International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:8, 2009

Principal Component Analysis-Ranking as a Variable Selection Method for the Simultaneous Spectrophotometric Determination of Phenol, Resorcinol and Catechol in Real Samples

Nahid Ghasemi, Mohammad Goodarzi, and Morteza Khosravi

***Abstract*—**Simultaneous determination of multicomponents of phenol, resorcinol and catechol with a chemometric technique a PC-ranking artificial neural network (PCranking-ANN) algorithm is reported in this study. Based on the data correlation coefficient method, 3 representative PCs are selected from the scores of original UV spectral data (35 PCs) as the original input patterns for ANN to build a neural network model. The results obtained by iterating 8000 .The RMSEP for phenol, resorcinol and catechol with PCranking-ANN were 0.6680, 0.0766 and 0.1033, respectively. Calibration matrices were 0.50-21.0, 0.50-15.1 and 0.50-20.0 μg ml^{-1} for phenol, resorcinol and catechol, respectively. The proposed method was successfully applied for the determination of phenol, resorcinol and catechol in synthetic and water samples.

***Keywords*—**Phenol, Resorcinol, Catechol, Principal component-ranking Artificial Neural Network, Chemometrics.

I. Introduction

PHENOL and substituted phenols are of special concern owing to the potential propagation of these compounds through the environment via leaching which comes from the industrial and petrochemical industries wastes. Some waterways can be contaminated for those phenols and hazard effects may occur to the people, also to aquatic organisms, fish and other life forms [1]. Furthermore, phenolic compounds are also formed during the natural decomposition of humic substances, tannins and lignins, and photolytic or metabolic degradation of herbicides and insecticides [2,3]. These compounds show toxicity values from moderate to higher, the toxicity level depends on the number, position and kind of substituent. The environmental aspects have became increasingly important in recent years and both the US Environmental Protection Agency (EPA) and the European Union (EU) have included the phenolic compounds as resorcinol, phenol and catechol in their list because they are considered dangerous pollutants [4,5]. The analysis established by EPA is based on liquid–liquid extraction (LLE), followed by gas chromatography (GC) using several detection methods (electron-capture detection (ECD) and mass spectrometry (MS) [6,7]. Thus, this method is complicated and it implies some disadvantages as time-consuming, high costs, also a large sample volume and toxic organic solvents are required to extract the analyte. A more recent extraction technique, solid-phase microextraction (SPME), coupled to high-performance liquid chromatography (HPLC) with UV and electrochemical detection (ED) [8,9], or coupled to gas chromatography–mass spectrometry [10,11] has been applied to the extraction of organic pollutants in environmental matrices, mainly in water samples, at trace levels. However, these methods are generally complex in nature and need expensive instruments and ultra pure solvents. In other hand, analysis of the clinical samples demands simple and fast analytical methods and therefore, finding an alternative analytical procedure or technique is crucial. Spectrophotometry combined with chemometric methods will be a simple analytical method for quantitative analysis [12, 13]. One of the chemometrics methods is multivariate calibration technique. Multivariate calibration is a collection of powerful mathematical tools that can be applied to resolve complexity in chemical analysis. It is useful in spectral analyses because the simultaneous inclusion of multiple spectral intensities can greatly improve the precision and applicability of quantitative spectral analysis of multicomponent mixtures that can not be resolved by conventional spectrometry. In recent years multivariate calibration has become an important tool in resolution of mixtures of components in many different fields including biomedical [14,15] environmental [16,17] and drug analysis [18,19] . This paper describes an analytical methodology for simultaneous determination of phenol, resorcinol and catechol using spectrophotometric method and a multivariate calibration technique (principal component analysis) with preprocessing by artificial neural network. Applications of ANNs in the field of chemistry and pharmacy have been reviewed [20-28]. The main of this work is to propose principal component-ranking artificial neural network (PCranking-ANN) method to resolve determining phenol, resorcinol and catechol in synthetic and real samples.

Nahid Ghasemi is with Islamic Azad University, Arak Branch, Arak, Iran (e-mail: N-Ghasemi@iau-arak.ac.ir).

Mohammad Goodarzi is with Young Researches Club, Islamic Azad University, Arak Branch, Arak, Iran (corresponding author phone : +98-861-3663041; fax: +98-861-3670017; e-mail: mohammad.godarzi@gmail.com).

Morteza Khosravi is with Islamic Azad University, North Tehran Branch, Tehran, Iran.

II. Experimental

A.. Reagents and Standard Solutions

All the chemicals used were of analytical reagent grade, sub-boiling, distilled water was used throughout. Stock solutions of phenol, resorcinol and catechol were purchased from Fluka. Standards of working solution were made by appropriate dilution daily as required.

B. Procedure

1. Linear Calibration Range

Individual calibration curves were constructed with several points as absorbance versus *phenol, resorcinol and catechol* concentrations. For constructing the individual calibration curves, the absorbances were measured at 270, 273 and 275 nm, against a blank for *phenol, resorcinol and catechol*, respectively. The linear regression equation for the calibration graph for *phenol* for the concentration range of 0.5–21.0 μg mL^{-1} was $A=0.0105C_{Phenol}+0.0923$ ($r^2=0.9952$) and for *resorcinol* for the concentration range 0.5–15.1 μg mL^{-1} was $A=0.0147C_{resorcinol}+0.0586$ ($r^2=0.9773$) and for *catechol* the concentration range 0.5–20.0 μg mL^{-1} was $A=0.0157C_{Catechol}+0.095$ ($r^2=0.9865$) and these were calculated according to calibration line characteristics.

2. Standard Calibration Set

A training set of 22 samples was taken (Table I). The concentrations of phenol, resorcinol and catechol were varied between 0.5–21.0, 0.5–15.1 and 0.5–20.0 μg ml^{-1}, respectively. The mixed standard solutions were placed in a 10 ml volumetric flask and completed to the final volume with deionized water (final pH 8.0). The absorption spectra were recorded between 245 and 700 nm against a blank of universal buffer. The spectral region between 245 and 700 nm, which implies working with 228 experimental points per spectra (as the spectra are digitized each 2.0 nm), was selected for analysis, because this is the zone with the maximum spectral information from the mixture components of interest. All absorption data are preprocessed by standard mean centring and scaling.

3. Prediction Set and Analysis of Real Samples

For prediction set, seven mixtures prepared, that were not included in the previous set were employed as an independent test (Table II). The real samples in this study were collected in surface waters from Gahar-Dorod (lakewater) from Venayee-Brojerd (wastewater) include Nitrate, Hydrazine, Fe(II), Cu(II) and all mineral materials.The range concentrations were added to be 0.5–21.0, 0.5–15.1 and 0.5–20.0 μg ml^{-1} for *phenol*, *resorcinol* and *catechol*, respectively.

4. Selection of the Optimum Number of Factors

The optimum number of factors (latent variables) to be included in the calibration model was determined by computing the prediction error sum of squares (PRESS) for cross- validated models using a high number of factors (half the number of total standard +1), which is defined as follows:

TABLE I
Concentration Data of the Different Mixtures used in the Calibration Set for the Determination of *Phenol*, *Resorcinol* and *Catechol* (μG ML^{-1})

Mixture	Phenol	Resorcinol	Catechol
M1	0.5	0.5	20
M2	16.5	0.5	0.5
M3	4.5	15.1	0.5
M4	0.5	4.9	10.25
M5	12.5	0.5	0.5
M6	1.5	4.9	2.5
M7	14	7	6
M8	19.5	14	12
M9	17	10.25	11.5
M10	8.5	0.5	10.25
M11	12.5	4.9	0.5
M12	0.5	15.1	0.5
M13	4.5	0.5	10.25
M14	4.5	10.25	0.5
M15	6	3	5
M16	7.5	7	3
M17	4	10.25	19
M18	0.5	0.5	15.1
M19	12.5	0.5	4.9
M20	8.5	10.25	0.5
M21	21	0.5	4.9
Validation			
M22	0.75	4.9	2.5
M23	10	3	5
M24	0.8	14	10
M25	11.5	10.25	11.5

$$PRESS = \sum (y_i - \hat{y}_i)^2$$

where y_i is the reference concentration for the *i*th sample and $\hat{y}_i$ represents the estimated concentration. The cross-validation method employed was to eliminate only one sample at a time and then PLS and PCR calibrate the remaining standard spectra. By using this calibration the concentration of the sample, left out was predicted. This process was repeated until each standard had been left out once. One reasonable choice for the optimum number of factors would be that number which yielded the minimum PRESS. Since there are a finite number of samples in the training set, in many cases the minimum PRESS value causes over fitting for unknown samples that were not included in the model. A solution to this problem has been suggested by Haaland et al. [29, 30] in which the PRESS values for all previous factors are compared to the PRESS value at the minimum. The F-Statistical test can be used to determine the significance of PRESS values greater than the minimum. The maximum number of factors used to calculate the optimum PRESS was selected as 13 and the

optimum number of factors obtained by the application of PLS and PCR models are summarized in Table II.

TABLE II
STATISTICAL PARAMETERS OF THE OPTIMIZED MATRIX USING THE PCRANKING-ANN, PLS AND PCR METHODS

parameter	PCranking-ANN		
	Phenol	Resorcinol	Catechol
RMSEP	0.668	0.0766	0.1033
RSEP(%)	0.818	0.9222	1.1549
MAE(%)	8.0812	8.9214	10.00
R^2	0.9999	0.9998	0.9998
PRESS			
NOF[a]			
PLS			
	Phenol	Resorcinol	Catechol
RMSEP	2.144	1.498	6.0914
RSEP(%)	26.27	18.03	68.10
MAE(%)	25.844	27.44	62.56
R^2	0.8613	0.8988	0.3599
PRESS	3.54	2.92	3.54
NOF[a]	5	6	5
PCR			
	Phenol	Resorcinol	Catechol
RMSEP	2.15	4.069	6.22
RSEP(%)	26.37	48.96	69.58
MAE(%)	18.75	35.15	62.29
R^2	0.8592	0.2063	0.3344
PRESS	3.63	3.63	3.64
NOF[a]	3	3	4

[a] Number of Factor

In all instances, the number of factors for the first PRESS values whose F-ratio probability drops below 0.75 was selected as the optimum. In Fig 1, the PRESS obtained by optimizing the calibration matrix of the absorbance data with PLS and PCR models is shown.

C. Instrumentation and Software

A Scinco (SUV-2120) spectrophotometer controlled by a Hewlett-Packard computer and equipped with a 1 cm path length quartz cell was used for UV–vis spectra acquisition. A Metrohm 692 pH-meter furnished with a combined glass-saturated calomel electrode was calibrated with at least two buffer solutions at pH 3.00 and 9.00. The back propagation neural network algorithm having three layers was used in Matlab (version 6.5, MathWork Inc.) using NNet toolbox. It's worth mentioning that PCA modeling was also written in the same software. It should be noted that all programs were run on a Pentium (IV), personal computer, with windows XP operating system. PLS and PCR calculus were carried out in the 'PLS Toolbox', version 2.0 (Eigenvectors Company).

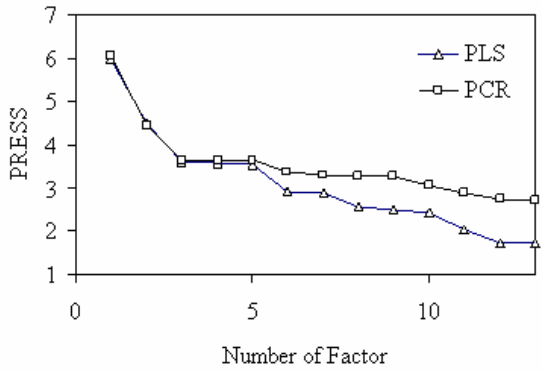

Fig. 1 Plots of PRESS versus number of factors by PLS and PCR Methods

D. Partial Least Squares

The determination of the *phenol, resorcinol and catechol* in mixtures by spectrophotometric using multivariate calibration involved constructing calibration and prediction set. According to the individual calibrations in section II.B.1, the calibration matrix designed. In Table I, the compositions of the ternary mixtures used in the calibration matrices are summarized.

TABLE III
ARCHITECTURE OF THE ANN MODELS AND THEIR SPECIFICATIONS

No. of nodes in the input layer	3+1[a]
No. of nodes in the hidden layer	3
No. of nodes in the output layer	3
Momentum	0.528
Learning rate	0.374
No. of iterations	8000
Transfer function	$Y(j) = 1/[1+\exp(-(\alpha_j Net(j) + \theta_j))]^b$

[a] Bias

[b] $Net(j) = \sum x_i w_{ij}$

For prediction set, seven mixtures prepared (see Table 4). To ensure that the prediction and real samples are in the subspace of training set, the score plot of first principal component versus second was sketched and all the samples are spanned with the training set scores.

TABLE IV
COMPOSITION OF SYNTHETIC MIXTURES AND PREDICTED VALUES FOR DETERMINATION OF PHENOL, RESORCINOL AND CATECHOL (μG ML-1)

Amount added (μg mL^{-1})							
Phenol	4.5	0.5	8.5	3	18.5	2.5	4
Resorcinol	0.5	10	4.9	3	7	14	10.25
Catechol	15.1	0.5	0.5	2.5	6	17	0.5
Determined by ANN, μg mL^{-1}							
Phenol	4.45	0.5	8.47	3.01	18.5	2.36	3.91
Resorcinol	0.54	9.87	4.76	3.02	7.01	14	10.2
Catechol	15.1	0.5	0.5	2.36	6.21	16.9	0.53
Determined by PLS, μg mL^{-1}							
Phenol	3.44	-0.846	8.245	0.67	15.9	2.64	8.413
Resorcinol	-0.01	8.695	5.261	6.28	6.71	14.6	11.80
Catechol	15.3	3.30	1.2	5.42	9.04	12.0	14.95
Determined by PCR, μg mL^{-1}							
Phenol	5.51	-0.53	8.18	3.57	14.8	0.611	7.59
Resorcinol	4.65	9.42	4.58	11.98	4.81	10.48	9.75
Catechol	14.0	3.05	1.65	4.21	9.57	12.926	15.66

III. RESULTS AND DISCUSSION

Fig. 2 shows the absorption spectra in aqueous solution of individual *phenol*, *resorcinol* and *catechol* at pH 8.0. With the aim of investigation the possibility of determining *phenol*, *resorcinol* and *catechol* in mixtures, the optimum working conditions were studied under the conditions previously established for each *phenol*, *resorcinol* and *catechol*. A universal buffer solution of pH 8.0 was selected. In order to select the optimum pH value at which the minimum overlap occurs, influences of the pH of the medium on the absorption spectra of *phenol*, *resorcinol* and *catechol* were studied over the pH range 1.0–10.0. The wavelengths used to generate calibration curves were 270, 273 and 275 nm for *phenol*, *resorcinol* and *catechol*, respectively. However, the system presents non-linearities in the signal-concentration relationship, which calls for a proper non-linear chemometric tool. Application of neural network in multivariate calibration is proposed when significant non-linearities are observed in the data. The principal component analysis (PCA) was applied to the data set and the scores of the principal components (PCs) were selected as input nodes for the input layer. It should be noted that the actual number of PCs , number of nodes in the hidden layer, the Learning rate, momentum and the number of epochs were selected based on the minimum value for the root mean square errors of the prediction set to prevent overfitting the model. The optimized parameters of PC-ANN architectures for the mixture are summarized in Table III. The data obtained from application of singular value decomposition on conventional spectra were processed by PC-ANN in order to increase determination range and to obtain wider dynamic ranges in simultaneous determination of *phenol, resorcinol and catechol*.

TABLE V
PCRANKING-ANN RESULTS APPLIED ON THE REAL MATRIX SAMPLES

Samples	Added		
	phenol	resorcinol	catechol
Tap Water	5.0	10	4
Venayee	12	5	4
Gahar	6	11	11
	Found		
Tap Water	5.27(2.42)[a]	9.57(0.7)[a]	3.99(2.86)[a]
Venayee	11.93(1.21)[a]	4.79(2.03)[a]	4.11(1.03)[a]
Gahar	6.6(2.05)[a]	11.70(1.08)[a]	1.66(3.36)[a]

[a] Relative standard of deviation (n=3)

A. Selection of Descriptors using PCA-ranking Approach

PCA-ranking technique was chosen as feature selection method. This method is an extremely useful explorative tool which maps samples through scores and individual variables through scores in a new vector space defined by the three high correlative principal components (PCs). The order of the PCs based on their decreasing correlation coefficients is PC2>PC1>PC6>PC3>PC5>PC8>…. It can be seen that among the PCs, the three components of PC2, PC1 and PC6, respectively, show the largest correlation coefficients with the chemical shifts. Although, PC1 demonstrates 90.3% of the variances in the space of the scores, but it shows a lower correlation with the concentration compared with the PC2. PC6 shows even lower variances in space of scores, but regarding the correlation with the concentration its rank is three among the PCs. However, in the PCA-ranking method the criterion for the selection of the space of the PCs containing the important respond matrix is correlation of the PCs with the concentration. Therefore, we have inspected the scores in the space of the correlated PCs for choosing the respond matrix with highest variances in this space.

B. Determination of Phenol, Resorcinol and Catechol in Synthetic Mixture

The predictive ability of method was determined using seven three-component *phenol*, *resorcinol* and *catechol* mixture (their compositions are given in Table IV). The results obtained by applying PCranking-ANN algorithm, PLS and PCR to seven synthetic samples are listed in Table IV. Also shows the recovery for prediction series of *phenol*, *resorcinol* and *catechol* mixture. As can be seen, the recovery was also quite acceptable. The root mean square error of prediction and relative standard error of prediction results are summarized in Table II. The plots of the predicted concentration versus actual values are shown in Fig. 3 for *phenol*, *resorcinol* and *catechol* (R^2 values are also shown).

C. Statistical Parameters

Four general statistical parameters were selected to evaluate the prediction ability of the constructed model. These parameters are root mean square error of prediction (RMSEP),

relative standard error of prediction (RSEP), mean absolute error (MAE) and square of correlation coefficient (R^2). These parameters are calculated as follows:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_{pred} - y_{obs})^2}{n}} \quad (1)$$

$$RSEP(\%) = 100 \times \sqrt{\frac{\sum_{i=1}^{n}(y_{pred} - y_{obs})^2}{\sum_{i=1}^{n}(y_{obs})^2}} \quad (2)$$

$$MAE(\%) = \frac{100}{n} \times \sqrt{\sum_{i=1}^{n}\left|(y_{pred} - y_{obs})\right|} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{pred} - y_{obs})^2}{\sum_{i=1}^{n}(y_{obs} - y_{mean})^2} \quad (4)$$

Where y_{pred} the predicted concentration in the sample is, y_{obs} is the observed value of the concentration in the sample and n is the number of samples in the validation set. The values for RMSEP, RSEP, MAE and R^2 are given in Table II.

D. Determination of Phenol, Resorcinol and Catechol in Real Samples

In order to test the applicability and matrix interferences of the proposed method to the analysis of real samples, the method was applied in a variety of situations. For this purpose, diverse spiked samples and reference materials were analyzed. Table V shows the results obtained for real matrix samples. Therefore, the PCranking-ANN model is able to predict the concentrations of each *phenol*, *resorcinol* and *catechol* in the real matrix sample.

IV. Conclusion

In general, using a neural network technique is time-consuming and difficult to be carried out with the data of an entire spectral signal for the simultaneous multicomponent determination. In this work, the correlation coefficient and the standard deviation methods are used as indicators to select only 3 PCs from 35 PCs of Scores of each original UV spectral signal, without loss of information in the original experimental data. Correspondingly, Fig. 3 shows the prediction values versus actual values of phenol, resorcinol and catechol materials, the results indicate that accurate estimation and prediction can be obtained by PCranking-ANN.

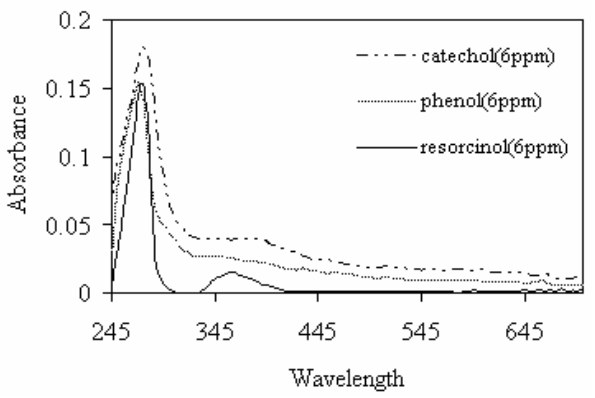

Fig. 2 Characteristic absorption spectrum of the individual *phenol*, *resorcinol* and *catechol* at pH 8.0

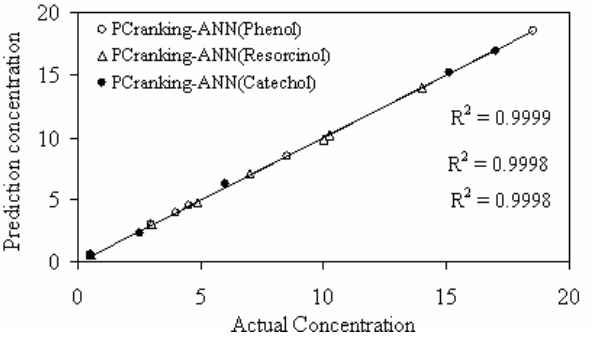

Fig. 3 Plots of predicted concentration versus actual concentration for *phenol*, *resorcinol* and *catechol* by PCranking-ANN method

References

[1] Y. Ni, L.Wang, S. Kokot, Anal. Chim. Acta. 431,101(2001).
[2] N. Masqué, E. Pocurull, R.M. Marcé, F. Borrull, , Chromatographia. 47, 176(1998).
[3] A.S. Narang, Ch.A. Vernoy, G.A. Eadon, J. Assoc. Off. Anal. Chem.66,1330(1983).
[4] US EPA, Code of Federal Regulations, title 40, part 136, Washington DC, 1992.
[5] US EPA, SW-846, Office of Solid Wastes, Washington DC, 1997.
[6] EPA Method 604, Phenols in Federal register, Friday, Environmental Protection Agency Part VIII. CFR Part 136, vol. 40, 1984, pp. 58–66.
[7] EPA Method 8041, Phenols by Gas Chromatography: Capillary Column Technique, Washington, DC, 1995, pp. 1–28.
[8] A. Peñalver, E. Pocurull, F. Borrull, R.M. Marce, J. Chromatogr. A., 953,79(2002).
[9] E. Gonzalez-Toledo, M.D. Prat, M.F. Alpendurada, J. Chromatogr., A,923, 45(2001).
[10] H. Bagheri, A. Saber, S.R. Mousavi, J. Chromatogr. A, 1046 27,(2004).
[11] M. Llompart, M. Lourido, P. Landýn, C. Garcýa-Jares, R. Cela, J. Chromatogr. A, 963, 137(2002).
[12] A. Niazi, M. Goodarzi, Spectrochimica Acta Part A. 69,1165(2007).
[13] M. Goodarzi, T. Goodarzi, N. Ghasemi, Ann. Chim., 97, 303(2007).
[14] G. Musmarra, D. F.Condorelli, S. Scire, A. S. Costa, Biochem. Pharmacol.62, 547(2001).
[15] M.Lapinsh, P.Prusis, A.Gutcaits, T. Lundstedt,J.E.S. Wikberg, Biochim. Biophys. Acta, 180,1525(2001).
[16] M.Jesús Gómez González, Olga Domínguez Renedo, M.Julia Arcos Martínez,Talanta, 68,67(2005).
[17] M. Jalali-Heravi, S. Masoum, P. Shahbazikhah, J. Magnetic Resonance, 171,176(2004).
[18] M.N. Taib, R. Andres, R, Narayanaswamy Anal. Chim. Acta, 330,31(1996).

[19] M.L.Luis, J.M.G.Fraga, F.Jiménez, A.I.Jiménez, J.J.Arias, Talanta, 53,761(2001).
[20] J. Zupan, Acta. Chim. Slov., 41,327(1994).
[21] J. Zupan, M. Novic, I. Ruisanchez, Chemom. Intell. Lab.Syst.38,1(1997).
[22] J. Bourquin, H. Schmidli, P. Van Hoogevest, H.Leuenberger, Pharm. Dev. Technol. 2,95(1997).
[23] K. Varmuza, Applied Chemometrics. Available from: <http:// ww.lcm.tuwien.ac.at>.
[24] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam, 1997.
[25] J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design, Wiley-VCH: New York, 1999, chapter 1.
[26] S. Agatonovic-Kustrin, R. Beresford, J. Pharm. Biomed.Anal.22,717 (2002).
[27] A. Eghbaldar, T. P. Forrest, D. Cabrol-Bass, Anal. Chim. Acta , 359, 283(1998).
[28] Y. Ni, C. Liu, Anal. Chim. Acta, 396,221(1999).
[29] D.M. Haaland, E.V. Thomas, Anal. Chem. 60,1193(1988).
[30] D.M. Haaland, E.V. Thomas, Anal. Chem. 62,1091(1990).