International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 34, No. 9 (2020) 2058010 (35 pages)

DOI: 10.1142/S0218001420580100

Practical Considerations on Nonparametric Methods for Estimating Intrinsic Dimensions of Nonlinear Data Structures

Jochen Einbeck*

Department of Mathematical Sciences
and Institute for Data Science
Durham University, UK
jochen.einbeck@durham.ac.uk

Zakiah Kalantan

Department of Statistics, King Abdulaziz University
Jeddah, Saudi Arabia
zkalanten@kau.edu.sa

Uwe Kruger

Department of Biomedical Engineering
Rensselaer Polytechnic Institute, Troy, NY, USA
krugeu@rpi.edu

Received 13 October 2015
Accepted 9 September 2019
Published 20 November 2019

This paper develops readily applicable methods for estimating the intrinsic dimension of multivariate datasets. The proposed methods, which make use of theoretical properties of the empirical distribution functions of (pairwise or pointwise) distances, build on the existing concepts of (i) correlation dimensions and (ii) charting manifolds that are contrasted with (iii) a maximum likelihood technique and (iv) other recently proposed geometric methods including MiND and IDEA. This comparison relies on application studies involving simulated examples, a recorded dataset from a glucose processing facility, as well as several benchmark datasets available from the literature. The performance of the proposed techniques is generally in line with other dimension estimators, specifically noting that the correlation dimension variants perform favorably to the maximum likelihood method in terms of accuracy and computational efficiency.

Keywords: Intrinsic dimension; projection technique; correlation dimension; charting manifold; maximum likelihood estimation; fractal dimension.

* Corresponding author.

1. Introduction

Recently, nonparametric concepts to extract m feature components embedded within a set of M recorded variables have gained interest in the scientific community.[23] In a nonparametric context, estimating the *intrinsic dimension* (ID), which can be integer- or real-valued, is challenging. The research literature has proposed several conceptual approaches for this problem, including fractal dimensions,[11,24] charting manifolds[6] and maximum likelihood (ML) dimension.[37] This paper develops methods on the basis of these existing concepts.

For more traditional parametric models, an often observed situation is that a particular variable may contain information that is encapsulated in other variables too. Thus, the variables are interrelated which allows describing them by a reduced set of $m \in \mathbb{N}$ *latent variables*, with m being the ID. Related (unsupervised) models, consequently, discriminate between significant and residual information and are, conceptually, of one of the following forms[27,57]:

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{r}, \tag{1a}$$

$$\mathbf{x} = \phi(\mathbf{s}) + \mathbf{r}. \tag{1b}$$

Here, $\mathbf{x} \in \mathbb{R}^M$ is the data vector, $\mathbf{s} \in \mathbb{R}^m$ stores the latent variables and $\mathbf{r} \in \mathbb{R}^M$ is a *noise* vector. The assumptions for the data models in Eq. (1) are as follows:

Assumption 1: $E\{\mathbf{x}\} = E\{\mathbf{r}\} = \mathbf{A}E\{\mathbf{s}\} = E\{\phi(\mathbf{s})\} = \mathbf{0}$;
Assumption 2: $E\{x_i^2\} > E\{r_j^2\}, \forall 1 \leq i, j \leq M$.

Here, $E\{\cdot\}$ is the expectation operator. Assumption 2 is required to ensure an insignificant loss of information,[8,21] with some works relying on the more restrictive assumption $E\{x_i^2\} \gg E\{r_j^2\}$.[35] The vector $\mathbf{s}$ describes common trends in $\mathbf{x}$. The assumption $E\{\mathbf{x}\} = \mathbf{0}$ is not a restriction of generality, as the offset term $\boldsymbol{\kappa} = -E\{\mathbf{x}\}$ can be added, such that $E\{\mathbf{x} + \boldsymbol{\kappa}\} = \mathbf{0}$. Equation (1a) describes a linear relationship between $\mathbf{s}$ and significant information in $\mathbf{x}$ through the use of a model subspace, defined by the column space of $\mathbf{A}$. Equation (1b) is a nonlinear extension of Eq. (1a), where the nonlinear transformation of $\mathbf{s}$ describes significant information in $\mathbf{x}$. Throughout this paper, we denote the density functions of $\mathbf{x}$ and $\mathbf{s}$ by f and g, respectively.

1.1. *Parametric intrinsic dimension estimation*

To estimate $m \in \mathbb{N}$ for Eq. (1a), a plethora of methods have been proposed over the past decades. The top left section in Table 1 summarizes a subset of methods that gained attention in the literature, most of which relate to the application of principal component analysis (PCA) to estimate the column space of $\mathbf{A}$ and rely on various assumptions. Depending on the assumptions imposed on $\mathbf{r}$, the variance of the reconstruction error[43] (VRE) and the equality of eigenvalues test for maximum likelihood PCA[20] provide consistent estimations of m. The eigen decomposition of the scaled covariance matrix $E\{\mathbf{x}\mathbf{x}^T\}$ gives a consistent estimation of the column space

Table 1. Overview of techniques to estimate m. Methods which are considered in this paper are printed in *italic* and novel contributions are additionally denoted in ***bold***.

Traditional Parametric Approaches	
Linear Eq. (1a)	Nonlinear Eq. (1b)

• *eigenvalue-based* (*linear projection*)[31,35,55]
  ◦ e.g. *PCA*[20]
• *cross-validation-based*[35] (e.g. *VRE*[43])
• information-based[35]
• *Velicer's partial correlation*[55]
• *Probabilistic/Bayesian PCA*[4,52]
	• eigenvalue-based (nonlinear projection)
	  ◦ KPCA[15,46]
	  ◦ autoassociative neural networks[2,34,58]
	• cross-validation-based[47]
	• residual-based[3,2]
	• H-principle[30]

More Recently Proposed Nonparametric Approaches	

Global approaches
• MDS
  ◦ via stress functions[8]
  ◦ ISOMAP[50]
• geometric/"correction" methods
  ◦ *IDEA*[45]
  ◦ *MiND*,[45] DANCo[13]
• *fractal-based concepts*
  ◦ box-counting[33]
  ◦ Taken's method[48]
  ◦ *correlation-dimension concepts*[11,51]
  – *slope method* (*log-log-plot*)[8]
  – ***intercept method***
  – ***polynomial method***
  – "*kernel*" *correlation integral*[29]
	Local approaches
	• local eigenvalues/PCA
	  ◦ Fukunaga–Olsen approach[21]
	  ◦ local eigenvalue algorithm[57]
	  ◦ topology representing networks[7,42]
	• near-neighborhood approaches
	  ◦ near-neighbor algorithm,[57] *kNN*[12]
	  ◦ *graph-based methods*[12,28]
	  ◦ localized representation learning[14,18]
	• *ML*[37]
	• *charting concepts*[6]
	  ◦ ***dip method***
	  ◦ ***regression method***

of $\mathbf{A}$, even if $\mathbf{s}$ does not follow a normal distribution.[22,38] Consistency for estimating m, however, is only guaranteed if $E\{\mathbf{s}\mathbf{r}^T\} = \mathbf{0}$.

Work on estimating the ID for the model structure in Eq. (1b) mainly relies on nonlinear extensions of PCA and includes autoassociative neural networks as well as kernel PCA (KPCA). The latter approach can extract nonlinear principal component scores using the same objective function as PCA.[36] Various approaches to estimate $m \in \mathbb{N}$ have been considered, including cross-validation, an analysis of the residual variance or the H-principle; see Table 1 (top, right). The suitability of nonlinear PCA for ID estimation, however, has been disputed.[41]

1.2. *Nonparametric intrinsic dimension estimation*

To develop nonparametric estimation methods, this paper considers concepts that do not assume, or make use of, the data model in Eq. (1). The underlying mechanism for generating data, however, may still follow Eq. (1). To provide a general framework for estimating m, we assume here that $m \in \mathbb{N}$ or $m \in \mathbb{R}^+$.

Multidimensional scaling (MDS)[5] is among the family of nonparametric approaches, listed in Table 1. Conceptually, MDS (just as PCA/NLPCA) is a *dimension reduction* rather than *dimension estimation* method, requiring *ad hoc*

rules, such as a "knee" in the stress function, to determine an integer-valued m.[34,36,58] Camastra[8] argued that estimating m in this way may be difficult since a distinct "knee" does not always exist.

It is, hence, preferable to have tailored ID estimation methods available. A suitable family of techniques is given by the *fractal dimension*, which includes, for example, box-counting and correlation-dimension. Associated concepts estimate $m \in \mathbb{R}^+$ directly by adapting methods from the chaos theory to determine the dimension of attractors of real datasets.[24] All methods discussed above rely on the entire dataset and are, therefore, *global methods*,[8] an overview of which is provided on the left side of Table 1.

Alternatively, the literature proposed *local methods*, which identify the *topological dimension* locally as the dimension of the tangent space along the data at a specific target point.[8,10] An early instance is the work by Fukunaga and Olsen,[21] where m is estimated to be the number of normalized nonzero eigenvalues of region-wise covariance matrices. A variety of alternative local methods have been developed since then; some of which are listed at the bottom right part of Table 1.

The presentation in Table 1 differs in some aspects from alternative categorizations. For instance, Camastra and Staiano[10] consider ISOMAP to be a *local* method, on the grounds that it makes use of a local variant of MDS. However, since it produces a single global ID estimate, we advocate considering it as a *global* method. More precisely, our classification criterion is as follows: Local methods produce (possibly, multiple) local ID estimates and global methods produce (a single) global ID estimates. It is important to note, however, that some of the local methods in Table 1 average over local ID estimates in order to derive an overall ID estimate.[28,37] Following this line of reasoning, our classification does not require the additional classification category of *pointwise*[10] methods.

Following the preceding discussion, the literature has reported substantial progress in estimating the ID of multivariate datasets in recent years, as evidenced by the methods and citations provided in Table 1. However, several of these techniques have, thus far, been proposed as concepts rather than a suite of tailored methods. This is, specifically, the case for the (global) correlation dimension, as well as the (local) charting technique. Related problems for their practical implementation include:

- fractal concepts require the computation of the correlation integral for a sphere with radius $r \to 0$, which is computationally nontrivial;
- charting manifold techniques struggle with multiple practical issues such as zeros in the denominator, multiple peaks in the objective function, and the aggregation of local estimates of m to produce a global estimate.

The aim of this paper is to address these problems by

(i) developing methods for the correlation dimension and the charting manifold concepts;

(ii) benchmarking their performance;
(iii) contrasting them with existing work, with particular focus not only on the MLE method, but also on other recent techniques, such as MiND or IDEA.[45]

We define the techniques related to correlation dimension as the *slope*, *intercept* and *polynomial* methods and those associated with the charting manifolds as *dip* and *regression* methods. While the *slope* method can be considered as a quantitative variant of the existing log–log plot technique,[8] the remaining methods are novel.

The paper is organized as follows: Sec. 2 summarizes the correlation dimension and the charting manifold concepts. Building on these concepts, Sec. 3 introduces the proposed global correlation dimension and the local charting manifold methods. This is followed by contrasting the developed methods with the MLE method in Sec. 4, which summarizes the application studies to simulated data, the analysis of a recorded dataset from an industrial glucose processing plant and three benchmark datasets that are available in the literature. This section also includes an analysis of the computational efficiency of the different techniques. Finally, Sec. 5 provides a concluding summary of this paper.

2. Concepts for Determining Intrinsic Dimensions

This section briefly revises the (global) correlation dimension and the (local) charting manifold concept[a] in Secs. 2.1 and 2.2, respectively. For the comparison in Sec. 4, Sec. 2.3 briefly summarizes the MLE approach.[37] For the remainder of this paper, $\mathbf{\Omega} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ denotes a M-variate dataset[b] containing n samples $\mathbf{x}_i = (x_{i1}, \ldots, x_{iM})^T$, $i = 1, \ldots, n$.

2.1. *Correlation dimension concept*

Correlation dimension is a fractal-based concept, which has been successfully employed to estimate the attractor dimension of dynamic systems.[24] It can be seen as a simple substitute of the box-counting dimension, which, in turn, corresponds to the Hausdorff dimension.[11] The concept is based on the empirical distribution function (EDF) of pairwise distances:

$$C_n(r) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathcal{I}\{\|\mathbf{x}_j - \mathbf{x}_i\| \leq r\}, \tag{2}$$

where $\mathcal{I}\{\cdot\}$ is the indicator function, that is 1 if $\|\cdot\| \leq r$ and 0 if $\|\cdot\| > r$, $\|\mathbf{x}_j - \mathbf{x}_i\|$ denotes the Euclidean distance between the samples $\mathbf{x}_j$ and $\mathbf{x}_i$ and the subscript for C_n refers to the number of samples in the reference set $\mathbf{\Omega}$. Essentially, Eq. (2) counts

[a] Brand's concept of "charting" has a broader scope and involves the construction of a patch-wise, low-dimensional coordinate system. This paper still uses the term "charting" for convenience.

[b] If the data are originally presented in the form of a time series, and the task is to identify the attractor dimension of the dynamical system underlying the time series, then the dataset $\mathbf{\Omega}$ needs to be firstly produced from the time series data via the "method of delays".[9, 45]

the number of pairs whose mutual distance is less than or equal to the radius r. The function $C_n(r)$ tends to 0 monotonically as $r \to 0$. Based on Eq. (2), the *correlation integral* is defined as

$$C(r) = \lim_{n\to\infty} C_n(r), \tag{3}$$

from which the correlation dimension can be extracted as follows:

$$m = \lim_{r\to 0} \frac{\log(C(r))}{\log r}. \tag{4}$$

Here, log denotes the natural logarithm function. Two important practical issues that immediately arise are the asymptotic limits in Eqs. (3) and (4). While these limits are necessary from a conceptual point of view, none of them are attainable in practice. We can practically assume, however, that if n is large enough, $C_n(r)$ can replace $C(r)$ in Eq. (4). As a guideline for selecting n, Eckmann and Ruelle[16] showed that estimation of m via correlation dimension requires at least $n = 10^{m/2}$.

The more serious issue is that the correlation integral needs to be evaluated for a sphere of radius $r \to 0$. For any finite dataset without replicated cases, $C_n(r) = 0$ as $r \to 0$. Consequently, the numerator of Eq. (4) is practically undefined for small enough radii. Hein and Audibert[29] addressed this problem by replacing the indicator function in (2) by a kernel function, and basing the dimension estimation on the speed of convergence of the correlation integral, rather than the correlation integral itself. When working with (2) directly, as in this paper, it is of practical importance to predefine a suitable range of values of r which allows an accurate estimation of m.[8,51]

Thus far, such "direct" algorithms to estimate m have been based mainly on the analysis of the log–log plot,[8] which attempts to determine the slope of $\log(C(r))$ as a function of $\log r$. This paper considers a quantitative version of the log-log technique, and also proposes two novel methods, which use appropriate modeling techniques to estimate m for a given dataset $\mathbf{\Omega}$ at a radius $r = 0$. These two methods, introduced in Sec. 3.1, which we refer to as the *intercept* and the *polynomial* methods, tackle the problem by exploiting features of the functions in Eqs. (3) and (4). The following remarks motivate their development.

Remark 1. To reflect why Eq. (4) is a correct relationship between the correlation integral and the ID, consider a structure which lies (perfectly) on some linear hyperspace or nonlinear surface of $\mathbf{\Omega}$. It then follows (further discussed in Remark 2) that $C(r) \propto r^m$ for a sufficiently small r. More precisely, the relationship between $C(r)$ and r for a sufficiently small r becomes

$$C(r) = c \cdot r^m,$$

where m is the ID and c is constant. Now, applying the logarithm to the above equality yields

$$\log(C(r)) = \log c + m \log r. \tag{5}$$

Next, substituting Eq. (5) into Eq. (4) gives rise to

$$\lim_{r\to 0}\frac{\log(C(r))}{\log r}=\lim_{r\to 0}\frac{\log c+m\log r}{\log r}=m. \tag{6}$$

Consequently, the correlation dimension asymptotically reveals the ID of the dataset.

Remark 2. We now justify the assumption $C(r)\propto r^m$. Consider a simple scenario in which n samples sit at predefined discrete positions (with distance 1) along a line:

• • • • … • •

For $r=0$, the double sum in the numerator of Eq. (2) is equal to 0. For $r=1$, this sum is $n-1$, and for $r=2$, it is $(n-1)+(n-2)$. Generally, this sum is $(n-1)+(n-2)+\cdots+(n-r)\approx nr\propto r$ for large n, confirming that $C(r)\propto r^m$ for the case $m=1$. For a one-dimensional curve, this statement would still hold for a sufficiently small r. These simple geometric considerations cannot be extended easily to higher dimensions $m\geq 2$. Even the case of $m=2$ requires complex graph theory. We, therefore, take a different line of reasoning. Recall that Eq. (2) describes an EDF of pairwise distances. According to the strong law of large numbers, the EDF will converge to the true distribution function (DF) of the pairwise distances for large n.

Considering the simple case of a uniform data distribution inside a sphere, it can be shown[1,25] that the corresponding DF is $a_m b(m,r)r^m+c(m,r)$, where $a_m>0$ is a constant, depending on m, and $a(m,r)$ and $b(m,r)$ are regularized incomplete Beta functions depending on m and r, with the properties $b(m,r)\approx 1$ and $c(m,r)\approx 0$ for small r. From this, it can be concluded that $C(r)\propto r^m$ if r is small enough. More recent research provides extensions to more general data distributions.[53]

2.2. *Charting manifold concept*

This concept estimates m by examining the rate of growth of samples in hyperspheres. Different from the previous concept, the charting manifold technique counts *points* rather than *pairs*, and does this *locally* instead of *globally*. Let $\mathbf{x}$ be an element of $\mathbf{\Omega}$, which is not assumed to be located on the boundary of the manifold. We refer to this point as a "target point" henceforth. The charting manifold relies on the proportion[c] of points that fall inside a sphere of radius r that is centered at $\mathbf{x}$

$$N_{\mathbf{x}}(r)=\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}\{||\mathbf{x}_i-\mathbf{x}||\leq r\}. \tag{7}$$

As before, $\mathcal{I}\{\cdot\}$ is the indicator function. The subscript for $N_{\mathbf{x}}$ emphasizes that this is a local estimate that depends on the center, $\mathbf{x}$, of the sphere. Brand[6] argued that if r falls below the *noise scale*, then $N_{\mathbf{x}}(r)\propto r^M$. Moreover, if the underlying manifold is sufficiently smooth, then there will be a scale r at which the manifold can be

[c]Brand does not use the constant $1/n$ preceding the sum in (7). It is, however, useful to interpret $N_{\mathbf{x}}(r)$, analogously to $C(r)$, as the EDF of distances to $\mathbf{x}$. This constant does not affect the developments which follow.

approximated by a *locally* linear hyperplane of dimension m. We may refer to this radius, say r_0, as the *signal level*, at which the points are distributed only in the directions of the local tangent space (hyperplane) of the manifold. Consequently, in a neighborhood of r_0, it follows that $N_{\mathbf{x}}(r) \propto r^m$. Increasing r further, the curvature of the manifold becomes significant so that $N_{\mathbf{x}}(r)$ rises at a rate between r^m and r^M. When reaching the boundary that encloses all data, $N_{\mathbf{x}}(r)$ eventually flattens and naturally approaches 1.

Brand[6] defined the statistic

$$G_{\mathbf{x}}(r) = \frac{d \log r}{d \log(N_{\mathbf{x}}(r))}, \tag{8}$$

for determining the radius r_0 and hence, reveals the intrinsic structure. It then follows from the above considerations that, for noise scales,[6]

$$G_{\mathbf{x}}(r) \approx \frac{1}{M} < \frac{1}{m}.$$

This, in turn, suggests that plotting $G_{\mathbf{x}}(r)$ versus r produces a maximum at the signal level of $1/m$. Hence, this *peak* gives the intrinsic (topological) dimension m. Although this concept is appealing, its implementation may be cumbersome and nontrivial. Practical applications may, for example, be hampered by the following:

(i) the choice of the range of r values investigated for this purpose;
(ii) the existence of the expression $\log(N_{\mathbf{x}}(r))$ in the denominator (possibly undefined for small r);
(iii) the choice of target points, $\mathbf{x}$;
(iv) the existence of multiple peaks in the graph $G_{\mathbf{x}}(r)$ versus r; and
(v) how to synthesize or average the individual estimates of m for the different target points.

While items (i) and (ii) have been noted in a similar form for the correlation dimension concept, items (iii)–(v) are intrinsic to the charting manifold concept. Section 3 proposes two novel variants of Brand's 6 conceptual algorithm, which implicitly address the above issues. To discriminate the charting manifold concept from the correlation dimension one, introduced in Sec. 2.1, we give the following remark.

Remark 3. As an alternative definition, let $C_{\mathbf{x}}(r)$ denote the number of **pairs** inside the sphere of radius r, centered at $\mathbf{x}$. Then, at the signal level, we get

$$C_{\mathbf{x}}(r) \propto \binom{N_{\mathbf{x}}(r)}{2} = \mathcal{O}(r^{2m}),$$

with $\mathcal{O}(\cdot)$ denoting the order. Hence, as the number of *data points* within the sphere of radius r increases with r^m, the number of *pairs* increases with $\mathcal{O}((r^m)^2) \propto r^{2m}$. The resulting ID estimates using $C_{\mathbf{x}}(r)$, therefore, would need to be divided by 2. In

light of this, the conclusions of Remark 2 may appear counter-intuitive. However, note that in the context of the correlation dimension, pairs are counted *globally*, leading to the order r^m, whereas here they are counted *locally*, resulting in the order r^{2m}. Although this paper does not utilize $C_{\mathbf{x}}(r)$, it is important to understand this fundamental conceptual difference between the two approaches.

2.3. *Maximum likelihood estimation*

Levina and Bickel[37] proposed this technique for estimating m. As for the charting manifold, a sphere of radius r is considered at a fixed point $\mathbf{x}$. It is assumed that the data stored in $\mathbf{\Omega}$ are independent, stem from the same underlying manifold, and that there exists an embedding of the form $\mathbf{x}_i = \phi(\mathbf{s}_i)$, where $\mathbf{s}_i \in \mathbb{R}^m$ is a sample drawn from the density function $g(\cdot)$, with both $\phi(\cdot)$ and $g(\cdot)$ being smooth functions operating on an m-variate space. These assumptions allow defining $\mathbf{x}_i$ as a homogeneous Poisson process.[37] The log-likelihood of this Poisson formulation yields a ML estimator based on the distances between close neighbors.

Let k be the number of nearest neighbors to the point $\mathbf{x}_i$, then the local ML estimator for m is

$$m_k(\mathbf{x}_i) = \left[\frac{1}{k-2}\sum_{j=1}^{k-1}\log\left(\frac{T_k(\mathbf{x}_i)}{T_j(\mathbf{x}_i)}\right)\right]^{-1}, \tag{9}$$

where $T_k(\mathbf{x}_i)$ and $T_j(\mathbf{x}_i)$ are the Euclidean distances between $\mathbf{x}_i$ and the kth and jth nearest neighboring samples, respectively. To guarantee an asymptotically unbiased estimate, the denominator of Eq. (9) must be $k-2$, as discussed in Sec. 3.1 in Ref. 37. The asymptotics here are $n \to \infty$, $k \to \infty$ and $k/n \to 0$. The local estimates in Eq. (9) need to be suitably combined to produce a global estimate. Levina and Bickel[37] argued that it is unnecessary to remove boundary points for this purpose and proposed utilizing the average

$$m(k) = \frac{1}{n}\sum_{i=1}^{n} m_k(\mathbf{x}_i), \tag{10}$$

as a suitable estimator for fixed k. However, it was subsequently suggested[19,40] that, from a ML perspective, the correct estimator to use is

$$m(k) = \left[\frac{1}{n}\sum_{i=1}^{n} m_k^{-1}(\mathbf{x}_i)\right]^{-1}. \tag{11}$$

In either case, the process is repeated for p values of k, say $k^{(1)}, \ldots, k^{(p)}$, within the data range, and the ID for a dataset $\mathbf{\Omega}$ can be obtained by averaging

$$m = \frac{1}{p}\sum_{j=1}^{p} m(k^{(j)}). \tag{12}$$

As a third variant, one may consider alleviating the bias incurred in (10), especially for small k,[37] by taking the median instead of the mean in (12). We consider all three options in this paper. It is finally noted that a ML point of view can also be adopted for the correlation dimension.[48]

3. Proposed Intrinsic Dimension Estimation Methods

This section gives a detailed description of the developed methods for the correlation dimension and the charting manifold concepts in Secs. 3.1 and 3.2, respectively, extending and formalizing preliminary ideas given in Ref. 17. We make the initial decision to normalize all columns of the dataset $\mathbf{\Omega}$ so that each of the M variables has a sample mean of zero and a sample standard deviation of 1. While this is not strictly necessary in order to apply the proposed methods, it is convenient for computational and comparative purposes, and it allows giving generic recommendations for the choice of range for radii in (2) and (7).

3.1. *Correlation dimension methods*

The three methods for estimating m are (i) the *slope* method, (ii) the *intercept* method and (iii) the *polynomial* method which Secs. 3.1.1–3.1.3 introduce, respectively. As the *slope* method is merely a further development of the log–log plot,[8] the paper only considers the *intercept* and *polynomial* methods as new contributions to knowledge. Each of these methods requires a set of monotonously increasing radii, defined by $\{r_j, j = 1, \ldots, s\}$, where the smallest radius $r_1 > 0$ is large enough to include at least 2 samples. The parameter s, which determines the number of grid points, is of little relevance as long as it is reasonably large, say $s \geq 20$. In majority of applications, one will have $0 < r_1 < r_s \leq 1$, and specific recommended settings of r_1 and r_s will be given for each of the three methods in the respective subsections. It is noted, however, that radii in a M-dimensional space scale with $\sqrt{M}$; that is, for very large M, the minimum radius which contains at least two data points may be considerably higher than the proposed boundaries. In such cases, we recommend to set r_1 equal to this minimum radius, and $r_s = 1.5 \times r_1$.

3.1.1. *Slope method*

According to Eqs. (3) and (4), the finite-sample estimator

$$\lim_{r \to 0} \frac{\log(C_n(r))}{\log r} \approx m \tag{13}$$

gives a good approximation of m for large n. Camastra[11] proposed to plot $\log(C_n(r))$ versus $\log r$ and graphically estimate the slope and, hence, m. To quantify this graphical approach, this paper estimates m from the simple linear regression model

$$\log(C_n(r)) = a + b \log r, \tag{14}$$

using the pairs $(\log r_j, \log(C_n(r_j)))$ for $j = 1, \ldots, s$. It follows from Remark 1 that the estimate of m is

$$\widehat{m}_S = b. \tag{15}$$

For further reference, the subscript S refers to the *slope* method. Through application studies, we found that $r_1 = 0.3$ and $r_s = 0.5$ yield satisfactory results in most situations, though problems may arise when the implicit linearity assumption in (14) fails. Smaller radii than $r_1 = 0.3$ may yield erroneous values for $C_n(r)$ or a flattening curve for $C_n(r)$, producing suboptimal estimates for m.

3.1.2. *Intercept method*

This method approximates the correlation integral directly for $r = 0$ instead of estimating the slope for some small values of r. For this, consider the graph $(r, D_n(r))$, where

$$D_n(r) = \frac{\log(C_n(r))}{\log r}. \tag{16}$$

The advantage of using $D_n(r)$ instead of $C_n(r)$ lies in its approximate linearity, as formulated in the following theorem (proven in Appendix A).

Theorem 1. *In the vicinity of* $r^* = e^{-2} \approx 0.14$, *the function* $D_n(r)$ *reduces to a linear function of the radius* r, *that is*,

$$D_n(r) = a + b(r - r^*). \tag{17}$$

The *intercept* method estimates m by extrapolating the linear regression line, fitted to the pairs $(r_j, D_n(r_j)), j = 1, \ldots, s$. More precisely, the estimate of m is the intercept of the fitted linear equation and the ordinate at $r = 0$. Finally, incorporating the constant $-br^*$ in the coefficient $a^* = a - br^*$ yields the regression equation $D_n(r) = a^* + br$, such that the estimate of m becomes

$$\widehat{m}_I = D_n(0) = a^*, \tag{18}$$

where the subscript I refers to the *intercept method*. The radius r should be $0.14 \leq r_j \leq 0.5$. If the minimum radius, $r_1 = 0.14$ does not result in the inclusion of at least two data points, r_1 needs to be increased. Note, however, that by construction the *intercept* method will not produce meaningful results if $r > 1$ (as otherwise $D_n(r) < 0$); hence, there is the additional restriction $r_s < 1$ and it is possible that for a very large M, the *intercept* method is not applicable. See Sec. 4.3 for some examples.

Notably, although the *intercept* method uses a range of small values of r to determine $D_n(r)$, it estimates the ID *at* the radius $r = 0$. The application studies in Sec. 4 show that the *intercept* method performs, generally, similar to the *slope* method.

3.1.3. *Polynomial method*

This method relies on an explicit model of the correlation integral $C(r)$ through a higher-order polynomial. The model, in conjunction with the property in Eq. (4), allows computing the correlation dimension directly. It is clear from the considerations in Remarks 1 and 2 that, if the data are well described by a manifold of some dimension m, then $C(r)$ will approach 0 as $r \to 0$. This, in turn, motivates the following intuitive condition:

Condition 1. $C(0) = 0$.

The preceding discussion gives rise to the following theorem, which Appendix B proves.

Theorem 2. *Expressing the correlation integral as a polynomial of order q*:

$$C(r) = \sum_{i=0}^{q} a_i r^i = a_0 + a_1 r + a_2 r^2 + \cdots + a_q r^q, \tag{19}$$

under Condition 1, *that is* $a_0 = 0$, *the correlation dimension m is as follows*:

$$\text{if } a_j = 0, j = 1, \ldots, q-1 \quad \text{and} \quad a_q \neq 0, \text{ then } m = q.$$

Theorem 2 suggests to carry out a series of hypothesis tests to estimate m. The parameters $a_1, \ldots, a_q$ can be obtained using multiple linear regression of $C_n(r)$ versus the the powers of r, for example, using the function `lm` in R.[44] To determine whether a parameter is zero, the standard t-test can be utilized. Based on application studies, it is recommended to leave the significance level of this test unspecified and to estimate m as the *most* significant parameter, that is,

$$\widehat{m}_P = \{j | a_j \text{ has minimal } p\text{-value among } a_1, \ldots, a_q\}. \tag{20}$$

The subscript P refers to the *polynomial* method. It is suggested to initially set $q = \min\{M, 4\}$ and increase the integer q successively if required.

Different to the *slope* and *intercept* methods, the *polynomial* method provides an integer estimation and not a "fractal dimension". The choice of the upper limit radius presents a trade-off between the radii being close enough to 0 and large enough to include a sufficient number of samples to guarantee an accurate estimation of the unknown parameters $a_1, \ldots, a_q$. We suggest to set r_1 such that the corresponding sphere contains at least one pair, and $r_s = 1$.

3.2. *Charting manifold methods*

This section develops two methods on the basis of the charting manifold concept, which we refer to as the *dip* method and the *regression* method, detailed in Secs. 3.2.1 and 3.2.2, respectively. Prior to their presentation, two issues need to be discussed.

Firstly, local methods require the selection of a set of target points over which the local ID estimates will be averaged.[51] The largest possible set for this purpose is $\mathbf{\Omega}$,

which is impractical for computational reasons. However, this is also not necessary, since the ID estimates for neighboring points can be expected to be very similar. It remains the issue of how to select such target points, considering that points close to the boundaries may result in underestimating m.[6] One approach could be to estimate the density f, for each $\mathbf{x}_i \in \mathbf{\Omega}$, via

$$\hat{f}(\mathbf{x}_i) = \frac{1}{nh_1 \dots h_M} \sum_{i=1}^{n} \prod_{j=1}^{M} K\left(\frac{x_{ij} - x_j}{h_j}\right), \tag{21}$$

where $K(\cdot)$ is a kernel function and h_j are component-wise bandwidths, and then select sample target points only from the set $\{\mathbf{x}|\hat{f}(\mathbf{x}) > c\}$, for some constant $c > 0$. However, computing this kernel density estimate for a (potentially large and/or high-dimensional) dataset can be computationally inefficient. Hence, we propose a simpler concept based on the notion of *isolated points*.[49] An isolated point is a point which is so far away from the rest of the data that the kernel density estimate at that point is only determined by itself. It is conceptually clear that isolated points are not able to contribute sensible ID estimates. According to Eq. (21), the density of an isolated point is given by $f^* = \frac{1}{nh_1 \dots h_M} K^M(0)$, which is independent of $\mathbf{x}$. If the kernel K has unbounded support, then f^* will rarely be attained exactly so that for our purposes we declare a point as isolated if $\hat{f}(\mathbf{x})/f^* < 2$. The specific choice of K is of little relevance as it does not impact on the ID estimation in itself, but only on the selection of target points. We have used the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$ in the application studies in Sec. 4.

Summarizing, in order to select b target points, we proceed as follows:

(1) Compute the bandwidths $h_j, j = 1, \dots, M$, as 10% of the range of the jth variable.
(2) Select a point, say $\mathbf{x}$, randomly from the dataset.
(3) If $\hat{f}(\mathbf{x})/f^* < 2$, dismiss the selected point.
(4) Iterate between 2 and 3 until b target points have been sampled. Denote the resulting set of points by $\mathcal{B}$.

This procedure is efficient, as only few kernel density estimates need to be computed. Our experiments have shown that $b = 50$ target points are usually sufficient to obtain good overall ID estimates.

Secondly, instead of utilizing the objective function in Eq. (8), it is advantageous for the development of both methods to consider the following alternative formulation:

$$H_{\mathbf{x}}(r) = \frac{1}{G_{\mathbf{x}}(r)} = \frac{\mathrm{d}\log(N_{\mathbf{x}}(r))}{\mathrm{d}\log r}. \tag{22}$$

The rationale behind the definition of $H_{\mathbf{x}}(r)$ is that it is more stable computationally, especially when r tends to 0 or 1, both of which would lead to an undefined

denominator in the case of $G_{\mathbf{x}}(r)$. Furthermore, as will be demonstrated in Sec. 3.2.1, it has the interpretational advantage that the ID can be directly read from its graph.

3.2.1. *Dip method*

Determining the *peak* of $G_{\mathbf{x}}(\cdot)$ is equivalent to obtaining the *dip* of $H_{\mathbf{x}}(\cdot)$. Denoting this extremal value by r_0, it follows from the discussion in Sec. 2.2 that $N_{\mathbf{x}}(r) = cr^m$ in a neighborhood of r_0. Applying the logarithm to this equality yields

$$\log(N_{\mathbf{x}}(r)) = \log c + m \log r. \tag{23}$$

Substituting this expression into $H_{\mathbf{x}}(\cdot)$ for $r = r_0$ gives rise to

$$H_{\mathbf{x}}(r_0) = \frac{\mathrm{d}(\log c + m \log r)}{\mathrm{d} \log r}\bigg|_{r=r_0} = \underbrace{\frac{\mathrm{d} \log c}{\mathrm{d} \log r}}_{=0} + m \frac{\mathrm{d} \log r}{\mathrm{d} \log r}\bigg|_{r=r_0} = m. \tag{24}$$

Therefore, if $H_{\mathbf{x}}(r)$ has a minimum, or a *dip*, for $r = r_0$, then m is given by $H_{\mathbf{x}}(r_0)$. To obtain the required derivative, consider a Taylor expansion of $\log(N_{\mathbf{x}}(r))$ as a function of $\log(r)$, for a value r' close to r:

$$\begin{aligned} \log(N_{\mathbf{x}}(r')) = \; & \log(N_{\mathbf{x}}(r)) + \frac{\mathrm{d}(\log(N_{\mathbf{x}}(r)))}{\mathrm{d}(\log r)} (\log r' - \log r) \\ & + \frac{1}{2} \frac{\mathrm{d}^2(\log(N_{\mathbf{x}}(\rho)))}{\mathrm{d}(\log \rho)^2} (\log r' - \log r)^2, \end{aligned} \tag{25}$$

where $\log \rho$ is in the interval between $\log r'$ and $\log r$. Now, defining kernel weights

$$w_h(r', r) = K\left(\frac{\log r' - \log r}{h}\right), \tag{26}$$

where h is a localization parameter,[26] we can get a smooth estimate of the derivative function $H_{\mathbf{x}}(r)$ for a fixed r-value based on the log-count of $N_{\mathbf{x}}(r_j)$ for the radii $r_1, \ldots, r_j, \ldots, r_s$ by minimizing

$$\sum_{j=1}^{s} w_h(r_j, r)(\log(N_{\mathbf{x}}(r_j)) - \alpha - \beta(\log r_j - \log r) - \gamma(\log r_j - \log r)^2) \tag{27}$$

with respect to α, β and γ, yielding least squares estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$. Comparing Eqs. (25) and (27), it follows that $\hat{\beta} = \hat{\beta}(r)$ is the required estimator for $H_{\mathbf{x}}(r) = \frac{d \log N_x}{d \log r}(r)$. Let us denote this estimator by $\tilde{H}_{\mathbf{x}}(\cdot)$. Notably, Eq. (27) can be evaluated for every fixed r, even if r is not part of the grid points $r_1, \ldots, r_s$.

It is noted that the kernel K used here does not necessarily need to be the same kernel as that one used in (21), but that, in (26), the choice of kernel is indeed important: An unsmooth kernel will impact on the smoothness of the derivative estimate, and hence on the reliability of the ID estimate. So, in (26), we strictly advise the use of a Gaussian kernel.

In practice, examining the function $\tilde{H}_{\mathbf{x}}(r)$ over r will often yield more than one dip, so the question arises which one too choose. Each of these dips can be argued to correspond to a local ID estimate at a different scale. We have settled on choosing the *maximal* dip (that is, among all the minima of $H_{\mathbf{x}}(\cdot)$, we choose that one of maximal value). The reasoning for this is twofold: Firstly, often, there will be some initial dips caused by little granularities in the data, and secondly, if there is evidence for different local dimensions at different scales, it is arguable that the larger dimension estimates supersede the smaller ones. Denoting the position of this "maximal dip" by $r = r_0$, one gets the local ID estimate $m_D(\mathbf{x}) = \tilde{H}_{\mathbf{x}}(r_0)$, with the subscript D denoting the *dip* method. Applying this procedure for each target point $\mathbf{x} \in \mathcal{B}$ allows determining the overall estimate $\widehat{m}_D$ as the median over each local estimate:

$$\widehat{m}_D = \text{med}\{m_D(\mathbf{x}) | \mathbf{x} \in \mathcal{B}\}. \tag{28}$$

3.2.2. *Regression method*

There is a similarity between Eq. (5) in Remark 1 and Eq. (23), and as a consequence also between Eqs. (6) and (24). This motivates applying a log–log analysis[8] on $N_{\mathbf{x}}(r)$ in a similar fashion to the *slope* method detailed in Sec. 3.1.1. For a given $\mathbf{x}$ and a range of r values, one initially computes the values $N_{\mathbf{x}}(r)$. Then, fitting the parameters of the equation

$$\log(N_{\mathbf{x}}(r)) = a + b \log r \tag{29}$$

using simple linear regression produces least squares estimates $\hat{a}$ and $\hat{b}$, yielding the local estimate

$$m_R(\mathbf{x}) = \hat{b}. \tag{30}$$

While this method does have the advantage of not requiring the selection of bandwidths or other tuning parameters, there is a caveat to this line of reasoning: unlike the case of the correlation integral, it is not possible to assume that Eq. (23) holds true for any small value of r. More precisely, Eq. (23) is only valid at the signal level, r_0. However, a possible strategy is to select a range of r values that contain r_0. This paper uses a range of radii such that the minimum radius contains at least two data points, and the maximum radius all data points.[d]

Again, this is a local method, which needs to be applied for each target point, and the local estimates need to be suitably averaged, using a median, to arrive at the overall estimate:

$$\widehat{m}_R = \text{med}\{m_R(\mathbf{x}) | \mathbf{x} \in \mathcal{B}\}. \tag{31}$$

The subscript R denotes the method used for the estimate, that is, the *regression* method in this case. The reference *regression* is to distinguish this method from the

[d]This wide range could possibly be finetuned in future research. However, as elaborated upon later on in this paper, this large range yields positive effects in terms of robustness to local granularities in the data.

conventional log–log plot approach. Experimental results for this method are provided in Sec. 4.

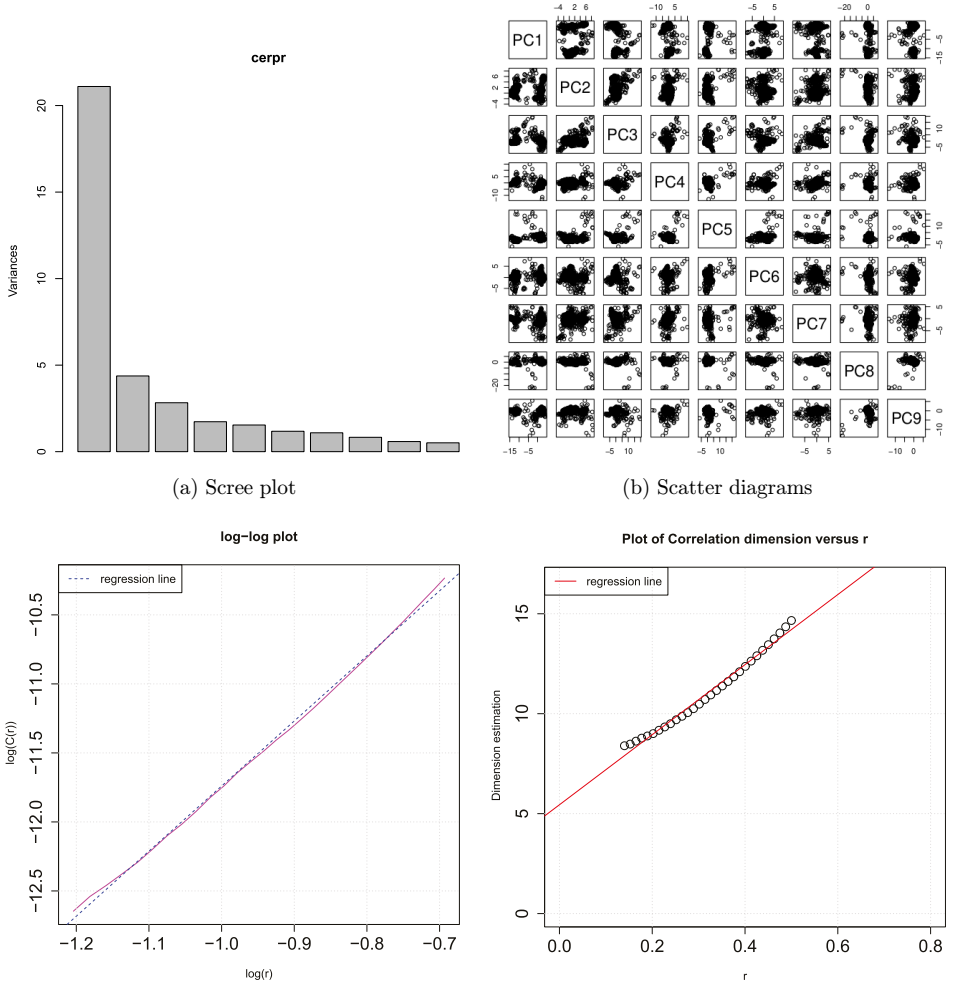
4. Comparing the Various Estimation Methods

This section illustrates, compares and benchmarks the methods developed in Sec. 3. The analysis is based on a recorded dataset that stems from an industrial glucose production facility in Sec. 4.1, a series of simulation examples in Sec. 4.2 serving as a "proof of concept", and a comparison with benchmark datasets and methods presented previously in the literature in Sec. 4.3. For each dataset analyzed here, the recorded variables are mean centered and subsequently scaled to unity variance. Section 4.4 critically examines the computational efficiency of each of the proposed methods.

Unless stated otherwise, the values of r_1 and r_s for the correlation dimension methods are chosen as described in Sec. 3. For the *intercept* and *slope* methods, the grid points r_j are placed with a spacing of 0.01 which e.g. implies that $s = 21$ when $r_1 = 0.3$ and $r_s = 0.5$. For the *polynomial* method, we use $r_s = 1$ and $s = 30$. For the charting manifold methods, the sequence of the radius r is selected such that the sphere with the minimum radius contains at least two data points and the sphere with the maximum radius includes all samples. This usually involves a much larger range of r values, up to $r_s = 10$, as compared to the correlation methods, so that a grid spacing of 0.1 is adequate for these techniques. For the dip method, we compute $\tilde{H}_{\mathbf{x}}(\cdot) = \hat{\beta}(\cdot)$ using the function `locpoly` in R[44] with bandwidth parameter $h = 0.1$, again unless stated otherwise. For the MLE method, we used the inverse averaging rule (11) as default, but provide, on some occasions, comparison to the other variants. For all other methods, parameter settings are as given in the respective source papers.[28,45]

4.1. *Industrial dataset*

For illustrative purposes, we consider data recorded from a glucose production facility containing $n = 28\,801$ observations for a total of $M = 39$ variables. The recorded variables include, among others, readings of various temperature, flow rates, pressure and pressure differences, measurements of viscosity, etc. A sample of each process variable was taken every 30 s. The recorded data cover four days of glucose production with two different grades and show a significant degree of variable correlation. The scree plot in Fig. 1 confirms the high degree of variable correlation. More precisely, the first principal component is dominant, as it explains 54% of the total variance in the data. Successive components take a non-negligible part of the variation as well. In fact, one needs nine principal components to capture at least 90% of the total variance in the data. To give additional insight, we produce a pairs plot of the first nine principal component scores (for the sake of visualization, only a sample of $n' = 5000$ scores have been plotted) in Fig. 1 (top right). We clearly



(a) Scree plot

(b) Scatter diagrams

(c) Slope method

(d) Intercept method

Fig. 1. Analysis of industrial dataset. (a) Scree plot; (b) matrix scatterplot of the first nine principal component scores; (c) slope method; (d) intercept method.

see that there remains some inner structure, indicating that $m < 9$ when taking this structure into account through nonparametric methods. The next subsections present the results of applying the methods proposed earlier to this dataset.

4.1.1. *Correlation dimension methods*

Slope method. The lower left plot in Fig. 1 shows the estimated linear regression curve of the computed values of $\log(C_n(r))$ versus $\log r$. The resulting linear equation is $y = -7.03 + 4.71 \log r$. Thus, the resulting estimate is $\widehat{m}_S = 4.71$.

Table 2. Results of fitting a polynomial of degree 6 for the industrial dataset. Left: summary table for linear model fit to $(r, C_n(r))$ using the full data. The * symbol indicates the chosen dimension; right: distribution of chosen ID for 50 random subsamples of sizes $n' = 2000$ and $n' = 4000$, respectively.

	Estimate	Std.Error	t-Value	$Pr(>\|t\|)$
a_1	6.84e−05	1.78e−05	3.84	7.82e−04
a_2	−1.15e−03	2.04e−04	−5.61	8.88e−06
a_3	6.42e−03	8.51e−04	7.54	8.84e−08
a_4	−1.53e−02	1.63e−03	−9.37	1.72e−09
a_5	1.61e−02	1.45e−03	11.06	6.59e−11*
a_6	−3.82e−03	4 90e−04	−7.79	4.99e−08

$\hat{m}_P$	1	2	3	4	5	6
$n' = 2000$	5	1	1	0	**33**	10
$n' = 4000$	2	0	0	0	**46**	2

Intercept method. The lower right plot in Fig. 1 shows the curve $D_n(r)$ versus r and the resulting estimated regression line. As expected, the function between $D_n(r)$ and r can be approximated by a linear function for r between 0.14 and 0.5, which follows from the discussion in Sec. 3.1.2. The estimated linear regression equation $D_n(r) = a^* + br = 5.45 + 17.53\, r$ produces the estimate $\widehat{m}_I = 5.45$, which is close to the estimate by the slope method.

Polynomial method. The estimate of m is determined by considering the significance of the estimated parameters of the polynomial in Eq. (19). We have carried out this estimation, for a polynomial of order $q = 6$, using the statistical software R.[44] The output is provided in Table 2 (left), where the column "Estimate" gives the estimated values of a_j in the jth row. The standard error of the estimate of a_j is given in the column "Std.Error" and the quotient of the first two columns gives the test statistic (t-value) displayed in the third column. The p-values in the fourth column are computed with reference to a t distribution with $s - q = 30 - 6$ degrees of freedom. The most significant parameter is a_5, implying that the estimate is $\widehat{m}_P = 5$. Generally, it is equivalent to look for the smallest p-value, or the largest absolute t-value. Venables and Ripley[56] provided a detailed discussion on how to interpret linear model outputs. It is finally noted that attempts using a polynomial degrees $q \geq 7$ did not produce reliable results.

4.1.2. *Charting manifold methods*

Dip method. We obtain $b = 50$ target points as outlined in Sec. 3.1. The derivative estimators are found using a local polynomial smoother with bandwidth $h = 0.08$ for a sample of size 50. The median of all 50 different estimates gives the overall estimate $\widehat{m}_D = 4.44$. Figure 2 presents exemplary derivatives $\tilde{H}_{\mathbf{x}}(r)$ for three of the target points.

Regression method. Utilizing the same target points as for the *dip* method, we consider the number of data points falling into hyperspheres of radius r. Next, fitting a linear regression of log-counts versus log-radii for each target point results in an overall estimate of $\widehat{m}_R = 3.05$.

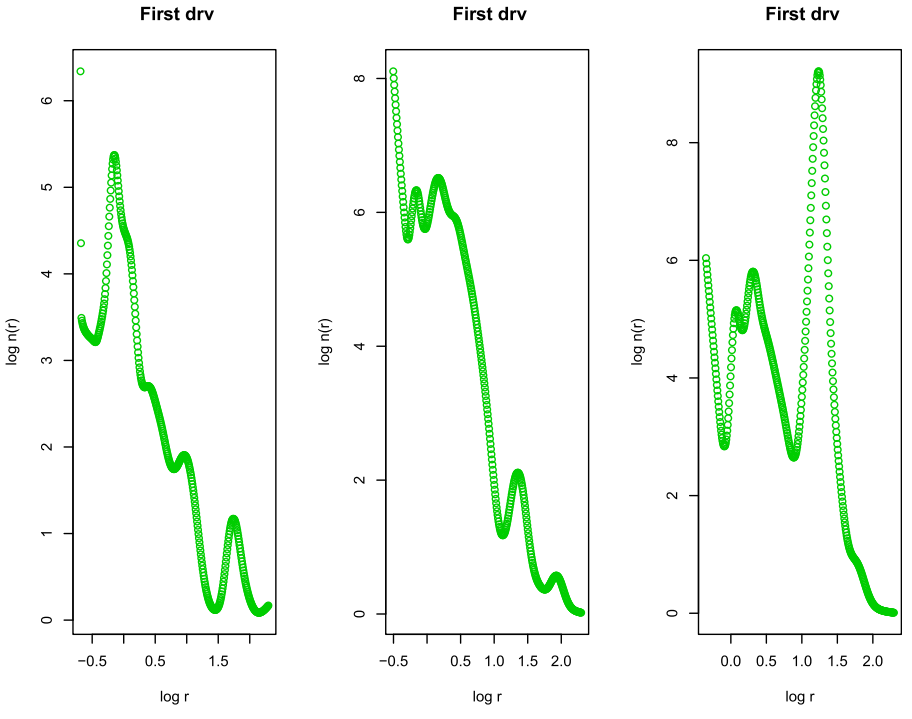


Fig. 2. Three exemplary curves $\tilde{H}_{\mathbf{x}}(r)$ used for ID estimation of the industrial dataset using the dip method.

4.1.3. *Other methods*

MLE method. The inverse-average version of the MLE estimator is not exactly computable for this dataset, since for most values of k, the quantity $m_k(\mathbf{x}_i)$ is exactly zero for four of the 28 801 observations $\mathbf{x}_i$. The other two variants for the computation of the MLE both yield the value 6.45.

VPC and VRE. For comparative purposes, we also provide in Table 3 the results of the parametric VRE,[54] which is a cross-validatory approach and Velicers Partial Correlation[55] methods mentioned in Sec. 1.1. These techniques are designed to determine the number of principal components, which are larger than the IDs

Table 3. Estimates of m for the industrial dataset.

	PCA					
Criterion	$> 70\%$	$> 80\%$	$> 90\%$	Broken stick	VRE	VPC
Estimate of m	3	5	9	4	7	8
	Nonparametric Techniques					
Methods	Reg.	Dip	Int	Slope	Poly.	MLE
Estimate of m	3.05	4.44	5.45	4.71	5	6.45

suggested by the correlation dimension and charting manifold techniques. This suggests that a linear subspace to capture the main variation in the data may not be adequate for this dataset.

4.1.4. *Discussion of results*

After inspection of the summary of results in Table 3, it is concluded that all non-parametric techniques, except the regression method, agree on an ID of ≈ 5 for this dataset, which is also sensible in the light of the parametric analysis via PCA. The reason for the possible failure of the regression method is not entirely transparent in this example, but as it appears that the local methods are potentially sensitive to granularities, such as local strings and clusters of low dimension, in the data. The pairs' plot of the principal component scores in Fig. 1 indicates that such granularities may exist for this dataset. The MLE value is close to the value obtained by the linear VRE method. This relatively high value is, we believe, influenced by two factors. Firstly, the MLE technique seems to show a slight tendency to overestimate the true ID when the sample size is large and/or the data are clustered, see also the further examples to follow. Secondly, manual inspection of the terms $m_k(\mathbf{x}_i)$ indicated that the inverse-average estimate of the ID would be in the region 5.5 if it was computable.[e]

4.2. *Simulation studies*

This subsection presents a number of simulation examples serving as a "proof of concept", which confirm that in some simple scenarios the expected results are obtained. The results for the individual methods are presented in the form of box-plots which show the median and distribution of the estimates for the MLE, *intercept*, *slope*, *regression* and *dip* method. In addition, the results of the *polynomial* method are shown in tabular form.

4.2.1. *First scenario*

Datasets containing four variables are generated from a multivariate normal distribution with the mean vector $\mu = (9, 5, 6, 4)^T$ and the diagonal covariance matrix $\mathbf{\Sigma} = 50\boldsymbol{I}_4$. Since these datasets do not possess any latent structure, related to Eq. (1), it follows that $M = m$. Two different choices of sample sizes $n = 200$ and $n = 2000$ are considered, and in either case a total of 100 datasets were generated. The top panels of Fig. 3 summarize the resulting estimates for the global *intercept* and *slope*, the local *dip* and *regression*, and the MLE method in its version (11). The results of the polynomial method are displayed in Table 4.

In summary, while the correlation methods for $n = 200$ featured a low bias but a high variance, the charting methods showed a very small variance at the expense of a negative bias. For the correlation methods, note that one observes a considerable

[e]Of course, one could at this point contemplate the development of a "robust" variant of the inverse-average method, but this is beyond the scope of this paper.

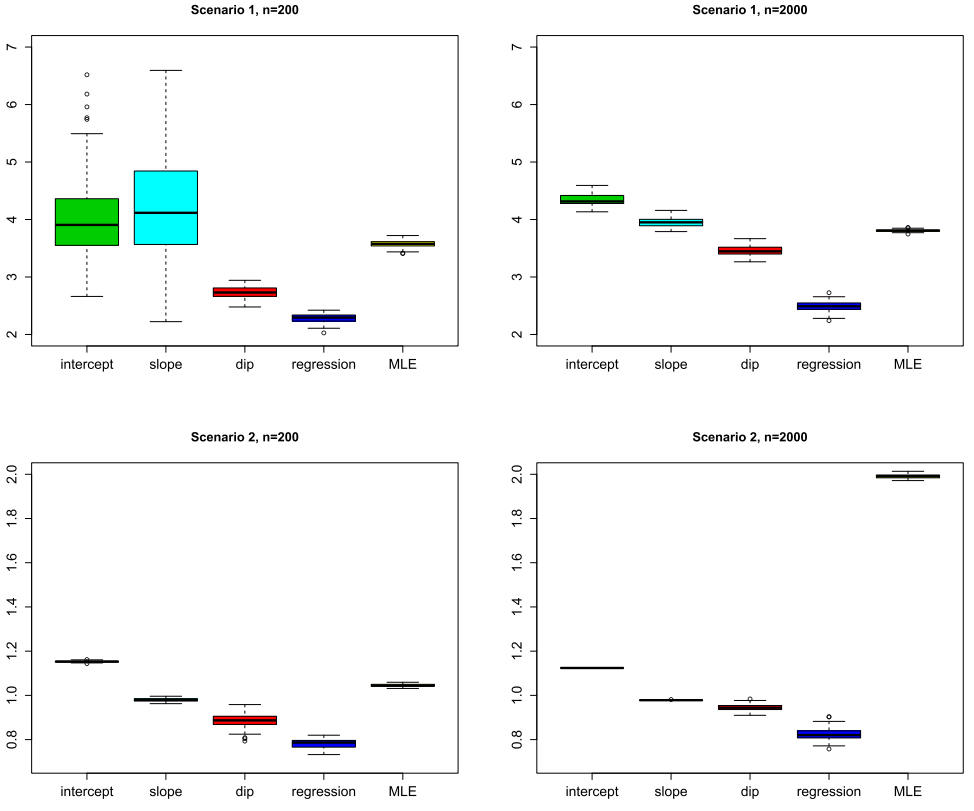


Fig. 3. Simulation study; boxplots of estimates for scenarios 1 (top) and 2 (bottom), using sample sizes $n = 200$ (left) and $n = 2000$ (right).

number of estimates that exceed $M = 4$, which could be argued to be implausible given how the datasets were generated. When the sample size is increased to $n = 2000$, the correlation methods improve strongly in terms of variance, and the *dip* method also improves strongly in terms of bias.

Table 4. Results of polynomial method for first (a), second (b) and third (c) simulation scenarios. Numbers marked in bold refer to the "correctly" identified estimates.

		$\hat{m}_P$			
Example	n	1	2	3	4
(a)	200	2	1	44	**53**
(a)	2000	0	0	63	**37**
(b)	200	**100**	0	0	0
(b)	2000	**100**	0	0	0
(c)	2000	0	**100**	0	0

From Fig. 3 top, the MLE demonstrates a consistently low bias and variance. Investigating in more detail, the MLE estimates obtained using the three different variants of MLE estimation are compared in Fig. 4(a), using $k^{(1)} = 5, k^{(p)} = 50$ and $p = 46$. We see that — in this particular example — the "incorrect" averaging scheme by Levina and Bickel seems to lead to a better result than the correct version using the inverse-averaging scheme. This may be due to the fact that in this example $m = M$, which will naturally force the MLE to lie below M.

4.2.2. *Second scenario*

The second scenario uses the data structure of type Eq. (1a) for a single latent variable, $m = 1$, which is uniformly distributed (between 0 and 10) and describes points on a straight line through a four-dimensional space. For each of the four variables, a zero mean noise variable that is independently and normally distributed with a variance of 0.0025 was added. As before, each method was contrasted using a total of 100 generated datasets, each containing $n = 200$ or $n = 2000$ samples.

Table 4 shows the results of the *polynomial* method, which, notably, correctly identifies $m = 1$ for each of the 100 datasets, irrespective of the sample size.

The lower panel in Fig. 3 bottom shows the resulting boxplots for the *intercept* and *slope*, the local *regression* and *dip* and the MLE methods. This demonstrates that all considered methods achieve good ID estimates of relatively low bias and variance, with two notable exceptions: the regression method delivers a slight negative bias, and, interestingly, the MLE becomes biased for the larger sample size $n = 2000$, as can be seen from the bottom right panel of Fig. 3. This is consistent with

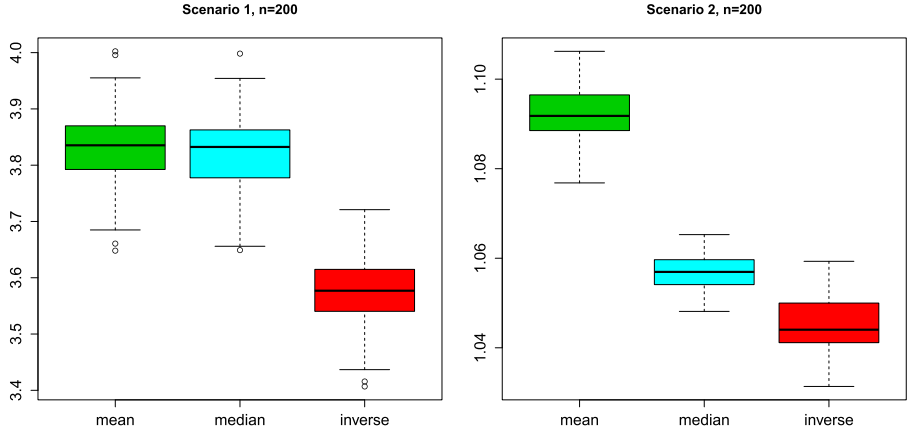


Fig. 4. Comparison of three ways of estimating the MLE for simulation scenarios 1 (a) and 2 (b), using sample size $n = 200$. Within each panel, the three boxplots give the MLE obtained via (from left to right): (i) Bickel and Levina's original estimator; (ii) a variant of the latter using the median in the averaging step (12); and (iii) the inverse-average version by McKay and Ghahramani.

Levina and Bickel's[37] observation "that $n = 100$ highly correlated points look like a line, but $n = 2000$ points fill out the space around the line". Higher values of $k^{(1)}$ and $k^{(p)}$ would be required to achieve a better estimate; changing the type of averaging does not help. The comparison of the different MLE variants is given again, for $n = 200$, in Fig. 4 (right). We see that the median version occupies a middle ground between the two other variants, and seems to have a slightly reduced variance as compared to these.

4.2.3. *Third scenario*

This example is based on the simulation setup presented in Liu *et al.*[38] The simulated process produces a five-variate data vector $\mathbf{x}$ that depends on the three latent variables s_1, s_2 and s_3:

- $s_1(i) = 2\cos(0.08\,i)\sin(0.06\,i)$;
- $s_2(i) = \text{sign}[\sin(0.03\,i) + 9\cos(0.01\,i)]$;
- $s_3(i) \sim \mathcal{N}\{0, 0.25\}$,

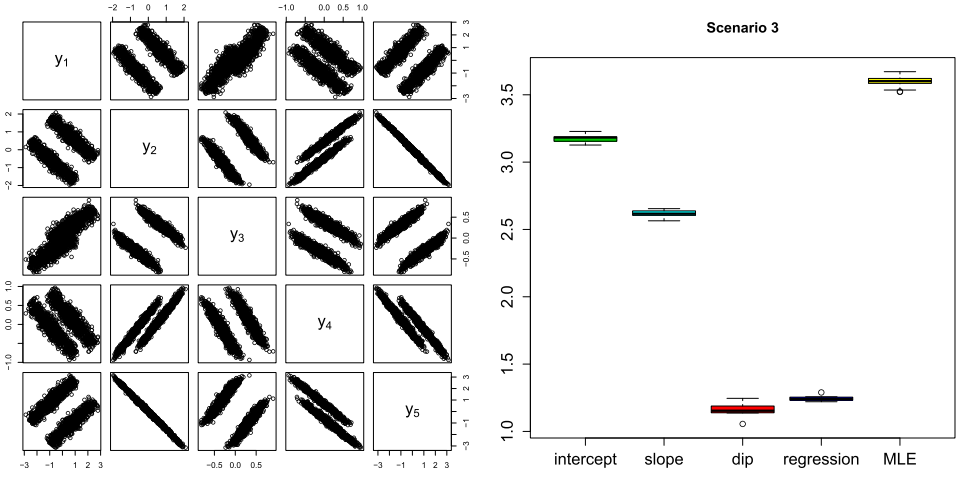
where i is a sampling index and $\mathcal{N}\{\cdot\}$ represents a normal distribution, here with zero mean and a variance of 0.25. The random vector is defined by $\mathbf{x}(i) = \mathbf{y}(i) + \mathbf{r}(i)$, where

$$\begin{pmatrix} y_1(i) \\ y_2(i) \\ y_3(i) \\ y_4(i) \\ y_5(i) \end{pmatrix} = \begin{bmatrix} 0.86 & 0.79 & 0.67 \\ -0.55 & 0.65 & 0.46 \\ 0.17 & 0.32 & -0.28 \\ -0.33 & 0.12 & 0.27 \\ 0.89 & -0.97 & -0.74 \end{bmatrix} \begin{pmatrix} s_1(i) \\ s_2(i) \\ s_3(i) \end{pmatrix} \tag{32}$$

and the random noise vector has a normal distribution $\mathbf{r}(i) \sim \mathcal{N}\{\mathbf{0}, 0.0025\mathbf{I}\}$. From this process, a total of 100 datasets, containing $n = 100$ samples each, were generated.

Figure 5 (right) displays the boxplot of the estimates of m for all methods. The results for the *polynomial* method are in Table 4. The correlation dimension approach produced median values of 3.17 and 2.60 for the *intercept* and *slope* methods, respectively. The application of the *polynomial* method yielded an estimate of 2 for each dataset. The charting manifold approach yielded median values of 1.25 and 1.17 for the *regression* and the *dip* method, respectively. Finally, a median value of 3.61 was determined for the inverse-averaged MLE technique (with the other two MLE variants delivering higher values). Especially when comparing to the first scenario, each method produced a considerably higher precision in estimating m for each of the 100 datasets.

According to Eq. (32), there are three latent variables implying that the ID estimate should not exceed 3. In light of this, one may consider the MLE overestimated. The global correlation-based results are more reasonable, while each of the local methods yield underestimates. This can be partially explained by



(a) (b)

Fig. 5. Third simulation scenario. Panel (a) gives one exemplary simulated dataset, and panel (b) the summarized ID estimates.

considering an exemplary simulated dataset as provided in the left-hand side of Fig. 5. The displayed scatter plots indicate that the data describe two noisy strings, each of which appears roughly one-dimensional, so that the obtained local ID estimates of ≈ 1 are plausible in this light.

4.3. *Comparison with reference data and methods*

In this subsection, we use three reference datasets, two of them synthetic and one of them real, which have been frequently employed in the recent literature to compare the performance of modern ID estimation routines. Specifically, the synthetic datasets are a Swiss roll and a 12-dimensional manifold in 72-dimensional space, which have been labeled as $\mathcal{M}_7$ and $\mathcal{M}_8$, respectively, in Rozza *et al.*[45] The real dataset is the "ISOMAP" face data $\mathcal{M}_{\text{Faces}}$ which is a collection of 698 gray-level sculpture images of dimension $64 \times 64 = 4096$. These datasets are among those proposed by Campadelli *et al.*[12] as benchmark datasets, with the latter one being highlighted as "particularly challenging due to its high curvature", and have also been examined in He *et al.*[28] We have generated the synthetic datasets using R package "manifgen"[32] based on methods developed by Hein and Audibert.[29] We followed the setup in Rozza *et al.*[45] and created 20 instances of datasets of size 2500 each. Average ID estimates over the 20 runs are provided for all methods in Table 5.

The alternative methods considered include PCA, Probabilistic PCA (PPCA[52]), Bayesian PCA (BPCA[4]), two versions of the "Minimum Neighbor Distance estimators",[39] namely a Maximum Likelihood version (MiND_{ML}[45]) and a variant based on the Kullback–Leibler divergence (MiND_{KL}[45]), the "Intrinsic Dimensionality Estimation Algorithm" (IDEA[45]), a fast graph-based variant of the

Table 5. Comparison of several methods using reference datasets. Results above the double horizontal line were obtained through own calculation; results below are extracted from the orignal sources.[28,45]

	$\mathcal{M}_7$	$\mathcal{M}_8$	$\mathcal{M}_{\text{Faces}}$
M	3	72	4096
m	2	12	3
int	2.23	—	—
slope	1.94	11.90	3.72
poly	2.00	(*)8.00	2.00
dip	1.80	9.69	1.79
reg	2.05	11.16	2.84
MLE (mean)	1.97	13.68	4.24
MLE (med)	1.97	13.67	4.22
MLE (inv)	1.81	12.50	3.75
PCA	3.00	24.00	21.00
PPCA	3.00	24.00	5.00
BPCA	2.00	24.00	4.00
Hein	2.00	12.00	3.00
MiND_{ML}	2.00	13.30	3.59
MiND_{KL}	2.00	16.50	3.90
IDEA	2.07	14.49	3.73
kNNG_1	1.97	13.87	3.60
NNG-He	1.81	4.55	6.07

Notes (*): The median of the 20 outcomes for the polynomial method is actually 11; the value 8 was caused through a few occurrences of the result $\hat{m}_P = 1$.

KNN method (kNNG_1[45]), as well as another graph-based NN-type algorithm proposed by He *et al.* (NNG-He[28]).

For dataset $\mathcal{M}_7$ we used the default settings of our methods; whereas for the high-dimensional datasets $\mathcal{M}_8$ and $\mathcal{M}_{\text{Faces}}$, the radii $(r_1, r_s) = (4, 6)$ and $(r_1, r_s) = (10, 20)$, respectively, were used for the slope method (with $s = 21$), reflecting that in higher dimensions, higher values of r are needed to obtain nonzero correlation integrals. For the polynomial method, we used $(r_1, r_s) = (1, 20)$. The intercept method could not be meaningfully applied on datasets $\mathcal{M}_8$ and $\mathcal{M}_{\text{Faces}}$ as it requires $r < 1$ for the entire grid. For the local methods, $b = 20$ target points were selected in each simulation run.

We find that our methods behave similarly to existing methods, noting some overestimation for the intercept method and underestimation for the dip method, where they were computable.

4.4. *Computational efficiency of proposed methods*

Examining the different ways in which the individual methods estimate m allows assessing their computational efficiency. Sections 4.4.1–4.4.3 discuss this issue for the correlation dimension, charting and MLE methods.

4.4.1. *Correlation dimension methods*

The main computational burden of the *slope*, *intercept* and *polynomial methods* is the determination of the correlation integral, which requires, according to Eq. (2), $n(n-1)/2 = \mathcal{O}(n^2)$ comparisons for typically $s = 20$–30 different radii. Additionally, these methods involve the estimation of a small set of linear model parameters, which follows from Eqs. (14), (17) and (19), respectively, which is an $\mathcal{O}(n)$ operation.

4.4.2. *Charting manifold methods*

Both charting manifold methods require the determination of typically $b = 50$ target points involving a random sampling and an outlier detection routine, which are of $\mathcal{O}(n)$ complexity. For each target point, Eq. (7) determines through n comparisons the number of points inside the sphere of radius r. The *dip method* then requires the application of a kernel smoother, followed by a search for the dips of the smoothed function in Eq. (27). The *regression method* is similar in approach to the *slope method* and requires the estimation of a set of parameters. In either case, this is an $\mathcal{O}(n)$ operation. This step needs to be repeated b times.

4.4.3. *MLE method*

The estimation of m using this method relies on Eqs. (9)–(12). The former equation involves a nearest neighborhood search and the determination of Euclidean distances for the kth and jth nearest neighbors of the sample $\mathbf{x}_i$. The integer k itself is not a fixed constant but includes, according to Eq. (12), a total of p different values, where p is typically between 10 and 50. The number of searches for Eq. (9) alone is of the order n^2. This is to be repeated for all p nearest neighbors, i.e. $k^{(1)}, \ldots, k^{(p)}$. By directly comparing the number of searches and floating point operations for the MLE method with the correlation dimension and the charting manifold methods, it is to be expected that the MLE method is, consequently, computationally inferior for large sample sizes.

4.4.4. *Direct comparison*

This subsection utilizes the second simulation scenario, discussed in Sec. 4.2.2, to compare the computational time consumed for each method to estimate m. For the six methods, Table 6 summarizes the median time consumed for a single run in seconds, calculated from 100 Monte Carlo experiments, each for $n = 200$ and $n = 2000$ samples. Each method was programmed in R, version 3.2.1, and executed using an Intel(R) Core(TM) i7-3770 CPU with 3.40 GHz.

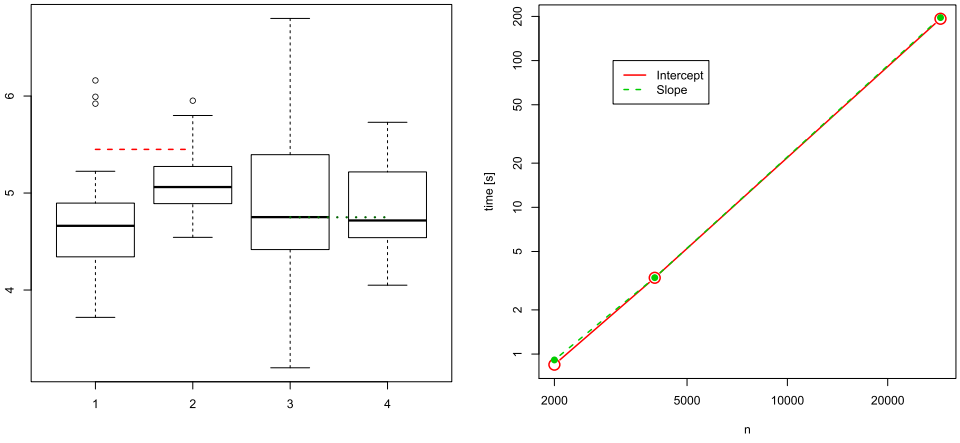
In both cases, $n = 200$ and $n = 2000$, the correlation techniques are the most efficient ones and are of the same order of magnitude. The increase in the sample size by a factor 10 resulted in an increase in computational time by around 100. This is to be expected, given that the number of searches/sums to determine $C_n(r)$ are of order $\mathcal{O}(n^2)$. The charting manifold methods are around 300 times slower in estimating m

Table 6. Computational comparison (median run time in seconds) of the 6 estimation methods.

Method					
Correlation Dimension			Charting Manifold		
slope	*intercept*	*polynomial*	*dip*	*regression*	MLE
$n = 200$					
0.004	0.005	0.007	1.374	1.381	0.423
$n = 2000$					
0.140	0.236	0.434	9.639	9.600	23.833

for $n = 200$ and around 30 times slower for $n = 2000$. Determining $N_{\mathbf{x}}(r)$ is, unlike the estimation of $C_n(r)$, of order n. The difference in computational burden between correlation dimension and charting manifold methods is therefore decreasing as n increases. The MLE technique was more efficient than the local methods for small sample sizes but expectedly lost this advantage for large sample sizes. The hundredfold increase in the computational time, resulting from the tenfold increase in the number of samples, is expected as the number of searches is of the order n^2.

Especially for large datasets, it is an appealing option to reduce the computational burden of the dimension estimation by using only a sample of the original dataset. Therefore, we give additionally some insight into the repeatability of ID estimates under subsampling. This is exemplified here using the industrial dataset for which 50 randomly selected datasets of the size 2000 and 4000 were constructed. Figure 6(a)



(a) ID estimates for subsamples

(b) Computational time

Fig. 6. (a) From left to right: Boxplots of 50 ID estimates for the industrial dataset, using the intercept method with samples of sizes 2000 and 4000, and using the slope method with samples of sizes 2000 and 4000, respectively. The dashed and dotted horizontal lines correspond to the respective full-sample estimates; (b) corresponding average running times (in seconds) for a *single* ID estimate as a function of the sample size, on log–log scale.

shows the resulting ID estimates for the correlation dimension method. The first two boxplots display the ID estimates using the *intercept* method and 50 subsets of size 2000 and 4000, respectively. The third and fourth boxplots present the results for the *slope* method. The full data estimates via the *intercept* and *slope* method are also provided through a dashed and dotted line, respectively. We see that the different estimates have a small variance that decreases as the sample size increases, and that the *intercept* method shows a stronger response to the subsampling than the *slope* method. In conformity with the considerations from Sec. 4.4.1, it is also evident that, on a log–log scale, the computational time of both slope and intercept method increases linearly with the sample size (Fig. 6 right). One finally finds from Table 2 (right) that the polynomial method becomes less variable in its decision when the subsample size increases.

5. Concluding Summary

This paper has summarized methods for estimating the ID of multivariate data. While parametric ID estimation methods have been intensively studied in the literature, only relatively recent work addressed the utilization of nonparametric methods. In fact, most methods proposed are difficult to implement in practice for large variable sets and numbers of samples or low data densities. Moreover, correlation dimension and charting manifold approaches have been proposed as concepts rather than tailored methods that can be readily applied in practice.

The correlation dimension concept relies on estimating the correlation integral as a function of a sphere of radius r engulfing pairs of samples. It follows from Eq. (2) that the correlation integral determines the distribution of the distances between points. The charting manifold concept uses a similar approach but instead of using each sample and obtain pairwise distances, this concept relies on counting the number of samples that are in the vicinity of some nonisolated target points.

The focus of this paper has, therefore, been the development of methods for the correlation dimension and the charting manifold concepts. More precisely, we have proposed three methods for the former, which have been referred to as *slope*, *intercept* and *polynomial* methods. While the *slope* method is a simple enhancement of the graphical-based log–log technique, only the *intercept* and *polynomial* methods have been considered here as novel methods. For the charting manifold concept, the paper has proposed the *dip* and *regression* method as new local-based methods.

By contrasting these five methods with the MLE method using a recorded set from a glucose production facility and three simulation examples, the paper has found that the correlation dimension methods have shown the best performance in terms of the estimation accuracy and the time consumed to produce an estimate. Particularly, the *polynomial* methods showed a consistently high degree of accuracy. However, it has to be noted that the polynomial method may run into difficulties for

larger m, which requires larger polynomial orders, q. More precisely, an increase in the order q of the polynomial is accompanied by a significant increase in the variability of the parameter estimates in the model for $C(r)$, rendering the resulting p-values uninformative. This follows from the well-known fact that the estimation variance tends to grow as the number of parameters to be estimated increases for a fixed number of samples.

While the MLE method gives generally very precise ID estimates, we observed a slight tendency to overestimate the ID in situations where the data are clustered, the sample size is large, and/or the inverse-average version cannot be computed. In contrast, the proposed charting manifold methods have had a tendency to underestimate the ID which was particularly visible for the *regression* method in the industrial data example. A more detailed analysis of our examples has indicated that, despite all methodological precautions, both local methods are affected by localized granularities, for example linear strings or small clusters, that are not representing the global structure of the data. A potential advantage of local methods, however, is their ability to identify such localized granularities. Based on the application studies, by directly comparing the performance of both charting manifold techniques, the *dip* method produced a more accurate estimation than the *regression* method. However, a notable advantage of the regression method is its simplicity and the absence of tuning parameters. A comparison with a wide range of recently proposed ID estimation techniques has demonstrated that our results are in line with, and competitive to, those methods; though not necessarily superior at all instances.

The question of local strings and clusters, which impact on ID estimates, relates to the question of how to deal with situations in which several disconnected manifolds coexist in a single dataset. The analysis of our third simulation example provides an illustration into the behavior of dimension estimation techniques in such a scenario. This topic, however, requires a further and a more thorough investigation involving more complex examples that include data structures consisting of multiple manifolds of different ID for example. It is clear that, by virtue of their construction, the global methods listed in the left-hand side of Table 1 are not able to deal with such disconnected data structures, as they are designed to produce a single ID estimate. Recent contributions to this problem have been based on finding sparse and local representations, which relate the neighborhood size directly to an ID estimate.[14,18]

The main computational burden for the correlation dimension methods is the estimation of the correlation integral, which, as explained, is of the order n^2. Though the computation of the local methods has been generally slower than the global methods in the examples examined in this paper, we have emphasized that the EDF of the distances of samples to the center of the sphere is only of order n. Hence, computationally, the larger the sample size, the more cumbersome is the estimate of the correlation integral for the global methods relative to the determination of the

EDF for the local methods, since $n^2 \gg n$. That is, for very large sample sizes, the local methods should turn out to be more efficient. While all datasets considered in this paper (except $\mathcal{M}_8$) fulfilled Eckmann and Ruelle's[16] rule that $n \geq 10^{m/2}$, the sample sizes considered in this paper were arguably still quite small. We regard it as a positive outcome that satisfactory dimension estimation has been possible under these conditions. Further research on the robustness of the estimation methods in the presence of outliers, very small or very large sample sizes, or excessive complexity, is nonetheless required. Significant challenges lie in the estimation of "large" IDs. We found that the *polynomial* method is of reduced reliability for polynomial degrees $q \geq 7$, hence restricting its use to dimensions $m \leq 6$. Future work should study this limitation, though it should be pointed out that the problem is more general: Already Eckmann and Ruelle[16] have stated that their rule makes dimension estimation for $m \geq 6$ or 7 virtually impossible. For instance, if $m = 20$, then the above rule would require 10 billion samples! Camastra and Vinciarelli[11] addressed this problem to some extent by providing a "reference curve" which corrects the bias when n is too small. While further advances in this direction, exploiting geometric properties of nearest neighbors, have recently been made,[13,45] further work on this problem would certainly be beneficial. A final, but very important, problem is to develop diagnostic tools or quantitative criteria which assess the goodness or reliability of ID estimates. The ability to quantify the accuracy of ID estimates is of considerable practical importance, as the ID is directly related to the information bottleneck in large-scale problems.

Acknowledgments

The authors wish to thank Alessandro Rozza and Claudio Ceruti for their help with accessing the benchmark datasets. We are further grateful to two anonymous referees for their constructive comments.

This project was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under Grant No. (DF-145-247-1441). The authors, therefore, gratefully acknowledge DSR technical and financial support.

Appendix A. Proof of Theorem 1

With $C_n(r) \sim cr^m$ and $D_n(r) = \frac{\log C_n(r)}{\log r}$, it follows that

$$D_n(r) \sim m + \frac{\log c}{\log r} \equiv m + f(r) \log c. \tag{A.1}$$

The next step is to develop a second-order Taylor expansion of $f(r) = \frac{1}{\log r}$ for $0 < r_0 < 1$, which has the following coefficients:

$$f(r) = \frac{1}{\log r}, \quad f'(r) = -\frac{1}{r \log^2 r}, \quad f''(r) = \frac{1}{r^2 \log^3 r}(\log r + 2). \tag{A.2}$$

Including a remainder in Lagrange form, $\frac{1}{6}f'''(\rho)(r-r_0)^3$, substituting these coefficients into Eq. (A.1) yields

$$D_n(r) = m + \log c(f(r_0) + f'(r_0)(r-r_0) + \frac{1}{2}f''(r_0)(r-r_0)^2 + \frac{1}{6}f'''(\rho)(r-r_0)^3) \tag{A.3a}$$

$$D_n(r) = \underbrace{m + \frac{\log c}{\log r_0}}_{a} - \underbrace{\frac{\log c}{r_0 \log^2 r_0}}_{b}(r-r_0) + \frac{1}{2}\frac{\log c}{r_0^2 \log^3 r_0}(\log r_0 + 2)(r-r_0)^2 + \frac{\log c}{6}f'''(\rho)(r-r_0)^3. \tag{A.3b}$$

Here, ρ is in the interval between r and r_0. Now, for $r_0 = e^{-2}$, the squared term vanishes and hence, Eq. (A.3b) reduces to

$$D_n(r) = a + b(r-r_0) + \frac{\log c}{6}f'''(\rho)(r-r_0)^3. \tag{A.4}$$

If the radius is in the vicinity of $r_0 = e^{-2} \approx 0.135$, the remainder is negligible and $D_n(r)$ is approximately a linear function of $r - r_0$.

Appendix B. Proof of Theorem 2

Assuming that $C(r)$ is a polynomial with degree $q \geq 1$ and considering the condition $C(0) = 0$

$$C(r) = \sum_{i=1}^{q} a_i r^i = a_1 r + a_2 r^2 + a_3 r^3 + \cdots + a_{q-1} r^{q-1} + a_q r^q.$$

For $a_1 \neq 0$, the estimate of m, according to Eq. (4), becomes

$$m = \lim_{r \to 0} \frac{\log(a_1 r + a_2 r^2 + a_3 r^3 + \cdots + a_{q-1} r^{q-1} + a_q r^q)}{\log r}. \tag{B.1}$$

Next, applying l'Hospital's rule yields

$$m = \lim_{r \to 0} \frac{r(a_1 + 2a_2 r + 3a_3 r^2 + \cdots + q a_q r^{q-1})}{a_1 r + a_2 r^2 + a_3 r^3 + \cdots + a_q r^q} = \lim_{r \to 0} \frac{a_1 r + 2a_2 r^2 + 3a_3 r^3 + \cdots + q a_q r^q}{a_1 r + a_2 r^2 + a_3 r^3 + \cdots + a_q r^q}.$$

Applying l'Hospital's rule again gives rise to

$$m = \lim_{r \to 0} \frac{a_1 + 4a_2 r + 9a_3 r^2 + \cdots + q^2 a_q r^{q-1}}{a_1 + 2a_2 r + 3a_3 r^2 + \cdots + q a_q r^{q-1}} \to 1. \tag{B.2}$$

Now, assuming $a_0 = a_1 = 0$ and $a_2 \neq 0$, produces the following estimate for m:

$$m = \lim_{r \to 0} \frac{\log(a_2 r^2 + a_3 r^3 + \cdots + a_q r^q)}{\log r}. \tag{B.3}$$

In a similar fashion to the derivation of Eq. (B.2), applying l'Hospital's rule three consecutive times to Eq. (B.3) yields

$$m = \frac{4a_2}{2a_2} = 2.$$

Similarly, under the assumption that $a_0 = a_1 = a_2 = 0$ and $a_3 \neq 0$, Eq. (B.1) reduces to

$$m = \lim_{r \to 0} \frac{\log(a_3 r^3 + \cdots + a_{q-1} r^{q-1} + a_q r^q)}{\log r}$$

and, as before, applying l'Hospital rule now a total of four consecutive times, produces

$$m = \frac{18a_3}{6a_3} = 3.$$

By induction, it is straightforward to show that if $a_0 = a_1 = \cdots = a_{q-1} = 0$ and $a_q \neq 0$, and consecutively applying l'Hospital's rule a total of q times, we get $m = q$ for $r \to 0$.

References

1. V. S. Alagar, The distribution of the distance between random points, *J. Appl. Probab.* **13** (1976) 558–566.
2. D. Antory, G. W. Irwin, U. Kruger and G. McCullough, Improved process monitoring using nonlinear principal component analysis, *Int. J. Intell. Syst.* **23**(5) (2008) 520–544.
3. D. Antory, U. Kruger, G. W. Irwin and G. McCullough, Fault diagnosis in internal combustion engines using non-linear multivariate statistics, *Proc. Inst. Mech. Eng. I: J. Syst. Control Eng.* **219**(4) (2005) 243–258.
4. C. M. Bishop, Bayesian PCA, in *Advances in Neural Information Processing Systems 11*, eds. M. J. Kearns, S. A. Solla and D. A. Cohn (MIT Press, 1999), pp. 382–388.
5. I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling* (Springer, New York, 2005).
6. M. Brand, Charting a manifold, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA, 2003), pp. 961–968.
7. J. Bruske and G. Sommer, Topology representing networks for intrinsic dimensionality estimation, in *Lect. Notes Comput. Sci.* **1327** (1997) 595–600.
8. F. Camastra, Data dimensionality estimation methods: A survey, *Pattern Recognit.* **36** (2003) 2945–2954.
9. F. Camastra and M. Filippone, A comparative evaluation of nonlinear dynamics methods for time series prediction, *Neural Comput. Appl.* **18**(8) (2009) 1021.
10. F. Camastra and A. Staiano, Intrinsic dimension estimation: Advances and open problems, *Inf. Sci.* **328** (2016) 26–41.

11. F. Camastra and A. Vinciarelli, Estimating the intrinsic dimension of data with a fractal-based method, *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(10) (2002) 1404–1407.
12. C. C. P. Campadelli, E. Casiraghi and A. Rozza, Intrinsic dimension estimation: Relevant techniques and a benchmark framework, *Math. Probl. Eng.* **2015** (2015) 759567.
13. C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi and P. Campadelli, Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration, *Pattern Recognit.* **47**(8) (2014) 2569–2581.
14. D. Chen, J. C. Lv and Z. Yi, A local non-negative pursuit method for intrinsic manifold structure preservation, in *Proc. 28th AAAI Conf. Artificial Intelligence* (AAAI-14, Quebec City, Canada, 2014), pp. 1745–1751.
15. P. Cui, J. Li and G. Wang, Improved kernel principal component analysis for fault detection, *Expert Syst. Appl.* **34**(2) (2008) 1210–1219.
16. J. Eckmann and D. Ruelle, Fundamental limitations for estimating dimensions and Lyanpounov exponents in dynamical systems, *J. Phys. D, Appl. Phys.* **56** (1992) 185–187.
17. J. Einbeck and Z. Kalantan, Intrinsic dimensionality estimation for high-dimensional data sets: New approaches for the computation of correlation dimension, *J. Emerg. Technol. Web Intell.* **5** (2013) 91–97.
18. E. Elhamifar and R. Vidal, Sparse manifold clustering and embedding, in *Advances in Neural Information Processing Systems*, eds. J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger, Vol. 24 (Curran Associates, Inc., 2011), pp. 55–61.
19. M. Fan, N. Gu, H. Qiao and B. Zhang, Intrinsic dimension estimation of data by principal component analysis, arXiv:1002.2050v1.
20. T. Feital, U. Kruger, L. Xie, U. Schubert, E. L. Lima and J. C. Pinto, A unified statistical framework for monitoring multivariate systems with unknown source and error signals, *Chemometr. Intell. Lab. Syst.* **104**(2) (2010) 223–232.
21. K. Fukunaga and D. R. Olsen, An algorithm for finding intrinsic dimensionality of data, *IEEE Trans. Comput.* **20**(2) (1971) 176–183.
22. Z. Ge, L. Xie, U. Kruger and Z. Song, Local ICA for multivariate statistical fault diagnosis in systems with unknown signal and error distributions, *AIChE J.* **58**(8) (2012) 2357–2372.
23. A. Gorban, B. Kégl, D. Wunsch and A. Zinovyev (eds.), *Principal Manifolds for Data Visualization and Dimension Reduction* (Springer, Heidelberg, 2008).
24. P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors, *Phys. D: Nonlinear Phenom.* **9** (1983) 189–208.
25. J. M. Hammersley, The distribution of distance in a hypersphere, *Ann. Math. Stat.* **21**(3) (1950) 447–452.
26. T. Hastie and C. Loader, Local regression: Automatic kernel carpentry, *Stat. Sci.* **8** (1993) 120–143.
27. T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2001).
28. J. He, L. Ding, L. Jiang, Z. Li and Q. Hu, Intrinsic dimensionality estimation based on manifold assumption, *J. Vis. Commun. Image Represent.* **25**(5) (2014) 740–747.
29. M. Hein and J.-Y. Audibert, Intrinsic dimensionality estimation of submanifolds in Rd, in *Proc. 22nd Int. Conf. Machine Learning, ICML '05* (ACM, New York, NY, USA, 2005), pp. 289–296.
30. A. Höskuldsson, H-methods in applied sciences, *J. Chemometr.* **22**(3–4) (2008) 150–177.
31. J. E. Jackson, *A Users Guide to Principal Components*, Wiley Series in Probability and Mathematical Statistics (John Wiley, New York, 2003).

32. K. Johnsson, manifgen: Data sets on manifolds (2012), R Package manifgen version 1.1.
33. B. Kégl, Intrinsic dimension estimation using packing numbers, in *Advances in Neural Information Processing 15* (MIT Press, 2003), pp. 681–688.
34. M. A. Kramer, Nonlinear principal component analysis using autoassociative neural networks, *AICHE J.* **37**(3) (1987) 233–243.
35. U. Kruger and L. Xie, *Statistical Monitoring of Complex Multivariate Processes* (John Wiley & Sons, Chichester, UK, 2012).
36. U. Kruger, Z. Zhang and L. Xie, Developments and applications of nonlinear principal component analysis — a review, in *Principal Manifolds for Data Visualization and Dimension Reduction*, Lecture Notes in Computational Science and Engineering, Vol. 58 (Springer, 2008), pp. 1–43.
37. E. Levina and P. J. Bickel, Maximum likelihood estimation of intrinsic dimension, in *Advances in Neural Information Processing Systems 17*, eds. L. K. Saul, Y. Weiss and L. Bottou (MIT Press, Cambridge, MA, 2005), pp. 777–784.
38. X. Liu, L. Xie, U. Kruger, T. Littler and S.-Q. Wang, Statistical-based monitoring of multivariate non-gaussian systems, *AICHE J.* **54**(9) (2008) 2379–2391.
39. G. Lombardi, A. Rozza, C. Ceruti, E. Casiraghi and P. Campadelli, Minimum neighbor distance estimators of intrinsic dimension, in *Machine Learning and Knowledge Discovery in Databases: European Conf. ECML PKDD 2011*, eds. D. Gunopulos, T. Hofmann, D. Malerba and M. Vazirgiannis (Springer, Berlin, Heidelberg, 2011), pp. 374–389.
40. D. J. MacKay and Z. Ghahramani, Comments on 'maximum likelihood estimation of intrinsic dimension' http://www.inference.phy.cam.ac.uk/mackay/dimension/.
41. E. C. Malthouse, Limitations of nonlinear PCA as performed with neural networks, *IEEE Trans. Neural Netw.* **9**(1) (1998) 165–173.
42. T. Martinetz and K. Schulten, Topology representing networks, *Neural Netw.* **7**(3) (1994) 507–522.
43. S. J. Qin and R. Dunia, Determining the number of principal components for best reconstruction, *Journal of Process Control* **10**(2–3) (2000) 245–250.
44. R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2015).
45. A. Rozza, G. Lombardi, C. Ceruti, E. Casiraghi and P. Campadelli, Novel high intrinsic dimensionality estimators, *Mach. Learn.* **89**(1) (2012) 37–65.
46. B. Schölkopf, C. J. C. Mika, S. Burges, P. Knirsch, K. R. Müller and G. Ratsch, Input space versus feature space in kernel-based methods, *IEEE Trans. Neural Netw.* **10**(5) (1999) 1000–1016.
47. M. Stone, Cross-validatory choice and assessment of statistical prediction (with discussion), *J. R. Stat. Soc. (Ser. B)* **36** (1974) 111–133.
48. F. Takens, On the numerical determination of the dimension of an attractor, in *Dynamical Systems and Bifurcations*, eds. B. Braaksma, H. Broer and F. Takens (Springer, 1985), pp. 99–106.
49. J. Taylor, *Strategies for Mean and Modal Multivariate Local Regression*, Ph.D. thesis, Durham University (2012).
50. J. Tenenbaum, V. deSilva and J. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* **12** (2000) 2319–2323.
51. J. Theiler, Estimating fractal dimension, *J. Opt. Soc. Am. A* **7**(6) (1990) 1055–1073.
52. M. E. Tipping and C. M. Bishop, Probabilistic principal component analysis, *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **61**(3) (1999) 611–622.
53. S.-J. Tu and E. Fischbach, Random distance distribution for spherical objects: General theory and applications to physics, *J. Phys. A: Math. Gen.* **35** (2002) 6557–6570.

54. S. Valle, W. Li and S. J. Qin, Selection of the number of principal components: The variance of the reconstruction error criterion compared to other methods, *Ind. Eng. Chem. Res.* **38** (1999) 4389–4401.
55. W. F. Velicer, Determining the number of components from the matrix of partial correlations, *Psychometrika* **41**(3) (1976) 321–327.
56. W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S* (Springer-Verlag, New York, 2002).
57. P. J. Verveer and R. P. W. Duin, An evaluation of intrinsic dimensionality estimators, *IEEE Trans. Pattern Anal. Mach. Intell.* **17** (1995) 81–86.
58. X. Wang, U. Kruger, G. W. Irwin, G. McCullough and N. McDowell, Nonlinear PCA with the local approach for diesel engine fault detection and diagnosis, *IEEE Trans. Control Syst. Technol.* **16**(1) (2008) 122–129.

Jochen Einbeck received his Ph.D. degree (Dr. rer. nat.) from the University of Munich, Germany, in 2003. Following a Postdoctoral position at NUI Galway, Ireland, he joined the Department of Mathematical Sciences at Durham, UK, in 2006, where he is now the Associate Professor (Reader).

Zakiah Kalantan received her Bachelor's degree in Statistics/ Computer Science and Master's degree in Mathematical Statistics from King Abdulaziz University, Saudi Arabia. She received her Ph.D. degree in Mathematical Sciences from the University of Durham, in 2014.

Uwe Kruger obtained his Master's degree in Mechanical Engineering from the University of Essen, Germany, in 1996, and a Doctoral degree in Engineering from the University of Manchester, UK, in the year 2000. Following Academic Posts in Belfast, Abu Dhabi, and Muscat, he is now the Professor of Practice at the Rensselaer Polytecnic Institute in Troy, NY.