Online Statistical Computation in the Feed Database.

1 Introduction to the Problem.

Each type of animal feed is composed by many components which are different types of nutrients and minerals. The amount of each component contained in a specific type of animal feed is measured from multiple samples of the feed and the collected measures are stored in the Feed Database. The number of measures is the major factor which determines their quality. However, since chemical analysis is a long and expensive process, only for a small number of components the number of measures is sufficiently large.

Example. *Let consider nutrients 'TSO' and 'RP' in barley from the given file 'Analyses of Barley 1992-2009.xlsx'. For the year 2005 the containment of both nutrients is represented by 16 measures which, in this case, is a sufficiently large number. In the year 2008 only the containment of 'RP' was measured and no measures for 'TSO' exist.*

Missing measures is the major problem in the Feed Database: for many scenarios, as examination of feed mixes for a given animal species, it is a requirement to consider all components, even, if they are not properly measured. It is possible to derive missing values based on a statistical method of regression which determines correlated components.

Example. *Let consider Figure 1 which illustrates regression analysis for 'RP' and 'TSO' nutrients. The dashed line represents estimated linear regression* $f(x) = m * x + b$ *with parameters* $m = 0.13$ *and* $b = 0.98$*. Now, each missed value of 'TSO' can be restored using the above equation by simply substituting the corresponding 'RP' values into it.*

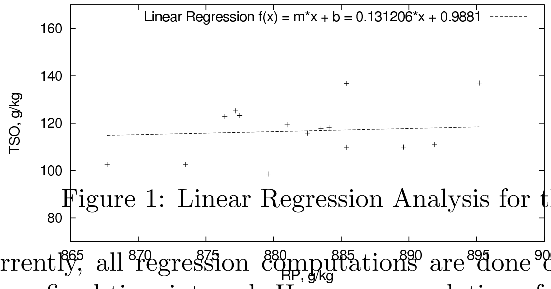

Figure 1: Linear Regression Analysis for the Year 2005.

Currently, all regression computations are done outside the database and only for a fixed time interval. However, correlation of components changes over the time: correlated components become uncorrelated and correlation parameters change. Our goal is to integrate regression analysis into Feed Database.

2 The Task

Your main task is to investigate, apply and compare kernel regression methods for identification of correlated components of barley. In particular, you will use linear and kernel regressions. Linear regression is capable to find out only linearly correlated components. Kernel regression does not assume any model and, thus, is capable to identify components with any type of correlation.

The task is divided into following steps:

- Implement linear regression and kernel regression methods. Use them to find out pairs of correlated components of barley. Feel free to use any environment as GnuPlot, R, Java o C++.
- Compare the results of both regression methods: does kernel regression find all correlated components as it does linear regression and why? Does kernel regression find new non-linear correlations between components and which? In your opinion, which regression method is better and why?
- Consider current design of the Feed Database in Figure 2 and evaluate the performance of the following query:

  select avg(Value) from Measures natural join Components
  where Date='2008' group by Id

  Assume, that for some components there are missing measures in the database. For these components all missing measures are restored by finding correlated components. Answer the following questions: how the the number of measures per component, total number of components and the number of components with missing values effect execution of the query; propose your design of the database which supports fast execution of the query.

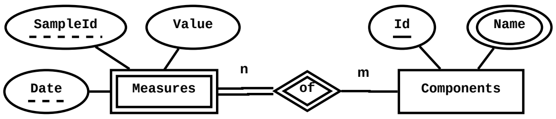

Figure 2: Partial ER diagram of the Feed Database.

3 Literature

All the literature is found at project's home page:
`http://www.ifi.uzh.ch/dbtg/Projects/feed_database/`