Probabilistic Principal Components and Mixtures, How This Works

Anna M. Bartkowiak[1,2(✉)] and Radoslaw Zimroz[3]

[1] Institute of Computer Science, Wroclaw University, 50-383 Wroclaw, Poland
aba@ii.uni.wroc.pl
[2] Wroclaw School of Information Technology, 54-239 Wroclaw, Poland
[3] Diagnostics and Vibro-Acoustics Science Laboratory,
Wroclaw University of Technology, 50-421 Wroclaw, Poland

Abstract. Classical Principal Components Analysis (PCA) is widely recognized as a method for dimensionality reduction and data visualization. This is a purely algebraic method, it considers just some optimization problem which fits exactly to the gathered data vectors with their particularities. No statistical significance tests are possible. An alternative is to use probabilistic principal component analysis (PPCA), which is formulated on a probabilistic ground. Obviously, to do it one has to know the probability distribution of the analyzed data. Usually the Multi-Variate Gaussian (MVG) distribution is assumed. But what, if the analyzed data are decidedly not MVG? We have met such problem when elaborating multivariate gearbox data derived from a heavy duty machine. We show here how we have dealt with the problem.

In our analysis, we assumed that the considered data are a mixture of two groups being MVG, specifically: each of the sub-group follows a probabilistic principal component (PPC) distribution with a MVG error function. Then, by applying Bayesian inference, we were able to calculate for each data vector x its a posteriori probability of belonging to data generated by the assumed model. After estimation of the parameters of the assumed model we got means - based on a sound statistical basis - for constructing confidence boundaries of the data and finding outliers.

Keywords: Probabilistic principal components · Multi-variate normal distribution · Mixture models · Un-mixing multivariate data · Condition monitoring · Gearbox diagnostics · Healthy state · Probabilities a posteriori · Outliers

1 Introduction

Classical Principal Components Analysis (PCA) is widely recognized as a method for dimensionality reduction and data visualization. However, PCA is a purely algebraic method, it considers just some optimization problem which fits exactly to the gathered data vectors with their particularities.

Yet, without a proper probability model it is impossible to formulate statistically significant statements.

K. Saeed and W. Homenda (Eds.): CISIM 2015, LNCS 9339, pp. 24–35, 2015.
DOI: 10.1007/978-3-319-24369-6_2