Identification of models for linear and nonlinear non-causal data structures

Part II: Intrinsic Dimensionality Estimation

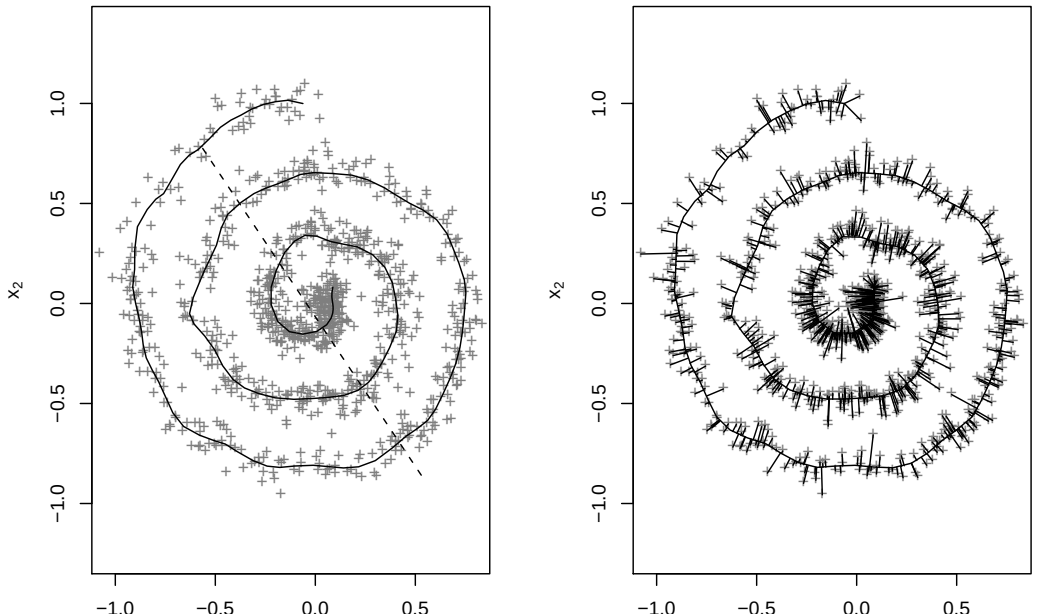
Zakiah Kalantan, Jochen Einbeck, Uwe Kruger

Intrinsic dimension (ID)

- Given data set $\Omega = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, with $\boldsymbol{x}_i \in \mathbb{R}^D$.
- Interested in the number of parameters, say $d \leq D$, required to describe the data without loss of information.
- Standard concept: Principal component analysis.
- However, this may not be adequate if the data do not reside on linear subspace of $\mathbb{R}^D$:

Intrinsic dimension (ID)

- Given data set $\Omega = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, with $\boldsymbol{x}_i \in \mathbb{R}^D$.
- Interested in the number of parameters, say $d \leq D$, required to describe the data without loss of information.
- Standard concept: Principal component analysis.
- However, this may not be adequate if the data do not reside on linear subspace of $\mathbb{R}^D$:
- For instance, does a bivariate spiral have $d = 1$ or $d = 2$?



Correlation dimension

- "Correlation integral" proposed by Grassberger–Procaccia (1983)

$$C(r) = \lim_{n\to\infty} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} I\left(\|\boldsymbol{x}_j - \boldsymbol{x}_i\| \leq r\right) \qquad (1)$$

- Correlation dimension:

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

Correlation dimension

- "Correlation integral" proposed by Grassberger–Procaccia (1983)

$$C(r) = \lim_{n\to\infty} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} I\left(\|\boldsymbol{x}_j - \boldsymbol{x}_i\| \leq r\right) \qquad (2)$$

- Correlation dimension:

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

- Is this plausible? If ID $= d$, we would expect $C(r) \sim r^d$.
- Then,

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

Correlation dimension

- "Correlation integral" proposed by Grassberger–Procaccia (1983)

$$C(r) = \lim_{n\to\infty} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} I\left(\|\boldsymbol{x}_j - \boldsymbol{x}_i\| \leq r\right) \qquad (3)$$

- Correlation dimension:

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

- Is this plausible? If ID $= d$, we would expect $C(r) \sim r^d$.
- Then,

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

- Problem: Correlation integral needs to be computed for a ball with radius tending to 0!

Three solutions

- Slope method: Estimate d_{cor} as slope, $\hat{b}$, of regression line

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

(Camastra, 2003).

Three solutions

- Slope method: Estimate d_{cor} as slope, $\hat{b}$, of regression line

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

(Camastra, 2003).

- Intercept method: Let $c(r) = \ln C(r)/\ln r$. Estimate d_{cor} as intercept, $\hat{a}$, of regression line

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

Three solutions

- Slope method: Estimate d_{cor} as slope, $\hat{b}$, of regression line

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

(Camastra, 2003).

- Intercept method: Let $c(r) = \ln C(r)/\ln r$. Estimate d_{cor} as intercept, $\hat{a}$, of regression line

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

- Polynomial method: Assume polynomial shape

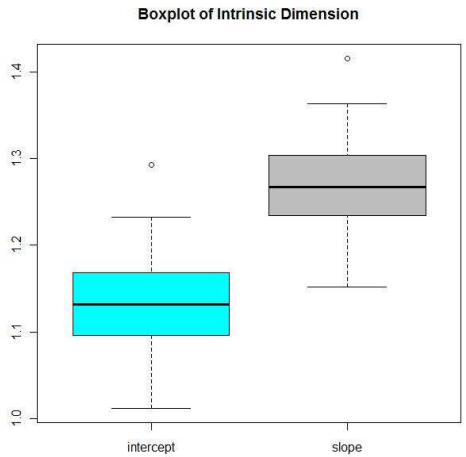
$$C(r) = a_p r^p + \cdots + a_2 r^2 + a_1 r + a_0 \text{ subject to } C(0) = 0;$$

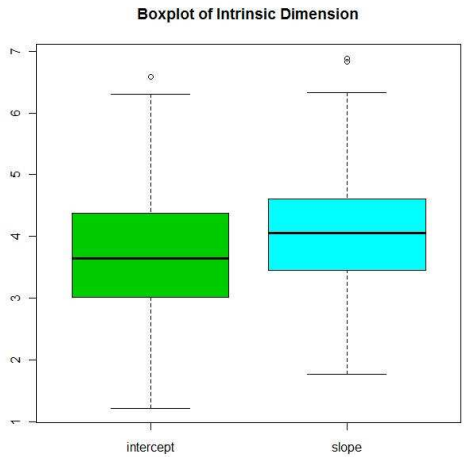
Then simple proof via L'Hôpital shows that

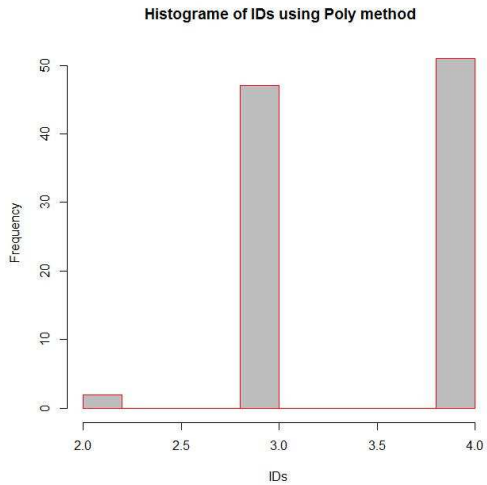
$$\text{If } a_p \neq 0 \text{ and } a_{p-1} = \cdots = a_2 = a_1 = 0, \text{ then } d_{cor} = p.$$

Simulation

- Simulate $r = 100$ data sets of size $n = 200$ from
  (i) a long noisy 'cigar' in 4D space (i.e., $D = 4$, $d = 1$)
  (ii) 4–variate Gaussian noise (i.e., $D = d = 4$)
- The resulting ID's are [left: (i), middle: (ii), right: polynomial (ii)]



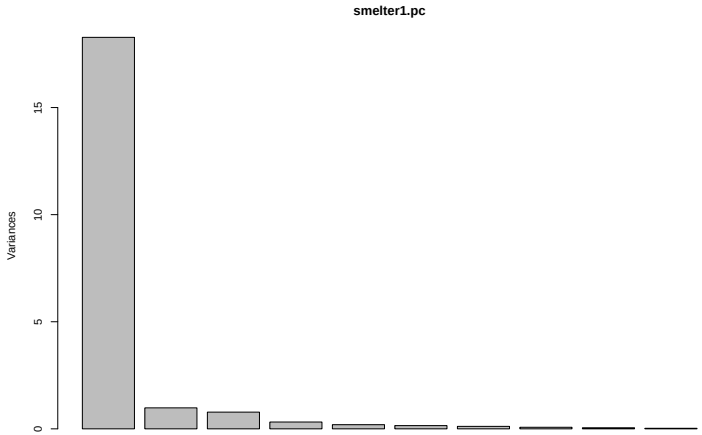


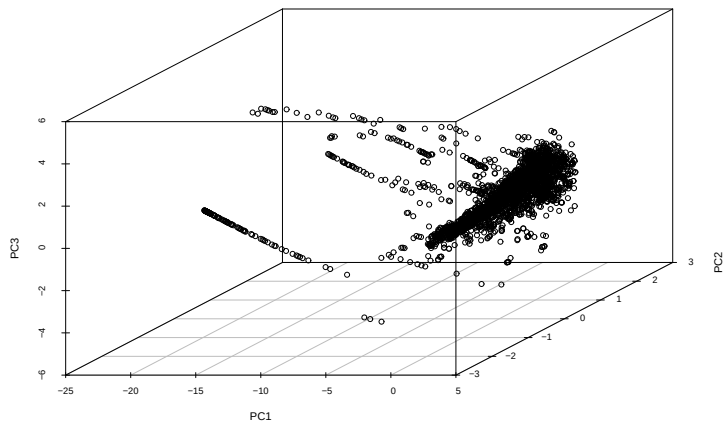


- For simulation (i), the polynomial method produced the correct ID=1 in 100% of the cases.

Industrial glass melter data

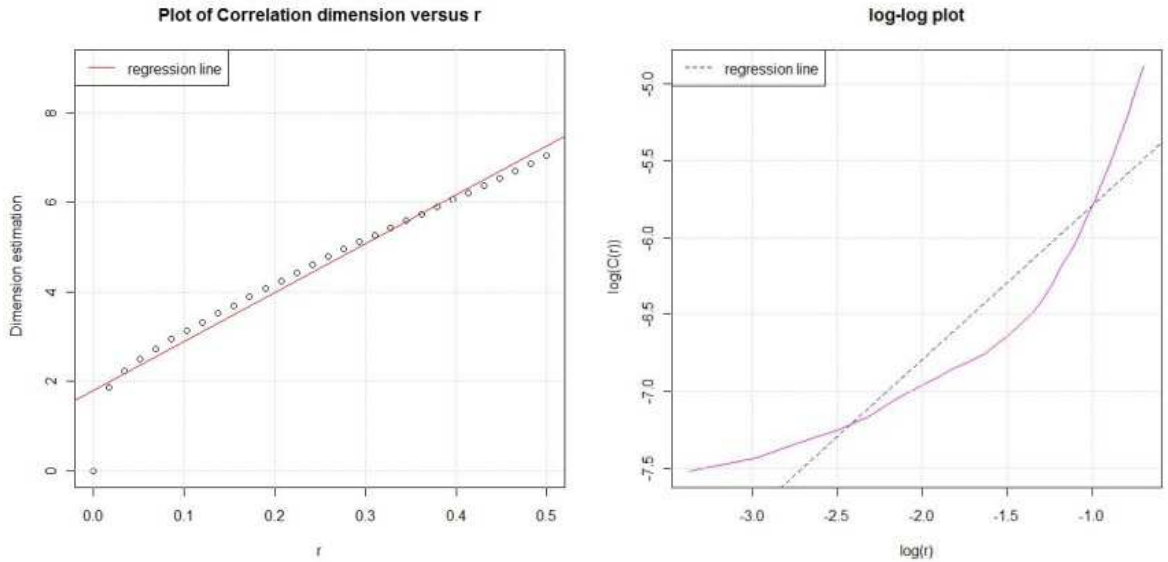
- 21-dim data from an industrial melter: 15 temperature sensors, 4 induction coils, viscosity, voltage.
- (Scaled) PCA suggests small ID:





Industrial glass melter data (cont'd)

- Intercept method and slope method



- Polynomial method

```
   Estimate  Std.Error   t-value    Pr(>|t|)
re          0.013943  0.001576    8.845      2.56e-09 ***
I(re^2)    -0.071003  0.016837   -4.217      0.000265 ***
I(re^3)     0.161690  0.056048    2.885      0.007769 **
I(re^4)    -0.030528  0.058397   -0.523      0.605558
```

- Overall, some evidence for ID=1 as well as ID=2.

Some remarks

- Compared to the (previously existing) "slope method", the intercept method produces comparable or favorable results.
- The polynomial method is theoretically appealing, but difficult to use for high data dimension D, since a polynomial degree $d \leq p \leq D$ needs to be chosen.
- The correlation dimension occupies some middle ground between purely linear methods (such as PCA) and purely topological methods (which average over localized IDs representing the dimension of the tangent space along the manifold). Indeed, the IDs obtained via the correlation dimension are generally equal or larger than what a scree plot (broken stick, etc...) would suggest, but smaller than the estimates obtained through local (topological) techniques, such as Brand's (2003) algorithm.
- Note: All this is *non-causal!* The value d may underestimate the number of variables needed for applications such as *regression*.

References

Brand (2003). *Charting a manifold*. In: Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, **15**, 961–968.

Camastra (2003). Data dimensionality estimation: A survey. *Pattern recognition* **36**, 2945–2954.

Grassberger & Procaccia (1983). Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, **9**, 189–208.

Liu, Xie, Kruger, Littler & Wang (2008). Statistical–based monitoring of Multivariate Non–Gaussian Systems. In *AIChE*, **54**, 2379–2391.

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

Einbeck & Kalantan (2013). Intrinsic Dimensionality Estimation for high–dimensional data sets: New approaches for the computation of correlation dimension. *Journal of Emerging Technologies in Web Intelligence*, Special Issue: *Novel Applications of Data Analytics*, in press.