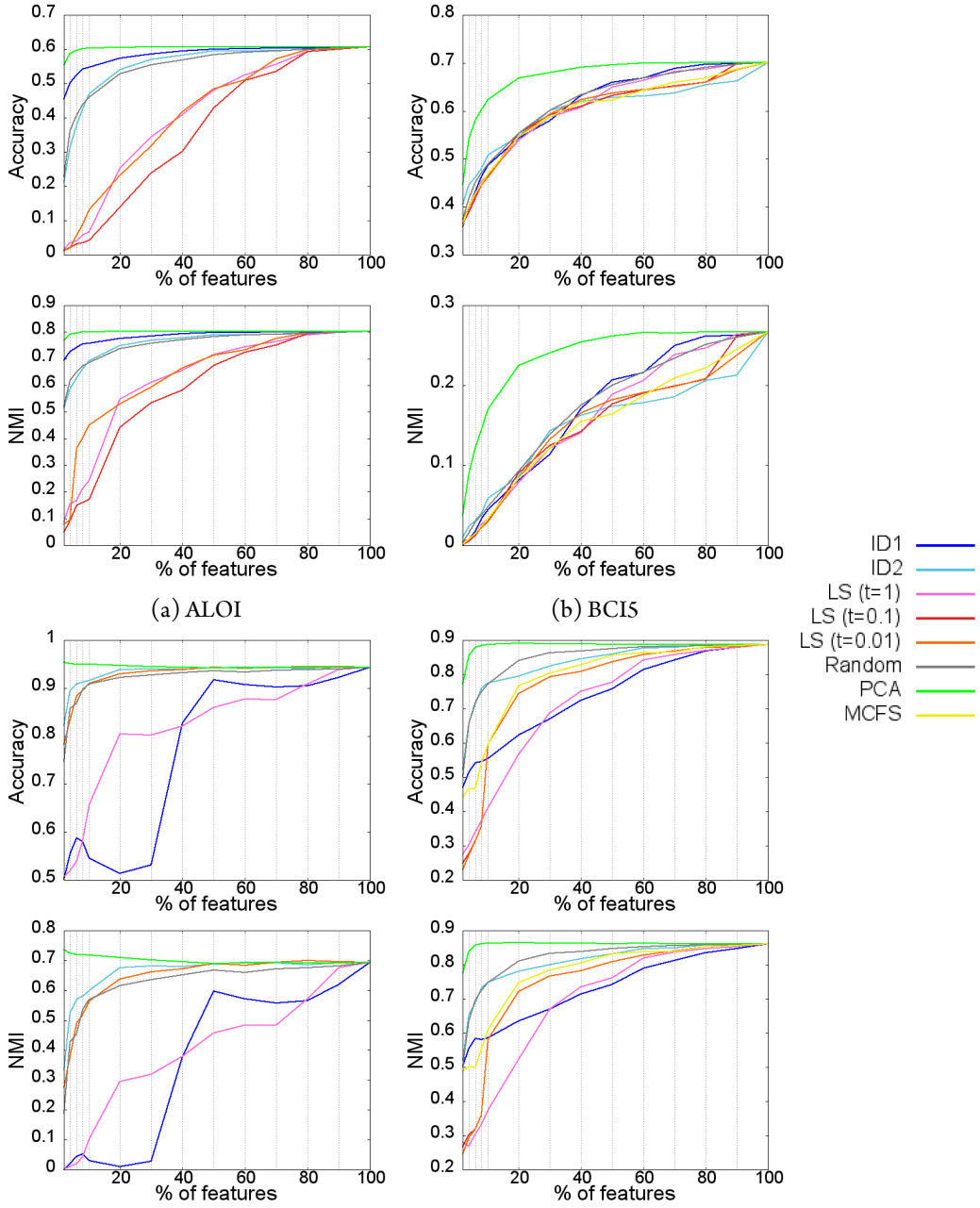


(a) ALOI (b) BCI5

(c) Gisette (d) Isolet

Figure 7.2: Quality of k-NN ($k = 10$) classification using features as ranked by the selection algorithms.

ecuted in parallel, which limits its usability on large datasets. Moreover, MCFS suffers from numerical instability, as it failed to converge to non-singular eigenvalues when attempting to select features from ALOI. The only case where this algorithm

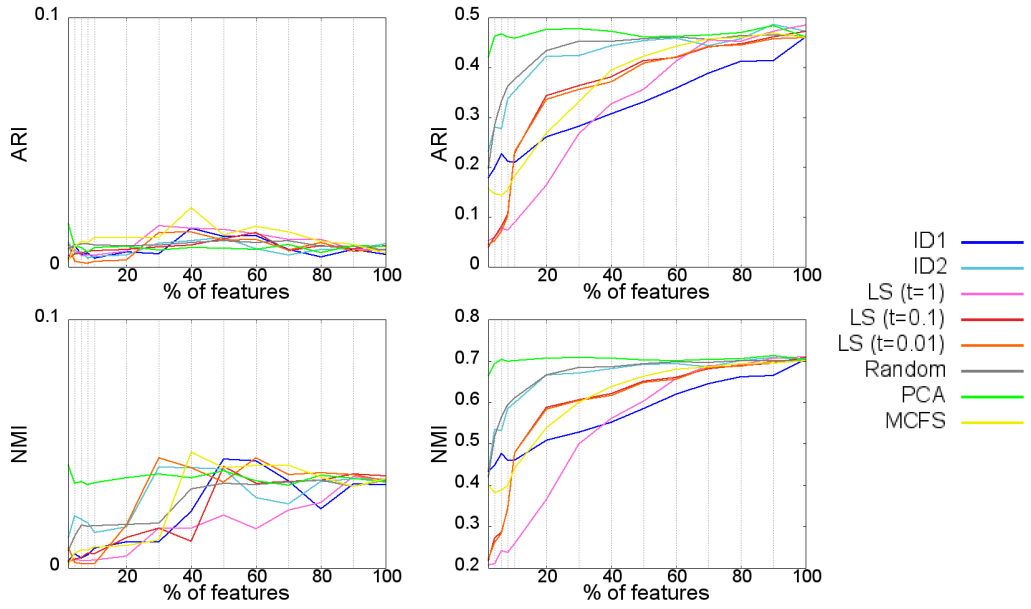


(a) ALOI (b) BCI5

(c) Gisette (d) Isolet

Figure 7.3: Quality of k-NN ($k = 100$) classification using features as ranked by the selection algorithms.

outperforms our methods is when selecting 20 and 30% of the features for 10-NN classification on BCI5, as seen in Figure 7.2b. We can conclude that MCFS is not well-suited for large high-dimensional data, mainly due to its prohibitively-high com-



(a) BCI5 (b) Isolet

Figure 7.4: Quality of K-means clustering using features as ranked by the selection algorithms.

putational complexity.

In all experiments on ALOI and Gisette, at least one of our methods clearly outperforms both Random and the state-of-the-art competitors LS and MCFS. For the indexing task on BCI5, IDFS-m outperforms the other selection methods. In the remaining experiments on BCI5 and Isolet, none of the methods clearly outperform the others. Our experimental results thus show that our methods have the potential for good performance on large, high-dimensional datasets.

Gisette and Isolet may seem to be similar as regards the number of samples and original features (cf. Table 7.1). However, in terms of class size, they are markedly different. While Isolet has classes of roughly 300 samples, Gisette has 3500 points per class, and is considerably denser than Isolet. The conditions for extreme-value-theoretic estimation are therefore more favorable with Gisette than with Isolet. Consequently, IDFS-m clearly outperforms LS on Gisette (cf. Figures 7.2c and 7.3c). The failure of IDFS on these sets could be explained by correlations among the features — note that the greedy algorithm IDFS-m has a multivariate approach capable of accounting for correlation between features. The failure of LS is due to the abundance of noise, as half of the Gisette features are known to be artificial.

ALOI and BCI5 have the size and continuity of feature values that satisfy our

extreme-value-theoretic modeling assumptions. In BCI5, there is little opportunity for improvement through feature selection, as a feature selection (spatial filtering) process had already been applied as part of the preprocessing of the set [Mil04]. Consequently, the seven algorithms performed almost equally well. However, when considering the first few features selected by the methods, IDFS-m outperforms its competitors, particularly in the indexing and 100-NN classification tasks (cf. Figure 7.3b). On ALOI, the performance of IDFS is unmatched by any other selection algorithm in the framework (cf. Figures 7.1a, 7.2a, and 7.3a), while the IDFS-m variant seems to lose some of its effectiveness due to its greedy approach.

In terms of execution time, LS required 23.71 hours to produce a ranking of ALOI's 641 features, while IDFS required 23.70 hours. Due to their low asymptotic complexity relative to that of their competitors (quadratic in the number of data instances and linear in the number of features), LS and IDFS are better suited for large data.

7.5 Discussion

7.5.1 Summary

The methods proposed in this paper differ from most other methods in the literature, in that they use neither information theoretic nor spectral measures to select features. Instead, they directly assess the discriminability of the features through the estimation of local intrinsic dimensionality. One major advantage held by LID is that other than continuity, it requires no knowledge or assumptions concerning the distribution of the data. Moreover, the methods are robust against noise. The main drawback of our LID-based methods is the fact that they require a relatively smooth distribution of distance values, which is more likely to occur when the data is dense, or high-dimensional, or both. Nevertheless, such high-dimensional situations are generally those where classical feature selection methods often fail.

As revealed by our experimental results, the use of clustering for the evaluation of feature selection can lead to questionable conclusions. As a rule, the performance of unsupervised learning algorithms is highly variable in terms of their final quality metrics. In clustering, for example, results often depend on initial settings that are selected randomly (such as in the popular K-means algorithm) or parameter choices made by the user (such as the number of clusters). Although running the task sev-

eral times may reduce the variance, other approaches may be both faster and more accurate.

Using a supervised task to assess feature quality does not contraindicate their later use in unsupervised learning tasks. Unlike clustering, classification tasks rely on labeled training data, and thus have more reliable information available to them than do clustering tasks. Although the ground truth information is not required in order to produce the actual ranking of features, it can be used to experimentally validate the selected attributes. There are many candidate classification algorithms; the choice of k-NN classification is motivated by its connection to similarity search, which is the basis of many machine learning tasks.

Comparing against a random baseline is inherent in several popular quality metrics, such as the NMI or ARI. Surprisingly however, despite being the most trivial and easily-implemented approach, uniform random feature selection is itself not always used as a baseline method for comparison. Our experimental results show that the classification results based on randomly ranked features are generally competitive with those using features selected by state-of-the-art algorithms. This phenomenon is a consequence of the high degree of feature redundancy and interdependence found in typical high-dimensional datasets. We believe that any future work on feature selection should systematically compare against random selection.

7.5.2 Future work

As the data volume increases, the cost of building an exact index often becomes prohibitively expensive. Instead of using exact nearest neighbors, it is possible to use approximation algorithms such as ϵ-NNS [IM98] or NN-Descent [DML11]. Using approximate nearest neighbor distances may adversely affect the ID estimation, and consequently change the feature ranking. However, trading off feature quality for computation time may be advantageous when the dataset is large or high-dimensional.

One possible extension of our feature selection approach is to consider other models of intrinsic dimensionality as an alternative to LID in the assessment of feature quality. Although other models may lack the theoretical guarantees regarding the discriminability of distance measures, or may have a higher computational complexity or sensitivity to noise [ACF+15] such models may still deserve consideration, particularly if they are more reliable for non-smooth distance distributions, or

for smaller data samples.

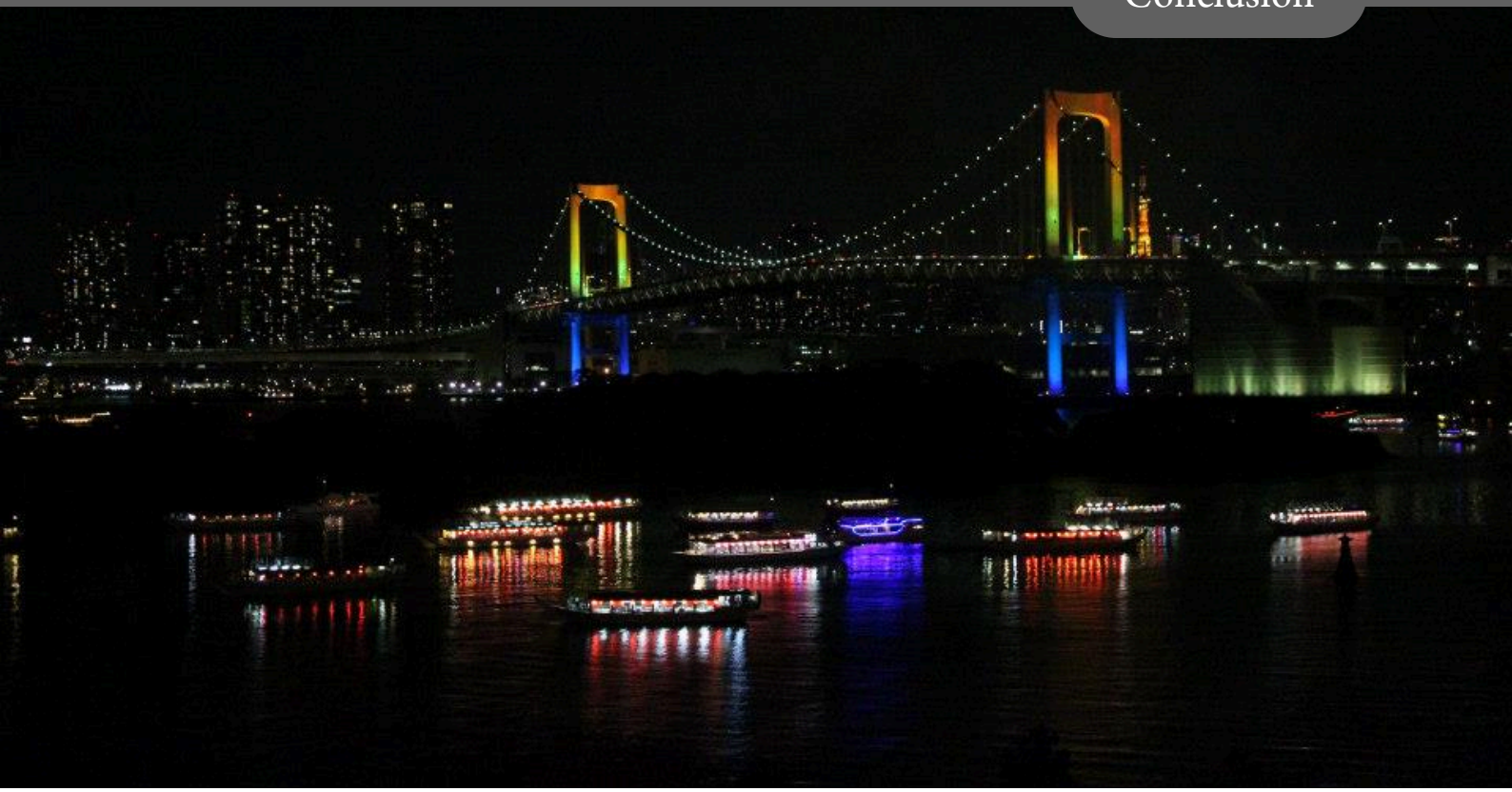
There are several other possible directions for future work on ID-based feature selection. First, it would be interesting to consider the problem of unsupervised tuning of the parameters of ID-based feature selection. Also, our approaches for unsupervised feature ranking could be extended to supervised applications, by restricting the ID estimation to training examples with common labels. Finally, ID-based feature evaluation could conceivably be applied to dimensional reduction through feature extraction, where the original features are linearly combined in order to obtain a new feature space in which distances are more discriminable than in the original.

Part IV

Conclusion

Chapter 8

Conclusion



In this Chapter we provide a short summary of the main contributions made in this thesis, and point out prospective directions for future research. We subsequently conclude the thesis by discussing its implications for data mining and machine learning applications, and by exploring the future research directions in which this work can be expanded.

8.1 Discussion

This work's main contribution was to provide new estimators for the LID. The family of estimators that was developed has empirically lower variance and smaller bias compared with the state of the art. In particular, with the use of auxiliary distances in ALID, our estimators can use more distance samples without increasing the neighborhood size. To the best of our knowledge, none of the existing models of local intrinsic dimensionality take advantage of distances other than those from a test point to its neighbors. The use of auxiliary distances gives our estimator a clear edge in terms of convergence. In addition, the ability of ALID to use a larger distance

sample without increasing the neighborhood size makes our estimator suitable when data consists of small groups of heterogeneous intrinsic dimensionality. This has the potential to improve the quality of algorithms where locality is an important factor, such as subspace clustering and local outlier detection. The simple assumptions that lead to our estimators make it easier to understand their limitations.

Our LID estimators have already seen interest from various research areas including: Sliced Inverse Regression [CGC14], nearest neighbor search [LZS+,Boy16, Cur16], similarity search [IAF16], text mining [Cla], outlier detection [vBHZ15], and dependency measures [RCN+16].

The extensive experimental framework on ID estimation that supports our results and discussions sets a standard for ID estimation in general and for local ID estimation in particular. Moreover, this framework shows the robustness of our estimators when approximate nearest neighbor distances are used instead of exact distances. This is an advantage over many state-of-the-art methods.

The feature selection algorithms proposed in this thesis differ from most other methods in the literature, in that they directly assess the discriminability of the features through the estimation of local intrinsic dimensionality instead of relying on information theoretic or spectral measures. The limited assumptions, namely the continuity of distance distributions which requires continuously distributed features make the methods applicable to various datasets. The main drawback of our LID-based methods is the fact that they require a relatively smooth distribution of distance values, which is more likely to occur when the data is high-dimensional, or dense, or both. Nevertheless, such high-dimensional situations are generally those where classical feature selection methods often fail.

In the experimental framework proposed to assess the quality of our feature selection algorithms, we compared against features selected uniformly at random. In fact, comparing against a random baseline is necessary in many data mining and machine learning contexts, and is inherent in several popular quality metrics, such as the NMI or ARI. Surprisingly however, uniform random feature selection is not always used as a baseline method for comparison. Our experimental results show that the classification results based on randomly ranked features are generally competitive with those using features selected by state-of-the-art algorithms. The competitive results obtained by randomly selected features against elaborately selected ones suggest that practitioners should systematically compare against random selection.

8.2 Future work

Possible directions for future work in LID estimation include the development of an auxiliary-distance estimator using the Method of Moments, instead of MLE. By applying in the estimation of the first two moments the same approach used to obtain an auxiliary-distances maximum likelihood estimator, a auxiliary-distance MoM estimation of LID can be obtained.

In cases where the neighborhood is very small, estimation of the LID can potentially be improved by using a limited number of points from outside the locality, at the cost of a slight increase in bias. In fact, points from outside the locality can shift the estimation towards the LID of their own localities. Moreover, violating the locality condition by using auxiliary distances to neighbors that belong outside the original locality can lead to additional computational costs associated with k-NN searches.

Second order LID [Hou15] introduced in [Hou15] can be viewed as a measure of inlierness that has potential applications in clustering, classification and outlier detection. Preliminary theoretical work on the estimation of second order LID is being done as a first step towards finding practical estimators based on Maximum Likelihood Estimation and the Method of Moments.

Using our notation, it is shown that the second order EVI can be expressed as $\rho = \mathrm{ID}_{|\mathrm{ID}_{\mathrm{ID}} f_X|}(0)$ [Hou15]. The numerical methods used to estimate the second order EVI have the potential to lead to estimators of the second order LID, since the two quantities are based on the same transformation, the former being applied to the probability density function while the latter to the cumulative distribution function. The EVT community has been working on estimating second order EVI for the past three decades, and the absence of closed-form second order EVI estimators could indicate the difficulty of finding second order LID ones. The sample size required for convergence of the second order EVI is larger than samples required for the convergence of first order EVI, suggesting that the same should hold in the case of second order LID. Estimators of the second order LID could also use auxiliary distances. Using these auxiliary distances would increase the number of distances available to the estimator without increasing the neighborhood size, and would potentially solve the numerical convergence issues.

One possible extension of our feature selection approach is to consider estimators or other models of intrinsic dimensionality as an alternative to LID in the as-

sessment of feature quality. ALID can potentially improve the quality of our methods. Other models could also be tested. Although these models may lack the theoretical guarantees regarding the discriminability of distance measures, or may have a higher computational complexity or sensitivity to noise [ACF+15] such models may still deserve consideration, particularly if they are more reliable for non-smooth distance distributions, or for smaller data samples.

There are several other possible directions for future work on ID-based feature selection. First, it would be interesting to consider the problem of unsupervised tuning of the parameters of ID-based feature selection. Also, our approaches for unsupervised feature ranking could be extended to supervised applications, by restricting the ID estimation to training examples with common labels. Finally, ID-based feature evaluation could conceivably be applied to dimensional reduction through feature extraction, where the original features are linearly combined in order to obtain a new feature space in which distances are more discriminable than in the original.

[ACF+15] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E Houle, Ken-ichi Kawarabayashi, and Michael Nett. Estimating local intrinsic dimensionality. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 29–38. ACM, 2015.

[AdL03] M.I. Fraga Alves, Laurens de Haan, and Tao Lin. Estimation of the parameter controlling the speed of convergence in extreme value theory. Mathematical Methods of Statistics, 12(2):155–176, 2003.

[AGd03] M. I. Fraga Alves, M. I. Gomes, and Laurens de Haan. A new class of semi-parametric estimators of the second order parameter. Portugaliae Mathematica, 60(2):193–214, 2003.

[AGdH03] MI Fraga Alves, M Ivette Gomes, and Laurens de Haan. A new class of semi-parametric estimators of the second order parameter. Portugaliae Mathematica, 60(2):193–214, 2003.

[AK00] Franz Aurenhammer and Rolf Klein. Voronoi diagrams. Handbook of Computational Geometry, 5:201–290, 2000.

[Ale76] Lawrence M Aleamoni. The relation of sample size to the number of variables in using factor analysis techniques. Educational and Psychological Measurement, 36(4):879–883, 1976.

[BCG11] Charles Bouveyron, Gilles Celeux, and Stéphane Girard. Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA. Pattern Recognition Letters, 32(14):1706–1713, 2011.

[BD99] Jock A Blackard and Denis J Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. Computers and electronics in agriculture, 24(3):131–151, 1999.

[BDG+05] Jan Beirlant, Goedele Dierckx, A Guillou, et al. Estimation of the extreme-value index and generalized quantile plots. Bernoulli, 11(6):949–970, 2005.

[BDH74] August A Balkema and Laurens De Haan. Residual life time at great age. The Annals of Probability, pages 792–804, 1974.

[Ben69] Robert Bennett. The intrinsic dimensionality of signal collections. IEEE Transactions on Information Theory, 15(5):517–525, 1969.

[BFF+01] N Boujemaa, J Fauqueur, M Ferecatu, F Fleuret, V Gouet, B LeSaux, and H Sahbi. IKONA for interactive specific and generic image retrieval. In Proceedings of International workshop on Multimedia Content-Based Indexing and Retrieval, 2001.

[BGRS99] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In International Conference on Database Theory, pages 217–235. Springer, 1999.

[BGT89] Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. Regular variation, volume 27. Cambridge University Press, 1989.

[BK81] Paul T Barrett and Paul Kline. The observation to variable ratio in factor analysis. Personality study and group behavior, 1(1):23–33, 1981.

[BKL06] Alina Beygelzimer, Sham Kakade, and John Langford. Cover trees for nearest neighbor. In International Conference on Machine Learning, pages 97–104. ACM, 2006.

[BMEWL11] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In ISMIR 2011, pages 591–596. University of Miami, 2011.

[BN03] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15(6):1373–1396, 2003.

[Boy16] Leonid Boytsov. Efficient and Accurate Non-Metric k-NN Search with Applications to Text Matching. PhD thesis, University of Massachusetts Amherst, 2016.

[Bra02] Matthew Brand. Charting a manifold. In Neural Information Processing Systems, pages 961–968, 2002.

[BS98] Jörg Bruske and Gerald Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(5):572–575, 1998.

[BVT+96a] Jan Beirlant, Petra Vynckier, Josef L Teugels, et al. Excess functions and estimation of the extreme-value index. Bernoulli, 2(4):293–318, 1996.

[BVT96b] Jan Beirlant, Petra Vynckier, and Jozef L Teugels. Tail index estimation, pareto quantile plots regression diagnostics. Journal of the American statistical Association, 91(436):1659–1667, 1996.

[BY95] Fred B Bryant and Paul R Yarnold. Principal components analysis and exploratory and confirmatory factor analysis. 1995.

[CA74] Chiu Kuan Chen and HC Andrews. Nonlinear intrinsic dimensionality computations. IEEE Transactions on Computers, 100(2):178–184, 1974.

[Cam03] Francesco Camastra. Data dimensionality estimation methods: a survey. Pattern Recognition, 36(12):2945–2954, 2003.

[CBR+14] Claudio Ceruti, Simone Bassis, Alessandro Rozza, Gabriele Lombardi, Elena Casiraghi, and Paola Campadelli. Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration. Pattern Recognition, 47(8):2569–2581, 2014.

[CBTD01] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. An introduction to statistical modeling of extreme values, volume 208. Springer, 2001.

[CC00] Trevor F Cox and Michael AA Cox. Multidimensional Scaling. CRC press, 2000.

[CDM85] Sandor Csorgo, Paul Deheuvels, and David Mason. Kernel estimates of the tail index of a distribution. The Annals of Statistics, pages 1050–1077, 1985.

[CF90] Ronald Cole and Mark Fanty. Spoken letter recognition. In Proceedings of the Third DARPA Speech and Natural Language Workshop, pages 385–390, 1990.

[CGC14] Alessandro Chiancone, Stephane Girard, and Jocelyn Chanussot. Collaborative sliced inverse regression. In Rencontres d'Astrostatistique, 2014.

[CHI04] Jose A Costa and Alfred O Hero III. Entropic graphs for manifold learning. In Asilomar Conference on Signals, Systems and Computers, volume 1, pages 316–320. IEEE, 2004.

[CKO+06] Li Juan Cao, S Sathiya Keerthi, Chong-Jin Ong, Jian Qiu Zhang, and Henry P Lee. Parallel sequential minimal optimization for the training of support vector machines. IEEE Transactions on Neural Networks, 17(4):1039–1049, 2006.

[CL92] Andrew L. Comrey and Howard B. Lee. A first course in factor analysis. Lawrence Eribaum Associates, Inc., Publishers, page 217, 1992.

[CL96] Thomas F Coleman and Yuying Li. An interior trust region approach for nonlinear minimization subject to bounds. SIAM Journal on Optimization, 6(2):418–445, 1996.

[Cla] Vincent Claveau. Dimensionalité intrinsèque dans les espaces de représentation des termes et des documents.

[CM85] Sándor Csörgoő and David M Mason. Central limit theorems for sums of extreme values. In Mathematical Proceedings of the Cambridge Philosophical Society, volume 98, pages 547–558. Cambridge Univ Press, 1985.

[CS16] Francesco Camastra and Antonino Staiano. Intrinsic dimension estimation: Advances and open problems. Information Sciences, 328:26–41, 2016.

[CT91] Thomas M Cover and Joy A Thomas. Elements of Information Theory. 1991.

[Cur16] Ryan R Curtin. Dual-tree k-means with bounded iteration runtime. arXiv preprint arXiv:1601.03754, 2016.

[CV98] Sándor Csörgő and László Viharos. Estimating the tail index. AsymptoticMethods in Probability and Statistics, pages 833–881, 1998.

[CV02] Francesco Camastra and Alessandro Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(10):1404–1407, 2002.

[CZH10] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 333–342. ACM, 2010.

[Dav84] Anthony C Davison. Modelling excesses over high thresholds, with an application. In Statistical extremes and applications, pages 461–482. Springer, 1984.

[DEdH89] Arnold LM Dekkers, John HJ Einmahl, and Laurens de Haan. A moment estimator for the index of an extreme-value distribution. The Annals of Statistics, pages 1833–1855, 1989.

[DFG99] Qiang Du, Vance Faber, and Max Gunzburger. Centroidal voronoi tessellations: applications and algorithms. SIAM Review, 41(4):637–676, 1999.

[DG03] David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Sciences, 100(10):5591–5596, 2003.

[DH97] Pierre Demartines and Jeanny Hérault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. IEEE Transactions on Neural Networks, 8(1):148–154, 1997.

[DHF07] Laurens De Haan and Ana Ferreira. Extreme value theory: an introduction. Springer Science & Business Media, 2007.

[DHM88] Paul Deheuvels, Erich Haeusler, and David M Mason. Almost sure convergence of the Hill estimator. In Mathematical Proceedings of the Cambridge Philosophical Society, volume 104, pages 371–381. Cambridge Univ Press, 1988.

[dHP98] Laurens de Haan and Liang Peng. Comparison of tail index estimators. Statistica Neerlandica, 52(1):60–70, 1998.

[dHR98] Laurens de Haan and Sidney Resnick. On asymptotic normality of the hill estimator. Stochastic Models, 14(4):849–866, 1998.

[DHS01] Richard O Duda, Peter E Hart, and David G Stork. Pattern Classification. 2001.

[DK98] Holger Drees and Edgar Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. Stochastic Processes and their Applications, 75(2):149–172, 1998.

[DM01] Ely Dahan and Haim Mendelson. An extreme-value model of concept testing. Management Science, 47(1):102–116, 2001.

[DML11] Wei Dong, Charikar Moses, and Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In International World Wide Web Conference, pages 577–586. ACM, 2011.

[DR84] Richard Davis and Sidney Resnick. Tail estimates motivated by extreme value theory. The Annals of Statistics, pages 1467–1487, 1984.

[dVCH12] Timothy de Vries, Sanjay Chawla, and Michael E Houle. Density-preserving projections for large-scale local anomaly detection. Knowledge and Information Systems, 32(1):25–52, 2012.

[ER92] J-P Eckmann and David Ruelle. Fundamental limitations for estimating dimensions and lyapunov exponents in dynamical systems. Physica D: Nonlinear Phenomena, 56(2-3):185–187, 1992.

[FJ13] T. Furon and H. Jégou. Using extreme value theory for image detection. Research Report RR-8244, INRIA, 2013.

[FJC06] Peter Filzmoser, Kristel Joossens, and Christophe Croux. Multiple group linear discriminant analysis: Robustness and error rate. Springer, 2006.

[FK94] Christos Faloutsos and Ibrahim Kamel. Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension. In Proceedings of the 13th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pages 4–13. ACM, 1994.

[FL12] Erik M Ferragut and Jason Laska. Randomized sampling for large data applications of svm. In ICMLA, volume 1, pages 350–355. IEEE, 2012.

[FO71] Keinosuke Fukunaga and David R Olsen. An algorithm for finding intrinsic dimensionality of data. IEEE Transactions on Computers, 100(2):176–183, 1971.

[FQZ09] Mingyu Fan, Hong Qiao, and Bo Zhang. Intrinsic dimension estimation of manifolds by incising balls. Pattern Recognition, 42(5):780–787, 2009.

[Fre02] Daniel Freedman. Efficient simplicial reconstructions of manifolds from their samples. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(10):1349–1357, 2002.

[FSA07] Amir M Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. Manifold-adaptive dimension estimation. In International Conference on Machine Learning, pages 265–272, 2007.

[FT28] Ronald Aylmer Fisher and Leonard Henry Caleb Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. 24(02):180–190, 1928.

[FXY12] Jiashi Feng, Huan Xu, and Shuicheng Yan. Robust PCA in high-dimension: A deterministic approach. In International Conference on Machine Learning, pages 249–256, 2012.

[GBS05] Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. The Amsterdam library of object images. International Journal of Computer Vision, 61(1):103–112, 2005.

[GDHP02] M Ivette Gomes, Laurens De Haan, and Liang Peng. Semi-parametric estimation of the second order parameter in statistics of extremes. Extremes, 5(4):387–414, 2002.

[GE03] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157–1182, 2003.

[GGBHD04] Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the NIPS 2003 feature selection challenge. In Neural Information Processing Systems, pages 545–552, 2004.

[Gin12] Corrado Gini. Variabilità e mutabilità. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi, 1, 1912.

[GKL03] Anupam Gupta, Robert Krauthgamer, and James R Lee. Bounded geometries, fractals, and low-distortion embeddings. In Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science, pages 534–543. IEEE, 2003.

[GLDW+03] Piet Groeneboom, HP Lopuhaä, PP De Wolf, et al. Kernel-type estimators for the extreme value index. The Annals of Statistics, 31(6):1956–1995, 2003.

[GMN07] M Ivette Gomes, M Joao Martins, and Manuela Neves. Improving second order reduced bias extreme value index estimation. Revstat, 5(2):177–207, 2007.

[Gne43] Boris Gnedenko. Sur la distribution limite du terme maximum d'une série aléatoire. Annals of Mathematics, pages 423–453, 1943.

[GP04] Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. In The Theory of Chaotic Attractors, pages 170–189. Springer, 2004.

[GR70] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. Numerische Mathematik, 14(5):403–420, 1970.

[Gri93] Scott D Grimshaw. Computing maximum likelihood estimates for the Generalized Pareto Distribution. Technometrics, 35(2):185–191, 1993.

[GS87] Charles M. Goldie and Richard L. Smith. Slow variation with remainder: Theory and applications. Quarterly Journal of Mathematics, 38(1):45–71, 3 1987.

[GV88] Edward Guadagnoli and Wayne F Velicer. Relation to sample size to the stability of component patterns. Psychological Bulletin, 103(2):265–275, 1988.

[H+75] Bruce M Hill et al. A simple general approach to inference about the tail of a distribution. The Annals of Statistics, 3(5):1163–1174, 1975.

[HA05] Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in R^d. In International Conference on Machine Learning, pages 289–296. ACM, 2005.

[Hal82] Peter Hall. On some simple estimates of an exponent of regular variation. Journal of the Royal Statistical Society. Series B (Methodological), pages 37–42, 1982.

[Har01] RI Harris. The accuracy of design values predicted from extreme value analysis. Journal of Wind Engineering and Industrial Aerodynamics, 89(2):153–164, 2001.

[HATW98] Joseph F Hair, Rolph E Anderson, Ronald L Tatham, and C William. Multivariate data analysis, 1998.

[Hau18] Felix Hausdorff. Dimension und äußeres maß. Mathematische Annalen, 79(1-2):157–179, 1918.

[HCN05] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In Neural Information Processing Systems, pages 507–514, 2005.

[HDJ+14] Jinrong He, Lixin Ding, Lei Jiang, Zhaokui Li, and Qinghui Hu. Intrinsic dimensionality estimation based on manifold assumption. Journal of Visual Communication and Image Representation, 25(5):740–747, 2014.

[HGV11] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PloS One, 6(12):e28210, 2011.

[HJZB11] Xiaofei He, Ming Ji, Chiyuan Zhang, and Hujun Bao. A variance minimization criterion to feature selection using laplacian regularization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(10):2013–2025, 2011.

[HKN12] Michael E Houle, Hideyuki Kashima, and Michael Nett. Generalized expansion dimension. In 12th International Conference on Data Mining Workshops, pages 587–594. IEEE, 2012.

[HMNO12] Michael E Houle, Xiguo Ma, Michael Nett, and Vincent Oria. Dimensional testing for multi-step similarity search. In 12th International Conference on Data Mining, pages 299–308. IEEE, 2012.

[HMOS14] Michael E Houle, Xiguo Ma, Vincent Oria, and Jichao Sun. Efficient algorithms for similarity search in axis-aligned subspaces. In International Conference on Similarity Search and Applications, pages 1–12. Springer, 2014.

[Hou13] Michael E Houle. Dimensionality, Discriminability, Density & Distance Distributions. In 13th International Conference on Data Mining Workshops, pages 468–473. IEEE, 2013.

[Hou15] Michael E. Houle. Inlierness, Outlierness, Hubness and Discriminability: an Extreme-Value-Theoretic Foundation. Technical Report 2015-002E, National Institute of Informatics, 2015.

[HS99] Mark A Hall and Lloyd A Smith. Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In FLAIRS Conference, volume 1999, pages 235–239, 1999.

[HT85] E Haeusler and Jozef L Teugels. On asymptotic normality of hill's estimator for the exponent of regular variation. The Annals of Statistics, pages 743–756, 1985.

[Hug68] Gordon P Hughes. On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory, 14(1):55–63, 1968.

[HW+85] Peter Hall, AH Welsh, et al. Adaptive estimates of parameters of regular variation. The Annals of Statistics, 13(1):331–341, 1985.

[HW87] Jonathan RM Hosking and James R Wallis. Parameter and quantile estimation for the Generalized Pareto Distribution. Technometrics, 29(3):339–349, 1987.

[IAF16] Ahmet Iscen, Laurent Amsaleg, and Teddy Furon. Scaling group testing similarity search. In ACM International Conference on Multimedia Retrieval 2016, 2016.

[Ike79] Kensuke Ikeda. Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system. Optics Communications, 30(2):257–261, 1979.

[IM98] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In ACM Symposium on Theory of Computing, pages 604–613. ACM, 1998.

[Ish93] Valerie Isham. Statistical aspects of chaos: a review, volume 3. Chap, 1993.

[Jol86] Ian T Jolliffe. Principal Component Analysis. New York, 487, 1986.

[JSF15] Kerstin Johnsson, Charlotte Soneson, and Magnus Fontes. Low bias local intrinsic dimension estimation from expected simplex skewness. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(1):196–202, 2015.

[JTDA11] Hervé Jégou, Romain Tavenard, Matthijs Douze, and Laurent Amsaleg. Searching in one billion vectors: re-rank with source coding. In International Conference on Acoustics, Speech and Signal Processing, pages 861–864. IEEE, 2011.

[Kég02] Balázs Kégl. Intrinsic dimension estimation using packing numbers. In Neural Information Processing Systems, pages 681–688, 2002.

[KG12] Andreas Krause and Daniel Golovin. Submodular function maximization. Tractability: Practical Approaches to Hard Problems, 3:19, 2012.

[KGKB03] George Kollios, Dimitrios Gunopulos, Nick Koudas, and Stefan Berchtold. Efficient biased sampling for approximate clustering and outlier detection in large data sets. IEEE Transactions on Knowledge and Data Engineering, 15(5):1170–1187, 2003.

[KJ94] Juha Karhunen and Jyrki Joutsensalo. Representation and separation of signals using nonlinear pca type learning. IEEE Transactions on Neural Networks, 7(1):113–127, 1994.

[Koh95] Teuvo Kohonen. Learning vector quantization. In Self-Organizing Maps, pages 175–189. Springer, 1995.

[KR96] Marie Kratz and Sidney I Resnick. The qq-estimator and heavy tails. Stochastic Models, 12(4):699–724, 1996.

[KR02] David R Karger and Matthias Ruhl. Finding nearest neighbors in growth-restricted metrics. In ACM Symposium on Theory of Computing, pages 741–750. ACM, 2002.

[Kru56] Joseph B Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical Society, 7(1):48–50, 1956.

[Kru64] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29(1):1–27, 1964.

[LB04] Elizaveta Levina and Peter J Bickel. Maximum likelihood estimation of intrinsic dimension. pages 777–784, 2004.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[LC00] Bernard H Lavenda and Elvio Cipollone. Extreme value statistics and thermodynamics of earthquakes: aftershock sequences. Annals of Geophysics, 43(5), 2000.

[LJM09] Anna V Little, Yoon-Mo Jung, and Mauro Maggioni. Multiscale estimation of intrinsic dimensionality of data sets. In AAAI Fall Symposium: Manifold Learning and its Applications, page 04, 2009.

[LL02] Pedro Larrañaga and Jose A Lozano. Estimation of distribution algorithms: A new tool for evolutionary computation, volume 2. Springer Science & Business Media, 2002.

[LMR16] Anna V Little, Mauro Maggioni, and Lorenzo Rosasco. Multiscale geometric methods for data sets i: Multiscale SVD, noise and curvature. Applied and Computational Harmonic Analysis, 2016.

[LMW79] J Maciunas Landwehr, NC Matalas, and JR Wallis. Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. Water Resources Research, 15(5):1055–1064, 1979.

[LS95] Huan Liu and Rudy Setiono. Chi2: Feature selection and discretization of numeric attributes. In Proceedings of the 7th International Conference on Tools with Artificial Intelligence, page 388. IEEE, 1995.

[LZ08] Tong Lin and Hongbin Zha. Riemannian manifold learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(5):796–809, 2008.

[LZS+] Wen Li, Ying Zhang, Yifang Sun, Wei Wang, Wenjie Zhang, and Xuemin Lin. Nearest neighbor search on high dimensional data experiments, analyses, and improvement.

[Mal98] Edward C Malthouse. Limitations of nonlinear PCA as performed with generic neural networks. IEEE Transactions on Neural Networks, 9(1):165–173, 1998.

[Man67] Benoıt B Mandelbrot. How long is the coast of britain? Science, 156(3775):636–638, 1967.

[Mar87] Jacques Marion. Mesures de hausdorff d'ensembles fractals. In Annales des sciences mathématiques du Québec, volume 11, pages 111–132. Université du Québec à Montréal, Département de mathématiques et informatique, 1987.

[Mas82] David M Mason. Laws of large numbers for sums of extreme values. The Annals of Probability, pages 754–764, 1982.

[MG05] David J.C. MacKay and Zoubin Ghahramani. Comments on 'maximum likelihood estimation of intrinsic dimension' by E. Levina and P. Bickel. http://www.inference.phy.cam.ac.uk/mackay/dimension/, 25, 2005.

[Mil04] José del R Millán. On the need for on-line learning in brain-computer interfaces. In Proceedings of the IEEE International Joint Conference on Neural Networks, volume 4, pages 2877–2882. IEEE, 2004.

[MM10] Philippos Mordohai and Gérard Medioni. Dimensionality estimation, manifold learning and function approximation using tensor voting. Journal of Machine Learning Research, 11(Jan):411–450, 2010.

[MS94] Thomas Martinetz and Klaus Schulten. Topology representing networks. IEEE Transactions on Neural Networks, 7(3):507–522, 1994.

[MS10] Erik Mooi and Marko Sarstedt. Cluster analysis. Springer, 2010.

[MSR+00] Peter J McNulty, Leif Z Scheick, David R Roth, Michael G Davis, and Michelle RS Tortora. First failure predictions for EPROMs of

the type flown on the MPTB satellite. IEEE Transactions on Nuclear Science, 47(6):2237–2243, 2000.

[MWZH99] Robert C MacCallum, Keith F Widaman, Shaobo Zhang, and Sehee Hong. Sample size in factor analysis. Psychological Methods, 4(1):84, 1999.

[Net14] Michael Nett. Intrinsic Dimensional Design and Analysis of Similarity Search. PhD thesis, University of Tokyo, 2014.

[NO09] Satoshi Niijima and Yasushi Okuno. Laplacian linear discriminant analysis approach to unsupervised feature selection. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 6(4):605–614, 2009.

[OC09] Jason W Osborne and Anna B Costello. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Pan-Pacific Management Review, 12(2):131–146, 2009.

[OGFA06] Orlando Oliveira, Maria Ivette Gomes, and Maria Isabel Fraga Alves. Improvements in the estimation of a heavy tail. Revstat, 4(2):81–109, 2006.

[Ott02] Edward Ott. Chaos in dynamical systems. Cambridge University Press, 2002.

[PBJD79] Karl W Pettis, Thomas A Bailey, Anil K Jain, and Richard C Dubes. An intrinsic dimensionality estimator from near-neighbor information. IEEE Transactions on Pattern Analysis and Machine Intelligence, (1):25–37, 1979.

[Pea01] Karl Pearson. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11):559–572, 1901.

[Pen98] L Peng. Asymptotically unbiased estimators for the extreme-value index. Statistics & Probability Letters, 38(2):107–115, 1998.

[Pes00] Vladimir Pestov. On the geometry of similarity search: dimensionality curse and concentration of measure. Information Processing Letters, 73(1):47–51, 2000.

[Pes08] Vladimir Pestov. An axiomatic approach to intrinsic dimension of a dataset. IEEE Transactions on Neural Networks, 21(2):204–213, 2008.

[PI75] James Pickands III. Statistical inference using extreme order statistics. The Annals of Statistics, pages 119–131, 1975.

[PLC13] Haoruo Peng, Ding Liang, and Chinchul Choi. Evaluating parallel logistic regression models. In Big Data, pages 119–126. IEEE, 2013.

[PLD05] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8):1226–1238, 2005.

[Pri57] Robert Clay Prim. Shortest connection networks and some generalizations. Bell System Technical Journal, 36(6):1389–1401, 1957.

[Ran71] William M Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66(336):846–850, 1971.

[Rao09] C Radhakrishna Rao. Linear statistical inference and its applications, volume 22. John Wiley & Sons, 2009.

[RCN+16] Simone Romano, Oussama Chelly, Vinh Nguyen, James Bailey, and Michael E Houle. Measuring dependency via intrinsic dimensionality. 2016.

[Rén59] Alfréd Rényi. On the dimension and entropy of probability distributions. Acta Mathematica Academiae Scientiarum Hungarica, 10(1-2):193–215, 1959.

[RL05] Maxim Raginsky and Svetlana Lazebnik. Estimation of intrinsic dimensionality using high-rate vector quantization. In Advances in Neural Information Processing Systems, pages 1105–1112, 2005.

[RLC+12] Alessandro Rozza, Gabriele Lombardi, Claudio Ceruti, Elena Casiraghi, and Paola Campadelli. Novel high intrinsic dimensionality estimators. Machine Learning Journal, 89(1-2):37–65, 2012.

[RLR+11] Alessandro Rozza, Gabriele Lombardi, Marco Rosa, Elena Casiraghi, and Paola Campadelli. Idea: intrinsic dimension estimation algorithm. In International Conference on Image Analysis and Processing, pages 433–442. Springer, 2011.

[RNI10a] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research, 11:2487–2531, 2010.

[RNI10b] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Time-series classification in many intrinsic dimensions. In Proceedings of the 2010 SIAM International Conference on Data Mining, pages 677–688. Citeseer, 2010.

[Rob00] Stephen J Roberts. Extreme value statistics for novelty detection in biomedical data processing. Proceedings of Science, Measurement and Technology, 147:363–367, 2000.

[RS00] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500):2323–2326, 2000.

[RSN72a] AK Romney, RN Shepard, and SB Nerlove. Multidimensionaling Scaling, volume I, Theory. volume 1. Seminar Press, 1972.

[RSN72b] AK Romney, RN Shepard, and SB Nerlove. Multidimensionaling Scaling, volume II, Applications. volume 2. Seminar Press, 1972.

[Sam69] John W Sammon. A nonlinear mapping for data structure analysis. IEEE Transactions on Computers, (5):401–409, 1969.

[SHH99] Colin Studholme, Derek LG Hill, and David J Hawkes. An overlap invariant entropy measure of 3d medical image alignment. Pattern Recognition, 32(1):71–86, 1999.

[SIL07] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19):2507–2517, 2007.

[Smi87] Richard L Smith. Estimating tails of probability distributions. The Annals of Statistics, pages 1174–1207, 1987.

[Smi88] Leonard A Smith. Intrinsic limits on dimension calculations. Physics Letters A, 133(6):283–288, 1988.

[SR06] Uri Shaft and Raghu Ramakrishnan. Theory of nearest neighbors indexability. ACM Transactions on Database Systems, 31(3):814–838, 2006.

[SSM98] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10(5):1299–1319, 1998.

[SSS78] Lynn Arthur Steen, J Arthur Seebach, and Lynn A Steen. Counterexamples in topology, volume 18. Springer, 1978.

[Tak85] Floris Takens. On the numerical determination of the dimension of an attractor. Springer, 1985.

[TB99] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(3):611–622, 1999.

[TC00] Robert G Tryon and Thomas A Cruse. Probabilistic mesomechanics for high cycle fatigue life prediction. Journal of Engineering Materials and Technology, 122(2):209–214, 2000.

[TDSL00] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500):2319–2323, 2000.

[The88] James Theiler. Lacunarity in a best estimator of fractal dimension. Physics Letters A, 133(4):195–200, 1988.

[The90] James Theiler. Statistical precision of dimension estimators. Physical Review A, 41(6):3038, 1990.

[TP07] Valentin Todorov and Ana M Pires. Comparative performance of several robust linear discriminant analysis methods. REVSTAT Statistical Journal, 5:63–83, 2007.

[vBHZ15] Jonathan von Brünken, Michael E. Houle, and Arthur Zimek. Intrinsic Dimensional Outlier Detection in high-dimensional data. Technical Report 2015-003E, National Institute of Informatics, 2015.

[VD95] Peter J. Verveer and Robert P. W. Duin. An evaluation of intrinsic dimensionality estimators. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(1):81–86, 1995.

[VK06] Jarkko Venna and Samuel Kaski. Local multidimensional scaling. IEEE Transactions on Neural Networks, 19(6):889–899, 2006.

[Wei78] Ishay Weissman. Estimation of parameters and large quantiles based on the k largest observations. Journal of the American Statistical Association, 73(364):812–815, 1978.

[WG14] Bo Wang and Anna Goldenberg. Gradient-based laplacian feature selection. CoRR, abs/1404.2948, 2014.

[WKN13] Randall Wald, Taghi Khoshgoftaar, and Antonio Napolitano. The importance of performance metrics within wrapper feature selection. In 14th International Conference on Information Reuse and Integration, pages 105–111. IEEE, 2013.

[WKV06] Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. A nearest-neighbor approach to estimating divergence between continuous random vectors. Convergence, 1000(1):11, 2006.

[WM+08] Xiaohui Wang, J Steve Marron, et al. A scale-based approach to finding effective dimensionality in manifold learning. Electronic Journal of Statistics, 2:127–148, 2008.

[WSB98] Roger Weber, Hans-Jörg Schek, and Stephen Blott. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In 24th International Conference on Very Large Data Bases, volume 98, pages 194–205, 1998.

[XXZC08] Jun-Ling Xu, Bao-Wen Xu, Wei-Feng Zhang, and Zi-Feng Cui. Principal component analysis based feature selection for clustering. In International Conference on Machine Learning and Cybernetics, volume 1, pages 460–465. IEEE, 2008.

[ZCC+08] Jing Zheng, Wenguang Chen, Yurong Chen, Yimin Zhang, Ying Zhao, and Weimin Zheng. Parallelization of spectral clustering algorithm on multi-core processors and gpgpu. In ACSAC, pages 1–8. IEEE, 2008.

[ZHMY13] Ye Zonglin, Cao Hui, Wang Miaomiao, and Zhang Yanbin. An improved density-based cluster analysis method combining genetic algorithm and data sampling for large-scale datasets. In CCC, pages 3552–3555. IEEE, 2013.

[ZL07] Zheng Zhao and Huan Liu. Spectral feature selection for supervised and unsupervised learning. In International Conference on Machine Learning, pages 1151–1157. ACM, 2007.

[ZZ04] Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. Journal of Shanghai University (English Edition), 8(4):406–424, 2004.