Water quality assessment using SVD-based principal component analysis of hydrological data

Petr Praus

Department of Analytical Chemistry and Material Testing, VSB-Technical University Ostrava, 17. listopadu 15, 708 33 Ostrava, Czech Republic

Abstract

Principal component analysis (PCA) based on singular value decomposition (SVD) of hydrological data was tested for water quality assessment. Using two case studies of waste- and drinking water, PCA via SVD was able to find latent variables which explain 80.8% and 83.7% of the variance, respectively. By means of scatter and loading plots, PCA revealed the relationships among samples and hydrochemical parameters which were also confirmed by factor analysis (FA).

In the case of wastewater, these latent variables clearly displayed changes of water composition over time. Drinking water samples were clustered into four groups which were characterised by their typical water composition. On the basis of these results PCA was found to be a suitable technique for water quality assessment.

Keywords: water quality, wastewater, drinking water, principal component analysis, singular value decomposition, factor analysis

Introduction

Real hydrochemical data sets contain not only important information useful for quality assessment and/or treatment technology but also confusing noise. Mostly, measured variables are not normally distributed, often co-linear or autocorrelated, containing outliers, erroneous or nonsense values. In order to reveal mutual dependence or logical structures of data, there are several chemometric procedures generally called as data mining techniques. Some of them are based on the reduction of data dimensionality, such as principal component analysis (Lavine, 2000; Jolliffe, 2002), factor analysis (Malinowski and Howery, 1980; Malinowski, 1991), independent component analysis (Comon, 1994), independent factor analysis (Attias, 1998), generative topographic mapping (Bishop et al., 1998), etc.

PCA is used to search new abstract orthogonal principal components (eigenvectors) which explain most of the data variation in a new coordinate system. Each principal component (PC) is a linear combination of the original variables and describes a different source of variation (information). The largest or 1st PC is oriented in the direction of the largest variation of the original variables and passes through the centre of the data. The 2nd largest PC lies in the direction of the next largest variation, passes through the centre of the data and is orthogonal to the first PC, and so forth.

Classical PCA is based on the decomposition of a covariance/correlation matrix (Geladi and Kowalski, 1986) by eigenvalue (spectral) decomposition (EVD) or by the decomposition of real data matrixes using SVD. Compared with EVD, SVD is a more robust, reliable, and precise method with no need to compute the input covariance/correlation matrix. From a numerical point of view, SVD is well known for its stability and convergence, even in the ill conditioned problems.

In general, SVD decomposes an arbitrary Matrix A (n x p) into three matrices:

$$A = U\, S\, V^T \qquad (1)$$

where:

U (n x n) and V^T (p x p) are orthogonal and normalised matrices, i.e., $U^T U=I$ and $V^T V=I$
S (n x p) is a diagonal matrix with singular values in decreasing order
U columns are the left singular vectors
V^T rows are the right singular vectors.

Computing the SVD consists of finding the eigenvalues and eigenvectors of A A^T and A^T A, respectively. The U columns are eigenvectors of A A^T and the V^T rows are the eigenvectors of A^T A. The powerful property of SVD is compressing the information contained in A into the first few singular vectors which are mutually orthogonal and their importance rapidly decreases after the first columns/rows. The importance of each singular vector is given by the squares of nonnegative diagonal (singular) values of S.

SVD has found a wide range of various applications in molecular dynamic and gene expression analysis (Wall et al., 2003), information retrieval in a technique called Latent Semantic Indexing (Berry et al., 1995), image processing (Zhang et al., 2005), hearing noise filtering (Maj et al., 2001), spectral analysis (Safavi and Abdollahi H., 2001), and so forth.

Multivariate statistical methods, encompassing cluster analysis, PCA, FA and discriminant analysis, have been successfully used in hydrochemistry for many years. Quality assessment of surface water (Simeonov et al., 2003; Vega et al., 1998; Wunderlin et al., 2001), groundwater (Reghunath et al., 2002), and environmental research (Ceballos et al., 1998; Lambrakis et al., 2004) employing multicomponent techniques are well described in the literature.

☎ +420-59-732-3370; fax: +420-59-732-3370;
e-mail: petr.praus@vsb.cz

Received 7 June 2005; accepted in revised form 12 August 2005.