ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Nearest neighbors and correlation dimension for dimensionality estimation. Application to factor analysis of real biological time series data.

J. Lapuyade-Lahorgue and A. Mohammad-Djafari [*]

Laboratoire des signaux et systèmes (L2S) - Supelec
UMR 8506 CNRS-SUPELEC-UNIV PARIS SUD
plateau de Moulon, 3 rue Joliot-Curie, 91192 GIF-SUR-YVETTE Cedex, France

Abstract. Determining the number of components in dimensionality reduction techniques is still one of the open problems of research on data analysis. These methods are often used in knowledge extraction of multivariate great dimensional data, but very often the number of components is assumed to be known. One of the classical methods to estimate this dimensionality is based on the Principal Components Analysis (PCA) eigenvalues [1, 2]. However, this method supposes that the model is linear and the signals are Gaussian. To be able to consider non-linear and non-Gaussian cases, we propose in this paper "measure based methods" as nearest neighbors dimension and correlation dimension. The comparaison between the three methods is evaluated both with simulated data and with real biological data, which are gene expression time series. The main goal of this study is to estimate the minimum number of factors.

1 Introduction

The estimation of the intrinsic dimension constitutes a preliminary treatment for dimensionality reduction. The dimensionality reduction is divided into two principal domains: the Factorial Analysis (FA) and the Low-Dimensional Representation (LDR). The FA includes all techniques which allows to estimate a reduced number of factors able to represent the data as conformly as possible. This number of factors appears to be the intrinsic dimension of the data. Amongst these FA techniques, one can quote Principal Component Analysis (PCA) [1], the Independent Component Analysis (ICA) [3] and the Linear and General discriminent analysis (LDA-GDA). Whereas the LDR aim to find a lower dimensional space which preserves some properties of the observed data. Amongst these methods, one can quote Isomap [4, 5], Laplacian Eigenmaps [6] or Diffusion Maps [7]. A good survey of the existing methods can be found in the paper of T. Lin [8]. In this paper, we propose to compare three methods of estimating the dimension used by FA and LDR. These methods are the classical Principal Component Analysis Eigenvalues proposed by [1, 2]. The two other methods are the nearest neighbors method introduced the first time by K.W. Pettis [9] and the correlation dimension introduced by A.M. Farahmand [10, 11]. The book [12] gives the principal details for non-linear dimensionality reduction.

[*] This work is a part of ERASYSBIO-C5Sys European project "circadian and cell cycle clock systems in cancer" http://www.erasysbio.net/index.php?index=272

This paper is structured as following. Firstly, we introduce the nearest neighbors and correlation dimension techniques. Secondly, we test the different methods of dimension estimation on simulated data. Finally, we use the methods on real biological data representing expression of genes.

2 Estimation of intrinsic dimension using nearest neighbors and correlation dimension

In the problem considered in this paper, we consider T realisations $\boldsymbol{g}^{(1)}, \dots, \boldsymbol{g}^{(T)}$ of a M-dimensional random vector $\boldsymbol{g} = (g_1, \dots, g_M) \in \mathbb{R}^M$. These realisations represent the observed data and $\mathbb{R}^M$ is the space of observations. The support of the distribution of $\boldsymbol{g}$ is not necessarily $\mathbb{R}^M$ but a subspace which is locally $\mathcal{C}^1$-diffeomorph to a vectorial space $\mathbb{R}^N$. Such a subspace, denoted $\mathcal{M}$, is called $\mathcal{C}^1$-differentiable manifold and the intrinsic dimension of the data is N. We call "parameterization" of a N-dimensional manifold an application $\boldsymbol{\varphi}$ from $\mathbb{R}^N$ to $\mathcal{M}$ which realizes a local diffeomorphism. Finally, the N-dimensional random vector $\boldsymbol{f} = (f_1, \dots, f_N)$ defined through the parameterization $\boldsymbol{g} = \boldsymbol{\varphi}(\boldsymbol{f})$ is called "source-vector" and the composantes $f_1, \dots, f_N$ are called "factors" or "sources". The model is linear if $\boldsymbol{\varphi}$ is a linear application. We propose in this section three methods to estimate the intrinsic dimension of the data which appears to be the dimension of the manifold where the data lie on. In the linear case, PCA consists in computing the covariance matrix of the observations $\boldsymbol{g}$ and to find its biggest eigenvalues. However, PCA eigenvalues method fails to consider the non-linearity. For the non-linear case, correlation dimension and nearest neighbors method are based on the following argument. Suppose that the random vector $\boldsymbol{g}$ has a density $\boldsymbol{g} \to p(\boldsymbol{g})$ whose the support is the N-dimensional manifold $\mathcal{M}$. Consider $B(\boldsymbol{x}_0, r)$ the ball of center $\boldsymbol{x}_0$ and radius r in the manifold $\mathcal{M}$, then $\mathbb{P}\left(\boldsymbol{g} \in B(\boldsymbol{x}_0, r)\right) = r^N \underbrace{\int_{z \in B(0,1)} p(rz + \boldsymbol{x}_0) d\sigma(z)}_{\eta(\boldsymbol{x}_0, r)}$, where σ is the Lebesgue measure of $\mathcal{M}$ defined from its volume form. For small values of r, one can suppose that $\eta(\boldsymbol{x}_0, r)$ is a constant η_0, consequently:

$$\log\left(\mathbb{P}\left(\boldsymbol{g} \in B(\boldsymbol{x}_0, r)\right)\right) = N \log(r) + \log(\eta_0). \tag{1}$$

Correlation dimension consists then in estimating $\mathbb{P}\left(\boldsymbol{g} \in B(\boldsymbol{x}_0, r)\right)$ by:

$$C_T(r) = \frac{2}{T(T-1)} \sum_{1 \le i < j \le T} 1_{\|\boldsymbol{g}^{(i)} - \boldsymbol{g}^{(j)}\| < r}, \tag{2}$$

where $\|.\|$ is the Euclidian norm of $\mathbb{R}^M$. The dimension is estimated using least-square method by:

$$\hat{N}_{corr} = \frac{\mathrm{Cov}\left(\log\left(C_T(\boldsymbol{r})\right), \log(\boldsymbol{r})\right)}{\mathrm{Var}(\log(\boldsymbol{r}))}, \tag{3}$$

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

where $\boldsymbol{r} = (r_1, \ldots, r_J)$ is a sample of small values of r. In the experiments we choose $(r_1 = 0, r_2 = 0.001, r_3 = 0.002, \ldots, r_J = 1)$.
Whereas, in the nearest neighbors technique, we fixe a number K of neighbors. If $\boldsymbol{x}_0$ is one of the realisations of $\boldsymbol{g}$ and r_K is the distance between $\boldsymbol{x}_0$ and its furthest neighbor, then $\mathbb{P}\left(\boldsymbol{g} \in B(\boldsymbol{x}_0, r_K)\right) \simeq \frac{K}{T}$. Similarly, if we consider a number $\frac{K}{2}$ of neighbors, $\mathbb{P}\left(\boldsymbol{g} \in B(\boldsymbol{x}_0, r_{K/2})\right) \simeq \frac{K}{2T}$. From the relations $\log\left(\frac{K}{T}\right) \simeq N\log(r_K) + \log(\eta_0)$ and $\log\left(\frac{K}{2T}\right) \simeq N\log(r_{K/2}) + \log(\eta_0)$, we deduce $\log(2) \simeq N\log\left(\frac{r_K}{r_{K/2}}\right)$, consequently we estimate N by:

$$\hat{N}_{near} = \frac{T\log(2)}{\sum_{i=1}^{T} \log\left(\frac{r_K(\boldsymbol{g}^{(i)})}{r_{K/2}(\boldsymbol{g}^{(i)})}\right)}. \tag{4}$$

3 Comparaison of the dimension estimation methods

In all experiments presented in this section, we simulate 1000 samplings of data. For this, we use four kinds of data. In all cases, the data lie on a bi-dimensional manifold embedded to $\mathbb{R}^5$, consequently, the observations are realisations of a random vector $\boldsymbol{g} = (g_1, \ldots, g_5)$. In the first model (Linear and Gaussian), the observations come from a linear combination of Gaussians variables f_1 and f_2. More exactly $g_1 = 0.8f_1 + 0.2f_2$, $g_2 = 0.2f_1 + 0.8f_2$, $g_3 = 0.5f_1 + 0.5f_2$, $g_4 = 0.9f_1 + 0.1f_2$ and $g_5 = 0.1f_1 + 0.9f_2$, and f_1 and f_2 are independent and centered-normalized Gaussian variables. In the second model (Linear and Non-Gaussian), we keep the same linear relations between sources and observations, but f_1 and f_2 are independent and uniformly distributed on $[0, 1]$. In the third (Non-Linear and Gaussian) and last experiments (Non-Linear and Non-Gaussian), the relation between sources and observations are $g_1 = \exp(-f_1)\cos(f_2)$, $g_2 = \exp(-2f_1)(\cos(f_2) + 1)$, $g_3 = \exp(-2f_1)(3\cos(f_2) + 5)$, $g_4 = \exp(-f_1)(6\cos(f_2) - 2)$ and $g_5 = \exp(-3f_1)(\cos(f_2) - 2)$. In the third experiment, the sources are Gaussian whereas in the last experiment, the sources are uniformly distributed on $[0, 1]$. For the estimation of dimension, we use 4 methods: PCA eigenvalues, correlation dimension, nearest neighbors dimension with respectively 4 and 10 neighbors. In PCA, we impose to recover 90% of the variance.
In the last series of experiments, we consider 3 sources f_1, f_2, f_3. In the first and the third experiments, the sources are independent and normalized-zero mean Gaussian and in the second and last experiments, the sources are independent and uniformly distributed on $[0, 1]$.
In the linear models, which are the first and second experiment, the relations are $g_1 = 0.8f_1 + 0.1f_2 + 0.1f_3$, $g_2 = 0.1f_1 + 0.8f_2 + 0.1f_3$, $g_3 = \frac{f_1}{3} + \frac{f_2}{3}$, $g_4 = 0.9f_1 + 0.05f_2 + 0.05f_3$ and $g_5 = 0.05f_1 + 0.9f_2 + 0.05f_3$ and in the non-linear models, they are $g_1 = \exp(-f_1)\cos(f_2)$, $g_2 = \exp(-2f_1)(\cos(f_2) + f_3)$, $g_3 = \exp(-2f_1)(3\cos(f_2) + 5f_3)$, $g_4 = \exp(-f_1)(6\cos(f_2) - 2f_3)$ and $g_5 = \exp(-3f_1)(\cos(f_2) - 2f_3)$. The Table 2 presents the estimation of dimension

	Linear Gaussian	Linear Non-Gaussian	Non-linear Gaussian	Non-linear Non-Gaussian
PCA	2	2	1	1
Correlation dimension	2 (1.98)	2 (1.76)	2 (1.62)	2 (1.69)
Nearest neighbors (4)	3 (2.59)	2 (2.43)	2 (2.39)	2 (2.46)
Nearest neighbors (10)	2 (2.35)	2 (2.20)	2 (2.21)	2 (2.09)

Table 1: Dimension estimation results. On brace: estimated value and without brace: the rounding value.

for the different cases. In PCA, we impose to recover 99% of the total variance of the data.

	Linear Gaussian	Linear Non-Gaussian	Non-linear Gaussian	Non-linear Non-Gaussian
PCA	3	3	2	2
Correlation dimension	3 (2.84)	2 (2.29)	2 (1.90)	2 (2.08)
Nearest neighbors (4)	3 (3.11)	3 (3.03)	3 (2.80)	3 (2.76)
Nearest neighbors (10)	3 (3.18)	3 (2.92)	3 (2.59)	2 (2.39)

Table 2: Dimension estimation results. On brace: estimated value and without brace: the rounding value.

One can see that if the relation is not linear, then PCA eigenvalues gives poor results compared to the two other methods. Moreover, the nearest neighbors method seems to give better results in the non-linear case than correlation dimension. The Table 2 shows that the correlation dimension method is very sensitive to the model and with three sources, the results are as poor as those got with PCA eigenvalues. However, it can be related to the choice of the sampling of the radius r. The Tables 1 and 2 show also that the distribution of the data on the manifold has an influence on the quality of estimation. The model influences the number of neighbors in the nearest neighbors method. In the experiments with two sources, chosing 10 neighbors gives better results than with 4 neighbors; whereas with three sources, we get better results by chosing 4 neighbors.

4 Experimentations with real biological time series

In this section, we test the proposed methods on the real biological data. These data represent the expression of genes taken at different times. The expression of a gene is the amount of RiboNucleic Acid (RNA) used for the transcription of the specific gene. In the first experiment, we measure 9 gene expressions, which are the expressions of three metabolism genes “CE2”, “UGT1A1” and “TOP1” in three different organs “liver”, “colon” and “ileum” . The size T of the sampling is of 72 samples. Even if these samples come from 3 classes and for each class from 3 mice measured every 3 hours (8 samples over $24h$), we used all these samples without distinction on classes and mice. Concerning, the estimation of dimension, PCA with 90% of variance gives 5, correlation dimension gives 3.43 and nearest neighbors with 4 neighbors gives 5.31, so the dimensionality of the observed data seems to be equal to 5. In Figure 1, we have estimated the sources using Factorial Analysis, which supposes that the mixing is linear. It can be interesting also to study the data organs by organs separately and genes by genes separately. The Table 3 represents, in its left part, the dimensionality analysis when we observe the three genes separately in the Liver, in the Colon end in the Ileum. In the second part of the table, we observe for the three genes separately, their expressions in the three organes. For instance, the first column "Liver" represents the dimension estimation when the observations are g_1 (CE2, Liver), g_2 (UGT1A1, Liver) and g_3 (TOP1, Liver), and for the fourth column "CE2", the observations are g_1 (CE2, Liver), g_4 (CE2, Colon) and g_7 (CE2, Ileum).

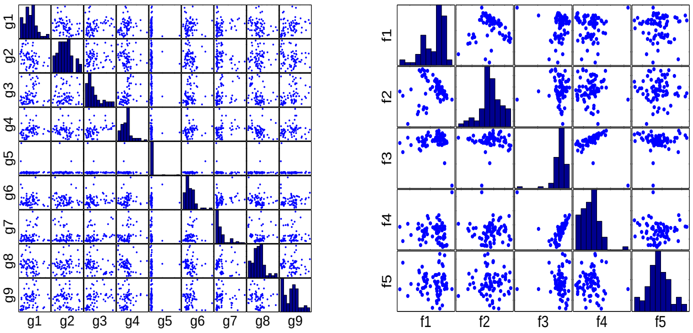

Fig. 1: Scatterplot of the genes data and of the five estimated sources, g_1: CE2, Liver; g_2: UGT1A1, Liver; g_3: TOP1, Liver; g_4: CE2, Colon; g_5: UGT1A1, Colon; g_6: TOP1, Colon; g_7: CE2, Ileum; g_8: UGT1A1, Ileum; g_9: TOP1, Ileum. $f_1, \dots, f_5$: sources.

As we can see from the Table 3, one can reduce the dimensionality of the three dimensional data representing genes' expressions to 2 if the measure have been realized in "Colon" and in "Ileum". However, the genes expression seems to be less dependent in the liver. If we study the correlation between organs in regard of the three genes separetely, it appears to have two factors in regard of UGT1A1 and TOP1.

	Liver	Colon	Ileum	CE2	UGT1A1	TOP1
PCA	3	2	2	3	2	2
Correlation dimension	3 (2.88)	2 (2.39)	2 (2.15)	3 (2.76)	2 (2.28)	2 (2.23)
Nearest neighbors (4)	2 (2.39)	2 (2.29)	3 (2.31)	3 (2.36)	3 (2.41)	2 (2.30)

Table 3: Dimensionality analysis in the three organs separately and for the three genes separetely.

5 Conclusion

In this paper, we have presented two non-linear methods of estimation of dimension. We have tested the efficiency of these methods both on simulated and biological data. The real data are more complex: there are missing and outliers data. Also, for now, we applied the method on only a homogeneous set of data (Gene expressions). We need to extend these methods for the cases where we want to apply them to heterogeneous data.

References

[1] Hotelling H. Analysis of a Complex of Statistical variables with Principal components. *Journal of Educational Psychology*, 1933.

[2] Kukunaga K. and Olsen D.R. An Algorithm for Finding Intrinsic Dimensionality of Data. *IEEE Transactions on computers*, 20(2):176–183, 1971.

[3] Hyvärinen A. and Oja E. Independent component analysis: Algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[4] Tenenbaum J., Silva V.D., and Langford J. A global geometric framework for non-linear dimensionality reduction. *Science*, pages 2319–2323, 2000.

[5] Silva V. and Tenenbaum J. Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems*, 15:705–712, 2003.

[6] Belkin M. and Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(1373-1396), 2003.

[7] Nadler B., Lafon S., Coifman R.R., and Kevrekidis I.G. Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. *Advances in Neural Information Processing Systems*, 18:955–962, 2005.

[8] Lin T., Hongbin Z., and Lee S.U. Riemannian Manifold Learning for Nonlinear Dimensionality Reduction. volume 3954, pages 44–55, Austria, May 2006. European Conference on Computer Vision.

[9] Pettis K. W., Bailey T. A., Jain A. K., and R. C. Dubes. An intrinsic dimensionality estimator from neighbor information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:25–37, 1979.

[10] Farahmand A., M., Szepesvári C., and Audibert J.-Y. Manifold-Adaptive Dimension Estimation. Proceedings of the 24th International Conference on Machine Learning, 2007.

[11] Hein M. and Audibert J.-Y. Intrinsic dimensionality estimation of submanifolds in Euclidian space. *ICML*, pages 289–296, 2005.

[12] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, 2007.