Summary

Sample size has always been an issue in statistical analyses. While for many years small sample sizes were a central issue, in recent years, large sample sizes too might confront the statistical analyst with serious problems. When large to huge sample sizes concur with complex models, prohibitive computation times, convergence issues, and the mere impossibility to analyze the data with the preferred inferential technique (e.g., full maximum likelihood) can ensue.

Furthermore, while in the simplest designs all measurements are independent, many designs lead to hierarchical (a.k.a., clustered, correlated, dependent) data. By 'cluster' we mean a set of measurements repeatedly collected for one single unit (e.g., subject, household, trial, etc.). Therefore, the number of repetitions refers to the number of measurements available per cluster. Examples include: patients measured repeatedly over time, several patients attending the same general practitioner or clinic, several links clicked by a specific user, genetic data available for a person, etc. Some important examples of clustered data are repeated measurements, longitudinal data, multi-center studies, and meta-analysis. Accounting for correlations remains a challenge today, when the total number of clusters is large and/or there are a large number of repetitions per cluster. This would sometimes result in prohibitive computations.

In this thesis, we propose methodologies to deal with such problems using the so-called data splitting. In a nutshell, data splitting means to split the difficult (sometimes infeasible) job of modelling the big dataset by splitting it into smaller chunks of manageable sub-data, analyze each of them separately, and then combine the results of these analyses using appropriate combinations rules. This would break a difficult task into several, often so-called embarrassingly parallel by computer scientists, simpler tasks. The data splitting approach consists of 3 steps:

1. Splitting the data.
2. Performing the analysis on each sub-dataset.
3. Efficiently combine the results from each sub-dataset to form an overall estimate with its precision.

We could summarize various splitting strategies into four main approaches:

- **Random horizontal data splitting.** This approach is useful when the number of clusters is large, and preferably the clusters are roughly of equal size, i.e., an (almost) balanced design. This will produce several sub-datasets with smaller sample sizes which are easier to handle.
- **Structured horizontal data splitting.** This approach is useful when some specific clusters are more convenient to deal with when they are together. Therefore, instead of taking random sub-samples, the sub-sampling can be done based on this pre-specified structure. An important example of this approach which is considered in this thesis is splitting based on the cluster sizes, i.e., forming sub-samples consisting of clusters of equal size. This would lead to balanced sub-samples. We have shown the existence of closed-form solutions, leading to faster convergence of model fitting procedures.

- **Random vertical data splitting.** This approach is useful when we are dealing with data with large number of measurements for each cluster per outcome. In this situation one can take sub-samples of each cluster with a manageable size.

- **Structured vertical data splitting.** This approach is useful when the number of outcomes of interest is large. While fitting a model to several outcomes could become heavy or infeasible, fitting models to all possible pairs (triples, etc.) of these outcomes is feasible. That means each sub-sample should consist of all of the measurements for each subject regarding the pair of interest. Therefore, the sub-samples are taken using a pre-specified structure.

For each of these splitting strategies appropriate and efficient combination rules are developed. We also have considered exceptional cases where the general mythologies would fail. Our findings are motivated and supported by various datasets. Furthermore, the methodologies we have developed are implemented in statistical computing language R.

Professional Background

Vahid Nassiri has obtained a Bachelor in Statistics (2002) from the Shahid Beheshti Univerity in Tehran, Iran, and a master in Mathematical Statistics (2008) from the Amirkabir University of Technology in Tehran, Iran. Afterwards, he has obtained a PhD in Applied Mathematics (2013) from Vrije Universiteit Brussel (VUB), Brussels, Belgium. His main research interests are in the field of Big Data and High-dimensional problems. While working on the sparse estimation of high-dimensional linear models in sense of least squares and quantile regression in his former PhD, his current work is concerned on the fast computation of models for correlated data with possible missing values. His research expertise is mainly concerned with analyzing incomplete big correlated data using mixed models, multiple imputation, penalized likelihood and related topics. He also has experiences in consultation involving projects from pharmacology, biology, medicine, engineering, psychology, sociology, linguistics, and market research. He is now working as a statistical consultant in Open Analytics NV, Antwerp, Belgium.