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 so-called data splitting. In a nutshell, data splitting refers to splitting the difficult (sometimes infeasible) job of modelling a big dataset by splitting it into smaller chunks of manageable sub-data, analyzing each of them separately, and then combining the results of these analyses using appropriate combination rules. This would break a difficult task into several, often so-called embarrassingly parallel, 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 established 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 numbers of measurements for each cluster per outcome. In this situation one can take sub-samples of each cluster of 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 methodologies would fail. Our findings are motivated and supported by various datasets. Furthermore, the methodologies we have developed are implemented in statistical computing language R.

Samenvatting

Steekproefgrootte is altijd een belangrijk thema geweest in de statistiek. Terwijl de problemen zich lange tijd situeerden ter hoogte van kleine steekproeven, is het nu zo dat ook (zeer) grote steekproeven voor problemen zorgen bij statistische analyse. Wanneer zeer grote steekproeven gebruikt worden samen met heel complexe modellen, dan kunnen er zich een reeks problemen voordoen, zoals bijvoorbeeld onrealistisch lange berekeningstijden, convergentieproblemen, tot en met de onmogelijkheid om de voorkeurstechniek (bijv. meest aannemelijke schatters) te gebruiken.

Verder, terwijl in de eenvoudigste studies alle metingen onafhankelijk zijn, zien we vandaag heel veel zogenaamd hiërarchische gegevens (ook gekend onder de termen: clusters, gecorreleerde gegevens, afhankelijke gegevens). Met de term 'cluster' verwijzen we naar metingen herhaald genomen aan eenzelfde eenheid (bijv. proefpersoon, huishouden, studie, enz.). Tegen die achtergrond verwijzen we met het aantal herhalingen naar het aantal metingen per cluster. Enkele voorbeelden: patiënten die herhaald doorheen de tijd worden gemeten; verscheidene patiënten van dezelfde huisarts of dezelfe kliniek, verscheidene web-links aangeklikt door dezelfde gebruiker, allerlei genetische informatie beschikbaar voor dezelfde persoon, enz. Belangrijke deelgebieden van cluster gegevens zijn herhaalde metingen, longitudinale gegevens, multi-centrische studies en meta-analyse. Ook vandaag is het in rekening brengen van correlatie binnen clusters een uitdagend probleem, vooral wanneer het totaal aantal clusters zeer groot is en/of wanneer een groot aantal metingen per cluster beschikbaar is. We worden dan veelal geconfronteerd met zo goed als onoverkomelijke problemen van berekenbaarheid.

In deze thesis stellen we methdologie voor om de bovenstaande problemen aan te pakken op basis van zogenaamde data-splitsing. In een notendop verwijst data-splitsing naar het opdelen van een moeilijke

(soms zelfs onmogelijke) taak zoals het modelleren van een zeer grote dataset, in kleinere deeltaken die elk wel overzichtelijk zijn. Een voorbeeld is het analyseren van delen van de dataset, één voor één, waarna de resultaten worden gecombineerd tot één enkel resultaat. Als het aantal deeltaken zeer groot wordt spreken we in het Engels van *embarrassingly parallel tasks.* Data-splitsing bestaat uit drie deeltaken:

1. Splitsen van de data;
2. Elk van de delen analyseren;
3. Het efficiënt combineren van de resultaten tot één enkel eindresultaat.

Er zijn verscheidene splitsings-strategieën. We kunnen ze in vier grote klasen opdelen:

- **Random horizontale splitsing.** Deze methode is nutting wanneer het aantal clusters groot is; idealiter zijn de clusters van ongeveer gelijke grootte, d.w.z. een (bijna) gebalanceerd steekproefopzet. Deze methode leidt tot een aantal sub-datasets van kleinere omvang, die elk makkelijker te manipuleren zijn.
- **Gestructureerde horizontale data splitsing.** Deze methode is nuttig wanneer sommige clusters makkelijker individueel te hanteren zijn dan in samenhang. Daarom, eerder dan willekeurige deelsteekproeven te nemen, maakt men hier gebruik van de vooraf aanwezige structuur. Een belangrijk voorbeeld, beschouwd in dit werk, is splitsing op basis van cluster-grootte. We vormen dus deel-datasets waarbinnen alle clusters dezelfde grootte hebben. Elk van deze deelsteekproeven is dan gebalanceerd. Voor dergelijke gebalanceerde deelsteekproeven hebben we het bestaan van gesloten schattingsvormen aangetoond. Dit leidt daarom tot bijzonder efficiënte procedures.
- **Random verticale data splitsing.** Deze methode is nuttig wanneer we een zeer groot aantal herhalingen hebben binnen een cluster. We splitsen clusters dus in deelclusters van een hanteerbare grootte.
- **Gestructureerde verticale data splitsing.** Deze aanpak is nuttig wanneer het aantal te bestuderen responsveranderlijken groot is. Het gemeenschappelijk analyseren van alle responsveranderlijken samen is vaak ondoenbaar, maar dat geldt niet voor alle mogelijke paren die we vormen uit de responsveranderlijken (of tripels, quadrupels, enz.). De deelsteekproeven bestaan dan uit alle gegevens beschikbaar voor het paar van veranderlijken dat wordt beschouwd. Deelsteekproeven respecteren dus deze specifieke structuur.

Voor elk van de splitsings-strategieën werden gepaste en efficiënte combinatieregels opgesteld. Daarbij werden ook uitzonderlijke gevallen waar de algemene methodologie niet werkt in ogenschouw genomen. Onze resultaten werden gemotiveerd en geïllustreerd aan de hand van een reeks datasets. Tenslotte hebben we de voorgesltede methodologie geïmplementeerd in de statistische softwaretaal 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 the 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 experience in consulting involving projects from pharmacology, biology, medicine, engineering, psychology, sociology, linguistics, and market research. He is now working as a statistical consultant at Open Analytics NV, Antwerp, Belgium.