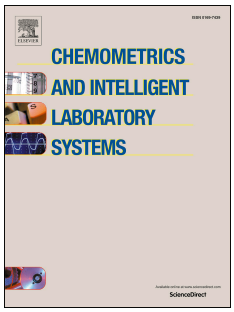
Accepted Manuscript



Robust and sparse estimation methods for high-dimensional linear and logistic regression

Fatma Sevinç Kurnaz, Irene Hoffmann, Peter Filzmoser

PII: S0169-7439(17)30124-7

DOI: 10.1016/j.chemolab.2017.11.017

Reference: CHEMOM 3550

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 17 February 2017

Revised Date: 22 November 2017

Accepted Date: 28 November 2017

Please cite this article as: Fatma.Sevinç. Kurnaz, I. Hoffmann, P. Filzmoser, Robust and sparse estimation methods for high-dimensional linear and logistic regression, *Chemometrics and Intelligent Laboratory Systems* (2017), doi: 10.1016/j.chemolab.2017.11.017.

Robust and sparse estimation methods for high-dimensional linear and logistic regression

Fatma Sevinç Kurnaz[a], Irene Hoffmann[b], Peter Filzmoser[b]

[a] *Department of Statistics, Yildiz Technical University, 34220, Istanbul, Turkey*
[b] *Institute of Statistics and Mathematical Method in Economics, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040, Vienna, Austria*

Abstract

Fully robust versions of the elastic net estimator are introduced for linear and logistic regression. The algorithms used to compute the estimators are based on the idea of repeatedly applying the non-robust classical estimators to data subsets only. It is shown how outlier-free subsets can be identified efficiently, and how appropriate tuning parameters for the elastic net penalties can be selected. A final reweighting step improves the efficiency of the estimators. Simulation studies compare with non-robust and other competing robust estimators and reveal the superiority of the newly proposed methods. This is also supported by a reasonable computation time and by good performance in real data examples.

Keywords: Elastic net penalty, Least trimmed squares, C-step algorithm, High-dimensional data, Robustness, Sparse estimation

1. Introduction

Let us consider the linear regression model which assumes the linear relationship between the predictors $\mathbf{X} \in \mathbb{R}^{n \times p}$ and the predictand $\mathbf{y} \in \mathbb{R}^{n \times 1}$,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are the regression coefficients and $\boldsymbol{\varepsilon}$ is the error term assumed to have standard normal distribution. For simplicity's sake, we as-

*Corresponding author
Email address: fskurnaz@yildiz.edu.tr, Tel: +902123837267 (Fatma Sevinç Kurnaz)