Analysis of Very Large Data Sets: Frequentist and Bayesian Regression Approaches

Leo N. Geppert

TU Dortmund University
Faculty of Statistics

Joint work with Katja Ickstadt, Alexander Munteanu, and Christian Sohler

Supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876, project C4

Setting

- Large number of observations n, small to moderate number of variables p ($n \gg p$), possibly data stream
- Conduct frequentist or Bayesian regression analysis with Y as dependent variable, X as matrix of independent variables, and β as parameter vector
- In the frequentist case, β is fixed, but unknown
- In the Bayesian case, β is a random vector with prior distribution $p(\beta)$

Recipe 1: Reduce the dimension

- Reduce the number of observations from n to k while retaining the original regression model
- Carry out regression analysis on the reduced data set

$$
\begin{array}{ccc}
[X, Y] & \xrightarrow{\Pi} & [\Pi X, \Pi Y] \\
\downarrow & & \downarrow \\
p_{\mathrm{post}}(\beta|X, Y) & \approx_{\varepsilon} & p'_{\mathrm{post}}(\beta|\Pi X, \Pi Y)
\end{array}
$$

- Trade-off between guaranteed goodness of approximation and data reduction can be adjusted using ε
- Π can be a subsampling matrix or a random projection (sketch)

Ingredients: Subsampling

- Choose a subsample of size $k < n$ of the original observations in X and Y
- Subsample should represent original data set with respect to certain properties
- Closely related to concept of coresets in computer science
- Uniform sampling does not lead to good results in general, sampling proportional to leverage scores often good in regression context
- Each entry of reduced data matrix is an entry of original data set, possibly weighted to correct for sampling probability

Ingredients: Random projections

- $\Pi \in \mathbb{R}^{k \times n}$ is a random matrix that can be stored implicitly
- Reduce number of observations by calculating random linear combinations
- Observations are typically not interpretable, but variables still are
- Finding suitable matrices Π for frequentist linear regression is a very active field of research in computer science
- Subspace embeddings differ in running time and target dimension k

Laying the foundation: linear regression

- In the case of linear regression, random projections are an excellent choice
- For frequentist linear regression, many random projections with theoretical guarantees are available
- We extended three random projections to the Bayesian case, also with theoretical guarantees ([Geppert et al. (2017)])

Generalisations of linear model

Generalisations for priors

- Hierarchical models (empirical, some theoretical support for guarantees of non-population parameters) [Rathjens (2015)]
- q-generalised normal distributions as prior ($q \in [1, 10]$) [Müller (2016)]
  - Bayesian version of the LASSO ($q = 1$)
  - Limiting case of $p \to \infty$ quickly approximated for $q > 2$

Generalisations for likelihood

- q-generalised normal distributions as likelihood ($q \in [1, 2]$) [Müller (2016)]
- Requires a combination of random projection and subsampling

Generalisations to different frequentist regression models

- Logistic regression [Munteanu et al. (2018)]
  - Reduction via subsampling/corsets
  - Difficult in worst case $\rightarrow$ introduced complexity parameter to deal with such cases
- Variable selection in presence of interactions
  - $n \ll p$ setting
  - subsampling approach based on leverage scores finds important main effects
  - additional sampling based on cross-leverage scores identifies variables involved in interactions

Recipe 2: Merge the models

- Split the data into blocks of size n_b
- Carry out regression analysis on each block
- Merge models along a tree structure
- Approach creates little overhead

Ingredients: Merge & Reduce

- General algorithmic principle
- Turns static data structures into dynamic ones
- Used mainly on coresets
- Our contribution: transfer principle from data structures to statistical (regression) models

Foundation and Toppings: Merge & Reduce

- We propose three different Merge & Reduce approaches [Geppert et al. (2020)]
  - One is suitable for general frequentist regression models
  - The second is suitable for general Bayesian regression models
  - The third is suitable for frequentist linear regression only
- Approaches 3 gives exact solution of regression model
- Approaches 1 and 2 offer no theoretical guarantee, but show convincing results empirically

Summary

- Random projections and subsampling offer good approaches for many regression models
- Theory approximations guarantees, especially for linear and logistic regression, mainly empirical results for further extensions of models
- R-package `RaProR` available on CRAN
- Merge & Reduce presents a different, rather general approach that is suitable for multiple regression models
- R-package `mrregression` soon available on CRAN

Literature I

LN Geppert, K. Ickstadt, A. Munteanu, J. Quedenfeld, C. Sohler
Random Projections for Bayesian Regression
Statistics and Computing **27**, 79–101 (2017)

LN Geppert, K. Ickstadt, A. Munteanu, C. Sohler
Streaming statistical models via Merge & Reduce
Journal of Data Science and Analytics, 1-17 (2020)

LN Geppert, K. Ickstadt, A. Munteanu, J. Quedenfeld, L. Sandig, C. Sohler
RaProR: Calculate Sketches using Random Projections to Reduce Large Data Sets, Version 1.1-5
2019

Literature II

A. Munteanu, C. Schwiegelshohn, C. Sohler, DP Woodruff
On Coresets for Logistic Regression
Proc. of NeurIPS (2018)

S. Müller
Bayes-Regression unter ℓ_p-Normen bei Einbettung großer Datensätze
Master's Thesis (2016)

J. Rathjens
Hierarchische Bayes-Regression bei Einbettung großer Datensätze
Master's Thesis (2015)