Insert here your thesis' task.

FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Bachelor's thesis

Probabilistic algorithms for computing the LTS estimate

Martin Jenč

Department of Applied Mathematics
Supervisor: Ing. Karel Klouda, Ph.D.

March 6, 2019

Acknowledgements

THANKS to everybody

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as school work under the provisions of Article 60(1) of the Act.

In Prague on March 6, 2019 ....

Citation of this thesis

Jenč, Martin. *Probabilistic algorithms for computing the LTS estimate*. Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2019.

Abstrakt

V několika větách shrňte obsah a přínos této práce v českém jazyce.

Klíčová slova LTS odhad, lineÃąrnÃŋ regrese, optimalizace, nejmenÅąÃŋ usekanÃľ ÄŊtvrece, metoda nejmenÅąÃŋch ÄŊtvercÅŕ, outliers

Abstract

The least trimmed squares (LTS) method is a robust version of the classical method of least squares used to find an estimate of coefficients in the linear regression model. Computing the LTS estimate is known to be NP-hard, and hence suboptimal probabilistic algorithms are used in practice.

Keywords LTS, linear regressin, robust estimator, least trimmed squares, ordinary least squares, outliers, outliers detection

Contents

List of Figures

Introduction

CHAPTER 1

Linear Regression

1.1 Description

1.2 Computation

1.3 Downfalls

Chapter 2

The Least trimmed squares

2.0.1 Objective function

2.0.1.1 Problems

CHAPTER 3

Algorithms

3.1 FAST-LTS

In this section we'll introduce FAST-LTS algorithm[1]. It's, as well as in other cases, iterative algorithm. We'll discuss all main components of the algorithm starting with its core idea called concentration step which authors simply calls C-step.

3.1.1 C-step

We'll show that from existing LTS estimate $\boldsymbol{\hat{w}_{old}}$ we can construct new LTS estimate $\boldsymbol{\hat{w}_{new}}$ which objective function is less or equal to old one. Based on this property we'll be able to create sequence of LTS estimates which will lead to better results.

Theorem 1: Consider dataset consisting of $\boldsymbol{x_1}, \boldsymbol{x_2} \ldots, \boldsymbol{x_n}$ explanatory variables where where $\boldsymbol{x_i} \in \mathbb{R}^p \,, \forall \boldsymbol{x_i} = (x_1^i, x_2^i, \ldots, x_p^i)$ where $x_1^i = 1$ and its corresponding $y_1, y_2, \ldots, y_n$ response variables. Let's also have $\boldsymbol{\hat{w}_0} \in \mathbb{R}^p$ any p-dimensional vector and $H_0 = \{h_i ; h_i \in \mathbb{Z}\,, 1 \leq h_i \leq n\}\,, |H_0| = h$. Let's now mark $RSS(\boldsymbol{\hat{w}_0}) = \sum_{i \in H_0} (r_0(i))^2$ where $r_0(i) = y_i - (w_1^0 x_1^i + w_2^0 x_2^i + \ldots + w_p^0 x_p^i)$. Let's take $\hat{n} = \{1, 2, \ldots, n\}$ and mark $\pi : \hat{n} \to \hat{n}$ permutation of $\hat{n}$ such that $|r_0(\pi(1))| \leq |r_0(\pi(2))| \leq \ldots \leq |r_0(\pi(n))|$ and mark $H_1 = \{\pi(1)\,, \pi(2)\,, \ldots \pi(h)\}$ set of h indexes corresponding to h smallest absolute residuals $r_0(i)$. Finally take $\boldsymbol{\hat{w}_1^{OLS(H_1)}}$ ordinary least squares fit on H_1 subset of observations and its corresponding $RSS(\boldsymbol{\hat{w}_1}) = \sum_{i \in H_1} (r_i^1)^2$ sum of least squares. Then

$$RSS(\boldsymbol{\hat{w}_1}) \leq RSS(\boldsymbol{\hat{w}_0}) \tag{3.1}$$

Proof. Because we took h observations with smallest absolute residuals r_0, then for sure $\sum_{i \in H_1} (r_0(i))^2 \leq \sum_{i \in H_0} (r_0(i))^2 = RSS(\boldsymbol{\hat{w}_0})$. When we we take into account that Ordinary least squares fit OLS_{H_1} minimize objective function of H_1 subset of observations, then for sure $RSS(\boldsymbol{\hat{w}_1}) = \sum_{i \in H_1} (r_i^1)^2 \leq$

$\sum_{i\in H_1}(r_i^0)^2$. Together we get $RSS(\hat{\boldsymbol{w}}_{\mathbf{1}}) = \sum_{i\in H_1}(r_i^1)^2 \leq \sum_{i\in H_1}(r_0(i))^2 \leq \sum_{i\in H_0}(r_0(i))^2 = RSS(\hat{\boldsymbol{w}}_{\mathbf{0}})$ $\square$

$F(x) := \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos\frac{k\pi x}{T} + b_k \sin\frac{k\pi x}{T}\,, \quad \forall x \in \mathbb{R}\,,$

```
Data: this text
Result: how to write algorithm with LaTeX2e
initialization;
while not at end of this document do
    read current;
    if understand then
        go to next section;
        current section becomes this one;
    else
        go back to the beginning of current section;
    end
end
```

In this section we'll describe FAST-LTS algorithm and it's main properties. The main idea of this algorthm is based on the fact that from one approximation of the algorithm we can compute another which can have lower objective function. TA DAAAAAAAAAAAAAAAAA Thoerem 1: [2] Let w0 ... wp be the LTS estimate. for each data sample we can compute —y-wx—

Hlavni myslenka tohoto algoritmu spociva ve faktu,

V ÄŊeskÃľ variantÄŻ naleznete Åąablony v souborech pojmenovanÃ · ch ve formÃątu prÃące_kÃşdovÃąnÃŋ.tex. Typ prÃące mÅŕÅ"e bÃ · t:

BP bakalÃąÅŹskÃą prÃące,

DP diplomovÃą (magisterskÃą) prÃące.

UTF-8 kÃşdovÃąnÃŋ Unicode,

ISO-8859-2 latin2,

Windows-1250 znakovÃą sada 1250 Windows.

V pÅŹÃŋpadÄŻ nejistoty ohlednÄŻ kÃşdovÃąnÃŋ doporuÄŊujeme nÃąsledujÃŋcÃŋ postup:

1. V opaÄŊnÃľm pÅŹÃŋpadÄŻ postupujte dÃąle podle toho, jakÃ · operaÄŊnÃŋ systÃľm pouÅ"ÃŋvÃąte:
    - v pÅŹÃŋpadÄŻ Windows pouÅ"ijte Åąablonu pro kÃşdovÃąnÃŋ Windows-1250,
    - jinak zkuste pouÅ"Ãŋt Åąablonu pro kÃşdovÃąnÃŋ ISO-8859-2.

V anglickÃľ variantÄŻ jsou Åąablony pojmenovanÃľ podle typu prÃące, moÅ"nosti jsou:

bachelors bakalÃąÅŹskÃą prÃące,

masters diplomovÃą (magisterskÃą) prÃące.

3.2 Exact algorithm

3.3 Feasible solution

3.4 MMEA

3.5 Branch and bound

3.6 Adding row

Chapter 4

Experiments

4.1 Data

4.2 Results

4.3 Outlier detection

Conclusion

Bibliography

[1] Rousseeuw, P. J.; Driessen, K. V. An Algorithm for Positive-Breakdown Regression Based on Concentration Steps. In *Data Analysis: Scientific Modeling and Practical Application*, edited by M. S. W. Gaul, O. Opitz, Springer-Verlag Berlin Heidelberg, 2000, pp. 335–346.

[2] Rybicka, J. *LaTeX pro začátečníky*. Brno: Konvoj, third edition, ISBN 80-7302-049-1.

Appendix A

Datasets

GUI Graphical user interface

XML Extensible markup language

Appendix B

Contents of enclosed CD

```
readme.txt ........................ the file with CD contents description
exe ..................................... the directory with executables
src ....................................the directory of source codes
  wbdcm ..................................... implementation sources
  thesis ..............the directory of LaTeX source codes of the thesis
text .......................................... the thesis text directory
  thesis.pdf ..........................the thesis text in PDF format
  thesis.ps.............................the thesis text in PS format
```