Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Sound and Vibration 280 (2005) 443–448
JOURNAL OF SOUND AND VIBRATION
www.elsevier.com/locate/jsvi

Short Communication

Distortion function and clustering for local linear models

G. Kerschen*, A.M. Yan, J.-C. Golinval

LTAS-Vibrations et Identification des Structures, Université de Liège, Chemin des Chevreuils 1 (B52), B-4000 Liège, Belgium

Received 28 August 2003; accepted 9 February 2004

1. Introduction

Principal component analysis (PCA) is a ubiquitous statistical technique for data analysis. PCA is however limited by its linearity and may sometimes be too simple for dealing with real-world data especially when the relations among variables are nonlinear. Recent years have witnessed the emergence of nonlinear generalizations of PCA, as for instance nonlinear principal component analysis (NLPCA) [1] or vector quantization principal component analysis (VQPCA) [2].

VQPCA involves a two-step procedure, namely a clustering of the data space into several regions and the application of PCA in each local region. In Ref. [3], VQPCA was applied for the reconstruction of dynamical response and it was shown that it is potentially a more effective tool than conventional PCA. The purpose of this technical note is to further investigate VQPCA and to have a closer look at the choice of the distortion function used for clustering the data space.

2. Vector quantizer—Euclidean or projection partition ?

Consider a set of observed n-dimensional data points $\mathbf{x}_i$. The first step of the VQPCA algorithm consists in dividing the data space into several regions using vector quantization. The vector

*Corresponding author. Fax: +32-4-366-48-56.
E-mail addresses: g.kerschen@ulg.ac.be (G. Kerschen), am.yan@ulg.ac.be (A.M. Yan).

0022-460X/$ - see front matter
doi:10.1016/j.jsv.2004.02.043

quantizer is based on an approach due to Lloyd [4] and is referred to as the generalized Lloyd algorithm [5].

A q-level vector quantizer is defined by a codebook $\mathscr{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_q)$, a partition $\mathscr{S} = (S_1, \ldots, S_q)$ and a distortion function $d(\mathbf{x}, \mathbf{c})$. The codebook vectors $\mathbf{c}_j$ and the regions S_j satisfy Lloyd's optimality conditions:

- each region S_j (with its corresponding codebook vector $\mathbf{c}_j$) corresponds to all $\mathbf{x}_i$ that lie closer to $\mathbf{c}_j$ than to any other codebook vector. Mathematically, $S_j = \{\mathbf{x}_i \mid d(\mathbf{x}_i, \mathbf{c}_j) < d(\mathbf{x}_i, \mathbf{c}_k), \forall k \neq j\}$;
- each codebook vector $\mathbf{c}_j$ is placed at the centroid of the corresponding region S_j. Mathematically,

$$\mathbf{c}_j = \min_{\mathbf{c}} E[\, d(\mathbf{x}, \mathbf{c}) \mid \mathbf{x} \in S_j]. \tag{1}$$

Accordingly, the generalized Lloyd algorithm is as follows:

(1) given q a number of regions, initialize the codebook $\mathscr{C}$ from randomly selected points in the data set;
(2) compute the corresponding optimal partition following the first optimality condition;
(3) compute the corresponding optimal codebook following the second optimality condition;
(4) iterate steps 2 and 3 until convergence.

It remains now to address the determination of the distortion measure $d(\mathbf{x}, \mathbf{c})$ for the vector quantizer. The choice is crucial since it will strongly affect the partition and hence the reconstruction error for the algorithm. In what follows, two distortion measures are discussed.

2.1. Euclidean partition (nearest-neighbor mapping)

The easiest way to build the partition is to consider a clustering based on Euclidean distance from the codebook vector:

$$d(\mathbf{x}, \mathbf{c}) = \|\mathbf{x} - \mathbf{c}\|. \tag{2}$$

In this case, the centroid is merely the mean of the data points in the region, i.e., $\mathbf{c}_j = E[\mathbf{x}_i \mid \mathbf{x}_i \in S_j] = \boldsymbol{\mu}_j$.

The procedure is illustrated in Fig. 1 for a two-dimensional sample $\mathbf{x}_i$ and two regions S_1 and S_2 characterized by their codebook vectors $\mathbf{c}_1$ and $\mathbf{c}_2$. The leading PCA mode is noted $\mathbf{p}_{11}$ in region 1 and $\mathbf{p}_{21}$ in region 2. As it can be seen, $\mathbf{x}_i$ belongs to region S_1 because the Euclidean distance from $\mathbf{c}_1$ is smaller than the distance from $\mathbf{c}_2$. It can also be observed that the boundary is the midperpendicular of segment $\mathbf{c}_1$–$\mathbf{c}_2$. Generally speaking, the regions are convex sets called Voronoi cells.

2.2. Projection partition

Although encouraging results have been obtained in Ref. [3] with the Euclidean distance, the clustering is constructed independent of the projection which follows. This is confirmed in Fig. 1 where the membership of $\mathbf{x}_i$ to region S_1 does not involve the PCA modes $\mathbf{p}_{11}$ and $\mathbf{p}_{21}$.

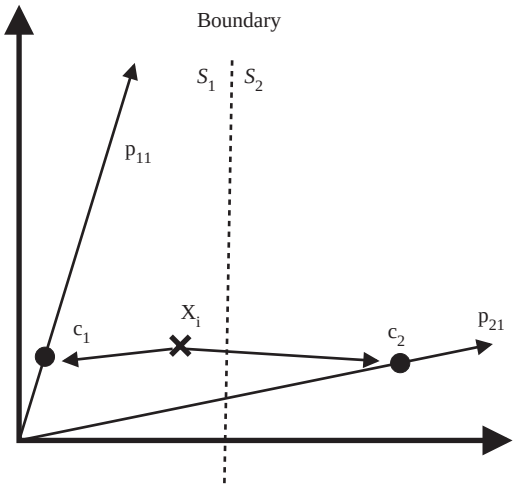

Fig. 1. Euclidean partition.

Accordingly, the reconstruction error may not be minimum by considering the Euclidean distance as the distance measure. Incidentally, this is the case in Fig. 1 because the distance between $\mathbf{x}_i$ and $\mathbf{p}_{21}$ is smaller than the distance between $\mathbf{x}_i$ and $\mathbf{p}_{11}$.

Instinctively, a better reduction should be performed by considering a distortion function based on the reconstruction error [2]:

$$d(\mathbf{x}, \mathbf{c}) = \|\mathbf{x} - \hat{\mathbf{x}}\| = \|(\mathbf{x} - \mathbf{c})[\mathbf{I} - \mathbf{P}\mathbf{P}^{\mathrm{T}}]\| \tag{3}$$

or equivalently,

$$d(\mathbf{x}, \mathbf{c}) = \|(\mathbf{x} - \mathbf{c})[\mathbf{P}\mathbf{P}^{\mathrm{T}}]^{\perp}\|, \tag{4}$$

where matrix $\mathbf{P} = [\mathbf{p}_{j1} \dots \mathbf{p}_{jr}]$ contains the leading r PCA modes in region S_j and $[\mathbf{P}\mathbf{P}^{\mathrm{T}}]^{\perp}$ is the space orthogonal to the projection matrix $\mathbf{P}\mathbf{P}^{\mathrm{T}}$. As for the Euclidean distance, it can be shown that the centroid for the reconstruction distance is the mean of the data points in the region, i.e., $\mathbf{c}_j = \boldsymbol{\mu}_j$.

Fig. 2 displays that when the distortion function is based on the reconstruction error, $\mathbf{x}_i$ is now mapped in region S_2 as expected. The boundary between regions S_1 and S_2 is the bisector of the angle formed by the two PCA modes $\mathbf{p}_{11}$ and $\mathbf{p}_{21}$.

It is interesting to note that the regions defined by the projection partition may not be connected sets. This can be explained by revisiting Fig. 2 for a collection of points (see Fig. 3). It has been mentioned that the boundary between regions S_1 and S_2 is the bisector of the angle formed by modes $\mathbf{p}_{11}$ and $\mathbf{p}_{21}$. Actually, two angles are formed by these modes and a second boundary, i.e., boundary 2 in Fig. 3, must also be taken into account. Due to the presence of this second boundary, regions S_1 and S_2 are no longer connected sets as it was the case for the Euclidean distance.

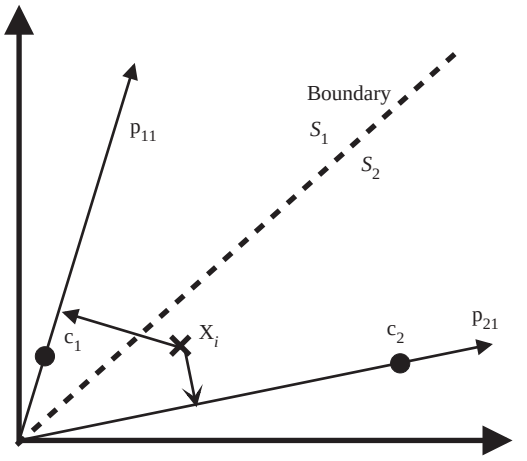

Fig. 2. Projection partition.

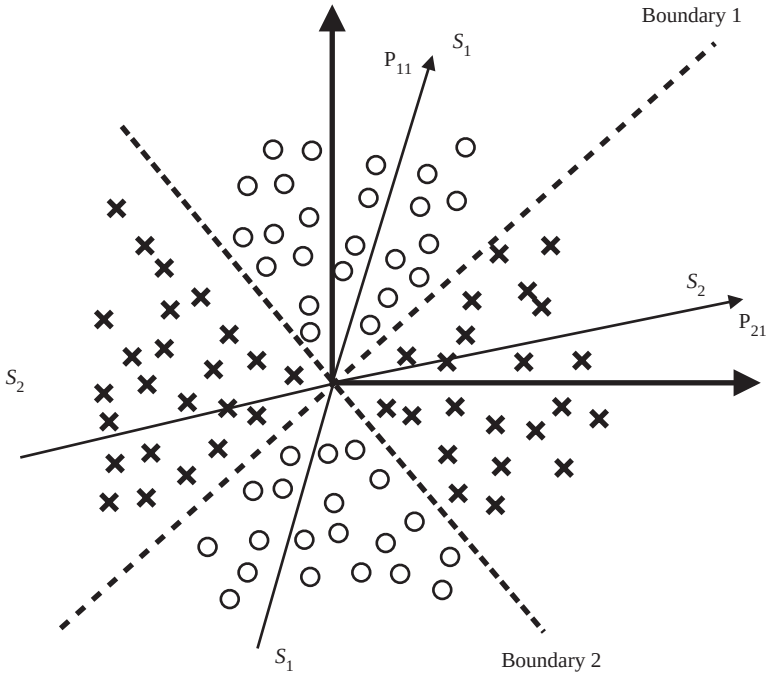

Fig. 3. Projection partition (revisited).

2.2.1. *Application example*

The superiority of the projection partition over the Euclidean partition may be illustrated using the following example. Consider 1000 samples obtained from the free response of a three-dimensional portal frame represented in Fig. 4. No beam makes the connection between nodes 6 and 7 but a cubic stiffness element is added between the translational degree-of-freedom. The free vibration of the beam is simulated with an initial displacement given by a static force applied at

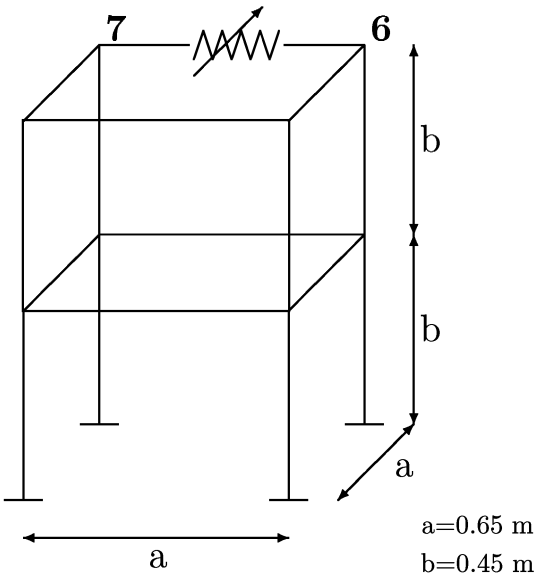

Fig. 4. Nonlinear portal frame.

Table 1
VQPCA applied to the portal frame (Euclidean partition)

Number of regions	MSE (%) 1 mode	MSE (%) 2 modes	MSE (%) 3 modes	MSE (%) 4 modes	MSE (%) 5 modes
2	35.18	24.22	15.75	10.37	6.34
3	29.80	19.91	12.54	7.92	4.63
4	26.58	17.09	10.61	6.22	3.59
5	23.63	14.88	7.82	5.71	2.15

Table 2
VQPCA applied to the portal frame (projection partition)

Number of regions	MSE (%) 1 mode	MSE (%) 2 modes	MSE (%) 3 modes	MSE (%) 4 modes	MSE (%) 5 modes
2	29.74	16.90	10.12	4.99	2.63
3	24.43	12.10	6.14	3.09	1.51
4	19.74	9.28	4.58	2.12	1.04
5	16.19	7.38	3.40	1.62	0.76

node 6. The data set consists of the translational accelerations in the three dimensions measured at each node of the portal frame.

Tables 1 and 2 contain the results of the reconstruction of the dynamical response given by the Euclidean and projection partitions, respectively. Table 3 presents the percentage of improvement

Table 3
Improvement in percent given by the projection partition

Number of regions	1 mode	2 modes	3 modes	4 modes	5 modes
2	15.46	30.22	35.75	51.88	58.52
3	18.02	39.23	51.04	60.98	67.39
4	25.73	45.70	56.83	65.92	71.03
5	31.49	50.40	56.52	71.63	64.65

given by the projection partition. The projection partition clearly offers a much better reconstruction of the dynamics of the portal frame than the Euclidean partition.

It should be noted that the application of VQPCA for damage diagnosis will be studied in a subsequent paper [6].

Acknowledgements

The author G. Kerschen is supported by a grant from the Belgian National Fund for Scientific Research (FNRS) which is gratefully acknowledged. This work was also sponsored by the Walloon Region government under "Convention Région Wallonne-ULg 9613419".

References

[1] M.A. Kramer, Nonlinear principal component analysis using autoassociative neural networks, *AIChE Journal* 37 (1991) 233–243.
[2] N. Kambhatla, Local Models and Gaussian Mixture Models for Statistical Data Processing, PhD Thesis, Oregon Graduate Institute of Science & Technology, 1996.
[3] G. Kerschen, J.C. Golinval, Nonlinear generalisation of principal component analysis: from a global to a local approach, *Journal of Sound and Vibration* 254 (2002) 867–876.
[4] S.P. Lloyd, Least squares quantization in PCM, Bell Laboratories Technical Note, 1957. Published in *IEEE Transactions on Information Theory* 28 (1982) 129–137.
[5] A. Gersho, R.M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Norwell, 1992.
[6] A.M. Yan, G. Kerschen, P. De Boe, J.C. Golinval, Structural damage diagnosis under changing environmental conditions—part II: local PCA for nonlinear cases, *Mechanical Systems and Signal Processing*, submitted.