A comparison between probabilistic and possibilistic models for data validation

V. Köppen, H.-J. Lenz

Freie Universität Berlin, Germany
Institute of Production, Information Systems and Operations Research
Garystr. 21
14195 Berlin
{koeppen, hjlenz}@wiwiss.fu-berlin.de

Summary. Data validation for models with errors in the variables is an important aspect for supporting decision making. In this context, several concepts have been employed. In this paper, we compare a possiblistic and a probabilistic approach. The DuPont Business model is chosen as an example for a controlling model with errors. Although the FuzzyCalc algorithm, representing the possiblistic approach, and the SamPro algorithm, representing the probabilistic approach, use different calculi, their results are quite similar, proving themselves suitable for data validation.

Key words: Data Validation, MCMC, Fuzzy Set

Introduction

With ever growing market demand, business management tools and techniques for decision making are becoming increasingly important. Business decisions are based on business figures. Presently, such business figures are being handled as crisp data, despite the fact that they are usually counted, estimated or measured. However, sampling, estimation, counting and measurement errors should be taken into consideration. This implies that the recorded figures need to be validated, since the data may not match a given numeric model chosen for data analysis.

There exist several concepts for data validation. We will compare two different approaches that take the aforementioned aspects into account: A probabilistic approach, which is represented by a simulation algorithm using the statistical programming language R, and a possibilistic approach, based on Fuzzy-Set-Theory.
Both approaches use different calculi, however they are used for data validation. [Dubois, Prade 2006] show that possibility delivers an upper bound for probability. They also show the transformations from possibility to probability. In our comparative study we are interested in the divergence of these approaches for numerical data validation.
We consider a model class $\mathcal{M}$ defined by numeric variables which are related by the arithmetic operators +, -, * and /. Note that multiplication and division lead to non-linear operations. We assume that for all $m \in \mathcal{M}$, the variables in the related equations are separable, i.e. each equation can be uniquely solved for each variable.

A Possibilistic Approach

Fuzzy set theory is used to solve an algebraic equation system given expert knowledge. FuzzyCalc [Lenz, Müller 2000; Müller, Lenz 2003], an Excel Add-In is employed for this purpose.
Before presenting the experimental results, let us refer to the theoretical background of Fuzzy logic, cf. [Zadeh 1965]: A membership function $\mu(x)$ represents a fuzzy variable (Fig. 1).

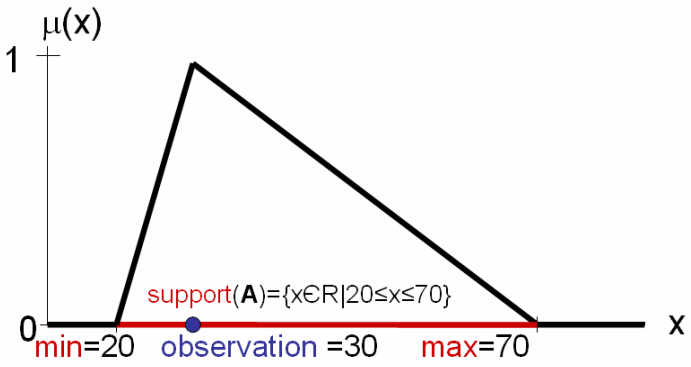

Fig. 1. Fuzzy-Set A

A variable may be represented by more than one membership function, if it occurs in more than one equation. In such a case, each pair of membership functions involved is combined and a renormalised.

$$\mu_{A_1 \cap A_2}(x) = \min\left\{\mu_{A_1}(x), \mu_{A_2}(x)\right\} \tag{1}$$

The following algorithm is used to solve non-linear fuzzy equation systems. It is implemented in FuzzyCalc.

FuzzyCalc Algorithm
Input: one observation per fuzzy variable, fuzzy equation system, fully specified membership functions (missing values allowed)
Output: adjusted fuzzy sets (variables)
Repeat
  For all equations
    For all variables in current equation
      Resolve (separation step) equation for each variable
      Compute (folding) new Fuzzy set using arithmetic as defined by Zadeh's extension principle
    End for
  End for
  For all fuzzy variables
    Compute the intersections
    If an intersection is empty **then** notify user: "System is inconsistent!"
    Else re-normalise the corresponding membership function
  End for
Until no changes for all fuzzy sets occur
End algorithm.

A fuzzy variable should have a restricted membership function. Interpreting a membership function with more than one peak is not a trivial task, therefore only convex membership functions are considered.
FuzzyCalc has a number of useful properties. Several of these properties are identical with properties of the probability theory under a Gaussian regime. These properties are:

- supports have a monotone contraction, when fuzzy variables are arithmetically combined
- the peak positions of adjusted fuzzy variables fulfil the equation system,
- invariance of adjusted data,
- shift of peak positions(values) dependent on length of support
- shift depends on support.

We illustrate the output of FuzzyCalc for one out of seven variables, linked by four equations, i.e. return on investment (ROI).

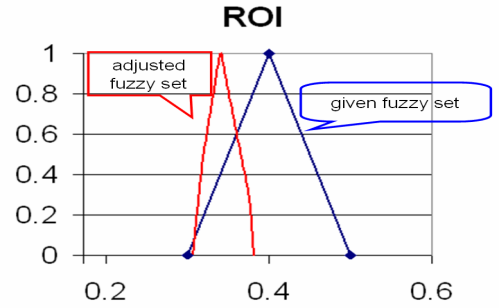

Fig. 2. FuzzyCalc results for ROI

SamPro – a MCMC algorithm

SamPro algorithm, [Köppen, Lenz 2005], can be applied for: (1) estimating missing values of variables and (2) improving the estimations of variables in case of an over-determined equation system.
Parametric probability distributions closed under all arithmetic operations do not exist. Therefore, we use the Metropolis-Hastings algorithm [Hastings 1970] in order to simulate stochastic equation systems. Interestingly enough, this algorithm does not need to know the normalisation factor of a candidate density function.
It should be noted, that candidate density functions (modulo normalisation) have to fulfil the L2-norm. A sufficient restriction is that the function is bounded. Obviously, MCMC sampling might introduce errors. However, increasing the number of simulation runs, these errors will be reduced according to the law of large numbers of probability theory.

The algorithm is given by:

SamPro-Algorithm

Input: a stochastic equation system, observation vector x, z with one observation per variable (missing values allowed)

Output: estimates for all variables

resolve (set LHS[1] ≡ RHS) for each variable in all equations

simulate samples for RHS

compute LHS by sampling from the joint density function of all RHS variables

estimate quantiles $\underline{q_{\max}}, \overline{q_{\min}}$ for each variable with

$\underline{q_{\max}} = \max\{\alpha\text{ - Quantiles for a variable }\}$ and,

$\overline{q_{\min}} = \min\{(1-\alpha)\text{- Quantiles for a variable }\}$

compute the distribution of $\hat{f}_{xz}$ restricted by the subspace x-z = 0

End algorithm.

SamPro has the following properties:

- Shift in mean
- Shift is dependent on variance,
- Variance-reduction,
- Invariance of variables with zero variance.

Fig. 3 shows the result of return on investment (ROI) using the SamPro – simulation from case one with Gaussian distributions.

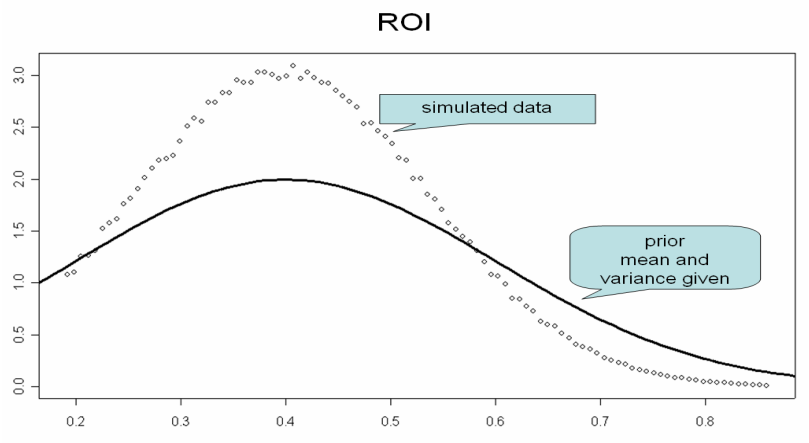

Fig. 3. SamPro results for ROI computed from a seven variable and four equation model

[1] z = x1 + x2 then z is left hand side variable (LHS) and the x's are right hand side variables (RHS).

Comparison of the approaches

We now compare both approaches by using the DuPont Business Model.

DuPont Business Model

The DuPont model consists of seven variables and four equations. The model graph is depicted in Fig. 4.
The structural equation system is given by:

- transaction volume = profit + costs
- capital turnover = transaction volume / capital
- profit margin = profit / transaction volume
- return on investment volume = profit / capital

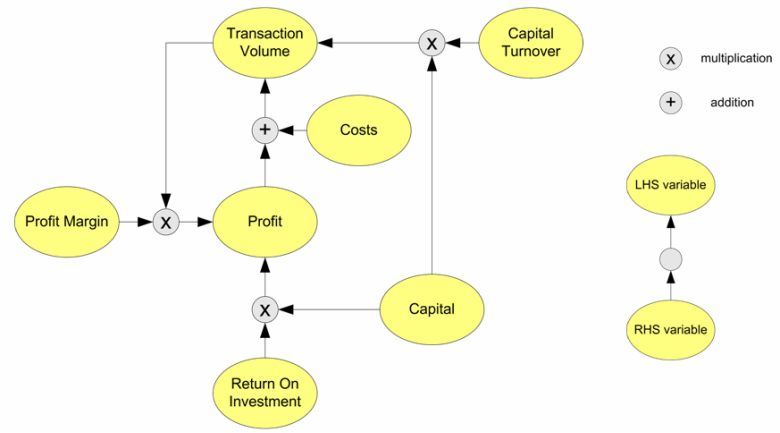

Fig. 4. Model graph of the DuPont System

We now present the results of our comparative study.
We developed 15 cases of different prior knowledge in the variables of the DuPont model. In all cases the results of FuzzyCalc and SamPro are very similar. Therefore we present here only one, but representing case.
Let the following apriori information be available:
Costs and capital turnover are unknown. In the possibilistic approach, all variables are represented by a triangular membership function, unless the variables have fixed values. In SamPro, a multivariate Gaussian distribution is initially assumed. To better evaluate its performance, we also use a

triangle function. In our experimental evaluation, FuzzyCalc, always uses a triangle function. Table 1 shows the results for two different cases. In case one, a higher variance is used, whereas in case two the variance is relative smaller.
Both algorithms, FuzzyCalc and SamPro, exhibit similar behaviours. The data sets are consistent. Consequently, we are able to reduce the support using the fuzzy logic approach and to decrease the standard deviations of the probability distributions involved. The differences in the values are negligible.

Table 1. Sampling and estimation results for the DuPont system

	FuzzyCalc		SamPro					
case 1	Input	Output	Gaussian	Mean	sd	Triangle	Mean	sd
transaction volume	(75, 100, 125)	(75, 102, 125)	$N(100,25^2)$	100.05	17.99	(75, 100, 125)	100.5	7.5
costs	?	(40, 72, 100)	?	70	25.5	?	70	10.4
profit	(25, 30, 35)	(25, 29.3, 35)	$N(30,5^2)$	29.99	2.9	(25, 30, 35)	30	1.3
capital	(60, 80, 100)	(60, 79, 100)	$N(80,20^2)$	80.01	14.39	(60, 80, 100)	79.98	6.4
profit margin	(0.15, 0.25, 0.35)	(0.2, 0.27, 0.35)	$N(0.25,0.1^2)$	0.26	0.06	(0.15, 0.25, 0.35)	0.264	0.023
ROI	(0.2, 0.4, 0.6)	(0.25, 0.38, 0.58)	$N(0.4,0.2^2)$	0.41	0.126	(0.2, 0.4, 0.6)	0.398	0.05
capital turnover	?	(0.75, 1.28, 2.08)	?	1.25	2.4	?	1.26	0.18
	FuzzyCalc		SamPro					
case 2	Input	Output	Gaussian	Mean	sd	Triangle	Mean	sd
transaction volume	(90, 100, 110)	(90, 101, 110)	$N(100,5^2)$	100	3.6	(90, 100, 110)	100.2	3
costs	?	(58, 72, 85)	?	70	5.4	?	71	4.3
profit	(25, 30, 32)	(25, 28.4, 32)	$N(30,2^2)$	30	1.2	(25, 30, 32)	29	0.9
capital	(72, 80, 88)	(72, 77.9, 88)	$N(80,4^2)$	80	2.9	(72, 80, 88)	80	2.6
profit margin	(0.2, 0.25, 0.3)	(0.22, 0.27, 0.3)	$N(0.25,0.05^2)$	0.27	0.02	(0.2, 0.25, 0.3)	0.259	0.01
ROI	(0.3, 0.4, 0.5)	(0.4, 0.367, 0.44)	$N(0.4,0.1^2)$	0.389	0.04	(0.3, 0.4, 0.5)	0.383	0.02
capital turnover	?	(1.02, 1.28, 1.53)	?	1.25	0.09	?	1.25	0.07

Conclusion

The results of the possibilistic and the probabilistic approaches are quite similar. Both FuzzyCalc and SamPro test whether or not the given data set is a possible solution of the corresponding equation system. Given better quality prior knowledge, both algorithms improve the accuracy of the given data. Thus, a variance reduction as well as a shift in level is achieved, whereas at the same time the adjusted data fulfils the equation system. While the SamPro algorithm is exact with respect to the selected probability distributions despite the sampling errors, the FuzzyCalc approach is restricted by the specifications of the membership functions.
In our study we did not consider multidimensional density functions that include dependencies in the variables due to the fact that this cannot be implemented in FuzzyCalc. We expect increased accuracy for the SamPro algorithm by using multidimensional density functions.

References

[DP06] Dubois, D., Prade, H.: Possibility Theory and its Applications: a Retro spective and Prospective view: Riccia, G., et. Al.: Decision Theory and Multi-Agent Planning, Springer, New York, 89- 109 (2006)

[Has70] Hastings, W.K.: Monte Carlo sampling methods using Markov Chains and their applications. Biometrica, 57 (1), 97-109 (1970)

[KL05] Köppen, V., Lenz, H.-J.: Simulation of Non-linear Stochastic Equation Systems, Proceeding of the Fifth Workshop on Simulation, St. Peters-burg, 373- 378 (2005)

[LM00] Lenz, H.-J., Müller, R.: On the Solution of Fuzzy Equation Systems: Ric cia, G. et.al.: Computational Intelligence in Data Mining, Springer, New York, 95-110 (2000)

[ML03] Müller, R.M., Lenz, H.-J.: FuzzyCalc – Analytische Prüfung mit Fuzzy methoden: Theorie, Anwendung, Beispiele. In: Geldermann, J., Rommel fanger, H.: Einsatz von Fuzzy Sets, Neuronalen Netzen und Künstlicher Intelligenz in industrieller Produktion und Umweltforschung, Fortschritt-Berichte VDI, 10, no. 725, 123-142 (2003)

[Zad65] Zadeh, L.A.: Fuzzy Set, Information and Control 8, 338-353 (1965)