Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE 2008
13–17 June 2008.

Adjust of Contradictory Information by a Fuzzy Method

Enrique Mérida-Casermeiro[1] and Domingo López-Rodríguez[1]

[1] Department of Applied Mathematics, University of Málaga

emails: merida@ctima.uma.es, dlopez@ctima.uma.es

Abstract

Many real-world engineering problems are expressed by imprecise terms, imprecision and errors are always present when measures are been taken in order to express the behavior of a concrete model. Moreover, analyzing such systems often involve the use of fuzzy information and so obtaining a specific set of consistent measures adjusting to the observations and verifying the model are necessary.

Key words: inconsistency, contradictory information

1 Introduction

Information obtained from real world is always imprecise. This is due to several possible reasons: tolerances in measurement systems, systematic alterations due to meteorological conditions (temperature, pressure, etc.), different calibrations of the measurement systems. Thus, there are always errors in data due to measurement accuracy, the method to obtain data, data coding, or even due to the intrinsic fuzzy nature of the information received.

Often, these inaccuracies produce values which are inconsistent with the underlying theory, and some of the measurements must be corrected in order to get values, differing as little as possible from obtained data, which agree with the theory.

A Bayesian procedure for data adjustment was proposed by Nir [2] in 1980. That method dealt with the treatment of uncertainties and discrepancies among different sets of data measuring the same variables, and was based on the concept of information entropy.

An linear programming procedure to tackle this problem was presented by Kikuchi [1]. In that work, his method was applied to adjust observed transportation data, concretely to adjust passenger counts on a transit line (bus, for example).

In this work, we extend Kikuchi's model and develop a general framework to incorporate fuzziness to observed data. Our method is able to generate more accurate estimations of the underlying model than the well-known Least Squares criterion.

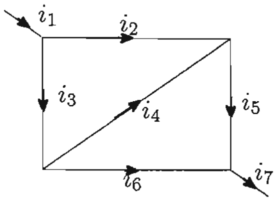

Figure 1: Circuit for Example 1.

2 Description of the Problem

In order to motivate the problem of adjusting inconsistent and contradictory information in some processes and systems, let us consider systems of linear equations coming from flux (electrical, hydraulic, traffic...) problems, where flux conservation law must hold. Real data must verify every equations. However, observed data may be inconsistent and should be debugged. Let us see the method with an example:

Example 1 Consider the electrical circuit given in Fig. 1. Flux conservation equations produce the following system of linear equations:

$$(S):\left\{\begin{array}{rcl} i_1 & = & i_2+i_3 \\ i_3 & = & i_4+i_6 \\ i_5 & = & i_2+i_4 \\ i_7 & = & i_6+i_5 \end{array}\right.$$

This system of equations constitutes, by now, an undetermined system and every real flux $\vec{i}$ circulating on the circuit must verify those equations.

Let us now suppose that we want to determine the accurate flux on the circuit. To this end, some measurements are taken in every segment ($\vec{i}^{\,obs}$), obtaining the following values: $i_1^{obs}=7$, $i_2^{obs}=4.7$, $i_3^{obs}=2.4$, $i_4^{obs}=2$, $i_5^{obs}=6.4$, $i_6^{obs}=0.5$ and $i_7^{obs}=6.8$. By a simple substitution of these values in the system (S), we can observe that the set of observed data is inconsistent. For example, the first equation yields: $i_1^{obs}=7\neq 7.1=i_2^{obs}+i_3^{obs}$.

We would like to adjust these values, obtaining a vector of adjusted and consistent measurements ($\vec{i}^{\,adj}$), such that it differs as less as possible from observed data, that is, minimizing the norm of the difference vector: $\|\vec{i}^{\,obs}-\vec{i}^{\,adj}\|$. This adjustment has been traditionally made by hand.

To obtain the minimum squared error $\|\vec{i}^{\,obs}-\vec{i}^{\,adj}\|^2$, as presented here, is straightforward. Since rank of system (S) is 4, we can express it in terms of 3 independent variables, for example $\{i_1, i_2, i_4\}$, obtaining the system :

$$(S') : \begin{cases} i_3 & = & i_1 - i_2 \\ i_5 & = & i_2 + i_4 \\ i_6 & = & i_1 - i_2 - i_4 \\ i_7 & = & i_1 \end{cases}$$

The adjusted solution $\vec{i}^{\,adj}$ is obtained by solving, in the minimum squared error sense, the new system (S^*), formed by substituting in (S') the value of the variables for its observed value: $i_1 = 7$, $i_2 = 4.7$, $i_3 = i_1 - i_2 = 2.4$, $i_4 = 2$, $i_5 = i_2 + i_4 = 6.4$, $i_6 = i_1 - i_2 - i_4 = 0.5$, $i_7 = i_1 = 6.8$.

$$\begin{pmatrix} i_1 \\ i_2 \\ i_3 \\ i_4 \\ i_5 \\ i_6 \\ i_7 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} i_1 \\ i_2 \\ i_4 \end{pmatrix} = \begin{pmatrix} 7 \\ 4.7 \\ 2.4 \\ 2 \\ 6.4 \\ 0.5 \\ 6.8 \end{pmatrix}$$

This system $Ax = \vec{b}$ is undetermined and one solution (may not be unique) minimizing the squared error $||Ax - \vec{b}||^2$ can be obtained by the QR orthogonalization method, or by the use of the pseudoinverse matrix (also called generalized inverse matrix): $\vec{i}^{\,adj} = \mathrm{pinv}(A) \cdot b$. Note that many numerical libraries, such as MatLab, have implemented the QR decomposition method to solve undetermined and overdetermined linear systems.

In this example, the solution minimizing the 2-norm of the residuum vector $||\vec{r}||_2 = ||A\vec{i}^{\,adj} - \vec{b}||_2$ is: $\vec{i}_1^{\,adj} = \frac{104}{15} \approx 6.9333$, $\vec{i}_2^{\,adj} = \frac{137}{30} \approx 4.5667$, $\vec{i}_4^{\,adj} = \frac{19}{10} = 1.9$. From this independent variables, we can deduce the other variables $\vec{i}_3^{\,adj} = \frac{104}{15} - \frac{137}{30} = \frac{71}{30} \approx 2.3667$, $\vec{i}_5^{\,adj} = \frac{137}{30} + 1.9 = \frac{97}{15} \approx 6.4667$, $\vec{i}_6^{\,adj} = \frac{104}{15} - \frac{137}{30} - 1.9 = \frac{7}{15} \approx 0.4667$, $\vec{i}_7^{\,adj} = \frac{104}{15} \approx 6.9333$. For these adjusted data, we obtain: $||\vec{r}||_2 \approx 0.238$.

The previous method can be easily generalized to the case of using a squared norm weighted by a vector $\vec{w}$, that is, minimizing $\sum_i w_i r_i^2 = \sum_i w_i (A\vec{i}^{\,\mathrm{adj}} - \vec{b})_i^2$, just by multiplying every row of the system of equations $A\vec{i}^{\,\mathrm{adj}} - \vec{b} = 0$ by $\sqrt{w_i}$. This way, the new residuum vector will be $r_i' = \sqrt{w_i} r_i$. If the new system is solved by the QR method, the solution will minimize $||\vec{r'}||_2 = \sqrt{\sum_i w_i r_i^2}$, as wanted.

3 Incorporating Fuzziness to Inconsistent Data

In this work, we develop a new model that allows to adjust multiple observations or informations, maybe inconsistent, about a variable.

3.1 Codification of Observed Data

Our proposal consists in considering that both observed and adjusted data come from a fuzzy set, with a triangular membership function characterized by 3 numbers a, b, c

$(a < b < c)$:

$$h^{\mathrm{Tr}}_{(a,b,c)}(x) = \begin{cases} \frac{x-a}{b-a} & x \in (a,b] \\ \frac{c-x}{c-b} & x \in [b,c) \\ 0 & (x \le a) \cup (x \ge c) \end{cases}$$

If m-th observation comes from variable x_i and takes the value x_i^{obs}, any estimate x of the actual value of x_i will have a triangular membership function $h(x) = \max\{0, 1 - k_m|x - x_i^{\mathrm{obs}}|\}$ with $k_m \le 0$. The parameter k_m indicates the reliability of the given measurement, $k_m = 0$ meaning that x_i^{obs} has no errors (in fact, it will represent a crisp number). The coding designed to represent this situation is: $b = x_i^{\mathrm{obs}}$, $a = x_i^{\mathrm{obs}} - k_m$, $c = x_i^{\mathrm{obs}} + k_m$ in $h^{Tr}_{(a,b,c,d)}(x)$.

Missing data: There is no need to have information about every variable of the system. Although there is no information about a given variable, our proposal is able to proportionate a valid value provided the existing relationships and the rest of data.

3.2 Formulation of Our Method

Let us consider a system described by the set of equations:

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0 \\ \vdots \\ f_n(x_1, \ldots, x_n) = 0 \end{cases}$$

Suppose that variable x_i can be obtained from i-th equation as $x_i = \varphi_i(\vec{x})$, where $\vec{x} = (x_1, \ldots, x_n)$. Then, we can re-write the system above as:

$$\vec{x} = \varphi(\vec{x}) \tag{1}$$

where $\varphi = (\varphi_1, \ldots, \varphi_n)$. Thus, a consistent observation or measurement of the variables x_i will be formed by a fixed point of the function φ. Let us denote by C the set of all possible fixed points of φ, $C = \{\vec{x} \in \mathbb{R}^n : \vec{x} = \varphi(\vec{x})\}$.

In most practical situations, the observed data $\vec{x}^{\,\mathrm{obs}}$ does not belong to C, that is, $\vec{x}^{\,\mathrm{obs}} = \varphi(\vec{x}^{\,\mathrm{obs}})$ does not hold.

Let us consider that M observations are taken in the system, that is M measurements of the variables are used. These observed data are $\{x^{\mathrm{obs}}_{m_1}, \ldots, x^{\mathrm{obs}}_{m_M}\}$, where $\{m_1, \ldots, m_M\} \subset \{1, \ldots, n\}$. Note that there may be more than one observation for some variables and missing data.

Let us assume that each observation $x^{\mathrm{obs}}_{m_k}$ comes from a fuzzy set whose membership function is h^{Tr}_k for given parameters a_k, b_k, c_k, as explained in the previous section.

Our aim is to find a new set of adjusted consistent measurements, $\vec{x}^{\,\mathrm{adj}} \in C$, close to the observed data in some sense.

Since $h^{\mathrm{Tr}}_k(x^{\mathrm{adj}}_{m_k})$ measures the closeness of the adjusted data to the observed ones, we have to study the multi-objective optimization problem of maximizing $\vec{h}(\vec{x}^{\,\mathrm{adj}}) = (h^{\mathrm{Tr}}_1(x^{\mathrm{adj}}_{m_1}), \ldots, h^{\mathrm{Tr}}_M(x^{\mathrm{adj}}_{m_M}))$ subject to $\vec{x}^{\,\mathrm{adj}} \in C$.

Several criteria can be adopted to obtain an optimum solution for this problem.

- Maxisum : The optimal adjusted variables correspond to the solution of the problem:

$$\max_{\vec{x}\in C} \sum_{j=1}^{M} h_j^{\mathrm{Tr}}(x_{m_j}) \tag{2}$$

It must be noted that this formulation is equivalent to minimize $\|\vec{h}(\vec{x})\|_1$ subject to $\vec{x} \in C$.

- Maximin: The optimal $\vec{x}^{\,\mathrm{adj}}$ is the solution to the problem:

$$\max_{\vec{x}\in C} \min_{j=1,\dots,M} h_j(x_{m_j}) \tag{3}$$

Since every membership function is bounded by 1, this is equivalent to maximize $\|1 - \vec{h}(\vec{x})\|_{\infty}$, subject to $\vec{x} \in C$.

- 2-norm : The solution considered is the result of maximizing the quantity $\|\vec{h}(\vec{x})\|_2 = \sqrt{\sum_{j=1}^{M}(h_j^{\mathrm{Tr}}(x_{m_j}))^2}$, subject to $\vec{x} \in C$.

Evidently, a lineal combination of these criteria can be used to obtain valid solutions. The two criteria most used in practice are Maxisum and Maximin.

If the Maxisum alternative is used, the optimization problem is formulated as follows:

$$\begin{aligned}
\text{Maximize} \quad & \sum_j h_j && (4)\\
\text{subject to:} \quad & 0 \le h_j \le 1 \quad \forall j && (5)\\
& h_j \le 1 - \frac{b_j - x_{m_j}^{\mathrm{adj}}}{b_j - a_j} \quad \forall j && (6)\\
& h_j \le 1 - \frac{x_{m_j}^{\mathrm{adj}} - b_j}{c_j - b_j} \quad \forall j && (7)\\
& \vec{x}^{\,\mathrm{adj}} \in C && (8)
\end{aligned}$$

The Maximin formulation is obtained by changing (4)-(8) into

$$\begin{aligned}
\text{Maximize} \quad & h && (9)\\
\text{subject to:} \quad & (5) - (8) && \\
& 0 \le h \le h_j \quad \forall j && (10)
\end{aligned}$$

In addition, condition (8), stating the consistency of the adjusted vector $\vec{x}^{\,\mathrm{adj}}$, can be removed if conditions (6) and (7) are rewritten as:

$$\begin{aligned}
h_j &\le 1 - \frac{b_j - \varphi_{m_j}(\vec{x}^{\,\mathrm{adj}})}{b_j - a_j} \quad \forall j\\
h_j &\le 1 - \frac{\varphi_{m_j}(\vec{x}^{\,\mathrm{adj}}) - b_j}{c_j - b_j} \quad \forall j
\end{aligned}$$

providing that $\vec{x}^{\,\mathrm{adj}} \in C$ is equivalent to $\vec{x}_i^{\,\mathrm{adj}} = \varphi_i(\vec{x}^{\,\mathrm{adj}})$ for all $i = 1, \dots, n$.

3.3 Reduction of the Dimensionality. Independent Variables

In general, our method reduces to solving an optimization problem with $n+M$ variables: x_i^{adj} $(i = 1, \ldots, n)$ and h_j $(j = 1, \ldots, M)$. Since the complexity of the optimization problem increases as the number of variables involved does, the following step is to consider a reduction in the number of independent variables.

To this end, let us suppose that there exists a subset $\vec{x}^* = (x_1^*, \ldots, x_{n'}^*)$ $(n' < n)$ of the variables x_i such that the set C of consistent observations can be expressed as $C = \{\vec{x} : \vec{x} = \overline{\varphi}(\vec{x}^*), \vec{x}^* \in \mathbb{R}^{n'}\}$ for an adequate function $\overline{\varphi}$: $\mathbb{R}^{n'} \to \mathbb{R}^n$.

This function $\overline{\varphi}$ can be obtained from φ by several methods: reduction, substitution... An example of a function of this kind can be seen in Example 1: system (S') is the representation of

$$(i_1, \ldots, i_7) = \overline{\varphi}(i_1, i_2, i_4)$$

Variables x_i^* are called independent variables, whereas the rest of variables are said to be dependent.

All the formulations of the previous section can be re-written in terms of the new function $\overline{\varphi}$, just by substituting $\varphi_{m_j}(\vec{x}^{\,\mathrm{adj}})$ by $\overline{\varphi}_{m_j}((\vec{x}^*)^{\mathrm{adj}})$.

4 Some Examples

Let us present some examples of use of our proposal to obtain consistent information from contradictory measurements of the variables of a system.

Example 2 Let us consider the circuit given in Fig. 1 again. In this case, suppose that there are inconsistent and missing data in our observations. These observations are coded with triangular membership functions.

Our observations are: $i_1 = 7$ $(k_1 = 0.5)$, $i_2 = 4.7$ $(k_2 = 0.5)$, $i_3 = 2.5$ $(k_3 = 2.5)$, $i_4 = 2.5$ $(k_4 = 2.5)$, $i_6 = 0.5$ $(k_6 = 0.5)$, $i_7 = 6.8$ $(k_7 = 0.5)$. Note that there is no measurement of variable i_5 and that measurements of variables i_3 and i_4 are less accurate than the others.

First, parameters a_k, b_k and c_k are computed for every observation, and stored in matrix form (matrix D consists of as many rows as measurements and 3 columns):

$$D = \begin{pmatrix} 6.5 & 7 & 7.5 \\ 4.2 & 4.7 & 5.2 \\ 0 & 2.5 & 5 \\ 0 & 2.5 & 5 \\ 0 & 0.5 & 1 \\ 6.3 & 6.8 & 7.3 \end{pmatrix}$$

As shown in Example 1, the system equations have 3 degrees of freedom, which are the variables i_1, i_2 and i_4. Then, the set of consistent states of the system (i.e. the set of all possible consistent measurements) can be expressed in terms of these variables, as $C = \{(i_1, i_2, i_1 - i_2, i_4, i_2 + i_4, i_1 - i_2 - i_4, i_1) / i_1, i_2, i_4 \in \mathbb{R}\}$.

A measure of the goodness of fit for each consistent vector $\vec{i}^{\,adj} \in C$ can be defined in terms of the pre-calculated membership functions: $h_k = h^{Tr}_{(a_k,b_k,c_k)}(i^{adj}_{m_k})$, where (a_k, b_k, c_k) represents de k-th row of matrix $\mathcal{D}$, and $\{m_k\} = \{1, 2, 3, 4, 6, 7\}$ is the set of indices of the measured variables.

As mentioned in the previous section, we obtain a vector $\vec{h} = (h_1, h_2, \dots, h_6)$. The adjusted data is computed as the solution to an optimization problem, following the criteria described in that section: Maximin, Maxisum, 2-norm, or a linear combination of them.

If we are interested in the Maximin version, then the following optimization formulation is obtained:

$$\begin{array}{lll}
\text{Maximize} & h & \\
\text{subject to:} & & \\
0 \leq h_k \leq 1, & \forall i & \\
h_k \leq 1 - \dfrac{b_k - i^{adj}_{m_k}}{b_k - a_k} & \forall i & \\
h_k \leq 1 - \dfrac{i^{adj}_{m_k} - b_k}{c_k - b_k} & \forall i & \\
0 \leq h \leq h_k, & \forall i &
\end{array}$$

This formulation corresponds to a linear programming problem with the variables $i^{adj}_{m_k}$, h_k and h. A very similar linear problem is obtained if the Maxisum criterion, ignoring the variable h, with the following objective function: Maximize $\sum_k h_k$.

In addition, a linear combination of these two formulations can be studied: maximize the quantity $(1-s)h + s\sum_k h_k$.

For a particular value $s = 0.1$, and taking into account that $i_3 = i_1 - i_2$, $i_5 = i_2 + i_4$, $i_6 = i_1 - i_2 - i_4$ and $i_7 = i_1$, we obtain:

$$\begin{array}{ll}
\multicolumn{2}{l}{\text{Maximize: } 0.9h + 0.1(h_1 + h_2 + h_3 + h_4 + h_5 + h_6)} \\
\multicolumn{2}{l}{\text{subject to:}} \\
h_1 \leq 1 - \frac{7 - i_1^{adj}}{0.5} & h_1 \leq 1 - \frac{i_1^{adj} - 7}{0.5} \\
h_2 \leq 1 - \frac{4.7 - i_2^{adj}}{0.5} & h_2 \leq 1 - \frac{i_2^{adj} - 4.7}{0.5} \\
h_3 \leq 1 - \frac{2.5 - (i_1^{adj} - i_2^{adj})}{2.5} & h_3 \leq 1 - \frac{(i_1^{adj} - i_2^{adj}) - 2.5}{2.5} \\
h_4 \leq 1 - \frac{2.5 - i_4^{adj}}{2.5} & h_4 \leq 1 - \frac{i_4^{adj} - 2.5}{2.5} \\
h_5 \leq 1 - \frac{0.5 - (i_1^{adj} - i_2^{adj} - i_4^{adj})}{0.5} & h_5 \leq 1 - \frac{(i_1^{adj} - i_2^{adj} - i_4^{adj}) - 0.5}{0.5} \\
h_6 \leq 1 - \frac{6.8 - i_1^{adj}}{0.5} & h_6 \leq 1 - \frac{i_1^{adj} - 6.8}{0.5} \\
0 \leq h \leq h_i \leq 1 \qquad \forall i &
\end{array}$$

By solving this linear programming problem, we obtain: $i_1^{adj} = 6.9125$, $i_2^{adj} = 4.5875$, $i_4^{adj} = 1.9375$, $h_1 = 0.8250$, $h_3 = 0.93$, $h_2 = h_4 = h_5 = h_6 = h = 0.7750$, from which the rest of variables can be deduced: $i_3^{adj} = 6.9125 - 4.5875 = 2.3250$,

$i_5^{adj} = 4.5875 + 1.9375 = 6.5250$, $i_6^{adj} = 6.9125 - 4.5875 - 1.9375 = 0.3875$, $i_7^{adj} = 6.9125$ and the objective function value is $f = 0.9h + 0.1(h_1 + h_2 + h_3 + h_4 + h_5 + h_6) = 1.183$.

It can be noted that adjusted values are consistent and close to the observed ones. In addition, non-measured variables (such as i_5 in this example) are assigned a consistent value, deduced from the equations of the system.

5 Least Squares Criterion Revisited

Suppose that a dataset $D = \{(x_i, y_i)\backslash i \in I\}$ (with two dimensions, for the sack of simplicity) is given. The objective is to fit a function $y = \phi(x)$ with general expression as follows:

$$\phi(x) = a_1\phi_1(x) + a_2\phi_2(x) + \ldots + a_m\phi_m(x) \tag{11}$$

to the data.

The classical least squares method tries to find the value of the parameters a_i such that the following expression is minimized:

$$F_1 = \min_{a_i} \sum_i (y_i - y_i^{adj})^2 = \min_{a_i} \sum_i \epsilon_i^2 \qquad y_i^{adj} = \phi(x_i) \tag{12}$$

Least Squares fit can be done due to two reasons:

1. Data come from a function (possibly unknown or with a difficult-to-manage expression) and one can approximate this function by a simpler one.

2. There is only a statistical dependence between x and y, and for a given x there may exist several values for y.

In both cases, input data are supposed to lack errors. In this work, we study the situation in which data contain errors.

The classical least squares method allows to consider two cases:

- Approximations y_i come from the real value of a unknown function plus an error term $y_i = \phi(x_i) + \epsilon_i$, therefore it does not consider any error in the observed x_i, thus obtaining the regression curve of Y given X, minimizing the expression (12).

- x_i values have errors, but y_i do not, obtaining the regression curve of X given Y, $x = \theta(y)$ minimizing expression (13):

$$F_2 = \min_{a_i} \sum_i (x_i - x_i^{adj})^2 \qquad x_i^{adj} = \theta(y_i) \tag{13}$$

However, the most frequent situation is that x_i and y_i contain errors. Thus, it seems reasonable to find a function $y = \phi(x)$ minimizing a linear combination of both cases above (F_1 and F_2), that is:

$$F_w = \min_{a_i} \{w \sum_i (y_i - y_i^{adj})^2 + (1 - w) \sum_i (x_i - x_i^{adj})^2\} \qquad 0 \le w \le 1 \tag{14}$$

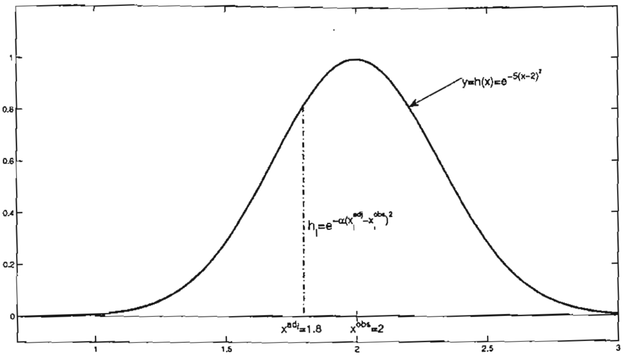

Figure 2: Graphical representation of a verisimilitude function.

verifying $y_i^{adj} = \phi(x_i^{adj})$, i.e. the point (x_i^{adj}, y_i^{adj}) lies on the regression curve $y = \phi(x)$.

It is possible to generalize, even more, this model, by assigning different weights to the observations, or by considering that measurements of the x_i are of different accuracy than those of y_i, leading to the following objective function to be minimized:

$$F = \min_{a_i} \sum_i \alpha_i (y_i - y_i^{adj})^2 + \sum_i \beta_i (x_i - x_i^{adj})^2 \tag{15}$$

5.1 Fuzzy Method

Let us consider functions $H(x)$ such that:

1. H is nonnegative: $H(x) \geq 0$.
2. H is symmetric: $H(x) = H(-x)$.
3. H is not increasing in $\mathbb{R}^+$.
4. Its value is 1 at $x = 0$: $H(0) = 1$.

As examples, we can consider $H_1(x) = e^{-\frac{x^2}{K}}$, $H_2(x) = e^{-\frac{|x|}{K}}$, $H_3(x) = max\{0, 1 - \frac{|x|}{K}\}$, $H_4 = max\{0, 1 - \frac{x^2}{K}\}$,... with $K > 0$, which are continuous functions and some of them differentiable.

Given a function $H(x)$ verifying the above conditions, for each observed datum x_i, the model aims to adjust a value x_i^{adj}, producing a verisimilitude value given by $h_{x_i} = H(x_i - x_i^{adj})$. On the other hand, given a function $\phi(x)$ with a expression such as the one given in (11), we obtain $y_i^{adj} = \phi(x_i^{adj})$, which produces a verisimilitude value for y_i given by $h_{y_i} = H(y_i - y_i^{adj})$. In Fig. 2, we can observe a graphical representation of $h = H_1(x - 2)$ with $K = 0.2$.

The aim of the model is to find the function $\phi(x)$ (actually, parameters a_i in (11)), maximizing the value of h_{x_i} and h_{y_i}.

There exist many possible elections for the function $G(h_{x_1},\dots,h_{x_N},h_{y_1},\dots,h_{y_N})$ to be maximized:

1. $G_1 = \sum_i \alpha_i h_{x_i} + \beta_i h_{y_i}$
2. $G_2 = \prod_i h_{x_i} h_{y_i}$
3. $G_3 = \min\{h_{x_1},\dots,h_{x_N},h_{y_1},\dots,h_{y_N}\}$
4. $G_4 = \sum_i \alpha_i h_{x_i}^2 + \beta_i h_{y_i}^2$
5. ...

By properly choosing the verisimilitude function $H(x)$ and the evaluation function $G(x)$, one can obtain a wide range of fits, among them, the least squares criterion, which can be obtained by combining H_1 and G_2, since:

$$\max \prod_i h_{x_i} h_{y_i} = \max \prod_i e^{-\frac{(x_i - x_i^{adj})^2}{K}} e^{-\frac{(y_i - y_i^{adj})^2}{K}} = \max e^{-\frac{\sum_i (x_i - x_i^{adj})^2 + (y_i - y_i^{adj})^2}{K}}$$

and since $y = e^{-x}$ is an increasing function and $K > 0$, it is equivalent to maximize:

$$\sum_i \left((x_i - x_i^{adj})^2 + (y_i - y_i^{adj})^2 \right)$$

By using different K_i for each datum, we can obtain the expression (15).

Suppose that α and β measure the error produced in variables x and y, respectively. Then, it seems reasonable to use a normalized verisimilitude function H_1, by making $K = \alpha^2$ for those observations of x_i and $K = \beta^2$ for those observations of y_i, leading to the minimization of the function:

$$F = \sum_i \left(\frac{(x_i - x_i^{adj})^2}{\alpha^2} + \frac{(y_i - y_i^{adj})^2}{\beta^2} \right)$$

and, by making $\gamma = \frac{\beta^2}{\alpha^2 + \beta^2}$, we obtain the equivalent minimization problem:

$$\min_{a_i} F = \min_{a_i} \sum_i \left(\gamma (x_i - x_i^{adj})^2 + (1 - \gamma)(y_i - y_i^{adj})^2 \right)$$

5.2 Example: Fitting a Parabola

As an example, let us analyze the case of parabolic regression

$$y = \phi(x) = ax^2 + bx + c$$

when x_i and y_i contain errors. For simplicity, we will consider that errors in the observations x_i and y_i have a Gaussian distribution with mean 0 and standard deviation σ_x (the same for all x_i) and σ_y (for all y_i).

Table 1: Improvement obtained by our method, compared against the Least Squares criterion, when measuring the distances between adjusted and real data.

	α=0.5	α=1	α=2	α=4
β=0.5	70.62	76.48	53.24	11.35
β=1	66.16	79.91	72.88	22.61
β=2	57.41	80.00	80.95	32.95
β=4	25.59	70.21	80.86	59.46

Table 2: Improvement obtained by our method, compared against the Least Squares criterion, when measuring the distances between adjusted and observed data.

	α=0.5	α=1	α=2	α=4
β=0.5	70.70	75.85	53.03	11.41
β=1	73.69	80.13	72.66	22.64
β=2	79.38	84.21	81.06	32.98
β=4	65.68	85.60	82.51	59.63

To test the goodness of the proposed fit, several datasets $\{(x_i, y_i)\}$ have been considered, formed by 10 random points, with $x_i \in [-10, 10]$, obtained from a parabola $y = ax^2 + bx + c$ whose coefficients a, b and c are also randomly chosen in the interval $[-10, 10]$. Each point in the parabola (real data) has been randomly perturbed following a Gaussian distribution of mean 0 and standard deviation σ_x and σ_y with values in $\{0.5, 1, 2, 4\}$ and rounded to 2 decimal digits. For each combination of values, 1000 different parabola have been studied, and for each of these parabola, 10 different sets of points. Results can be observed in Tables 1 and 2.

Different measures of the goodness of fit, used in Tables 1 and 2, are:

1. Sum of the squared distances between real data (without perturbation) and adjusted data, for 100 random validation points, built in the same way that the 10 points used for the fit. See Table 1.

2. Sum of the squared distances between observed and adjusted data, for 100 random validation points, built in the same way that the 10 points used for the fit. See Table 2.

In both cases, the measure given in the Tables is the percentage of relative improvement with respect to the least squares criterion:

$$\text{improvement} = \left(1 - \frac{\epsilon}{\epsilon_{LS}}\right) \cdot 100$$

where ϵ is the measure of the goodness of fit obtained by our proposal (in each of the two cases above), and ϵ_{LS} is the goodness of fit obtained by the Least Squares method.

Observe that our proposal is able to achieve better results than the Least Squares Criterion in both situations:

1. First, our method is able to recover the original model better than by using Least Squares, as shown in Table 1, since the improvement is, in most cases, over 55-60%.

2. Finally, this method is able to get a better representation of observed data, as shown in Table 2, with an improvement comparable to the obtained for real data.

In addition, another measure can be applied to show that our model is actually closer to the original one than the proposed by Least Squares method. It consists on measuring the difference (in $\|\cdot\|_2$ and $\|\cdot\|_1$) between the original parabola coefficients a, b and c, and the obtained by our method a_1, b_1 and c_1, with respect to the obtained by Least Squares (a_{LS}, b_{LS} and c_{LS}). In our experiments, we have found that the relative improvement achieved by our method is over 30% when considering both $\|\cdot\|_2$ and $\|\cdot\|_1$ of the difference of the coefficients.

It must be noted that when errors produced in both variables x and y are of the same magnitude ($\alpha \approx \beta$), we obtain $\gamma \approx 0.5$, and the objective function to be minimized is:

$$2F = \min_{a_i} \sum_i (y_i - y_i^{adj})^2 + \sum_i (x_i - x_i^{adj})^2 = \sum_i d_i^2$$

where $d_i^2 = (x_i - x_i^{adj})^2 + (y_i - y_i^{adj})^2$. This expression can be interpreted as the sum of the squared distances between observed data and the fitted parabola. That is, the parabola obtained after the minimization process is the one minimizing the sum of squared distances $\sum_i d_i^2$.

In Fig. 3, we can see the situation of a real parabola, a fitted parabola, some observed points and its corresponding adjusted point in the fitted parabola, depending on whether $\gamma = 0.5$ (if errors in both variables are considered comparable) and $\gamma = \frac{16}{17}$, for concrete values of $\alpha = 1$ and $\beta = 4$.

6 Conclusions

Real world observations usually are inconsistent with the underlying theory, and some of the measurements must be corrected in order to get values differing as little as possible from obtained data, and agreeing with the theory.

In this work, we develop a general framework to incorporate fuzziness to observed data. Our method is able to generate more accurate estimations of the underlying model than the well-known Least Squares criterion.

The theoretical foundations of this model are explained, and several examples of data recovering and fitting are presented.

Concretely, our simulations have shown that our model is able to recover the original model of the experiments, and to get a better representation of the observed data, obtained by the original model with some perturbations (errors in measurements).

Future work covers aspects such as the possibility to incorporate fuzzy observations to the model (that is, a variable x_1 has a 'high' value whereas x_2 is 'low'), and to

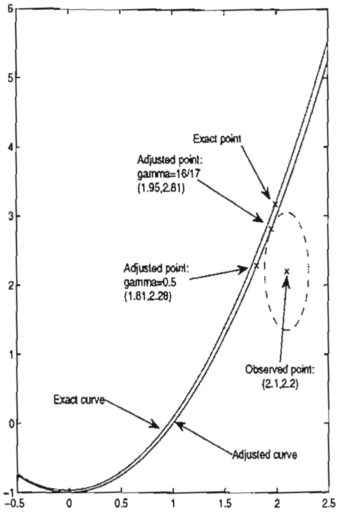

Figure 3: Example of a fitted parabola.

deal with more complicated data representations (trapezoidal instead of triangular, for example).

Acknowledgments

This work has been partially supported by Junta de Andalucía (Spain) under contract TIC-01615, project name Intelligent Remote Sensing Systems.

References

[1] S. Kikuchi and D. Mijkovic, A Method to Adjust Observed Transportation Data: Application to Passenger Counts on a Transit Line, Joint 9th IFSA World Congress and 20th NAFIPS International Conference (2001) 2888–2893.

[2] I. Nir, Bayesian Approach to Data Adjustment and its Application in Reactor Physics, Ph. D. Thesis, Carnegie-Mellon University, 1980.