Journal of Chromatography A, 1251 (2012) 134–140

Contents lists available at SciVerse ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

A principal component analysis approach for developing retention models in liquid chromatography

P. Nikitas[a,*], A. Pappa-Louisi[a], S. Tsoumachides[a], A. Jouyban[b]

[a] *Laboratory of Physical Chemistry, Department of Chemistry, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece*
[b] *Drug Applied Research Center and Faculty of Pharmacy, Tabriz University of Medical Sciences, Tabriz 51664, Iran*

ARTICLE INFO

Article history:
Received 11 April 2012
Received in revised form 7 June 2012
Accepted 14 June 2012
Available online 20 June 2012

Keywords:
PCA
LSER
ANN
Retention models
Liquid chromatography

ABSTRACT

Three retention models for liquid chromatography are developed using principal component analysis (PCA). It is shown that they exhibit features similar to that of the model based on linear solvation energy relationship (LSER). However, the fitting performance of the PCA models is better than that of the LSER model, the performance of which can be considerably improved by the use of artificial neural networks. In addition, the possibility of using the proposed models as well as the LSER model to predict the retention times of solutes under chromatographic conditions at which these solutes have never been studied is also examined by means of three data sets of analytes consisting of non-polar compounds to polar compounds with a variety of functional groups.

1. Introduction

The free energy of transfer of a solute between the stationary and the mobile phases in the liquid chromatography can be described as the linear sum of contributing processes. This results in the linear solvation energy relationship (LSER) expression of the logarithmic capacity factor, which using recent Abraham's notation is given by [1–5]

$$\ln\ k = c + eE + sS + \alpha A + bB + \nu V \tag{1}$$

where the letters E, S, A, B, V are analyte parameters (descriptors) representing its polarizability, dipolarity, hydrogen bond donating ability, hydrogen bond accepting ability, and molecular size, respectively. Descriptors for more than 4000 compounds are available [6] plus a software program to estimate analyte descriptors from structure [7].

The coefficients e, s, a, b, ν and the constant c reflect properties of the mobile phase and the stationary phase of the chromatographic column and they are determined by multi-parameter linear least squares fit to experimental data. Note that in literature and especially in exo-thermodynamic studies log k is used instead of ln k. However, in this article for reasons of coherence we use the natural logarithm of k for all expressions of the retention models.

When the column/mobile phase system changes, the column parameters e, s, a, b, ν, and c change as well. Thus, in the same column but at different mobile phase compositions the above coefficients become functions of φ, the volume fraction of the organic modifier in the mobile phase. Wang et al. proposed a linear dependence of these parameters upon φ [8], while for wider ranges of solvent compositions Torres-Lapasió et al. proposed, among others, the quadratic dependence [9]. Thus we may in general write

$$\begin{aligned}\ln k = {} & (c_0 + c_1\varphi + \cdots + c_r\varphi^r) + (e_0 + e_1\varphi + \cdots + e_r\varphi^r)E \\ & + (s_0 + s_1\varphi + \cdots + s_r\varphi^r)S + (\alpha_0 + \alpha_1\varphi + \cdots + \alpha_r\varphi^r)A \\ & + (b_0 + b_1\varphi + \cdots + b_r\varphi^r)B + (\nu_0 + \nu_1\varphi + \cdots + \nu_r\varphi^r)V \end{aligned} \tag{2}$$

The above relationship with $r = 1$ or $r = 2$ has been used for the prediction of the elution time under isocratic conditions and its performance in comparison to other models has been recently evaluated [10]. Similar expressions have been proposed for gradient elution [11–13].

Eq. (1) resembles those obtained from either the principal component analysis (PCA) or the factor analysis (FA). For example, if we consider a matrix of centered observations X_{ij} ($i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$), then each observation is a linear combination of factor scores F_{ik} plus noise [14–16]

$$X_{ij} = \sum_{k=1}^{q} w_{kj}F_{ik} + \varepsilon_{ij} \tag{3}$$

[*] Corresponding author. Tel.: +30 2310 997765; fax: +30 2310 997709.
E-mail address: nikitas@chem.auth.gr (P. Nikitas).

0021-9673/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chroma.2012.06.049

Here, the weights w_{kj} are called the factor loadings of the observable features, ε_{ij} is the noise term, and $q \leq m$ is the number of factor scores (and factor loadings) used to express each observation. PCA is usually adopted to extract and visualize patterns in large data matrices since by PCA the number of variables in a data set can be reduced by finding linear combinations of variables explaining most of the variability. Thus, the PCA method has been used for the classification of large sets of solutes based on HPLC retention data, usually in a QSRR (quantitative structure–retention relationships) context [17–24]. Alternatively, PCA has been combined with target transformation factor analysis to derive retention models [25–28].

In the present paper we also examine the possibility of using PCA to develop models, like the one of Eq. (2), based on the general expression of Eq. (3). Moreover, taking into account that in Eq. (3), as in Eq. (1), w_{kj} are functions of the of the mobile and the stationary phase of the chromatographic column and F_{ik} depend exclusively on the solute properties, we examine whether the PCA retention models can be used to predict the retention time of solutes under conditions that they have never been previously studied.

2. Theory

2.1. General relationships

The theory of PCA is described in many textbooks [14–16]. The basic relationships that we are going to use in developing retention models are the following. Consider an original **X** matrix with X_{ij} (i = 1, 2,. . ., n; j = 1, 2,. . ., m) elements. The basis of PCA is the singular value decomposition of **X**. However, in most statistical packages, the input of a conventional PCA is the normalized matrix **Z** calculated from

$$z_{ij} = \frac{X_{ij} - \overline{X}_j}{\sigma_j}, \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m \tag{4}$$

where $\bar{X}_j$ is the mean value of the jth column of **X** and σ_j is the corresponding standard deviation. The output of PCA is among others the $n \times q$ score matrix **S** and the $m \times q$ loadings matrix ***V***, which are related to the **Z** matrix through the following relationship

$$\mathbf{Z} = \mathbf{S}\mathbf{V}^T + \mathbf{E} \tag{5}$$

where **E** is the residual or noise matrix. From Eq. (5) we readily obtain that the elements z_{ij} of **Z** are given by

$$z_{ij} = v_{1j}P_{i1} + v_{2j}P_{i2} + \cdots + v_{kj}P_{iq} + e_{ij} \tag{6}$$

The most interesting property of this expression is that we can directly calculate the loadings matrix **V** from matrices **Z** and **S**, or the score matrix **S** from matrices **Z** and **V** by applying in both cases multivariate linear regression. In terms of matrix algebra, this can be done by means of the following two equations

$$\mathbf{V}^T = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{Z} \quad \text{and} \quad \mathbf{S} = \mathbf{Z}\mathbf{V}(\mathbf{V}^T\mathbf{V})^{-1} \tag{7}$$

If we take into account the definition of z_{ij} and the noise term in Eq. (6) is removed, we obtain

$$\frac{X_{ij}}{\sigma_j} \approx \frac{\bar{X}_j}{\sigma_j} + v_{1j}P_{i1} + v_{2j}P_{i2} + \cdots + v_{kj}P_{iq}. \tag{8}$$

2.2. First model

Consider that matrix **X** consists of n rows and m columns that contain $\ln k(\varphi_1, \varphi_2,\ldots, \text{pH})$ values. In particular, each row corresponds to the $\ln k$ value of a certain solute and each column to certain experimental conditions, $(\varphi_1, \varphi_2,\ldots, \text{pH})$, used for the determination of $\ln k$. If matrix **X** is normalized according to Eq. (4) and this normalization matrix is used as an input of a conventional PCA without rotation, then we obtain the retention model

$$\ln k_{ij}(\text{calc}) = \overline{\ln k_j} + \sigma_j v_{1j}P_{i1} + \sigma_j v_{2j}P_{i2} + \cdots + \sigma_j v_{qj}P_{iq} \tag{9}$$

Here, $\ln k_{ij}(\text{calc})$ is the predicted value of $\ln k$ that corresponds to the ith solute measured under experimental conditions of the jth column and $\overline{\ln k_j}$ is the mean value of the $\ln k$ values of the jth column. In this expression the score factors P_{iq} depend exclusively on solute i. In contrast, the quantities $\sigma_j v_{qj}$ as well as the average value $\overline{\ln k_j}$ depend on the experimental conditions $\varphi_1, \varphi_2,\ldots$, pH. In the simple case that $\ln k$ depends only on the single factor φ, $\sigma_j v_{qj}$ and $\overline{\ln k_j}$ may be approximated by polynomials and therefore the retention model may be expressed as

$$\begin{aligned}\ln k &= (c_{00} + c_{01}\varphi + \cdots c_{0r}\varphi^r) + (c_{10} + c_{11}\varphi + \cdots c_{1r}\varphi^r)P_1 \\ &\quad + (c_{20} + c_{21}\varphi + \cdots c_{2r}\varphi^r)P_2 + \cdots + (c_{q0} + c_{q1}\varphi + \cdots c_{qr}\varphi^r)P_q\end{aligned} \tag{10}$$

where for simplicity we have omitted subscript i. However, at this point we should stress that the approximation of $\sigma_j v_{qj}$ and $\overline{\ln k_j}$ by polynomials may not be always the best choice. Alternative solutions, like the use of rational expressions [29–32], may be tested.

It is seen that the expression of Eq. (10) is at least formally equivalent to Eq. (2) provided q = 5. However, Eq. (10) is much more flexible than Eq. (2). There is no need to know the factors $P_1, P_2,\ldots, P_q$ in advance in order to apply Eq. (10) and their number should not be constant. We may test several numbers of factors having as a criterion the balance between simplicity and accuracy of prediction.

2.3. Second model

Consider that matrix **X** consists of the elements $x_{ij} = \delta \ln k_{ij}$ (i = 1, 2,. . ., n; j = 1, 2,. . ., m) defined from

$$\begin{aligned}x_{ij} &= \delta \ln k_{ij} = \ln k_i(\varphi_1, \varphi_2, \ldots, \text{pH}) - \ln k_i(\varphi_1 = a_1, \\ \varphi_2 &= a_2, \ldots, \text{pH} = a_g)\end{aligned} \tag{11}$$

where $\ln k_i(\varphi_1 = a_1, \varphi_2 = a_2,\ldots, \text{pH} = a_g)$ is a reference value of $\ln k$ and, in particular, the $\ln k$ value of the ith solute measured when $\varphi_1 = a_1$, $\varphi_2 = a_2,\ldots$, and pH = a_g. If we use as an input of PCA the centered matrix $\mathbf{X}_c$ with elements $x_{ij} - \bar{x}_j$, then Eq. (8) is still valid but with σ_j = 1. That is,

$$x_{ij} = \bar{x}_j + v_{1j}P_{i1} + v_{2j}P_{i2} + \cdots + v_{kj}P_{iq} \tag{12}$$

Moreover, since for a certain solute i the quantity x_{ij} takes positive and negative values, there will be a set of P_{ip} values, which will be denoted by P_{0p}, such that Eq. (12) yields

$$0 = \bar{x}_j + v_{1j}P_{01} + v_{2j}P_{02} + \cdots + v_{qj}P_{0q} \tag{13}$$

Below we show that such a set of P_{0q} values always exists. From Eqs. (12) and (13) we obtain

$$x_{ij} = v_{1j}P^*_{i1} + v_{2j}P^*_{i2} + \cdots + v_{qj}P^*_{iq} \tag{14}$$

where $P^*_{iq} = P_{iq} - P_{0q}$. If we compare this equation to Eq. (6), we readily conclude that P^*_{iq} can be directly computed by PCA provided that the input matrix **Z** has been replaced by the matrix **X** with elements $x_{ij} = \delta \ln k_{ij}$. That is

$$\mathbf{S}^* = \mathbf{X}\mathbf{V}(\mathbf{V}^T\mathbf{V})^{-1} \tag{15}$$

where the elements of matrix $\mathbf{S}^*$ are the values of P^*_{iq}. This proves also that the set of P_{0q} exists under all circumstances.