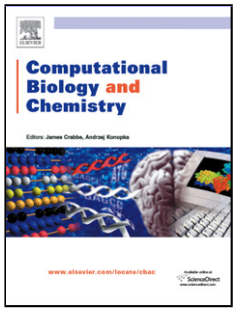

Title: A random version of principal component analysis in data clustering

Author: Luigi Leonardo Palese

PII: S1476-9271(17)30064-6
DOI: https://doi.org/doi:10.1016/j.compbiolchem.2018.01.009
Reference: CBAC 6780

To appear in: *Computational Biology and Chemistry*

Received date: 2-2-2017
Revised date: 5-10-2017
Accepted date: 23-1-2018

Please cite this article as: Luigi Leonardo Palese, A random version of principal component analysis in data clustering, *<![CDATA[Computational Biology and Chemistry]]>* (2018), https://doi.org/10.1016/j.compbiolchem.2018.01.009

A random version of principal component analysis in data clustering

Luigi Leonardo Palese[a,*]

[a] *University of Bari "Aldo Moro", Department of Basic Medical Sciences, Neurosciences and Sense Organs (SMBNOS), Bari, 70124, Italy*

Abstract

Principal component analysis (PCA) is a widespread technique for data analysis that relies on the covariance/correlation matrix of the analyzed data. However, to properly work with high-dimensional data sets, PCA poses severe mathematical constraints on the minimum number of different replicates, or samples, that must be included in the analysis. Generally, improper sampling is due to a small number of data respect to the number of the degrees of freedom that characterize the ensemble. In the field of life sciences it is often important to have an algorithm that can accept poorly dimensioned data sets, including degenerated ones. Here a new random projection algorithm is proposed, in which a random symmetric matrix surrogates the covariance/correlation matrix of PCA, while maintaining the data clustering capacity. We demonstrate that what is important for clustering efficiency of PCA is not the exact form of the covariance/correlation matrix, but simply its symmetry.

Keywords: Principal Component Analysis, Random Projection, Dimensionality Reduction, Data Clustering, Protein Structure, Structural Bioinformatics

1. Introduction

Science today is surrounded by large amounts of data. These are produced by techniques and instruments able to measure a huge number of variables on a large number of samples, or are deposited in an increasing number of online databases that grow exponentially [1, 2]. Also modern numerical simulations can produce very large and high-dimensional outputs [3]. The challenge of the growing size of data concerns all fields, but the one in which we have seen the most spectacular growth is probably that of life sciences, where the advancement of genomics, proteomics and other high-throughput technologies has produced an overwhelming amount of data, more and more often freely available to all

*Corresponding author
Email address: luigileonardo.palese@uniba.it (Luigi Leonardo Palese)