Development of a New Regression Analysis Method Using Independent Component Analysis

Hiromasa Kaneko, Masamoto Arakawa, and Kimito Funatsu*

Department of Chemical System Engineering, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku,
Tokyo 113-8656, Japan

Received July 12, 2007

In this paper, independent component analysis (ICA) and regression analysis are combined to extract significant components. ICA is a method that extracts mutually independent components from explanatory variables. A relationship between the independent components and an objective variable is constructed by the least-squares method. This method is named ICA-MLR (MLR = multiple linear regression). We verified the superiority of ICA-MLR over partial least squares (PLS) with simulation data and tried to apply this method to a quantitative structure−property relationship analysis of aqueous solubility. We constructed models between aqueous solubility and 173 molecular descriptors. PLS and genetic algorithm PLS models were constructed for a comparison of ICA-MLR. R^2, Q^2, and $R_{\mathrm{pred}}{}^2$ values of the PLS model are 0.836, 0.819, and 0.848, respectively. These values of the ICA-MLR model are 0.937, 0.868, and 0.894, respectively. ICA-MLR achieved higher predictive accuracy than PLS. ICA-MLR could extract effective components from explanatory variables and construct the regression model with high predictive accuracy. In addition, the information of regression coefficients $\mathbf{b}_{\mathrm{ICA-MLR}}$ indicates the magnitude of contribution of each descriptor in the analysis of aqueous solubility.

INTRODUCTION

In many fields of chemistry, attempts are made to apply informatics methods to various problems. Recently, "chemoinformatics"[1] has become a generic term used for referring to these fields. Chemoinformatics is a field in which problems of chemistry are solved by using an informatics method, and there is much research in the field, revolving around topics such as quantitative structural−property relationships (QSPRs), quantitative structure−activity relationships, reaction design, and drug design.

Multivariate techniques such as multiple linear regression (MLR), principal component regression (PCR),[2] and partial least squares (PLS)[3] are powerful tools for handling several problems in chemoinformatics. It is possible to construct an accurate model by using these methods, for example, PLS, but is difficult to construct a predictive model. As for a problem we often face, there is the possibility of statistical problems occurring, such as overfitting.

It is important to indicate the magnitude of contribution of each variable to a model. However, it is dangerous to simply consider regression coefficients as important for each variable because there are correlations in explanatory variables. Thus, it is desirable that a prediction model be constructed that has high predictive power and that is easy to interpret in various fields of science.

In this paper, we have applied the independent component analysis (ICA)[4] method to a regression analysis to extract significant components from explanatory variables and construct a model that has high predictive power and that is easy to interpret. ICA is a method that is used in many fields such as signal processing. Through making full use of the high-order statistical characteristics of the source, that is, the fourth-order central moment, ICA can effectively resolve the independent components from the measured mixed signals without any additional information about the source signals. It has been widely applied in fields such as spectral analysis[5,6] and statistical process control.[7] A regression method based on ICA, independent component regression (ICR), was proposed by Chen and Wang.[5] In their work, the NIR spectra of water, starch, and protein mixtures were investigated.

We propose the ICA-MLR method, which combines ICA and MLR. After extracting independent components from explanatory variables, a relationship between the components and an objective variable is constructed by the least-squares method. Basically, the ICA-MLR method is the same as the ICR method. In this paper, we prove that the ICA-MLR method and ICA-PLS method, which combines ICA and PLS, are essentially the same. By using the ICA method effectively, we can construct a model that has high predictive power and that is easy to interpret.

First, we showed the superiority of ICA-MLR over PLS with the simulation data and, next, tried to apply this method to QSPR. QSPR is a technique utilized to quantitatively correlate the chemical structure and the properties that the chemical structure has. We constructed a model between aqueous solubilities based on the experimental data for 1290 molecules[8] and 173 molecular descriptors. Aqueous solubility is one of the most important physicochemical properties, which plays a significant role in various physical and biological processes and has a marked impact on the design and pharmaceutical formulation development. Numerous in silico based methods for the prediction of solubility of organic compounds have been developed.[9−14] We verified the superiority of ICA-MLR by comparing the results of PLS, genetic algorithm PLS (GAPLS),[15] and ICA-MLR. By using the ICA-MLR method, we constructed a simple and highly predictive model.

* Corresponding author e-mail: funatsu@chemsys.t.u-tokyo.ac.jp.

10.1021/ci700245f CCC: $40.75 © 2008 American Chemical Society
Published on Web 03/06/2008

METHODS

PLS. Modeling a relationship between **X** and **y** is done by using MLR, which works well as long as **X** variables are few and uncorrelated. However, it is impossible to construct a regression model when the number of **X** variables is more than the number of samples. Thus, attempts to pretreat **X** have been proposed. One of these methods is PCR. It is a method that constructs a regression model of **y** by means of the score matrix **T** calculated by principal component analysis (PCA). Since **T** is mutually orthogonal, a stable model would be constructed by the PCR method.

PLS is a method for relating **X** and **y**, by a linear multivariate model, but goes beyond traditional regression methods in that it models also the structures of **X** and **y**. In PLS modeling, the covariance between score vector $\mathbf{t}_i$ and **y** is maximized. A PLS model has higher predictive power than those of MLR and PCR.

A PLS model consists of two equations, as follows:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}$$
$$\mathbf{y} = \mathbf{Tq} + \mathbf{f} \qquad (1)$$

where **P** is an **X**-loading matrix, **q** is a **y**-loading vector, **E** is a matrix of **X** residuals, and **f** is vector of **y** residuals. The PLS-regression model is as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathrm{const}$$
$$\mathbf{b} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{q} \qquad (2)$$

where **W** is an **X**-weight matrix and **b** is a vector of regression coefficients. The number of components must be appropriately decided to construct a highly predictive model. R^2 and Q^2 values are used as the measure and defined as follows:

$$R^2 = 1 - \frac{\sum(y_{\mathrm{obs}} - y_{\mathrm{calc}})^2}{\sum(y_{\mathrm{obs}} - \bar{y})^2}$$
$$Q^2 = 1 - \frac{\sum(y_{\mathrm{obs}} - y_{\mathrm{pred}})^2}{\sum(y_{\mathrm{obs}} - \bar{y})^2} \qquad (3)$$

where y_{obs} is the actual **y** value, y_{calc} is the calculated **y** value, and y_{pred} is the predicted **y** value in a procedure of cross-validation such as leave-one-out. The optimum number of components is determined by the best Q^2, the first local maximum of Q^2, and so on. However, if the rise in Q^2 value is gentle, the number of components gets too large by these methods. In this paper, the number of components that is just before the point where the ratio of rise of Q^2 is less than 0.03 is used as the optimum number of components.

It is important to indicate the magnitude of contribution of each variable to a model. However, it is dangerous to simply consider the vector of regression coefficients **b** as important as the variables. For example, we assume that there are is an objective variable $\mathbf{y}^{\mathrm{T}} = [\ 2\ 4\ 6\ 8\ ]$ and explanatory variables $\mathbf{x}_1^{\mathrm{T}} = [\ 1\ 2\ 3\ 4\ ]$ and $\mathbf{x}_2^{\mathrm{T}} = [\ 2\ 4\ 6\ 8\ ]$. In this case, there are infinite solutions of **b** such as $\mathbf{b}^{\mathrm{T}} = [\ 2\ 0\ ]$, $\mathbf{b}^{\mathrm{T}} = [\ 0\ 1\ ]$, and $\mathbf{b}^{\mathrm{T}} = [\ 1\ 0.5\ ]$. It is inaccurate to consider **b** as the magnitude of contribution of each variable for a model. In other words, when there is a correlation between **X** variables, the value of **b** cannot be trusted.

GAPLS. One of the methods that selects important variables from **X** variables is GAPLS. A GA[16] is an optimization method by which a principle of a natural evolution in biology is modeled. Species having a high level of fitness under some environmental conditions can prevail in the next generation, and the best species may be reproduced by crossover together with random mutations of chromosomes in the surviving ones. The solution space around superior individuals is searched preferentially, and then a solution that is close to the optimum is discovered.

GAPLS is a variable selection method applying GA. Each of the **X** variables is assigned to bit of chromosome, and a set of variables that is able to construct the optimum PLS model is searched. The Q^2 value calculated with a cross-validation method such as leave-one-out is used as an evaluation function of the chromosome. Therefore, a model that has high predictive power is obtained.

A GAPLS model is simpler than one of PLS because the number of **X** variables decreases. However, if a variable is excluded, the mean of it is not reflected on the model. For example, when there is a pair of variables whose correlation coefficient is close to +1 or −1, one of them may be excluded by the GAPLS method. But it is not desirable because the information of the variable is considered useful.

ICA. ICA is a method for transforming observed multivariate data into statistically independent components expressed as the linear combinations of observed variables. Statistical independence is a different concept from decorrelation. Denote $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$ as random variables with a joint probability density function (pdf) of $\mathrm{p}(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n)$, and assume that these variables have zero mean; then, they are said to be mutually statistically independent if the following condition holds:

$$p(\mathrm{s}_1,\mathrm{s}_2,...,\mathrm{s}_n) = \prod_{i=1}^{n} p_i(\mathrm{s}_i) \qquad (4)$$

where $p_i(\mathbf{s}_i)$ (for $i = 1, 2, ..., n$) denotes a marginal pdf of $\mathbf{s}_i$, that is, the pdf of $\mathbf{s}_i$ when it is considered alone. For performing ICA, **X** is first transformed into mutually uncorrelated variables. By defining the sphering matrix as **M**, transformed matrix **Z** is given as

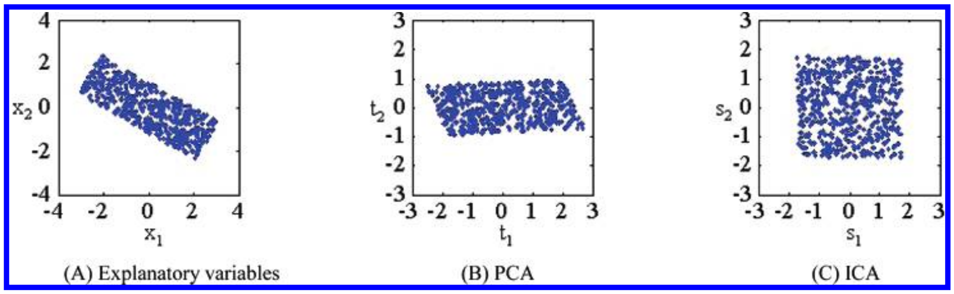
$$\mathbf{Z} = \mathbf{XM} \qquad (5)$$

Generally, this pretreatment can be accomplished by singular value decomposition. One dimension of **X** is reduced at a time. Next, **Z** is transformed into mutually independent variables as follows:

$$\mathbf{S} = \mathbf{ZB} \qquad (6)$$

where **S** is an independent component matrix and **B** is a transformation matrix. Several techniques that calculate **B** have been proposed. In this paper, FastICA[17] is used. The relationship between **X** and **S** is given as

$$\mathbf{S} = \mathbf{XMB} = \mathbf{XW}$$
$$\mathbf{W} = \mathbf{MB} \qquad (7)$$



(A) Explanatory variables (B) PCA (C) ICA

Figure 1. Comparison between PCA and ICA.

where **W** is a separating matrix. Mutually independent components underlying in **X** can be extracted by **W**.

Figure 1 shows a comparison between PCA and ICA. Figure 1A is a plot of explanatory variables $\mathbf{x}_1$ and $\mathbf{x}_2$. PCA and ICA methods are applied to this data set. A plot of principal components $\mathbf{t}_1$ and $\mathbf{t}_2$ extracted by PCA method is shown in Figure 1B, and a plot of independent components $\mathbf{s}_1$ and $\mathbf{s}_2$ extracted by the ICA method is shown in Figure 1C.

Mutually uncorrelated components whose dispersion is large are extracted in order by the PCA method. On the one hand, the distribution of $\mathbf{s}_1$ and $\mathbf{s}_2$ is like that in Figure 1C. In other words, even if a value of one of these components is obtained, the information of another one is not obtained, because $\mathbf{s}_1$ and $\mathbf{s}_2$ are mutually independent.

ICA-MLR. By using the ICA method, independent components were extracted from **X** variables. These components can resolve the correlation problem because they are mutually independent. Therefore, we propose the ICA-MLR method, which combines ICA and MLR. First, to extract independent component matrix **S**, the ICA method is applied to **X**:

$$\mathbf{S} = \mathbf{XW} \tag{8}$$

The next step is to calculate vector of regression coefficients $\mathbf{b}_{S\rightarrow y}$ from **S** by the MLR method as follows:

$$\mathbf{b}_{S\rightarrow y} = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{y}$$
$$\mathbf{y} = \mathbf{S}\mathbf{b}_{S\rightarrow y} \tag{9}$$

It is possible to convert $\mathbf{b}_{S\rightarrow y}$ to regression coefficient vector $\mathbf{b}_{X\rightarrow y}$, which estimates **y** from **X**.

$$\mathbf{y} = \mathbf{XW}\mathbf{b}_{S\rightarrow y}$$
$$= \mathbf{X}\mathbf{b}_{X\rightarrow y}$$
$$\mathbf{b}_{X\rightarrow y} = \mathbf{W}\mathbf{b}_{S\rightarrow y} \tag{10}$$

By the use of eq 10, it is possible to estimate **y** directly from **X** and to verify a relationship between **X** and **y** by analyzing $\mathbf{b}_{X\rightarrow y}$. However, it is dangerous to simply consider $\mathbf{b}_{X\rightarrow y}$ as the magnitude of contribution to a model because there is a correlation between variables. On the one hand, it is possible to consider $\mathbf{b}_{S\rightarrow y}$ as the magnitude of contribution to a model because the independent components are mutually independent.

It is conceivable that PLS is used as a method to construct a regression model instead of MLR, but these are essentially the same. The proof is shown in Appendix A.

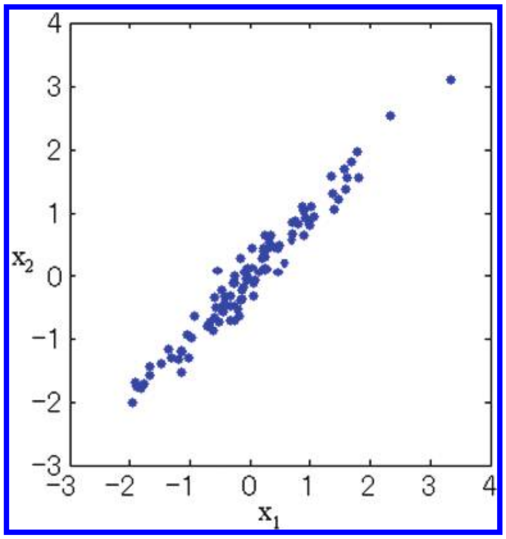


Figure 2. $\mathbf{x}_1-\mathbf{x}_2$ plot.

RESULTS AND DISCUSSION

We verified the superiority of ICA-MLR over PLS with the simulation data and tried to apply this method to QSPR. We constructed a model between aqueous solubility and 173 molecular descriptors. PLS and GAPLS models were constructed for a comparison of ICA-MLR.

Modeling of the Simulation Data. The ICA-MLR method was compared with the PLS method using the simulated data to verify the superiority of ICA-MLR over PLS. Explanatory variables $\mathbf{x}_1$ and $\mathbf{x}_2$, whose correlation is fairly high, were calculated. Figure 2 shows the distribution of $\mathbf{x}_1$ and $\mathbf{x}_2$. An objective variable $\mathbf{y}_1$ was calculated as

$$\mathbf{y}_1 = (\mathbf{x}_1 - \mathbf{x}_2) \tag{11}$$

$\mathbf{x}_1$ is vector of random numbers from normal distribution given a standard deviation of 1 and a mean of 0, and $\mathbf{x}_2$ is a vector of sum of $\mathbf{x}_1$ and one-fifth of the random numbers that do not correlate with $\mathbf{x}_1$. A total of 25 sets of pairs of these explanatory variables and the objective variable were calculated. The **X** variables were these 50 variables, and the **y** variable were calculated as

$$\mathbf{y}_i = \mathbf{x}_{2i-1} - \mathbf{x}_{2i}$$
$$\mathbf{y} = \sum_{i=1}^{25}\mathbf{y}_i \tag{12}$$

The number of samples is 100. PLS and ICA-MLR methods were applied to this data set.

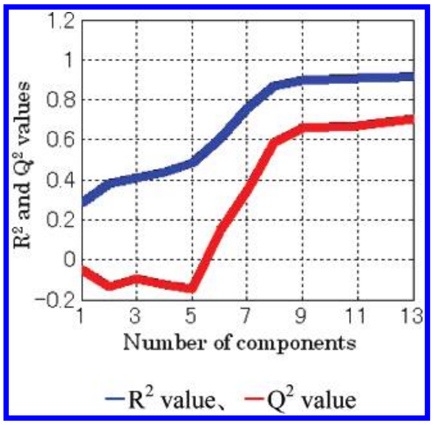


Figure 3. R^2 and Q^2 values in PLS modeling.

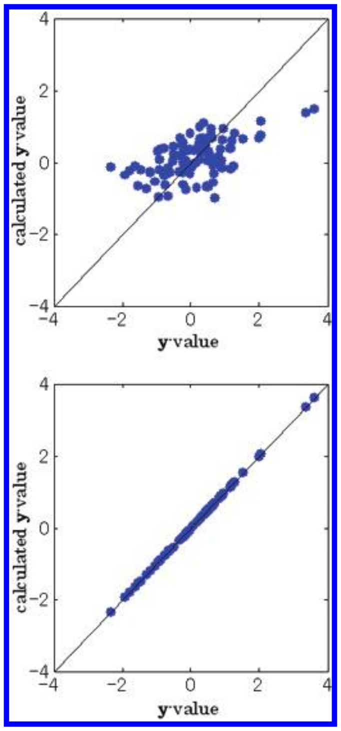


Figure 4. The relationship between **y** value and calculated **y** value in PLS (upper) and ICA-MLR (lower) modeling.

Figure 3 shows R^2 and Q^2 values in PLS modeling. They are small, while the number of components is small. Correlation between **y** and score vector $\mathbf{t}_i$ were small because the dispersion of the direction of $\mathbf{x}_1 = \mathbf{x}_2$ is too large.

On the one hand, R^2 and Q^2 values were 1.00 in ICA-MLR modeling. Figure 4 shows a plot of **y** values and calculated **y** values. It was verified that the ICA-MLR model has higher predictive power than that of PLS with one component.

QSPR Study on Aqueous Solubility. We have constructed QSPR models on aqueous solubility using PLS, GAPLS, and ICA-MLR methods, and compared the results. In this paper, we analyzed aqueous solubility data investigated by Hou.[8] Aqueous solubility is expressed as logS, where S is the solubility at a temperature of 20−25 °C in moles per liter. The Hou data set includes 1290 diverse compounds and has been analyzed by several groups.[9−14]

Molecular descriptors of 346 of them were calculated on this data set by means of the ModelBuilder software.[18] The descriptors with zero variance are meaningless; thus, they were removed. When the correlation coefficient of a pair of descriptors was greater than 0.9, either descriptor was removed because the existence of variables whose correlation coefficients are high decreases the predictive accuracy of a model. We reduced the descriptors as above, and 173 descriptors remained for constructing prediction models. This data set was divided into a training set of 878 molecules and a test set of 412 molecules just as Hou did. The 878 molecules were used to construct the prediction model, and the 412 molecules were used to test the predictive accuracy of the obtained model.

Table 1. Prediction Profiles and Statistical Data for Each Model

	number of explanatory variables	number of components	R^2	Q^2	R_{pred}^2
PLS	173	3	0.836	0.819	0.848
ICA-MLR	169[a]		0.937	0.868	0.894
GAPLS	80	4	0.866	0.849	0.867

[a] The number of independent components.

Table 1 shows the results of PLS, GAPLS, and ICA-MLR modeling. The R_{pred}^2 value is an R^2 value that is calculated with the test set. The PLS method showed the model with three components. Statistics of this model are $R^2 = 0.836$, $Q^2 = 0.819$, and $R_{\text{pred}}^2 = 0.848$. By using the PLS method, a rather accurate predictive model could be constructed. On the one hand, we applied the ICA method to descriptors to construct the ICA-MLR model. As a result, 169 independent components were extracted because the rank of **X** was 169. The relationship between the components and logS was modeled by using MLR with the least-squares method. R^2, Q^2, and R_{pred}^2 values were 0.937, 0.868, and 0.894, respectively. The ICA-MLR model has higher predictive power than PLS. We verified that the simple model that has good predictive performance can be constructed by using ICA as a pretreatment of explanatory variables.

A GAPLS model was constructed to compare the result with that of ICA-MLR. We used the Genetic Algorithm Optimizing Toolbox for MATLAB5[19] for the calculation of GAPLS. A Q^2 value calculated with a 10-fold cross-validation was used as an evaluation function of the chromosome. The number of generations was set to 1000, and the number of populations was set to 300. The probability of crossover and the probability of mutation were given as default values, 0.6 and 0.05, respectively. As a result of the optimization by GA, we obtained a PLS model with four components using 80 variables. The number of descriptors was reduced, but the number of components increased. R^2, Q^2, and R_{pred}^2 values were 0.866, 0.849, and 0.867, respectively. These values are larger than those of PLS but smaller than those of ICA-MLR. Therefore, we verified the superiority of ICA-MLR over GAPLS.

Figure 5 shows a plot of experimental and predicted logS values in PLS and ICA-MLR modeling. The plot shows a much tighter clustering of predicted values along the diagonal in ICA-MLR modeling, reflecting the higher prediction of logS. Furthermore, when values of logS are large, the predictive accuracy of ICA-MLR is higher than that of PLS. The experimental solubility values can differ by ∼1.0 log unit, especially for compounds with a low logS value. Therefore, the larger the value of logS is, the more desirable a predictive accuracy is.

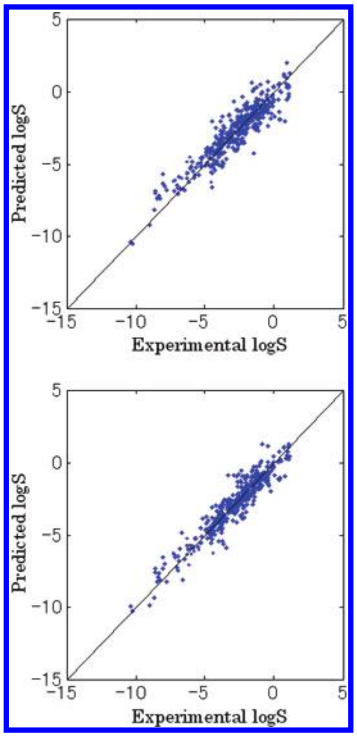


Figure 5. The relationship between experimental and predicted log S in PLS (upper) and ICA-MLR (lower) modeling.

It is important to invoke the magnitude of contribution of each descriptor to a model. Figure 6 shows values of $\mathbf{b}_{X\to y}$ corresponding to each descriptor in PLS and GAPLS modeling. The values of $\mathbf{b}_{X\to y}$ corresponding to only selected descriptors in GAPLS modeling are shown. Many descriptors are similar in the value of $\mathbf{b}_{X\to y}$, and the figure is complicated. Furthermore, it is dangerous to simply consider regression coefficients $\mathbf{b}_{X\to y}$ as important for each descriptor because there are correlations in **X** variables. The difference of the values of $\mathbf{b}_{X\to y}$ in the PLS and GAPLS methods does not mean there is any unity of the values of $\mathbf{b}_{X\to y}$. On the one hand, it is possible to consider $\mathbf{b}_{S\to y}$ as the magnitude of contribution to the model because independent components are mutually independent.

Figure 7 shows values of $\mathbf{b}_{S\to y}$ corresponding to each independent component. The larger the absolute value of $\mathbf{b}_{S\to y}$ is, the larger the magnitude of contribution of each component to logS. There are three components corresponding to $\mathbf{b}_{S\to y}$ whose absolute value is larger than 0.2, and their numbers are 52, 101, and 105. This shows that these components have more contribution to logS than the other components. Figure 8 shows the value of **W** corresponding to these components. **W** gives information on the magnitude of contribution of each descriptor to each independent component. Table 2 shows descriptors corresponding to **W** whose absolute values are larger than 1.0. This shows that these descriptors have a greater contribution to each independent component than the other descriptors. The value of $\mathbf{w}_{52}$ corresponding to SSS(12-1_═CH−) is larger than those of any other descriptors, as shown in Figure 8 and Table 2. This shows that SSS(12-1_═CH−) and the other descriptors are almost mutually independent. The value of $\mathbf{w}_{52}$ corresponding to SSS(12-1_═CH−) is positive, and that of $\mathbf{b}_{S\to y}$ corresponding to $\mathbf{s}_{52}$ is negative, as shown in Figure 7. This descriptor contributes to logS negatively, because a positive value times a negative value equals a negative value. We can perceive that a molecule is not soluble in water while the number of aromatic bonds in the molecule is large, because substructure ═CH− of SSS(12-1_═CH−) is included mostly in aromatic rings. The absolute values of $\mathbf{w}_{101}$ and $\mathbf{w}_{105}$ corresponding to WTPT3 are relatively large, as shown in Figure 8 and Table 2. WTPT3 is the descriptor that represents like reciprocal size of molecules. The value of $\mathbf{w}_{101}$ corresponding to WTPT3 and that of $\mathbf{b}_{S\to y}$ corresponding to $\mathbf{s}_{101}$ are positive, as shown in Table 2 and Figure 7. On the one hand, the value of $\mathbf{w}_{105}$ corresponding to WTPT3 and that of $\mathbf{b}_{S\to y}$ corresponding to $\mathbf{s}_{105}$ are negative.

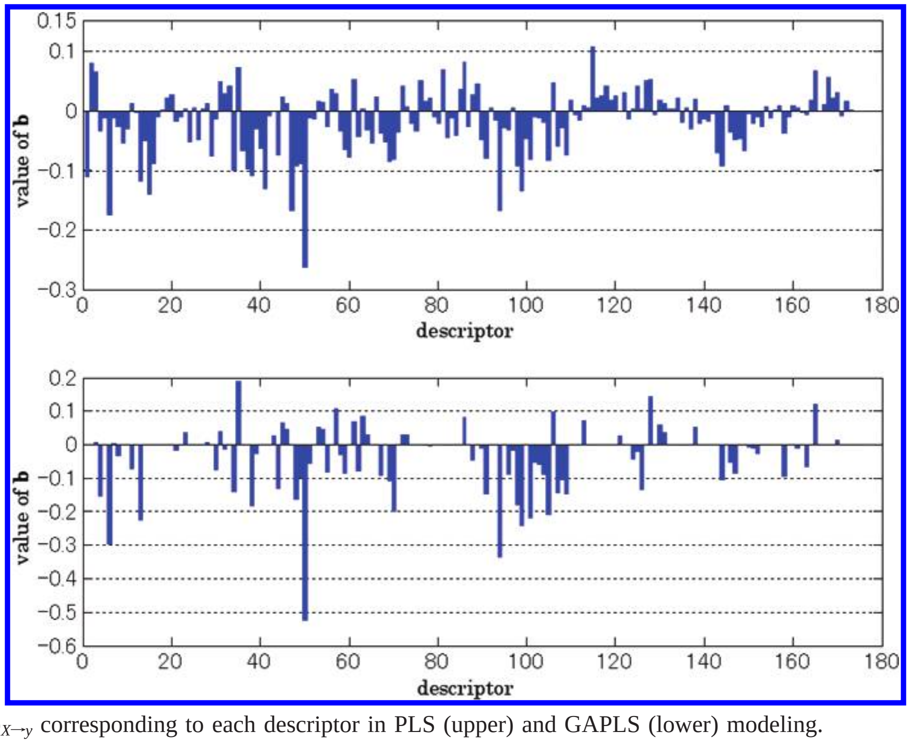


Figure 6. Value of $\mathbf{b}_{X\to y}$ corresponding to each descriptor in PLS (upper) and GAPLS (lower) modeling.

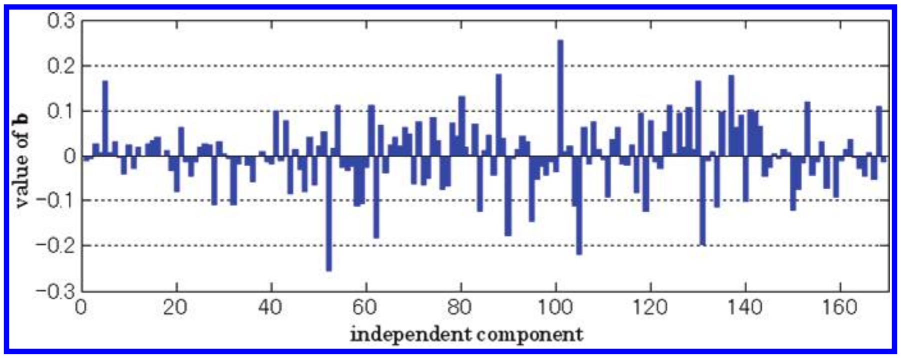


Figure 7. Value of $\mathbf{b}_{S\rightarrow y}$ corresponding to each independent component.

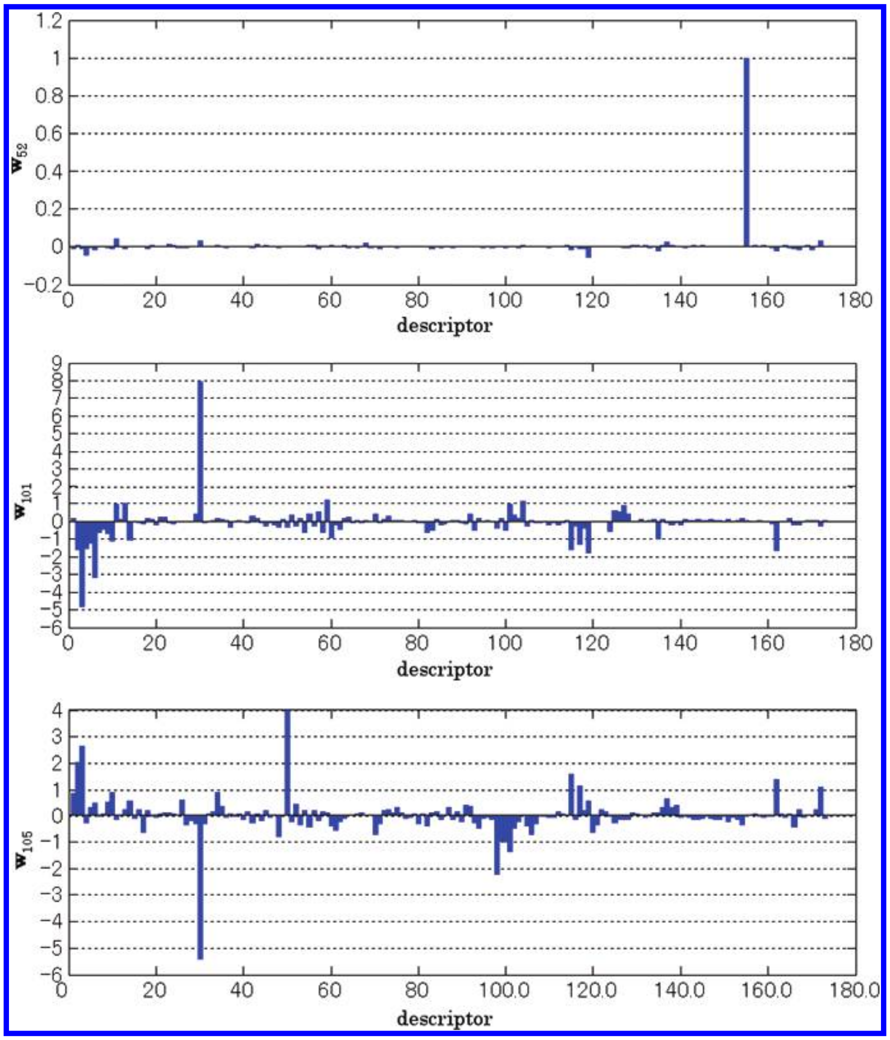


Figure 8. Value of **W** corresponding to independent components on that absolute value of $\mathbf{b}_{S\rightarrow y}$ is larger than 0.2.

This descriptor contributes to logS positively, because a positive value times a positive value equals a positive value and a negative value times a negative value equals a positive value. We can perceive that a molecule is soluble in water while the size is small. The value of $\mathbf{w}_{105}$ corresponding to logP shows that logP also contributes to logS significantly.

Analyzing independent components indicates the additional advantages. For example, Figure 9 shows a plot of $\mathbf{s}_{101}$ and $\mathbf{s}_{105}$, which are discussed in the previous paragraph. There seems to be some clusters in it. It is thought that information on molecules is included in $\mathbf{s}_{101}$ and $\mathbf{s}_{105}$, and we can obtain the information on molecules effectively by analyzing the clusters. In fact, values of logP of molecules located in the lower cluster in the figure are low. It is difficult to obtain such a specialty, even if we investigate the clusters formed with the components that are extracted by using PCA and PLS methods.

CONCLUSION

To construct a simple and highly predictive model, ICA was applied to a regression analysis as a pretreatment. By using the ICA method, independent components were extracted from explanatory variables and then used to construct a prediction model of an objective variable. We use the MLR method as a regression method. It is conceiv-

Table 2. Descriptors Corresponding to **W** Whose Absolute Values Are Larger than 1.0

abbrev.	descriptor name	$\mathbf{w}_{52}$	$\mathbf{w}_{101}$	$\mathbf{w}_{105}$
NO	number of oxygens		−1.6	2.0
NN	number of nitrogens		−4.8	2.6
NS	number of sulfurs		−1.5	
NF	number of fluorines		−1.2	
NCl	number of chlorines		−3.2	
NSB	number of single bond		−1.1	
NBR	number of basis rings		−1.0	
WTPT3	sum of path lengths starting from heteroatoms		7.9	−5.6
logP	calculation of log P			4.0
MDE34	molecular distance edge between all tert quat C		1.2	
2SP2	doubly bound carbon bound to two other carbons			−2.2
2SP3	singly bound carbon bound to two other carbons			−1.4
SSS(−C)	count of substructure		1.1	
SSS(≡O)	count of substructure		−1.6	1.6
SSS(−O−)	count of substructure		−1.3	1.1
SSS(−C(O)−)	count of substructure		−1.8	
SSS(12-1_≡CH−)	count of substructure	1.0		
SSS(7.1−SO2−)	count of substructure		1.6	1.4
SSS(43.1=O)	count of substructure			1.1

able that PLS is used as a method to construct a regression model instead of MLR, but we show that these are essentially the same.

Modeling a relationship between descriptors and logS by using ICA-MLR, PLS, and GAPLS methods showed that the ICA-MLR model has higher predictive power than those of PLS and GAPLS. The GAPLS method might select descriptors to overfit to a training set. The GAPLS model is simpler than that of PLS because the number of **X** variables decreases. However, it is not desirable because the information of eliminated variables is considered useful. On the one hand, by using the ICA-MLR method, the information of all variables can be included in a regression model.

The information of $\mathbf{b}_{S\rightarrow y}$ suggests that it is possible to indicate the magnitude of contribution of each descriptor in the analysis of aqueous solubility. It would be expected that a more simple model could be constructed by selecting independent components. Moreover, ICA-MLR could be applied to not only the analyses of aqueous solubility but also the analyses of other properties, and the independent components and the information of $\mathbf{b}_{S\rightarrow y}$ could contribute to these analyses.

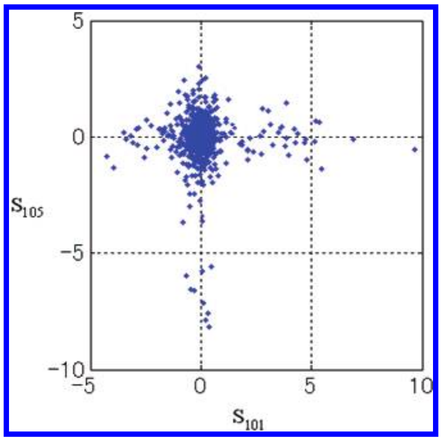


Figure 9. The relationship between two independent components.

APPENDIX A

Here, we show that ICA-MLR and ICA-PLS methods are essentially the same by representing regression coefficients with **X** and **y** by MLR, PLS, ICA-MLR, and ICA-PLS methods. ICA-PLS is a method that combines ICA and PLS. After extracting independent components from **X** variables by using the ICA method, a relationship between the components and the **y** variable is constructed by using the PLS method. It is assumed that **X** and **y** are mean-centered and scaled.

MLR. In MLR modeling, regression coefficient $\mathbf{b}_{\mathrm{MLR}}$ is calculated by the least-squares method as follows:

$$\mathbf{b}_{\mathrm{MLR}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{A.1}$$

PLS. A PLS model with one component is described as follows:

$$\begin{aligned} \mathbf{t}_1 &= \mathbf{X}\mathbf{w}_1 \\ \mathbf{y} &= \mathbf{t}_1\mathbf{q}_1 + \mathbf{f} \end{aligned} \tag{A.2}$$

where $\mathbf{t}_1$ is the score vector, $\mathbf{w}_1$ is the **X** weight vector, $\mathbf{q}_1$ is the **y**-loading vector, and **f** is the vector of **y** residuals. With the PLS algorithm, $\mathbf{w}_1$ is given as

$$\mathbf{w}_1 = \frac{\mathbf{X}^T\mathbf{y}}{||\mathbf{X}^T\mathbf{y}||} \tag{A.3}$$

Then, $\mathbf{q}_1$ is calculated by the least-squares method as follows:

$$\mathbf{q}_1 = \frac{\mathbf{y}^T\mathbf{t}_1}{||\mathbf{t}_1{}^T\mathbf{t}_1||} \tag{A.4}$$

In PLS modeling with one component, $\mathbf{y}_{\mathrm{calc}}$, which represents the calculated **y** value, is given as

$$\mathbf{y}_{\mathrm{calc}} = \mathbf{t}_1\mathbf{q}_1 \tag{A.5}$$

By using eqs A.2−A.5, $\mathbf{y}_{\mathrm{calc}}$ is represented with **X** and **y** as follows:

$$\mathbf{y}_{\mathrm{calc}} = \frac{\mathbf{X}\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y}}{||\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{y}||} \tag{A.6}$$

where $\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ and $\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ are scalars. Thus, $\mathbf{y}_{\mathrm{calc}}$ is as follows:

$$\begin{aligned} \mathbf{y}_{\mathrm{calc}} &= \mathbf{X}c\mathbf{X}^T\mathbf{y} \\ c &= \frac{\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y}}{\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{y}} \end{aligned} \tag{A.7}$$

Therefore, regression coefficient $\mathbf{b}_{\mathrm{PLS}}$ is as follows:

$$\mathbf{b}_{\mathrm{PLS}} = c\mathbf{X}^T\mathbf{y} \tag{A.8}$$

ICA-MLR. ICA-MLR is a method in which ICA is used as a pretreatment of MLR. **X** in eq A.1 is replaced by independent component matrix **S**.

$$\mathbf{b}_{\mathrm{ICA-MLR}} = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{y} \tag{A.9}$$

where $\mathbf{b}_{\mathrm{ICA-MLR}}$ is regression coefficients in ICA-MLR modeling. Every component in **S** is mutually independent, and each norm of the component is the same. Thus, $\mathbf{S}^{\mathrm{T}}\mathbf{S}$ is as follows:

$$\mathbf{S}^T\mathbf{S} = ||\mathrm{S}||^2\mathbf{I} \tag{A.10}$$

Therefore, regression coefficients $\mathbf{b}_{\mathrm{ICA-MLR}}$ are transformed as:

$$\mathbf{b}_{\mathrm{ICA-MLR}} = \frac{\mathbf{S}^T\mathbf{y}}{||\mathrm{S}||^2} \tag{A.11}$$

ICA-PLS. ICA-PLS is a method in which ICA is used as a pretreatment of PLS. **X** in eqs A.6 and A.7 is replaced by **S**.

$$\begin{aligned}\mathbf{b}_{\mathrm{ICA-PLS}} &= c\mathbf{S}^T\mathbf{y}\\ c &= \frac{\mathbf{y}^T\mathbf{S}\mathbf{S}^T\mathbf{y}}{\mathbf{y}^T\mathbf{S}\mathbf{S}^T\mathbf{S}\mathbf{S}^T\mathbf{y}}\end{aligned} \tag{A.12}$$

By using eq A.10, c is transformed as

$$\begin{aligned}c &= \frac{\mathbf{y}^T\mathbf{S}\mathbf{S}^T\mathbf{y}}{\mathbf{y}^T\mathbf{S}||\mathrm{S}||^2\mathbf{I}\mathbf{S}^T\mathbf{y}}\\ &= \frac{1}{||\mathrm{S}||^2} \times \frac{\mathbf{y}^T\mathbf{S}\mathbf{S}^T\mathbf{y}}{\mathbf{y}^T\mathbf{S}\mathbf{S}^T\mathbf{y}}\end{aligned} \tag{A.13}$$

Therefore, regression coefficients $\mathbf{b}_{\mathrm{ICA-PLS}}$ are as follows:

$$= \frac{1}{||\mathrm{S}||^2}$$

$$\mathbf{b}_{\mathrm{ICA-PLS}} = \frac{\mathbf{S}^T\mathbf{y}}{||\mathrm{S}||^2} \tag{A.14}$$

This shows that $\mathbf{b}_{\mathrm{ICA-MLR}}$ and $\mathbf{b}_{\mathrm{ICA-PLS}}$ are equal.

REFERENCES AND NOTES

(1) Gasteiger, J.; Engel, T. *Chemoinformatics-A Textbook*; Wiley-VCH: Weinheim, Germany, 2003.

(2) Wold, S. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37−52.

(3) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(4) Comon, P. Independent component analysis, A new concept? *Signal Process.* **1994**, *36*, 287−314.

(5) Chen, J.; Wang, X. Z. A New Approach to Near-Infrared Spectral Data Analysis Using Independent Component Analysis. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 992−1001.

(6) Shao, X.; Wang, W.; Hou, Z.; Cai, W. A new regression method based on independent component analysis. *Talanta* **2006**, *69*, 676−680.

(7) Kano, M.; Tanaka, S.; Hasebe, S.; Hashimoto, I. Monitoring Independent Components for Fault Detection. *AIChE J.* **2003**, *49*, 294−298.

(8) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266−275.

(9) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E. Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 643−651.

(10) Sun, H. A Universal Molecular Descriptor System for Prediction of LogP, LogS, LogBB, and Absorption. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748−757.

(11) Wegner, J. K.; Fröhlich, H.; Zell, A. Feature Selection for Descriptor Based Classification Models. 1. Theory and GA-SEC Algorithm. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 921−930.

(12) Bergström, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1477−1488.

(13) Clark, M. Generalized Fragment-Substructure Based Property Prediction Method. *J. Chem. Inf. Model.* **2005**, *45*, 30−38.

(14) Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386−393.

(15) Hasegawa, K.; Miyashita, Y.; Funatsu, K. GA Strategy for Variable Selection in QSAR Studies: GA-Based PLS Analysis of Calcium Channel Antagonists. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 306−310.

(16) Whitley, D. A Genetic Algorithm Tutorial. *Stat. Comput.* **1994**, *4*, 65−85.

(17) Hyvärinen, A.; Karhunen, J.; Oja, E. A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Comput.* **1997**, *9*, 1483−1492.

(18) *ADMEWORKS/ModelBuilder*, version 2.1; Fujitsu Kyushu System Engineering Limited: Fukuoka, Japan.

(19) Houck, C. R.; Joines, J. A.; Kay, M. G. *A Genetic Algorithm for Function Optimization: A Matlab Implementaion;* NCSU-IE TR 95-09; Meta-heuristic Research and Applications Group: North Carolina State University: Raleigh, NC, 1995.

CI700245F