ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 339-344

# Extraction of intrinsic dimension using CCA - Application to blind sources separation

N. Donckers[1,*], A. Lendasse[2], V. Wertz[2], M. Verleysen[1,**]

[1]Université catholique de Louvain, Electricity Dept., 3 pl. du Levant,
B-1348 Louvain-la-Neuve, Belgium, {donckers, verleysen}@dice.ucl.ac.be.
[2]Université catholique de Louvain, CESAME, 4 av. G. Lemaître
B-1348 Louvain-la-Neuve, Belgium, lendasse@auto.ucl.ac.be.

**Abstract.** A general-purpose useful parameter in data analysis is the intrinsic dimension of a data set, corresponding to the minimum number of variables necessary to describe the data without significant loss of information. The knowledge of this dimension also facilitates most non-linear projection methods. We will show that the intrinsic dimension of a data set can be efficiently estimated using Curvilinear Component Analysis; we will also show that the method can be applied to the Blind Source Separation problem to estimate the number of sources in a mixing.

## 1. Introduction

In real-world (industrial, economical…) problems, the state of an unknown process is known through measurements. The variables to measure are usually selected according to some heuristics or experience on the problem. Nevertheless, this selection can be somewhat arbitrary, including the number of variables to measure, which is usually fixed with a certain error margin in order to be sure to capture all the necessary information about the state variables. In many situations, measurement variables are linear or non-linear mixings of state variables. It is important to determine the minimum number of state variables needed to characterize a process, for example for a better selection of measurement variables.

If the mixings are linear, well known techniques such as Independent Component Analysis (ICA) and Principal Component Analysis (PCA) [1,2] can be used. However without any information on the process a more general method is needed. We will use a powerful data analysis method known as Curvilinear Component Analysis (CCA) [3] to extract relevant information from data; this method makes it possible to determine the so-called intrinsic dimension of a set of data's, i.e. the minimum number of state variables. We will also show that this method, applied to the blind source separation problem (BSS), is able to determine the number of sources present in mixings.

[*] Nicolas Donckers is working towards the Ph. D. degree under F.R.I.A. fellowship.
[**] Michel Verleysen is a Research Associate of the Belgian FNRS.

## 2. Intrinsic dimension

Assume that $x_i$ is a set of $n$-dimensional input data vector. The intrinsic dimension is defined as the smallest number $m$ of variables that are needed to describe the set of data without any significant loss of information. Using CCA, these variables can be chosen on curvilinear axes. As an example, consider a circle with a fixed radius. Each point has 2 Cartesian coordinates, but can be fully described by only one angular variable $\alpha$. Its intrinsic dimension is 1.

The well-known distribution shown in Fig. 1 can be described using 2 state variables: one corresponds to the linear vertical axis, the other one being a curvilinear axis in the horizontal plan.

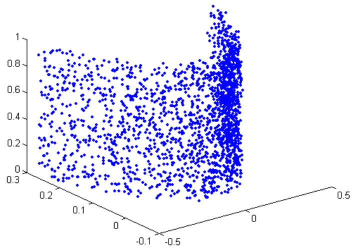


Fig. 1: A typical 3-D data set with intrinsic dimension equal to 2.

Fig. 1 also shows that the intrinsic dimension can be a non-integer value. If the set of data comes from a non-ideal process, as it is generally the case, data are noisy. The shape of Fig. 1 will have a non-zero thickness: it cannot be described with only 2 state variables without loss of information (about thickness). In the above example, the third state variable due to noise was negligible.

## 3. CCA: a generalized ICA algorithm

ICA methods are aimed to search independent directions in set of points [4]. The result of these methods is a certain number of vectors representing the directions. Single vectors cannot describe curvilinear axes: other methods are needed to perform non-linear CCA. One of these methods is the well-known Kohonen's maps (or SOM – Self-Organizing Maps) algorithm. Fig. 2 (a and b) show the result of a CCA on the previous data's using SOM.

It is clear that this algorithm makes it possible to find principal directions in a data set, even if they are curvilinear. Nevertheless it suffers from a number of drawbacks. For our purpose, the main one is the fact that the intrinsic dimension must be known in advance before using the method. Another disadvantage is the fact that the shape of the map must also be chosen a priori. Finally, computation complexity highly increases with the output space dimension.

For these reasons we used another algorithm called VQP (Vector Quantization and Projection) [5,6], which will be the “core” of the CCA method. This algorithm performs separately the vector quantization and the projection. Its principle is described bellow.

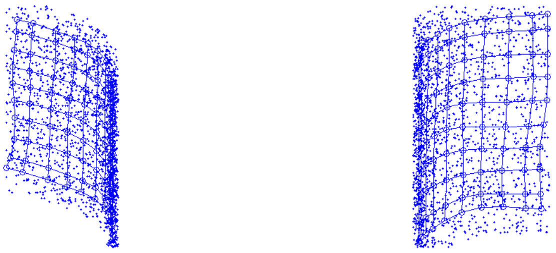


Fig. 2a and Fig. 2b: Result of SOM algorithm on Fig.1 data set.

## 4. The VQP algorithm

Kohonen’s maps perform a vector quantization under the constraint of an a priori neighborhood. In VQP, neurons are looking for their position in the output space by fitting (at least locally) the topology of the input space. This is done in two main steps. First, a vector quantization is performed on the data set. Any classical algorithm like competitive learning or frequency sensitive learning can be used. Secondly, a projection on an output space can be achieved.

Remind that the positions of the points in the input space are $x_i$ and assume that they are projected on $y_i$ in the output space. *Let* $X_{ij}=d(x_i,x_j)$ be the Euclidean distance between points $i$ and $j$ in the input space and $Y_{ij}$ the corresponding distance computed in the output space. The goal of the projection is to make $Y_{ij}$ similar to $X_{ij}$. This will preserve the local topology of the input space. The function to minimize can be written as:

$$E = \frac{1}{2} . \sum_{i} \sum_{j \neq i} \left(X_{ij} - Y_{ij}\right)^2 . F\left(Y_{ij}\right) , \tag{1}$$

where $F$ is a positive, monotone and decreasing function emphasizing the local topology. Minimizing this function leads to the following adaptation rule on the position $y_i$ of the centroids in the output space:

$$\Delta y_i = \alpha . \sum_{j \neq i} \frac{X_{ij} - Y_{ij}}{Y_{ij}} . \left[2.F\left(Y_{ij}\right) - \left(X_{ij} - Y_{ij}\right) F'\left(Y_{ij}\right)\right] , \tag{2}$$

where $\alpha$ is an adaptation factor.

This rule suffers from important drawbacks:

- the complexity of each iteration is O(n²);
- the process can fall in a local minimum of the error function;
- the sum of all contributions around a point produces a mean effect that slows down the convergence.

They can be avoided by using the following empirical rule [6]:

$$\Delta y_i = \alpha . \frac{X_{ij} - Y_{ij}}{Y_{ij}} . \left[ 2.F(Y_{ij}) - (X_{ij} - Y_{ij}) F'(Y_{ij}) \right] (y_j - y_i) \quad \forall j \neq i \tag{3}$$

The only constraint to use this last rule is to fix in advance the dimension of the output space. As we will see below, the intrinsic dimension can be estimated using this method by computing the projection on several output spaces with different dimensions.

## 5. Estimation of the intrinsic dimension

The estimation of the intrinsic dimension [7] is based on the following assumption: when a data set is projected on a space whose dimension is greater or equal to the intrinsic dimension, there remains enough information to describe correctly the data structure. On the contrary, when they are projected on a space whose dimension is smaller than the intrinsic dimension, there is a loss of information characterised by a deformation of the global topology, which can be measured through the increase of error *E* (computed on the centroïds) in equation (1). We will choose *F(x)=1* to obtain a global error.

The main steps of the algorithm can be summarised as follows:

- vector quantization on the data's;
- projection of the centroïds in the output spaces of dimension 1,… *n*;
- plot of the error term as a function of the dimension of the output space;
- the intrinsic dimension is the smallest dimension of the output space for which the error do not increase significantly.

Applying this principle to the database shown in Fig. 1 leads to the following results. We have performed the vector quantization using the competitive learning algorithm. Then the projection of the centroïds into spaces of dimension 1, 2 or 3 gives the errors presented at Table 1. The error for an output dimension of 2 represents the loss of information (due to noise) about thickness.

| Dimension of output space | 1 | 2 | 3 |
|---|---|---|---|
| Error | 0.56 | $2.10^{-4}$ | $< 10^{-7}$ |

Table 1: The error term for Fig.1 distribution

Notice that the error computed using (1) has been divided by the square of the number of centroïds to obtain a mean error. As expected, the error term increases significantly when the dimension of the output becomes smaller than or equal to the intrinsic dimension.

## 6. Application to Blind Sources Separation

Blind Sources Separation (BSS) is a well-known problem in the field of neural networks and adaptive filtering [8,9]. It exists a large number of algorithms able to solve this problem. Nevertheless, most of them require the knowledge of the number of sources present in the mixings. We will consider VQP as a pre-processing to estimate this number of sources. Fig. 3 shows the two sources used in the experiment.

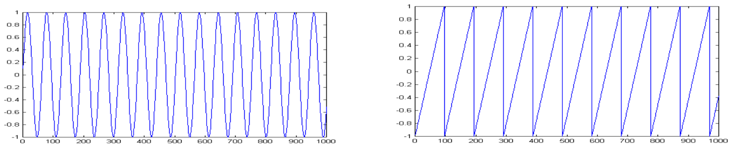
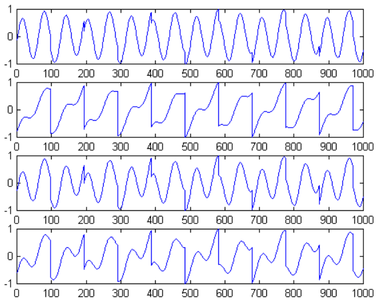


Fig. 3: The sources used in the BSS experiment

We consider two different problems to illustrate the VQP potential. In the first one, we use 4 linear mixtures of the sources; in the second one, we use 4 non-linear mixtures. Fig. 4 illustrates the linear mixings and Fig. 5 the non-linear one's.



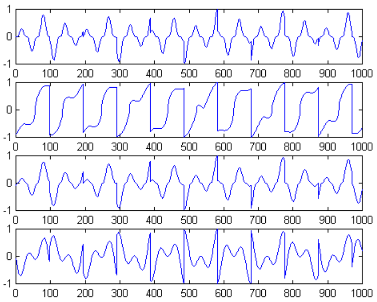


Fig. 4: The linear mixings

Fig. 5: The non-linear mixings

Table 2 shows the value of the error computed in both experiments.

| Dimension of the output space | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Linear mixings | 1.67 | $< 10^{-7}$ | $10^{-7}$ | $10^{-6}$ |
| Non-linear mixings | 1.61 | 0.02 | $10^{-6}$ | $10^{-6}$ |

Table 2: Results of VQP projection applied to BSS.

As expected, a non-significant loss of information occurs when the data are projected on a space of dimension 2 (in the non-linear case); it seems clear however that the intrinsic dimension of the database is 2.

## 7. Conclusion

The preliminary results presented in this paper show that the CCA method can be used to estimate the intrinsic dimension of a data set, and that VQP is a powerful tool to perform it. VQP does not require any information about the shape of the distribution, and can be used for linear, non-linear, and noisy mixings. It has also been shown that this method can be applied to the BSS problem as a useful pre-processing to determine the number of sources present in mixtures.

## References

[1] Oja E., *"Principal Components, Minors Components, and Linear Neural Networks."* Neural Networks, Vol. 5, pp. 927-935, 1992

[2] Karhunen J., *"Neural Approaches to Independent Component Analysis and Source Separation"* European Symposium on Artificial Neural Networks (ESANN 96), 1996, Bruges, Belgium, pp.249-266

[3] Demartines P., Herault J., *"Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets"* IEEE Transaction on Neural Networks 8: (1) 148-154 January 1997

[4] Common P.*"Independent Component Analysis, A new concept?"*, Signal Processing 36 (1994) pp. 287-314

[5] Demartines P. and Hérault J. *"Vector Quantization and Projection"* In A. Prieto, J. Mira, J. Cabestany, editor, International Workshop on Atificial Neural Networks, Vol. 686 of Lecture Notes in Computer Sciences, pp. 328-333. Springer-Verlag, 1993.

[6] Demartine P., *"Analyse de données par réseaux de neurones auto-organisés".* Thèse présentée en vue de l'obtention du grade de Docteur de l'Institut National Polytechnique de Grenoble. 1994.

[7] Lendasse A., de Bodt E., Verleysen M., *"Estimation de la dimension intrinsèque d'une série temporelle et prédiction par une méthode de projection"* ACSEG'98, Association Connectioniste en Sciences Economiques et de Gestion, Louvain-la-Neuve (Belgium), November 20, 1998, pp. D-37 - D-46.

[8] Jutten C., Hérault J., *"Blind Separation of Sources, Part I: An adaptative algorithm based on neuromimetic architecture"* Signal Processing 24, 1991

[9] Pajunen P., Hyvärinen A., Karhunen J., *"Non-linear Blind Source Separation by Self-Organizing Maps."* Process on Neural Information ICONIP 96, Vol. 2, Hong-Kong 1996