Chapter 18
Principal Component Analysis of Hydrological Data

Petr Praus
VSB - Technical University of Ostrava, Czech Republic

ABSTRACT

In this chapter the principals and applications of principal component analysis (PCA) applied on hydrological data are presented. Four case studies showed the possibility of PCA to obtain information about wastewater treatment process, drinking water quality in a city network and to find similarities in the data sets of ground water quality results and water-related images. In the first case study, the composition of raw and cleaned wastewater was characterised and its temporal changes were displayed. In the second case study, drinking water samples were divided into clusters in consistency with their sampling localities. In the case study III, the similar samples of ground water were recognised by the calculation of cosine similarity, the Euclidean and Manhattan distances. In the case study IV, 32 water-related images were transformed into a large image matrix whose dimensionality was reduced by PCA. The images were clustered using the PCA scatter plots.

Motto: Variation is information

INTRODUCTION

The hydrological data set is typically represented by a matrix of water samples (objects), which are characterized by many physical, chemical, microbiological, and biological parameters (variables) of different magnitude and units. The data contain specific parameters, such as nitrate, ammonium, chloride, individual heavy metals and organic compounds, and group parameters, which give the total concentrations of similar compounds: for instance, total nitrogen, conductivity, total organic carbon, mesophilic bacteria etc. The parameters are determined by various analytical methods, including microbiological and biological tests.

Except of important information, the real data contain also useless or even confusing informa-
Principal Component Analysis of Hydrological Data

The main objective of PCA is looking for new latent (hidden) variables of \( n \) samples, which are not correlated to each other. Each latent variable \( t_j \) (principal component) is a linear combination of \( p \) variables and describes a different source of total variation

\[
\begin{align*}
t_1 &= w_{1,1}x_{1,1} + w_{1,2}x_{1,2} + \ldots + w_{1,p}x_{1,p} \\
t_2 &= w_{2,1}x_{2,1} + w_{2,2}x_{2,2} + \ldots + w_{2,p}x_{2,p} \\
t_n &= w_{n,1}x_{n,1} + w_{n,2}x_{n,2} + \ldots + w_{n,p}x_{n,p} 
\end{align*}
\]  
(1)

where \( w_{ij} \) and \( x_{ij} \) \((1 \leq i \leq n, 1 \leq j \leq p)\) are component weight (loading) and original variable (parameter), respectively. The component loadings are the contribution measures of a particular variable to the principal components. It also holds

\[
w_{1,1}^2 + w_{1,2}^2 + \ldots + w_{1,p}^2 = 1  
\]  
(2)

The variability of the principal components is ordered as follows \( \text{Var}(t_1) > \text{Var}(t_2) > \ldots \text{Var}(t_p) \). The condition of \( t_j \) with maximum variance implies to look for the vector \( w_j = (w_{j,1}, w_{j,2}, \ldots, w_{j,p}) \), which maximises the variation \( \text{var}(Xw_j) = w_j^TCw_j \), where \( C \) is a covariance matrix and the normalisation condition holds \( w_j^T w_j = 1 \). The product \( w_j^TCw_j \) equals the largest eigenvalue \( \lambda_j \), \( w_j \) is the eigenvector. Similarly, the maximum variance of the second component is computed if \( \text{var}(Xw_j) = w_j^TCw_j \) and also it holds \( w_j^T w_j = 1 \) and \( w_j^T w_i = 0 \) if \( j < i \). Generally, the \( k \)-th component \( t_k = Xw_k \) has maximum when \( \text{var}(Xw_j) = w_j^TCw_j \) and on the conditions of \( w_j^T w_j = 1 \) and \( w_j^T w_i = 0 \) if \( j < i \).

The Equation (1) can be rewritten for a data matrix \( X(n \times p) \)

\[
X = T W^T  
\]  
(3)

where \( T (n \times k) \) and \( W (k \times p) \) are the matrices of the \( k \) principal components (score matrix) and their loadings (loading matrix), respectively. PCA

BACKGROUND

Principal component analysis is a basic multivariate statistical method. The method was firstly introduced by Karl Pearson (1901) and subsequently developed by Hotelling (1933a, b). Until the 1950s, the method had limited applications due to the lack of computational equipment.
23 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product’s webpage:

www.igi-global.com/chapter/principal-component-analysis-hydrological-data/45454?camid=4v1


www.igi-global.com/e-resources/library-recommendation/?id=15

Related Content

Identification of Associations between Clinical Signs and Hosts to Monitor the Web for Detection of Animal Disease Outbreaks

www.igi-global.com/article/identification-of-associations-between-clinical-signs-and-hosts-to-monitor-the-web-for-detection-of-animal-disease-outbreaks/163316?camid=4v1a

Role of Organic Soil Amendments in Controlling Ground Water Pollution Due to Pesticides: An Effective Approach

www.igi-global.com/chapter/role-of-organic-soil-amendments-in-controlling-ground-water-pollution-due-to-pesticides/196792?camid=4v1a

Foundations and Applications of Computer Based Material Flow Networks for Environmental Management

www.igi-global.com/chapter/foundations-applications-computer-based-material/18549?camid=4v1a

Pyroxene: A Territorial Decision Support System Based on Spatial Simulators Integration for Forest Fire Risk Management

www.igi-global.com/article/pyroxene-territorial-decision-support-system/55953?camid=4v1a