

Number of components using modified PCA scree plot in spectroscopy

Tomáš Syrový, Milan Meloun

Department of Analytical Chemistry, Faculty of Chemical Technology,

University of Pardubice,

Naměstí Čs. Legií 565, 532 10 Pardubice, Czech Republic

e-mail: tomas.syrový@upce.cz, milan.meloun@upce.cz

The determination of the number of components in a mixture is an important tool for qualitative and quantitative analysis in spectroscopy. The accuracy of selected indices for an estimation of the number of components that contribute to a set of spectra was critically tested on experimental data sets of protonation equilibria of drugs using the INDICES algorithm in S-Plus. Methods are classified into precise methods and approximate methods. Methods are based on the first criterion concerning the procedure on finding the point where the slope of the indicator function $PC(k) = f(k)$ changes so called scree plot. Besides the first criterion applied, indicator function $PC(k)$ of precise methods are also based on a comparison of an actual index $PC(k)$ of method used with the experimental error of the instrument used, $s_{\text{inst}}(A)$. The precise methods are *Kankare's residual standard deviation*, $sk(A)$, *Residual standard deviation*, $RSD(k)$, *Average error criterion*, $AE(k)$, *Bartlett χ^2 criterion*, $\chi^2(k)$. Improved identification uses the second or third derivative function for some indices, namely when the number of component in the mixture is higher than three and when, due to large variations in the indicator values even at logarithmic scale, the indicator curve does not reach an obvious point where the slope changes. The *derivative criteria* $SD(k)$ are based on the point where the slope changes and reaches a maximum. The *third derivative* $TD(k)$ value crosses zero and reaches a negative minimum which can be used as a criterion. The change in slope can also be found by calculating the *derivatives ratio* $ROD(k)$. Ideally $ROD(k)$ should have a maximum at the point where $k = p$. A more difficult problem is to deduce the number of components without relying on an estimation of the instrumental error of absorbance, $s_{\text{inst}}(A)$; then the first criterion only remains. Empirical functions are *Exner function* $\psi(k)$, *Scree test*, $RPV(k)$, *Imbedded error function*, $IE(k)$, *Factor indicator function*, $IND(k)$, *Ratio of eigenvalues calculated by smoothed PCA and those by ordinary PCA*, $RESO(k)$.

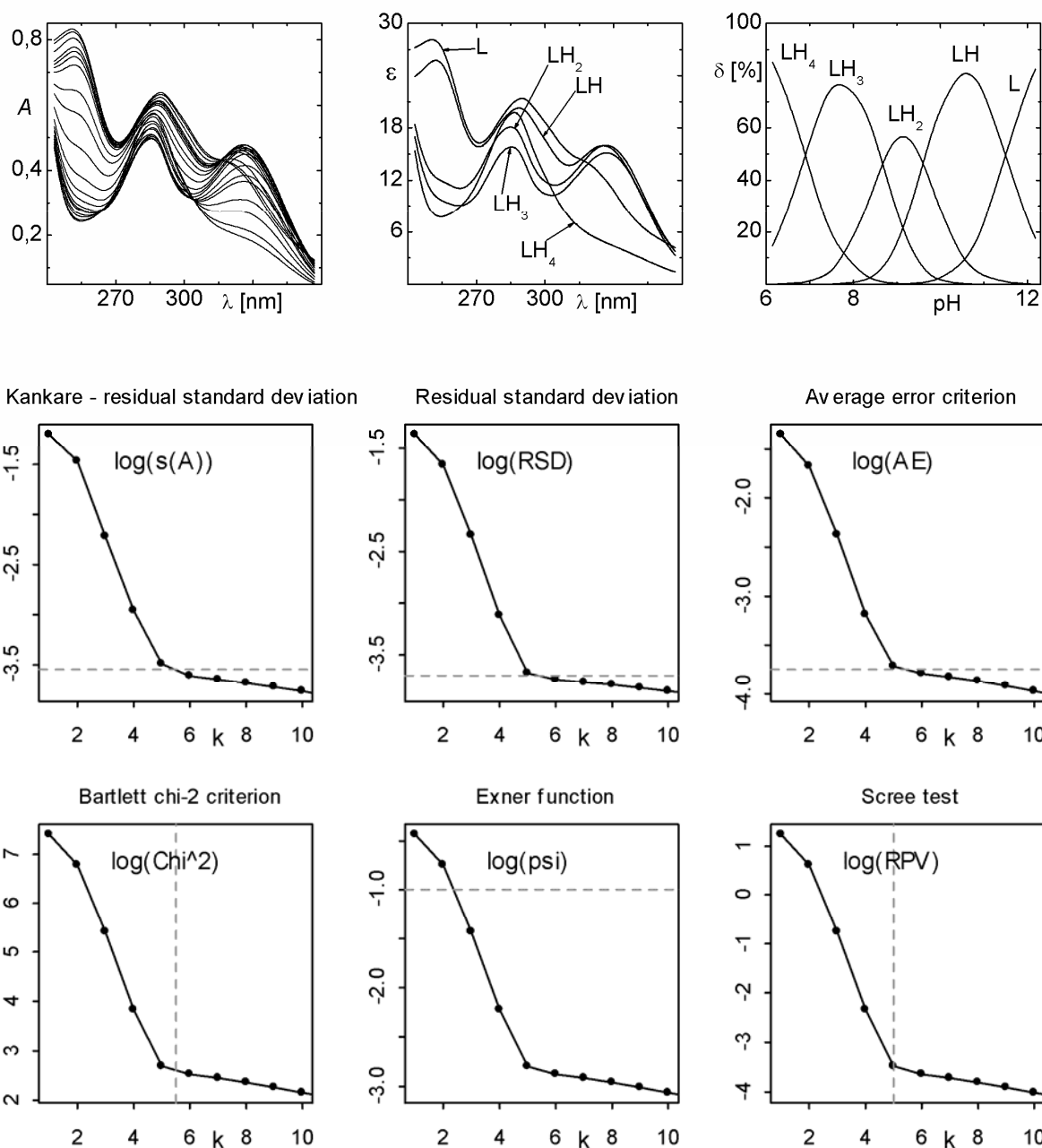


Fig. 1 Protonation equilibria of silybin is presented on pH-absorption spectra at 25°C; The spectra of molar absorptivities vs. wavelengths for all variously protonated species; Distribution diagram of the relative concentrations of all of the variously protonated species. The logarithm dependence of 6 indices methods as a function of the number of principal components k for the pH-absorbance matrix: 2nd row: Kankare's residual standard deviation $s_k(A)$, Residual standard deviation RSD , Average error criterion AE , 3rd row: Bartlett χ^2 criterion, Exner ψ function, Scree test RPV .

Keywords: Factor analysis; Number of species; Determining the number of components

[1] M. Meloun, J. Čapek, P. Mikšík, R.G. Brereton, *Critical comparison of methods predicting the number of components in spectroscopic data*, Anal. Chim. Acta 423 (2000) 51–68.

[2] M. Meloun, T. Syrový, A. Vrána, *Determination of the number of light-absorbing species in the protonation equilibria of selected drugs*, Anal. Chim. Acta, 489 (2003) 137 - 151.