Chapter 2
Principal Graphs and Manifolds

Alexander N. Gorban
University of Leicester, UK

Andrei Y. Zinovyev
Institut Curie, Paris, France

ABSTRACT

In many physical, statistical, biological and other investigations it is desirable to approximate a system of points by objects of lower dimension and/or complexity. For this purpose, Karl Pearson invented principal component analysis in 1901 and found 'lines and planes of closest fit to system of points'. The famous k-means algorithm solves the approximation problem too, but by finite sets instead of lines and planes. This chapter gives a brief practical introduction into the methods of construction of general principal objects (i.e., objects embedded in the 'middle' of the multidimensional data set). As a basis, the unifying framework of mean squared distance approximation of finite datasets is selected. Principal graphs and manifolds are constructed as generalisations of principal components and k-means principal points. For this purpose, the family of expectation/maximisation algorithms with nearest generalisations is presented. Construction of principal graphs with controlled complexity is based on the graph grammar approach.

INTRODUCTION

In many fields of science, one meets with multivariate (multidimensional) distributions of vectors representing some observations. These distributions are often difficult to analyse and make sense of due to the very nature of human brain which is able to visually manipulate only with the objects of dimension no more than three.

This makes actual the problem of approximating the multidimensional vector distributions by objects of lower dimension and/or complexity while retaining the most important information and structures contained in the initial full and complex data point cloud.

DOI: 10.4018/978-1-60566-766-9.ch002

The most trivial and coarse approximation is collapsing the whole set of vectors into its *mean* point. The mean point represents the 'most typical' properties of the system, completely forgetting variability of observations.

The notion of the mean point can be generalized for approximating data by more complex types of objects. In 1901 Pearson proposed to approximate multivariate distributions by *lines* and *planes* (Pearson, 1901). In this way the Principal Component Analysis (PCA) was invented, nowadays a basic statistical tool. Principal lines and planes go through the 'middle' of multivariate data distribution and correspond to the first few modes of the multivariate Gaussian distribution approximating the data.

Starting from 1950s (Steinhaus, 1956; Lloyd, 1957; and MacQueen, 1967), it was proposed to approximate the complex multidimensional dataset by several 'mean' points. Thus *k-means algorithm* was suggested and nowadays it is one of the most used *clustering methods* in machine learning (see a review presented by Xu & Wunsch, 2008).

Both these directions (PCA and K-Means) were further developed during last decades following two major directions: 1) linear manifolds were generalised for non-linear ones (in simple words, initial lines and planes were bended and twisted), and 2) some links between the 'mean' points were introduced. This led to the appearance of several large families of new statistical methods; the most famous from them are Principal Curves, Principal Manifolds and Self-Organising Maps (SOM). It was quickly realized that the objects that are constructed by these methods are tightly connected theoretically. This observation allows now to develop a common framework called "Construction of Principal Objects". The geometrical nature of these objects can be very different but all of them serve as *data approximators of controllable complexity*. It allows using them in the tasks of *dimension* and *complexity reduction*. In Machine Learning this direction is connected with terms 'Unsupervised Learning' and 'Manifold Learning.'

In this chapter we will overview the major directions in the field of principal objects construction. We will formulate the problem and the classical approaches such as PCA and *k*-means in a unifying framework, and show how it is naturally generalised for the Principal Graphs and Manifolds and the most general types of principal objects, Principal Cubic Complexes. We will systematically introduce the most used ideas and algorithms developed in this field.

Approximations of Finite Datasets

Definition. *Dataset* is a finite set X of objects representing N multivariate (multidimensional) observations. These objects $\mathbf{x}^i \in X$, $i = 1 \ldots N$, are embedded in $\mathbf{R}^m$ and in the case of complete data are vectors $\mathbf{x}^i \in \mathbf{R}^m$. We will also refer to the individual components of $\mathbf{x}^i$ as x^i_k such that $\mathbf{x}^i = (x^i_1, x^i_2, \ldots, x^i_m)$; we can also represent dataset as a *data matrix* $X = \{ x^i_j \}$.

Definition. *Distance function* dist(**x**,**y**) is defined for any pair of objects **x**, **y** from X such that three usual axioms are satisfied: dist(**x**,**x**) = 0, dist(**x**,**y**) = dist(**y**,**x**), dist(**x**,**y**)+dist(**y**,**z**) ≤ dist(**x**,**z**).

Definition. *Mean point* $\mathbf{M}_F(X)$ for X is a vector $\mathbf{M}_F \in \mathbf{R}^m$ such that $\mathbf{M}_F(X) = \arg\min_{\mathbf{y} \in R^m} \sum_{i=1..N} (\mathrm{dist}(\mathbf{y}, \mathbf{x}_i))^2$.

In this form the definition of the mean point goes back to Fréchet (1948). Notice that in this definition the mean point by Fréchet can be non-unique. However, this definition allows multiple useful generalisations including using it in the abstract metric spaces. It is easy to show that in the case of complete