Intrinsic Dimensionality estimation for high-dimensional data sets: New approaches for the computation of correlation dimension

Jochen Einbeck[a] Zakiah Kalantan[a,b],
[a] Department of Mathematical Sciences, University of Durham, Durham, UK
Email: jochen.einbeck@durham.ac.uk
[b] Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia
Email:zkalanten@kau.edu.sa

***Abstract*— The analysis of high–dimensional data is usually challenging since many standard modelling approaches tend to break down due to the so–called "curse of dimensionality". Dimension reduction techniques, which reduce the data set (explicitly or implicitly) to a smaller number of variables, make the data analysis more efficient and are furthermore useful for visualization purposes. However, most dimension reduction techniques require fixing the intrinsic dimension of the low-dimensional subspace in advance.**

The intrinsic dimension can be estimated by fractal dimension estimation methods, which exploit the intrinsic geometry of a data set. The most popular concept from this family of methods is the correlation dimension, which requires estimation of the correlation integral for a ball of radius tending to 0. In this paper we propose approaches to approximate the correlation integral in this limit. Experimental results on real world and simulated data are used to demonstrate the algorithms and compare to other methodology. A simulation study which verifies the effectiveness of the proposed methods is also provided.

***Index Terms*— intrinsic dimensionality, fractal-based methods, correlation dimension**

I. Introduction

Many real-life applications deal with very high dimensional data. In order to handle those data in a proper way, we need to investigate whether they can be represented in some lower dimensional space. This step is very important since it alleviates the curse of dimensionality [2] and other issues such as increased computing time and data storage space. Dimension reduction is the mapping of high dimensional data into a lower dimension in which they have a meaningful representation. Adequately, the reduced dimensionality should be compatible with the intrinsic dimension (ID) of the data set. Dimension reduction methods can be categorized as linear or nonlinear methods. Linear methods, such as principal component analysis, seek a globally flat subspace . Nonlinear methods try to search a locally flat subspace, such as multidimensional scaling methods and ISOMAP. Most dimension reduction methods, whether linear or non–linear, require fixing the intrinsic dimension of the low-dimensional subspace in advance. To fix terms throughout the manuscript, we are given a data set $\Omega = \{x_1, \ldots, x_n\} \in \mathbb{R}^D$ which we assume to be scaled; i.e. each variable has been divided by its standard deviation. Let the intrinsic dimension of Ω be given by a value $d \leq D$, which gives effectively the minimum number of variables necessary to describe the data without much loss of information [4] [8]. According to Funkanaga's definition, the ID is equal to d when the points lie entirely within an d-dimensional subspace of $\mathbb{R}^D$ [4] [8]. It should be noted that, while Fukanaga's concept of "subspace" clearly had to be understood as that of a "linear subspace", we have in this paper a more general notion in mind which comprises linear as well as nonlinear manifolds.

ID estimation methods can be classified into two groups: **local** methods divide the data into small subregions, or provide a series of local ID estimates at several target points, in order to arrive at a suitably averaged overall ID estimator. Examples for such methods, which do not form the focus of this manuscript, include Levina–Bickel's Maximum Likelihood estimator [15], Brands' concept of 'charting' [3], among others [4]. On the other hand, **global** methods try to estimate the dimension using the whole data set, imposing the implicit assumption that the intrinsic dimension is constant over the data set. This family includes purely linear methods based on linear approximation (such as the "broken stick method" and many other stopping rules for principal component analysis [11] [14]), but also non–parametric approaches such as fractal-based methods. The term "fractal" is used since under this sort of approach, the intrinsic dimensionality d does not need to be an integer.

Fractal techniques [16] provide a useful tool for a variety of scientific fields [23] [18]. For instance, fractals are used to produce realistic natural objects, as moons or planets, by using computer graphics. The most important properties of fractals are self-similarity and symmetry. To put the analogy to statistics short: while fractals can be considered as mathematical *sets* with non–integer dimension, in fractal dimension estimation we deal with

Manuscript received ...2013; revised January ....; accepted .....

The first author is supported by the JSP program at King Abdulaziz University.

doi:10.4304/jetwi.5.2.91-97

data sets of non–integer intrinsic dimension.

Fractal dimension is a measure that describes the geometry of an irregular object (here: a data set) by an estimated real number. It describes the filling of the fractal object's space, which can be used to construct ID estimators. Various fractal-based methods have been proposed, as quantization estimator [20], kernel correlation method [10], horizontal structuring element, box-counting and correlation dimension [23]. Camastra presented a good survey on intrinsic dimension estimation methods focusing on fractal-based methods [4] [5].

The most common route to fractal dimension estimation is via correlation dimension. The method requires the construction of a so–called correlation integral, from which the ID is extracted using appropriate techniques. This step is not straightforward, since it requires counting the number of data pairs within a ball of radius tending to 0. This paper presents new techniques which address this problem. In Section II, the concept of correlation dimension is briefly reviewed. The improved methods — Intercept method, Slope method and Polynomial method — are discussed in Section III. In Section IV, we provide case studies on real data sets and a simulation study, which are used to state the effectiveness of the methods. Finally, conclusions are drawn in Section V. Most results in this article are based on an earlier paper presented at the ICSSBE2012 conference [12], but a substantial new case study using data from the astrophysical space mission Gaia [1] has been included in Section IV-C.

II. Correlation dimension

The idea of the correlation dimension method is to estimate the intrinsic dimension via a pairwise distances algorithm which counts the number of point pairs that are closer to each other than a given radius. Let $\Omega = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^D$ denote a set of data points, and r any positive number. The correlation integral, according to the Grassberger–Procaccia (GP) method [5], is defined as

$$C(r) = \lim_{n\to\infty} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} I\left(\|x_j - x_i\| \leq r\right) \tag{1}$$

where $I(.)$ is an indicator function, and $\|x_j - x_i\|$ denotes the Euclidean distance between data points, x_j and x_i. In practice, when $r \to 0$, then $C(r)$ is monotonically decreasing to zero. Note also that the number of data pairs which can be formed from n points is given by $\begin{pmatrix} n \\ 2 \end{pmatrix} = \frac{1}{2}n(n-1)$, which is just the inverse normalizing constant, so that clearly $0 \leq C(r) \leq 1$. Now, the correlation dimension is defined by:

$$d_{cor} = \lim_{r\to 0} \frac{\ln\left(C(r)\right)}{\ln(r)}. \tag{2}$$

Therefore, for small r, the dimensionality can be obtained as the slope of the (linear part of) the "log-log" curve of $\ln(C(r))$ versus $\ln(r)$ [4].

Generally, the fractal dimension of a data set is affected by several factors; the relationship among variables, the data dimensionality, the intrinsic dimension of the data set, the portion of distance pairs that are used for calculation, and the sample size n [17]. Notably, the definition (1) of the correlation integral would require an infinitely sized data set. Compared to the box–counting dimension, the correlation dimension is in practice less demanding on the sample size, and has a larger dynamical range of $O(n^2)$. Furthermore, it can be evaluated for smaller values of r [22] [9].

The main problem with the practical implementation of the correlation dimension is that the correlation integral needs to be estimated for a ball of radius tending to 0. Clearly, the radius r can not be equal to zero because this implies that there are no data points in the circle, yielding "NAN" at $C(0)$. Hence, one needs to decide on a suitable range of values of r which is used to arrive at an estimate of the ID [22].

With our techniques, we try to capture the distance pairs of $C(r)$ in a more effective way and consistent with the GP method. The algorithms achieve the estimation of the ID of a given data set at radius $r = 0$. The developed algorithms are Intercept method, Slope method and Polynomial method. While the Slope method is effectively an implementation of the log-log technique described above, which makes use of the approximately linear part of the correlation integral curve, the other two methods are entirely new and tackle the problem by direct exploitation of features of the function $\ln C(r)/\ln r$ and $C(r)$, respectively. All three approaches are based on the idea of linear regression. The improved methods are described in the next sections.

III. Practical computation of correlation dimension

A. Intercept method

The intercept method estimates the fractal dimension not through direct evaluation of $C(r)$ at $r \approx 0$, but through linear extrapolation of the graph $(r, c(r))$, where $c(r) = \ln C(r)/\ln r$. In practice, the curve $c(r)$ is plotted versus the radius r. Then a grid of values of r, say $r_j, j = 1, \ldots, s$ is chosen which is positioned close to 0 and contains a sufficient number of data points. In practice choices like $0.3 \leq r \leq 0.5$, with a grid size of $s = 30$, work well. Hence, it is only necessary to compute the correlation integral for a portion of data pairs which reduces the computational time.

This approach is motivated through similar ideas proposed by Rummel [21], who suggested backwards extrapolation to obtain regression estimates under covariate measurement error ("SIMEX"). Following this idea, we predict the intrinsic dimension by extrapolating a linear regression line (fitted to the values $(r_j, c(r_j)), j = 1, \ldots, s$) to $r \to 0$. The intrinsic dimension is then obtained as the intercept of the fitted linear equation. Specifically, consider a linear regression with least squares

estimator a (intercept) and c (slope), then the correlation dimension can be approximated as

$$c(r) = a + cr,$$

which at $r = 0$ gives

$$d_{cor} = c(0) = a.$$

B. Slope method

In this section, we exploit the previously stated properties of the log-log curve of the correlation integral. Hence, suppose the high-dimensional data set Ω has an intrinsic dimension d. If the sample size is large enough then the number of distance pairs will increase due to the increase of r, and since $C(r)$ is a function of r, then as r increases $C(r)$ will increase proportionally with r^d. Thus, the $C(r)$ curve can be described by;

$$C(r) \propto r^d,$$

or we can write

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

where d is the intrinsic dimension and c is constant. Applying the logarithm to the above equality, we get

$$\ln(C(r)) = \ln(c) + d\ln(r).$$

By substituting into equation (2), the correlation dimension can be formulated as

$$d_{cor} = \lim_{r\to 0} \frac{\ln(C(r))}{\ln(r)} = \lim_{r\to 0} \frac{\ln(c) + d\ln(r)}{\ln(r)}.$$

$$d_{cor} = \lim_{r\to 0} \frac{\ln(c)}{\ln(r)} + \frac{d\ln(r)}{\ln(r)}.$$

At $r \to 0$, the term $\frac{\ln(c)}{\ln(r)}$ approaches 0. Therefore, $d_{cor} = d$, which means that the correlation dimension is a good estimate of the intrinsic dimension of the corresponding data set.

Now, to obtain the estimate of intrinsic dimension, we plot the curve of $\ln(C(r))$ versus $\ln(r)$ and the slope value is computed using a simple linear regression method which fits a line on the curve of $\ln(C(r))$, this is done by assuming that the equation of regression line is;

$$\ln(C(r)) = b\ln(r) + a;$$

where, a is the intercept, and the slope of the equation (b) is the estimate of the intrinsic dimension. For the choice of interval in which the linear regression is fitted, we recommend again $0.3 \leq r \leq 0.5$.

C. Polynomial method

This section provides a potential model for the correlation integral based on the relationship between the correlation integral $C(r)$ and the radius r. We develop an approach in which $C(r)$ is explicitly modelled through a higher–order polynomial, considering the following condition;

- at $r = 0, \Rightarrow C(0) = 0$.

We state the following general result (see appendix for proof): For a polynomial with degree p, let $C(r) = a_p r^p + \cdots + a_2 r^2 + a_1 r + a_0$, and subject to constraint $C(0) = 0$; one has,

1) If a_1 exists then $d = 1$,
2) For $a_1 = 0$, then $d = 2$,
3) For $a_2 = a_1 = 0$, then $d = 3$,
4) For $a_{p-1} = \cdots = a_2 = a_1 = 0$, then $d = p$.

The correlation dimension can be obtained using multiple linear regression (e.g., function `lm` in `R`), and as a default we assume that $C(r) = a_4 r^4 + a_3 r^3 + a_2 r^2 + a_1 r$ (the polynomial degree would need to be increased in order to detect IDs with $d \geq 5$). Then, one examines the significances of parameters by t-test, and the first significant parameter corresponds to the ID. In practice, we recommend to leave the significance level of this test unspecified, but to determine the ID by the *most* significant parameter.

It is important to emphasize that, in difference to the intercept and slope methods, the polynomial method provides an integer ID estimator (so, the estimated ID is not really 'fractal' in a strict sense).

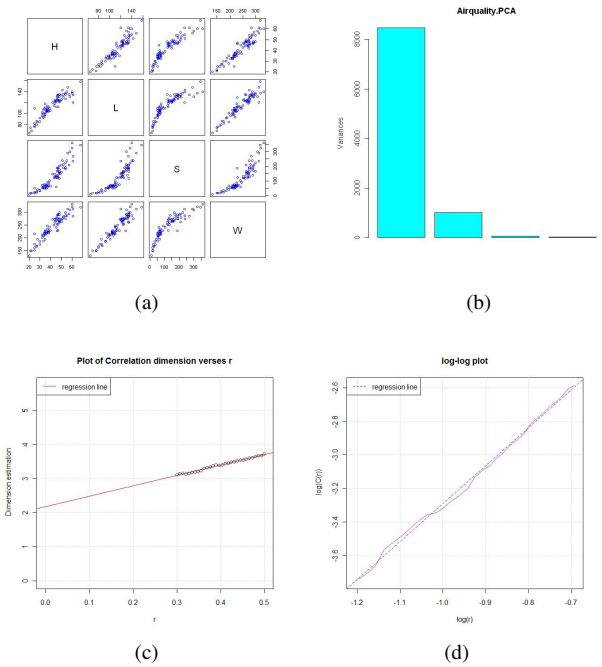
IV. Experimental result

In this section, we verify our methods on real data sets; the horse mussels data ($D = 4$), airquality data ($D = 4$), and the gaia data ($D = 19$), all of which are available in R packages [19]. In order to implement our methods, the data is scaled to zero mean and unit standard deviation as the first step. Practically, for the correlation dimension method (Intercept method and Slope method), the sequence of r is 0.3 to 0.5. This choice of the lower bound guarantees that a sufficient number of data pairs is included in the computation of $C(r)$ [12]. Comparison is made with principal component analysis (PCA).

A. Horse mussels data

We discuss horse mussel data (sampled from the Marlborough Sounds, NZ) with 82 observations on four variables; shell width (W), height (H), length (L), and mass (S). Figure 1b illustrates the result of a principal component analysis on the (scaled) data set. One finds that the first and second PC explain 94% and 3%, respectively, of the total variance. Clearly, when performing linear dimension reduction via PCA, users decide the dimension by how much variance they want to preserve. Hence, depending on this choice (common default choices would be 90% or 95%), one will conclude that the (linear) ID for this data set is 1 or 2, which matches the visual impression from Figure 1a.

Intercept method. We use this technique to estimate the intrinsic dimension via the correlation dimension method. We start the implementation by studying the correlation dimension curve with radius r. Here, figure 1c illustrates that the curve is given by a grid on the right side, and the curve looks to be reasonably linear from 0.3 to 0.5. Figure 1c displays the fitted linear regression $c(r) = a + cr$ on



(a) (b) (c) (d)

Figure 1: Horse mussels data; (a) Scatter plot, (b) Scree plot, (c) Correlation dimension curve with range of r from 0.3 to 0.5, (d) Log-log plot of correlation integral versus radius.

```
Coefficients:
        Estimate Std. Err t value Pr(>|t|)
re      -0.06009 0.01720 -3.495   0.00172 **
I(re^2)  0.61647 0.17471  3.529   0.00158 **
I(re^3) -0.75786 0.55602 -1.363   0.18457
I(re^4)  0.73413 0.55925  1.313   0.20075
---
Signif.codes:0 *** 0.001 ** 0.01 * 0.05 . 0.1
```

TABLE I.: Mussels data: summary table of the output of Polynomial method.

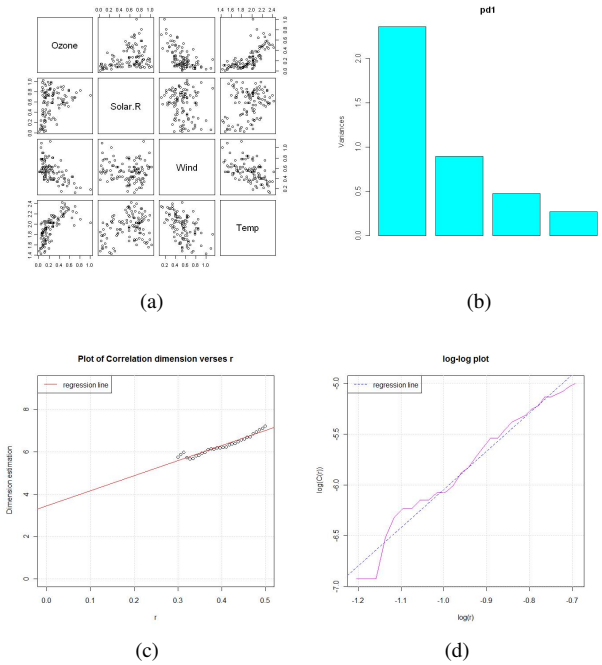
the correlation dimension curve. Therefore, the intrinsic dimension estimation is equal $a = 2.17461$ which is the intercept value in the linear equation of $y = 2.17461 + 3.06748(r)$.

Slope method. The figure 1d displays the plotted curve of $\ln(C(r))$ versus $\ln(r)$ with a fitted linear regression. Then the estimated intrinsic dimension is equal $b = 2.264904$, this value is close to the dimension value estimated by intercept method.

Polynomial method. We test the significance of parameters using a polynomial fit to $C(r)$ with degree 4. The results of the polynomial regression are provided in table I. The most significant parameter is a_2, and hence, ID $= 2$, though the significance of a_1 is of similar magnitude, so there may also be evidence for ID $= 1$.

B. Air Quality data

The air quality data, displayed in figure 2a with 111 observations, consists of; mean ozone (Ozone), solar radiation (Solar.R), average wind speed (Wind), and maximum daily temperature (Temp) recorded in New York, May to September 1973. In figure 2b, the scree plot shows that three components explain 93% of the total variance of the scaled data, so depending on where one places the cut point one would decide for IDs of 3 or 4. This result is intuitive when considering the data, which do not possess a very pronounced inner structure.



(a) (b) (c) (d)

Figure 2: Airquality data; (a) Pairwise plots, (b) Scree plot of four measurements of airquality data, (c) $c(r)$ curve versus r, which is roughly linear for a reasonable range of r. (b) Log-log plot of correlation integral versus r

Next, the ID is obtained via the correlation dimension: *Intercept method.* We plot $c(r)$ versus versus r. Figure 2c shows that the curve of correlation dimension is mostly linear in the chosen range of r. Figure 2c displays the fitted regression line $c(r) = a + cr$ on the correlation dimension curve. Therefore, ID = 3.438883 which is the intercept value in the linear equation of $y = 3.438883 + 7.127591\ (r)$.

Slope Method. The linear regression is fitted through the curve of $\ln(C(r))$ in the log-log plot as shown in figure 2d. The linear equation is $y = -2.279512 + 3.764282\ln(r)$, so the intrinsic dimension is equal $b = 3.764282$. The result is reasonably close to the intercept method.

Polynomial method. We test the significance of parameters using a polynomial with degree 4. From provided $*$ symbols in the summary (table II) we see immediately that the most significant parameter is a_3, and, hence, the estimated ID is equal to 3.

We find that the techniques arrive at sensible results which broadly agree with each other, and are consistent with the visual impression and the scree plot.

```
Coefficients:
          Estimate Std. Err t value Pr(>|t|)
re         0.01320  0.00389   3.395 0.00221 **
I(re^2)  -0.13403  0.03765 -3.559 0.00146 **
I(re^3)   0.44958  0.11535   3.898 0.00061 ***
I(re^4)  -0.35947  0.11263 -3.192 0.00368 **
---
Signif.codes:0 *** 0.001 ** 0.01 * 0.05 . 0.1
```

TABLE II.: Airqualty data: the result of fitting a polynomial of degree 4.

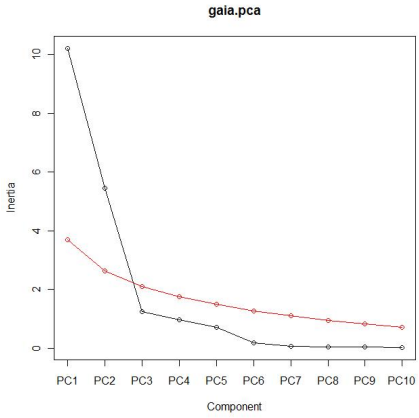
C. Gaia data

Gaia is a space observatory mission of the European Space Agency (ESA). The mission aims to collect data from about 1 billion stars in our Galaxy and extragalactic objects. Gaia will provide comprehensive astrophysical information for each star, including mass, temperature, chemical composition, among others. One of its major goals is to determine the distances, the positions, and annual proper motions of stars [1]. Gaia consists of two telescopes providing two observing directions with a fixed, wide angle between them. This will sample the spectral energy distribution at 96 points across the optical and near-infrared wavelength range (3301000nm). The measurements themselves are photon counts (energy flux). Therefore, each star can be represented as a point in a 96-dimensional data space.

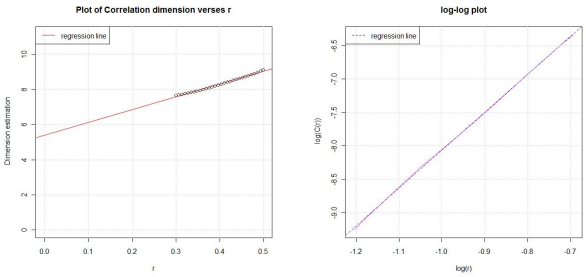
We are going to analyze a simplified version of such data, which is generated by computer models. Our data consist of photon counts measured in 16 (rather than 96) wavelength bands with 8286 observations. Additionally we include the three astrophyical parameters temperature, metallicity, and gravity (which form the input space of the computer model) in our data set, giving a total of $D = 19$ dimensions for the raw data. We begin our analysis by providing a scree plot in figure 3a. The quickly falling curve starting in the left top provides the share of total variance explained by the respective principal component. The usual way of interpreting this plot is to identify sudden breakpoints, which separate the informative from the noise-carrying components. One finds here that there are two possible interpretations for this data set: There is a first breakpoint at about 3 components, and a second (weaker) breakpoint between 5 and 6 components. Commonly, when performing linear dimension reduction via PCA, users decide the dimension by how much variance they want to preserve. In the first case, 89% of the total variance are explained, while in the second case about 98% is explained. Note that the result $d = 3$ is backed up by the broken stick method (this technique compares variation explained by the $j-$th PC with the expected length of the j-th largest segment if the total variance was randomly distributed into D parts), represented by the flatter (red) curve.

We now compare these results to the estimated dimensionality via the correlation dimension.

Intercept method. We study the correlation dimension



(a)



(b) (c)

Figure 3: Gaia data; (a) Scree plot of 19 variables, (b) The implementation of Intercept method "$c(r)$ curve versus r", (c) Log-log plot of correlation integral versus r

curve $c(r)$ as a function of radius r. As shown in figure 3b, the curve of correlation dimension looks to be reasonably linear in the chosen range of r. Figure 3b also displays the fitted regression line $c(r) = a + c\,r$ on the correlation dimension curve. Then, ID = 5.401008 which is the intercept value in the linear equation of $y = 5.401008 + 7.298104\ (r)$.

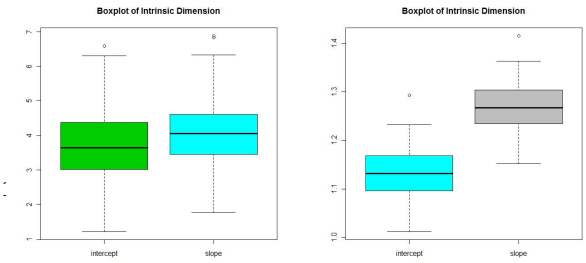
Slope method. The plot in figure 3c displays the curve of $\ln(C(r))$ versus $\ln(r)$ with a fitted linear regression. Therefore, the estimated intrinsic dimension is equal $b = 5.657659$; this value is close to the dimension value estimated by intercept method.

Polynomial method. The ID is derived by considering the significances of parameters. For a polynomial of degree 5, one observes from Table III that the most significant parameter is by far a_5, so the intrinsic dimension of 5 is clearly identified. We should note that a further increase of the degree of the polynomial for this data set provides a somewhat less clear picture, since the higher–degree polynomials correlate in a complex manner with each other, which dilutes the distinctiveness with which the intrinsic dimension is identified.

We find that our approaches indicate that the estimated intrinsic dimension for the Gaia data could be between 5 and 6, which is a sensible result, and agrees with one of the two possible interpretations from PCA. Without providing the results explicitly, we note at this occasion

```
Coefficients:
          Estimate Std. Err  t value Pr(>|t|)
re        0.000814 0.000501   1.625   0.1166
I(re^2) -0.011956 0.006337 -1.887   0.0709 .
I(re^3)  0.062008 0.029009  2.138   0.0425 *
I(re^4) -0.149304 0.057140 -2.613   0.0150 *
I(re^5)  0.190286 0.041011  4.640 9.47e-05 ***
------------
Signif.codes:0 *** 0.001 ** 0.01 * 0.05 . 0.1
```

TABLE III.: Gaia data; the result of fitting a polynomial of degree 5.



(a) (b)

Figure 4: Simulation study; box plots of ID estimates via Intercept and Slope method of 100 data sets generated from multivariate Gaussian distribution. a: first simulation, b: second simulation.

that (our variant of) Brand's method [3] [13], representing a *local* ID estimation technique, produces an ID value of about 2.4 for this data set; hence, favoring the alternative PCA–based interpretation. In general, local methods will provide smaller IDs than global methods, since they are able to resolve the local data structure more flexibly [13].

It should also be noted that both results have a plausible physical interpretation. Since the input space is three-dimensional, and since the remaining 16 variables are generated from this input space, there is a strong argument for an intrinsic dimension of 3. On the other hand, the 16-dimensional data cloud of photon counts, which has been simulated in some complex manner from the APs, will arguably increase the ID of the whole data set at least to some extent, where it is known that this increase should be less than three since the first three principal component scores of the 16–dimensional photon counts are strongly correlated [7]. This is reflected in the ID of 5 obtained through the correlation dimension technique.

D. Simulation studies

The purpose of this section is to present the precision of our approaches. We generate data sets of known ID and try to identify its ID through Intercept and Slope methods. We consider two cases; firstly (a), data set of size $n = 200$ with dimension $D = 4$ is generated from a multivariate Gaussian distribution with parameters $\mu = (9, 5, 6, 4)$, where the diagonal of the covariance matrix is equal to $(50, 50, 50, 50)$. Since these data do not possess any inner structure, we would assume the ID to be equal (or close to) 4 in this case. We generate 100 data sets in this manner, and for each sample we calculate the ID estimate. For illustration, a boxplot is provided which shows the median and distribution of ID estimates (figure 4a). These results indicate that both methods provide reasonable ID estimates. In fact, the slope method gives a result very close to $D = 4$ with a median slope estimate of $d = 3.854845$, while the median of the IDs obtained via the intercept method is 3.497044.

Secondly (b), the data is generated by adding four–variate Gaussian noise, with zero means and identity matrix serving as covariance, to data distributed uniformly on a straight line (think of a long cigar–like object in 4D space). We would assume these data to have ID roughly equal to 1. Again, we provide a box plot of the ID estimates for 100 simulated data sets. Figure 4b illustrates that both the Intercept and the Slope methods yield good ID estimates, with the Intercept method achieving results (median: 1.162404) which are closer to 1 than the Slope method (median: 1.291982).

V. Conclusion

The estimation of intrinsic dimension is very useful to deal with real-life data with high dimension. In this paper we develop new approaches for calculation of the intrinsic dimensionality via correlation dimension. We have investigated three techniques, two of which are novel, to implement fractal ID estimation via the correlation integral. Both Intercept and Slope methods provide non-integer ID estimates while the Polynomial method provides an integer value.

All three methods could be classified as nonparametric methods, as opposed to linear methods such as PCA. Conceptually, the "linear" intrinsic dimension should provide an upper bound for IDs achieved via nonlinear methods, and in fact, we have observed the values suggested by PCA-based ID to be often larger than those obtained by nonparametric ID estimation methods. To be even more precise, within the nonparametric methods, we found that global methods tend to produce larger IDs than local methods.

The proposed techniques require relatively few data points and are not demanding on the sample size. For the Intercept and Slope method, the chosen range of r is motivated by the part of the respective curve that looks approximately linear. These regions of linearity may differ between different data sets, but we have provided default choices, which, according to our experience, work well for a wide range of data sets. The concepts introduced in this paper are not restricted to a particular type of application. We have given three examples – from the environmental and physical sciences – where the methods turned out to be useful, but they could be applied onto data sets of any kind, including, for instance, data (bases) which are created and collected in the internet.

Examples with real data verify the concept of estimating correlation dimension at exactly $r = 0$. All proposed methods are easy to implement and to apply, and the experimental analysis indicates that the methods are able to deal with various types of data, including linear and non–linear structures. A simulation study has confirmed that the Intercept and Slope method provide ID estimates which, in average, are close to the underlying "true" ID. The Polynomial method is of theoretical appeal, though we have not attempted a simulation study since the result needs to be extracted manually from the regression output.

REFERENCES

[1] C.A.L. Bailer-Jones, Determination of stellar parameters with GAIA, *Astrophysics and Space Science*, **280**, 2002, 21-29.

[2] R.E. Bellman, R.E., Adapaptive Control Processes, In *Princeton University Press*, 1961.

[3] M. Brand, Charting a manifold, In *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, MA, **15**, 2003, 961–968.

[4] F. Camastra, Data dimensionality estimation methods:a survey, *Pattern Recognition*, **36**, 2003, 2945–2954.

[5] F. Camastra and A. Vinciarelli, Estimating the intrinsic dimension of data with a fractal-based method, *IEEE Trans. Patt. Anal. Mach. Intell.*, **24**, 2002, 1404-1407.

[6] J. Einbeck, G. Tutz and L. Evers, Local principal curves, *Statistics and Computing*, **15**, 2005, 301-313.

[7] J. Einbeck, L. Evers and C. Bailer-Jones, Representing complex data using localized principal components with application to astronomical data. In: Gorban, A, Kegl, B, Wunsch, D, & Zinovyev, A: Principal Manifolds for Data Visualization and Dimension Reduction; *Lecture Notes in Computational Science and Engineering* **58**, 2008, 180–204.

[8] K. Fukunaga, and D.R. Olsen, An Algorithm for Finding Intrinsic Dimensionality of Data, *IEEE transaction on computers*, **20**, 1971, 176–183.

[9] P. Grassberger and I. Procaccia, I., *Measuring the strangeness of strange attractors*, Physica D: Nonlinear Phenomena, **9**, 1983, 189-208.

[10] M. Hein, and J.Y. Audibert, *Intrinsic dimensionality estimation of submanifolds in* R^d, In: Proceedings of the 22nd International Conference on Machine Learning (ed. Morgan Kaufmann), 2005, 289-296.

[11] D.A. Jackson, Stopping rules in principal component analysis: A comparison of heuristical and statistical approaches. *Ecology* **74**, 1993, 2204–2214.

[12] Z. Kalantan, and J. Einbeck, On the computation of the correlation integral for fractal dimension estimation, *International Conference on Statistics in Science, Business, and Engineering (ICSSBE2012)*, IEEE conference publications, doi 10.1109/ICSSBE.2012.6396531, 2012, pages 80–85.

[13] Z. Kalantan and J. Einbeck, 2012. An overview of intrinsic dimension estimation techniques. Proceedings of the 1st ISM International Statistical Conference 2012, Johor, Malaysia, 2012, 516 – 524.

[14] U. Kruger and L. Xie, Statistical Monitoring of Complex Multivariate Processes: With Applications in Industrial Process Control, Wiley, 2012.

[15] E. Levina and P. Bickel, Maximum likelihood estimation of intrinsic dimension, In: *Advances in NIPS* **17**, Eds. L. K. Saul, Y. Weiss, L. Bottou, MIT Press, 2005.

[16] B. Mandelbrot, *Fractal Geometry of Nature*, Freeman, San Francisco, 1982.

[17] D. Mo and S. Huang, Fractal-based intrinsic dimension estimation and its application in dimensionality reduction, *IEEE Transactions on knowledge and data engineering*, **24**, 2012.

[18] A. Pentland, *Fractal -based description of natural scenes*, IEEE Trans. Patt. Anal. Mach. Intell., **6**, 1984, 661-674.

[19] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`, 2012.

[20] M. Raginsky, and S. Lazebnik: *Estimation of intrinsic dimensionality using high-rate vector quantization*, Advances in Neural Information Processing Systems, **19**, 2005, 352-356.

[21] D. Rummel, The relevance vector machine under covariate measurement error, In: C. Weihs & W. Gaul (eds), *Classification—The Ubiquitous Challenge*, Springer, 2005, 296-303.

[22] J. Theiler, Estimating Fractal Dimension, *J. Optical Soc. of Am.*, **7**, 1990, 1055–1073.

[23] D. Zhang, A. Samal, and J.r. Brandle, *A method for estimating fractal dimension of tree crowns from digital images*, CSE Journal Articles, Paper 97, 2007.

Jochen Einbeck obtained an undergraduate degree in Mathematics and Physics from University of Munich in 1999, and a Ph.D. in Statistics from the same institution in 2003. After a postdoctoral position at NUI Galway (Ireland), he is working since 2006 as Lecturer at Durham University, UK. His research interests include dimension reduction, smoothing, and statistical modelling in general.

Zakiah Kalantan received her B.Sc. Degree in Statistics and Computer Science in 1997 from King Abdulaziz University, Jeddah, Saudi–Arabia. She received her M.Sc. in Statistical Science from the same instutition in 2001. She is currently employed as Lecturer at King Abdulaziz University, and studying for a Ph.D. degree at Durham University, UK.