Fuzzy Linear Regression Method in Ultratrace Analysis and Analytical Interpretation

Takuya SHIMOSAKA, Takehiko KITAMORI, Akira HARATA and Tsuguo SAWADA†

Department of Industrial Chemistry, Faculty of Engineering, the University of Tokyo, Hongo, Tokyo 113, Japan

Fuzzy sets theory was introduced for quantitative treatment of measured values in ultratrace analysis. The data obtained in extremely low concentration have uncertainties such as errors of measurement or errors arising from chemical handling of the samples. In place of the least squares method, a fuzzy linear regression method was applied to measured values expressed by the fuzzy numbers, and instead of a calibration curve, a calibration area was obtained. The validity and propriety of the calibration area were shown from simulation and experimental data of ultratrace analyses done using laser induced photothermal spectrometry. Analytical meanings of the shape of the calibration area were discussed.

Keywords Fuzzy, regression, calibration curve, error, confidence interval

Uncertainties in experimental data sometimes become a serious problem when very sophisticated analytical methods are employed. In ultra-sensitive and ultratrace analyses, for example, unexpected error factors can invade every step of the analytical procedures. The uncertainties are derived from various sources, such as foreign noises, instrumental uncertainties and chemical handling procedures. Furthermore, they sometimes hinder data processing, for example, estimating concentration or recovering a spectrum.

In analytical chemistry, statistical methods based on normal distribution have been used to deal with measured values and their variances. Calibration curves, for example, are obtained by the least squares method. Confidence limits are often defined as two or three times the standard deviation of residuals, though they are exactly defined by using a t-distribution. When a sample concentration is estimated by a calibration curve, the true value is between the lower and higher confidence limits, and the probability that the true value is in the interval is 95% if twice the standard deviation, 99% if threefold. The lower limit of determination (LOD) is given on the basis of confidence limits and the limit is widely accepted and used.

Recently, however, laser spectrometries have allowed LODs three orders lower than conventional methods, and uncertainties in measured values have become more varied and complicated.

In laser fluorescence methods or photothermal methods, sensitivity is much higher than conventional fluorometry or spectrophotometry employing an incoherent light source, and then signal reproducibility is more seriously influenced by various factors, such as instability of instruments, slight differences of composition between lots of solvents, differences in residual amounts of reagent in extraction solvents, contamination of reaction systems by foreign materials, obstruction by admixtures adsorption of analytes on a glass instrument, and so on. They are all negligibly small when determinations are made at ordinary concentrations. Consequently the values measured by such highly sensitive methods, including uncertainties, do not satisfy the suppositions of conventional statistical methods, or if they satisfy the suppositions, complicated models and calculations are required to deal with these uncertainties. Therefore, another theory and method are needed.

In this study, these uncertainties of data at an ultra-low concentration may be able to be treated more easily and flexibly by fuzzy sets theory than by ordinary statistical methods. Fuzzy sets theory can quantitatively deal with uncertainty in experimental data or ambiguity in human perception, and so it has been applied to various fields in which uncertainty and/or ambiguity have a serious influence. The theory does not need strict assumptions of probability functions as in the statistical methods, such as the normal distribution described above, and it can deal with the uncertainty more easily and more flexibly.

In analytical chemistry, the theory has been mainly used for reasoning and judging, for example, peak tracking for high performance liquid chromatography1,2 and multicomponent spectral analysis by infrared spectroscopy.3 Application of fuzzy sets theory to analytical chemistry is reviewed in refs. 4 and 5. However, in this paper, fuzzy sets theory was introduced to deal with uncertainties in experimental data quantitatively and to estimate quantities such as concentration. We tried to

† To whom correspondence should be addressed.
derive interpretations of the obtained results from the viewpoints of analytical chemistry. Variances and errors caused by various factors in experimental data were quantitatively treated as an ambiguity in fuzzy sets theory. These data were expressed by fuzzy numbers, which are often used to express such ideas as "about". By using fuzzy numbers, the representative value and a dispersion level of experimental data could be simultaneously expressed. For estimation of, for example, a concentration from signal intensity, fuzzy linear regression (FLR) in place of the least squares method was introduced to process the data expressed by fuzzy numbers.

First, the basic idea of fuzzy treatment of experimental data with uncertainties is overviewed and introduced to analytical data. Then, it is applied to calibration curves of determination colorimetry. A new idea for calibration curves called an "estimation area" is derived. Finally, the propriety of this idea is confirmed by simulation data and the estimation area is applied to experimental data of ultratrace analyses such as photoacoustic spectroscopy (PAS) and photothermal immunoassay analysis (PIA). Information involved in estimation area is interpreted from the viewpoints of analytical chemistry.

Theoretical and Experimental

In this paper, a fuzzy number is used to simultaneously express experimental data and their uncertainties according to a general theory of fuzzy mathematics, and FLR is introduced to process data expressed by the fuzzy numbers. In this section, the necessary bases of fuzzy sets theory are briefly presented, and then a fuzzy treatment method for experimental data is given. For more detailed theoretical comprehension, the readers should refer to elsewhere and to references presented in ref. 5.

Conventionally, expression of uncertainty included in experimental data is done by an error bar. The bar implies that the true value is on it with some probability. But, is the reliability of the estimation value the same all over the error bar? The answer to that question is "no", reliability is the highest at a representative value, and monotonously decreases on moving away from that value. Figure 1(a) represents an example of the reliability of experimental data when samples of the same concentration are repetitively measured. The average value of the measured values is \( a \). The abscissa shows signal intensity, and the reliability is numerically expressed as a function of it. The relative reliability is defined to be 1 at the average value \( a \), and decreases on moving away from it. The reliability curve, expressed as a function of the signal intensity, can be called a membership function in fuzzy sets theory. In our use of the theory, experimental data including uncertainties are expressed by a fuzzy number which is expressed by a convex membership function.

\[
\{ y \mid L(y) \geq h \},
\]

Figure 1(b) shows an \( h \)-level set; it is defined as follows:

where \( L(y) \) is a membership function. Elements of the sets are signal intensities of \( y \) whose membership function values are higher than \( h \). In the case that the membership function represents reliability for the experimental data, an \( h \)-level set means the set of measured values whose reliability is higher than \( h \), that is, the thick line expressed in Fig. 1(b). Hence the parameter \( h \) means the degree of reliability. The detailed meaning will be discussed later.

A fuzzy number is used for expression of "about 2" in the theory, and then it is not an ordinal single numerical number, namely a crisp number, but a set itself which involves uncertainty and/or ambiguity. In analytical language, a fuzzy number is a datum and its accom-
panying uncertainty expressed by a membership function. Therefore, operations for a crisp number cannot be used for a fuzzy number, and the extended operations for fuzzy sets theory are required for fuzzy numbers.\(^8\)

The triangles in Fig. 2(a) represent colorimetric experimental data expressed by fuzzy numbers in three dimensions: concentration, absorbance and reliability. The membership function is assumed to be an equilateral triangle function for simple discussion. As the ordinary least squares method cannot process fuzzy number data, FLR based on fuzzy sets theory is introduced to calculate a calibration curve for the data. Figure 2(b) represents experimental data (dashed line triangles) and estimated values (solid line triangles) expressed by fuzzy numbers. The triangles drawn by solid lines are estimation values of absorbance which is also expressed by fuzzy numbers.

FLR based on fuzzy sets theory, is expressed as

\[ Y^* = A_0 \oplus A_1 x, \]

where \( A_0 \) and \( A_1 \) are fuzzy numbers, \( Y^* \) is an estimated value (fuzzy number), \( x \) is a concentration which is a crisp number, and \( \oplus \) represents fuzzy addition. A measured absorbance is expressed by a fuzzy number \( Y \) whose mean value and spread are \( y \) and \( e \), respectively. The estimated value of absorbance is also expressed by a fuzzy number \( Y^* \), whose mean value and spread are \( y^* \) and \( e^* \), respectively. As \( A_0 \) and \( A_1 x \) are fuzzy numbers, \( A_0 \oplus A_1 x \) is also a fuzzy number.

The fuzzy number coefficients \( A_0 \) and \( A_1 \) are obtained by the following two restrictions represented by Eq. (3) and Eqs. (4)-(6).

\[
\min J(d^*) = \min \sum_{i} (d_{ai}^* + d_{oi}^* |x_i|),
\]

(3)

\[
Y_i \subseteq h_0 Y^* = A_0 \oplus A_1 x,
\]

(4)

\[
[Y]_{h_0} \subseteq Y^* = \left[y, y^* + e^* \cdot L^{-1}(h_0)\right],
\]

(5)

where \( d_{ai}^* \) and \( d_{oi}^* \) are spreads of fuzzy coefficients \( A_0 \) and \( A_1 \), respectively, \([Y]_{h_0}\) means an \( h_0 \)-level set of \( Y \), \( \subseteq \) and \( \subseteq \) represent the inclusion relation of \( h_0 \)-level set and of ordinary set theory, and \( L^{-1}(h) \) is an inverse function of a membership function \( L(x) \) of fuzzy number. Restrictions (3)-(6) give linear programming\(^8\), and \( A_0 \) and \( A_1 \) are obtained by solving this programming.

The spreads of the estimated values \( (d_{ai}^* + d_{oi}^* |x_i|) \) correspond to residuals in the least squares method. Therefore Eq. (3) represents the same restrictions as the minimized residuals in the least squares method. Equation (4) means that the line segment at \( h = h_0 \) in the dashed line triangle, which represents a certain reliability to be used in the estimation, is involved in the estimated fuzzy number value (solid line triangle) in Fig. 2(b). As Eq. (3) is an objective function and Eq. (4) contains restrictions for linear programming, two fuzzy number parameters, \( A_0 \) and \( A_1 \) are determined by solving it, and then the estimated fuzzy absorbance \( Y^* \) is determined.

The \( h_1 \)-level sets of \( Y^* \) give a region, as shown in Fig. 2(b), which is indicated as an area. In this paper, we called this area an "estimation area". It is defined as follows,

\[
\{(x, y) | y^* - e^* \cdot L^{-1}(h_1) \leq y \leq y^* + e^* \cdot L^{-1}(h_1)\}.
\]

Calculation and simulation

A program for FLR was written in C language and executed on a PC-9800 series personal computer (NEC, Japan) equipped with a numeric coprocessor.

At first, the propriety of the FLR was examined using two simulation data sets of the most common analytical measurement and colorimetry in very low concentration. Then, the developed method was applied to measured
data for ultratrace analyses by photoacoustic spectroscopy and photothermal immunoassay.

The dispersion of the simulation data was given by normally distributed random numbers generated by the Box-Muller method. For the most common analytical data, the relative standard deviation (RSD) value is constant; the deviations of the measured data are therefore proportional to the average values. This is because the systematic factor of variation is constant, so its ratio is constant.

**Ultratrace determination**

In the experiment which provided the data, an ultratrace amount of cobalt, at the ppt level, in water also containing the foreign metals, iron, nickel, chromium and copper, was extracted into m-xylene and the optical absorbance was measured by the photoacoustic method. Another data set was obtained by a photothermal beam deflection method for a small amount of carcinoembryonic antigen (CEA) which is a cancer marker antigen. Anti-CEA antibodies, labeled by colloidal gold ultrafine particles, formed an antigen-antibody bonding with CEA which was adsorbed on polystyrene latex microparticles (0.9 µm in size) in a sample serum. The polystyrene microparticles were collected on a nitrocellulose membrane, and the quantity of the gold ultrafine particles was determined by the photothermal beam deflection method. Their concentrations were from 80 ng/ml to 720 ng/ml. The details of the method and the experimental conditions have been reported elsewhere.

**Results and Discussion**

When experimental data are expressed by symmetrical fuzzy numbers, their spread is determined by standard deviations which are statistically given by a large number of data. On the other hand, spreads can be given by a priori information which an expert has, even if data numbers are small. The shape function for the fuzzy numbers should be determined by the error distribution of the experimental data. It is not easy to determine the shape function exactly because determination of the profile requires many data. Therefore, it can be determined by a priori information and expert experiences.

First of all, we applied our method to the most common analytical data, that is, a linear calibration line having constant RSD values. When RSD values of data is constant, its standard deviation is proportional to the averaged values, and such data is made by simulation as mentioned above.

As shown in Fig. 2(b), the \( h_i \)-level sets of \( Y^* \) give an area which is given by Eq. (7). We named this area an "estimation area". The area corresponds to the calibration line of the conventional least squares method for crisp number data. The width of the estimation area means the uncertainty at each concentration. The data at each concentration were expressed by fuzzy numbers, whose spread was determined by a standard deviation of random numbers. The results of FLR for the simulation data are shown in Fig. 3. An estimation area was also wider at a high concentration and narrow at a low concentration, and therefore the area well agreed with the character of the simulation data. Confidence intervals of the ordinal least squares method have almost constant widths, as discussed later.

In the conventional theory of analytical chemistry, RSD value is constant at each concentration, and therefore, the deviation of the data at one concentration is wider for a higher concentration. The variation due to the chemical procedure and measurement is well assumed to be in proportion to the concentration. In such a region, unanticipated factors in the ultratrace region can be neglected. A case identical to the conventional region can also be treated by FLR, and it will be discussed along with the practical data afterward, with Fig. 7.

Figure 4(a) shows the result of FLR for the simulation data when a larger standard deviation was given to the data at a lower concentration, because unanticipated factors, such as adsorption on glass equipment, disturb a quantitative reaction more seriously for tracer analyte. Hence, quantitative chemical reactions, e.g. masking, chelate formation and extraction, in the ultratrace region with enormous quantities of foreign elements are always difficult. The results of Fig. 4(a) successfully represent this experience. In the high concentration region, the width of the estimation area is narrow (uncertainty is low), while in the low concentration region the width is wide (uncertainty is high).

Then we could estimate a concentration using this estimation area. In Fig. 4(a), the estimated value \( x_a \) for the measured absorbance data \( y_a \) is accompanied by a wider region than the estimated value \( x_b \) for the higher concentration.
absorbance data $y_b$. This agrees with the change of uncertainty of simulation data along the concentration axis. Figure 4(b) represents a conventional calibration curve and its confidence interval for the simulation data. The estimated concentration of the ordinary least squares method for the simulation data had the same confidence interval all over the concentration region. From the simulation results, it can be understood that the estimation area can well express the degree of dispersion at each concentration more practically than the conventional method can, and therefore the area not only represents the relation between concentration and absorbance but also has information on reliability, while the conventional calibration curve represents only the relation between concentration and absorbance.

Estimation areas with parameter $h_0$ as 0.7 and 0.1 are represented by solid and dashed lines respectively, in Fig. 5. The width of the estimation area for the lower $h_0$-level ($h_0=0.1$) is narrower than that for the higher one. This means that the width of the estimation area becomes wider when a higher reliability is taken for estimation. This is the same as with the conventional statistical method. For example, when the confidence interval is taken to be $2\sigma$ ($\sigma$: standard deviation), the reliability of the estimation value is 95.5%, and for $3\sigma$, it is 99.7%. Therefore, the $h_0$-level corresponds to the reliability of the conventional statistical method and the parameter $h_0$ is decided according to required reliability.

The change of width of estimation area along the abscissa corresponds well with standard deviations of the simulation data, while the conventional calibration curve for the least squares method does not express this characteristics of the data. Therefore, the validity and propriety of the estimation area are demonstrated. The proposed method was applied to ultratrace analyses. In ultratrace analyses, errors are the dominant factor for determining reliability of the estimated value obtained by a calibration curve. Experimental data measured by photoacoustic spectroscopy (PAS) and the photothermal immunoassay method (PIA) were used here.

The estimation area obtained for photoacoustic determination data of ultratrace cobalt is represented in Fig. 6. Its parameter $h_0$ is equal to zero. The width of the estimation area decreases a little as concentration increases. The width for the blank is about 1.6 times wider than that for the sample at 20 ng. The uncertainty in experimental data is caused by variation in the chemical reaction procedures such as extraction efficiency and instrument instability when making highly sensitive measurements, etc. Then the estimation area is considered to express the uncertainties of the analytical
procedures and the measurements in ultratrace analysis. The shape of the estimation area corresponds with the one shown in Fig. 4(a). This fact suggests that some unanticipated factor, as mentioned before, is operating in the analytical procedure of the determination. The adsorption on the glass wall might be suspicious. Then a lower LOD would be obtained if the factor can be identified and eliminated.

The estimation area for PIA whose $h_0$ equals to zero is shown in Fig. 7. Samples whose concentration ranged from blank to 720 ng/ml were plotted on a paper, and the six papers were measured. The sensitivity and offset of the signal intensities were different for each paper. Therefore, the data shown in Fig. 7 were corrected for the sensitivity and offset.

In an immunoassay method which uses biological affinity, dispersion of antigen-antibody reaction rate is dominated by pH, temperature, or a stirring method, and their influence is greater at a higher concentration. This is the reason why the standard deviation of data at a higher concentration is larger than that at a lower concentration. This trend is just the same as the conventional theory, that is, the RSD value is constant. Actually, the RSD value of the raw data was 15% and was almost constant for each concentration. Therefore, in this case, these systematic factors made the analytical data varied. Furthermore, factors which disturb quantitation are based on an interfering reaction. Such a reaction is usually a stoichiometric reaction of the analyte, and the degree of the disturbance is in proportion to the concentration. Hence, the width of the estimation area increases as the concentration increases, and the change in width along the concentration axis quantitatively represents the degree of the influence. In this case, unlike the former one, the result indicates that more precise control of the immunoreaction conditions would make LOD lower. Thus our method proved to be able to treat both cases.

Using the simulation and experimental data, the validity of the estimation area was shown. The proposed method could process the change of the dispersion in the data of the determinations, while the ordinary least squares method could not treat it sufficiently. Hence, this method is considered to be more suitable for processing data having a different dispersion at each concentration. In particular, such an error treatment is important for data analysis at very low concentrations and with very small absolute amounts of samples. Furthermore, the calculation time to obtain an estimation area is less than a few seconds according to the personal computer used, and the proposed method is suitable for laboratory use.

In our procedure for calculating the estimation area, one problem about fuzzy mathematics still remains. There are some cases in which the uncertainty of the estimation area, namely the width, decreases as concentration increases, as shown in Figs. 4(a), 5 and 6. When an estimation area is given by the usual FLR, the width of the estimation area at high concentration must be wider than that at low concentration if $d_l$ is equal or more than zero, because fuzzy addition and fuzzy submission make the spread wider. To avoid this problem, we allow $d_l$ to be not only positive, but also negative. The fuzzy number whose spread is negative is not defined in fuzzy sets theory, but such a negative spread is convenient to calculate the estimation area; a negative value of $d_l$ must be taken to express the estimation area whose width is narrower at a higher concentration. Even in this case, the result obtained was consistent with that of the usual FLR theory, because the result of the estimation, namely $y^*$, had a positive width. The problem of the negative width has been
discussed as "ambiguity explosion" or an inverse operation for fuzzy addition in fuzzy sets theory, and introducing a fuzzy number whose width is negative is one way to avoid an "ambiguity explosion".

The presented method quantitatively represents uncertainty of experimental data, and by extending FLR to non-linear curves, the method can be applied to non-linear calibration curves. Coupling the extended FLR with other data processing methods, such as spectrum recovering and treating the inverse integral problem encountered in three-dimensional determination by PAS and the attenuated total reflection method, the data processing methods could flexibly treat individual errors included in experimental data.

References


(Received December 18, 1995)
(Received March 25, 1996)