Spreadsheet Computation with imprecise and uncertain Data

Hans - J. Lenz, Freie Universität Berlin
hans-j.lenz@fu-berlin.de

We consider universal relations as flat tables having numeric data types which can be simply viewed as spreadsheets. The underlying stochastic model is the finite dimensional metric (Euclidean) space $\boldsymbol{R}^p$ backed up by the Borel set $\boldsymbol{B}$ as an adequate σ-algebra, and a related uncertainty (see Fuzzy Logic) or probability measure P, i.e. $(\boldsymbol{R}^{p+q}, \boldsymbol{B}, P)$ as the corresponding possibility or probability space. You may consider P as the multivariate distribution on $\boldsymbol{B}$ specified by a multivariate Gaussian distribution. According to existing integrity constraints P may be lossless decomposed into a mixture of partial and marginal distributions (cf. Generalized Markov Theorem).

Ex. 1: Linear component Models Aitkinson (1996)
As completely tackled in science, experts in chemistry for instance sample materials, and measures the fractions (percentages) of S, SO, Fe,… Due to measurement errors such fractions never add-up to one (100%). Can one find "best" estimates of those fractions $\hat{\xi}_i$ (i = 1, 2,…, p) with constraints $\hat{\zeta} = \sum \hat{\xi}_i = 1$?

Ex. 2: Business Indicators (Lenz& Rödel (1991), Köppen & Lenz (2006))
Business indicator systems became popular under the parole "Business Score Cards". The intention is to measure all relevant facts as a status quo (current state) of a firm on one "Bierdeckel". Partial information of more or less loosely coupled departments as well as counting or measurement errors may lead to severe problems of semantic inconsistency of data. For instance, fixing period the indicator "Profit" can be alternatively (non-exclusively!) computed in three ways: Using data from the accountancy dept. as $profit_1$= sales –cost, from the dept. of finance as $profit_2$=return-on-capital x capital, and from the marketing dept. as $profit_3$=margin x sales. Excel sheets as used by firms simplify facts to crisp numbers irrespective of errors of variables. But is it true that $\hat{\zeta} = \sum \hat{\xi}_j$?

Ex. 3: Main Economic Indicators (Köppen & Lenz (2008))
About 220 UNO membership countries annually submit their national reports on the main economic indicators (e.g. GDP, national income, prices,…) to the National Accounting Group of the Statistical Division of the UNO, New York, Such reports have an extra cell for each indicator to represent the so called "statistical discrepancy" allowed. Any further deviation must be footnoted. The balance equation system of the above type consists of about 500-600 equations or rules (edits) on the national and international level. Again, the challenge for uncertainty management is to find contradictions (error location step) and/or to impute erroneous values (error correction step) of those reports, cf. Bertossi (2005).
More formally, we have a data set $(x,z) \in \boldsymbol{R}^{p+q}$, and a fully specified, nonlinear model with errors in the variables. The variables are exclusively related by the arithmetic operators only.

$$\mathbf{P}: \quad \begin{aligned} x &= \xi + v \\ z &= H(\xi) + u \\ (u\ v)' &\sim N(0, \Sigma_{uv}) \text{ and jointly Gaussian distributed.} \end{aligned}$$

For the linear case (the operator set is limited to {+, -}), we can derive under mild mathematical conditions the analytic solution of **P** as the GLS estimator fulfilling the famous Gauss-Markov Theorem:

$$\hat{\xi}_{GLS} = \arg\min\{(u\ v)'\Sigma_{uv}^{-1}\begin{pmatrix} u \\ v \end{pmatrix}\}$$

We discuss data of "bad quality", efficiency and CPU complexity for large data sets (n→∞).