Contributions to nonlinear Principal component analysis, manifold learning and projection pursuit

Stéphane Girard

INRIA Rhône-Alpes, projet Mistis, Inovallée,
655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France.
Stephane.Girard@inria.fr

Principal component analysis (PCA) is a well-known method for extracting linear structures from high-dimensional datasets. It computes the subspace best approaching the dataset from the Euclidean point of view. This method benefits from efficient implementations based either on solving an eigenvalue problem or on iterative algorithms. We refer to [12] for details. In a similar fashion, multi-dimensional scaling [1, 14, 19] addresses the problem of finding the linear subspace best preserving the pairwise distances. More recently, new algorithms have been proposed to compute low dimensional embeddings of high dimensional data. For instance, Isomap [21], LLE (Locally linear embedding) [18] and CDA (Curvilinear distance analysis) [4] aim at reproducing in the projection space the structure of the initial local neighborhood. These methods are mainly dedicated to visualization purposes. They cannot produce an analytic form of the transformation function, making it difficult to map new points into the dimensionality-reduced space. Besides, since they rely on local properties of pairwise distances, these methods are sensitive to noise and outliers. We refer to [16] for a comparison between Isomap and CDA and to [23] for a comparison between some features of LLE and Isomap.

Finding nonlinear structures is a challenging problem. An important family of methods focuses on self-consistent structures. The self-consistency concept is precisely defined in [20]. Geometrically speaking, it means that each point of the structure is the mean of all points that project orthogonally onto it. For instance, it can be shown that the k-means algorithm [10] converges to a set of k self-consistent points. Principal curves and surfaces [3, 11, 15, 22] are examples of one-dimensional and two-dimensional self-consistent structures. Their practical computation requires to solve a nonlinear optimization problem. The solution is usually non robust and suffers from a high estimation bias. In [13], a polygonal algorithm is proposed to reduce this bias. Higher dimensional self-consistent structures are often referred to as self-consistent manifolds even though their existence is not guaranteed for arbitrary datasets. An estimation algorithm based on a grid approximation is proposed in [9]. The fitting criterion involves two smoothness penalty terms describing the elastic properties of the manifold.

In my work, auto-associative models are proposed as candidates to the generalization of PCA. We show in [7] that these models are dedicated to the approximation of the dataset by a manifold. Here, the word "manifold" refers to the topology properties of the structure [17]. The approximating manifold is built by a projection pursuit algorithm presented in [5]. At each step of the algorithm, the dimension of the manifold is incremented. Some theoretical properties are provided in [7]. In particular, we can show that, at each step of the algorithm, the mean residuals norm is not increased. Moreover, it is also established that the algorithm converges in a finite number of steps. The note [6] is devoted to the presentation of some particular auto-associative models. They are compared to the classical PCA and some neural networks models. Implementation aspects are discussed in [2]. We show that, in numerous cases, no optimization procedure is required. Some illustrations on simulated and real data are presented in [8].

References

[1] J.D. Carroll and P. Arabie. Multidimensionnal scaling. *Annual Rev. of Psychology*, 31:607–649, 1980.

[2] B. Chalmond and S. Girard. Nonlinear modeling of scattered multivariate data and its application to shape change. *IEEE Pattern Analysis and Machine Intelligence*, 21(5):422–432, 1999.

[3] P. Delicado. Another look at principal curves and surfaces. *Journal of Multivariate Analysis*, 77:84–116, 2001.

[4] P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. on Neural Networks*, 8(1):148–154, 1997.

[5] S. Girard. A nonlinear PCA based on manifold approximation. *Computational Statistics*, 15(2):145–167, 2000.

[6] S. Girard, B. Chalmond, and J-M. Dinten. Position of principal component analysis among auto-associative composite models. *Comptes-Rendus de l'Académie des Sciences, Série I*, 326:763–768, 1998.

[7] S. Girard and S. Iovleff. Auto-associative models and generalized principal component analysis. *Journal of Multivariate Analysis*, 93(1):21–39, 2005.

[8] S. Girard and S. Iovleff. Auto-associative models, nonlinear principal component analysis, manifolds and projection pursuit. In A. Gorban et al., editor, *Principal Manifolds for Data Visualisation and Dimension Reduction*, volume 28, pages 205–222. LNCSE, Springer-Verlag, 2007.

[9] A. Gorban and A. Zinovyev. Elastic principal graphs and manifolds and their practical applications. *Computing*, 75(4):359–379, 2005.

[10] J.A. Hartigan. *Clustering algorithms.* Wiley, New-York, 1995.

[11] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84 (406):502–516, 1989.

[12] I. Jolliffe. *Principal Component Analysis.* Springer-Verlag, New York, 1986.

[13] B. Kégl, A. Krzyzak, T. Linder, and K. Zeger. A polygonal line algorithm for constructing principal curves. In *Proceedings of 12h NIPS*, pages 501–507, Denver, Colorado, 1998.

[14] J.B. Kruskal and M. Wish. *Multidimensional scaling.* Sage, Beverly Hills, 1978.

[15] M. LeBlanc and R. Tibshirani. Adaptive principal surfaces. *Journal of the American Statistical Association*, 89(425):53–64, 1994.

[16] J.A. Lee, A. Lendasse, and M. Verleysen. Curvilinear distance analysis versus isomap. In *European Symposium on Artifical Neural Networks*, pages 185–192, Bruges, Belgium, 2002.

[17] J. Milnor. *Topology from the differentiable point of view.* University press of Virginia, Charlottesville, 1965.

[18] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[19] R.N. Shepard and J.D. Carroll. Parametric representation of nonlinear data structures. In P.R. Krishnaiah, editor, *Int. Symp. on Multivariate Analysis*, pages 561–592. Academic-Press, 1965.

[20] T. Tarpey and B. Flury. Self-consistency: A fundamental concept in statistics. *Statistical Science*, 11(3):229–243, 1996.

[21] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[22] R. Tibshirani. Principal surfaces revisited. *Statistics and computing*, 2:183–190, 1992.

[23] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas. Non-linear dimensionality reduction techniques for classification and visualization. In *Proceedings of 8th SIGKDD*, pages 23–26, Edmonton, Canada, 2002.