Bachelor Thesis

Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

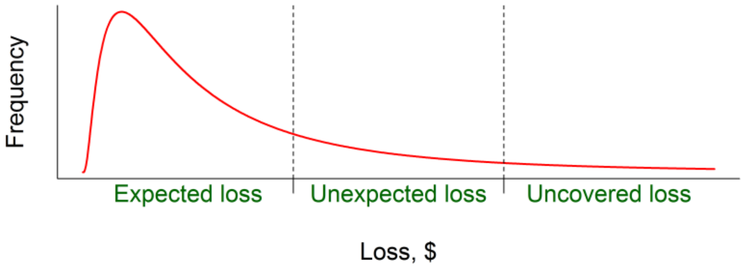
Bachelor Thesis

Comparing Scaling Models to Estimating Risks in Finance

This work focuses on one specific aspect of well-established and powerful Machine Learning techniques such as generalized additive models, generalized linear models, quantile and simple linear regressions, and their application to the specific domain of quantifying Operational Risk data. The successful result of the work also includes ready-to-use R package.

Banks or insurance companies want to measure their operational losses. These include losses due to fraud, mistakes of employees or disasters. To this end, the institutions estimate the distribution of annual losses and quantify unexpected losses (see the figure).

For these calculations to be stable, one needs a large amount of data. This is especially crucial for the precision in the area of high losses (right hand part of the plot).



However, a single bank usually does not have any or only little data on high losses. Hence, they use data from other banks for this estimation. It is natural to assume, that the distribution parameters of the losses depend on characteristics of the company size such as the average gross income or the number of employees. Various models have been proposed to estimate this dependence, starting with simple OLS regression to more sophisticated GAMLSS model. However, we are not aware of any comparison of the models to each other. This may be due to the different data used for modelling. But the most crucial reason in our view is the absence of the general quality measure, since the real distribution of the loss is unknown.

You are asked to come up with a meaningful comparison based on various synthetic data sets which you are supposed to generate. Doing so, one knows the true distribution and is able to calculate unexpected loss. The generating process should be as general as possible. That is, the synthetic data can follow different distribution, can be observed only over some threshold and have a bias towards higher values, contain “outliers” etc.

The thesis implies the following tasks:

- Reviewing proposed models and implementing some of them in R (if not readily available).
- Compiling assumptions on data generating processes behind the models envisioned
- Designing and implementing a general artificial data generator
- Generating data and comparing the models. Explaining the results.
- Creating an annotated R package of models, data generator and function for comparisons.

Contact

Vadim Arzamasov
vadim.arzamasov@kit.edu
Room: 340

Am Fasanengarten 5
76131 Karlsruhe
Building: 50.34