Journal of Statistical Planning and Inference 142 (2012) 48–64

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Shrinkage and variable selection by polytopes

Sebastian Petry, Gerhard Tutz *

Ludwig-Maximilians-University, Department of Statistics, Ludwigstraße 33, 80539 Munich, Germany

ARTICLE INFO

Article history:
Received 14 July 2010
Received in revised form
17 June 2011
Accepted 20 June 2011
Available online 29 June 2011

Keywords:
Constraint regions
Polytopes
LASSO
Elastic net
OSCAR

ABSTRACT

Constrained estimators that enforce variable selection and grouping of highly correlated data have been shown to be successful in finding sparse representations and obtaining good performance in prediction. We consider polytopes as a general class of compact and convex constraint regions. Well-established procedures like LASSO (Tibshirani, 1996) or OSCAR (Bondell and Reich, 2008) are shown to be based on specific subclasses of polytopes. The general framework of polytopes can be used to investigate the geometric structure that underlies these procedures. Moreover, we propose a specifically designed class of polytopes that enforces variable selection and grouping. Simulation studies and an application illustrate the usefulness of the proposed method.

1. Introduction

We consider the linear normal regression model

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon}\sim \mathrm{N}_n(\mathbf{0},\sigma^2\mathbf{I}),$$

where the response $\mathbf{y}=(y_1,\ldots,y_n)^T$ and the design $\mathbf{X}=(\mathbf{x}_1|\ldots|\mathbf{x}_p)$ are based on n iid observations. Since the methods considered are not equivariant we will use standardized data. Therefore, $\mathbf{y}=(y_1,\ldots,y_n)^T$ is the centered response and $\mathbf{x}_j=(x_{1j},\ldots,x_{nj})^T$ the j-th standardized predictor, $j\in\{1,\ldots,p\}$, so that

$$\sum_{i=1}^{n} y_i=0, \quad \sum_{i=1}^{n} x_{ij}=0, \quad \sum_{i=1}^{n} x_{ij}^2=1, \ \forall j\in\{1,\ldots,p\}$$

holds.

In normal distribution regression problems one typically uses the *ordinary least squares estimator* $\widehat{\boldsymbol{\beta}}_{OLS}$. The underlying loss function is the *quadratic loss* or *sum of squares*:

$$\mathrm{Q}(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) := \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2$$

and $\widehat{\boldsymbol{\beta}}_{OLS}$ minimizes the unconstrained regression problem:

$$\widehat{\boldsymbol{\beta}}_{OLS}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ \mathrm{Q}(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}).$$

When c is appropriately chosen the contours of the quadratic loss

$$S_c(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})=\{\boldsymbol{\beta}\in\mathbb{R}^p : \mathrm{Q}(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})\le c\}$$

* Corresponding author.
E-mail addresses: sebastian.petry@stat.uni-muenchen.de (S. Petry), gerhard.tutz@stat.uni-muenchen.de (G. Tutz).

0378-3758/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2011.06.020

form hyperellipsoids centered at $\widehat{\boldsymbol{\beta}}_{OLS}$. Moreover, $Q(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})$ is an upper semicontinuous and strictly convex, which are properties that guarantee a unique solution of constrained estimates.

Constraining the domain of $\boldsymbol{\beta}$ can be motivated by non-sample information given by some scientific theory. For example in economical input–output-systems it is assumed that the inputs have a positive influence on the output. Then the domain of the estimate is restricted by $\beta_{input} > 0$. More general, there is a mathematical motivation to constrain the parameter domain of a regression problem. James and Stein (1961) proposed the first *shrinkage estimator* which became known in the literature as James–Stein-estimator. The expression "shrinkage" is due to the geometrical interpretation of Hoerl and Kennard (1970). Hoerl and Kennard (1970) described that the length of the OLS-vector $|\widehat{\boldsymbol{\beta}}_{OLS}|$ tends to be longer than the length of the true parameter vector $|\boldsymbol{\beta}_{true}|$. This effect can be overcome by restricting the parameter domain to a centrosymmetric region around the origin of the parameter space.

Hoerl and Kennard (1970) used centered *p*-dimensional spheres with radius *t* which yields *ridge regression*. Centrosymmetric regions around the origin are a general concept to compensate for the "$|\boldsymbol{\beta}_{true}| < |\widehat{\boldsymbol{\beta}}_{OLS}|$-effect" since the properties of the loss function $Q(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})$ together with compactness and convexity of the domain guarantee existence and uniqueness of the solution. In the following we will call regions with the three properties convexity, compactness, and centrosymmetry *penalty regions*.

The term penalty region is commonly used when the problem is represented in its penalized form. For some constrained regression problems there exist alternative formulations which have equivalent solutions. For example, the *constrained version* of the ridge estimator is

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2, \quad \text{s.t.} \sum_{j=1}^{p} \beta_j^2 \le t, \;\; t \ge 0. \tag{1}$$

For fixed *t* the corresponding *penalized regression problem* has the form

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \quad \lambda \ge 0. \tag{2}$$

The proof of the equivalence is based on the theory of Lagrangian multipliers and can be found in Luenberger (1969) where the equivalence for a set of constraints is shown by using a vector $\boldsymbol{\lambda}^T \in \mathbb{R}^p$. It should be noted that not every constrained regression problem can be given as a penalized regression problem.

It is intuitively clear that a penalty region determines the properties of the estimate beyond tackling the "$|\boldsymbol{\beta}_{true}| < |\widehat{\boldsymbol{\beta}}_{OLS}|$-problem". Therefore the penalty regions should be carefully designed. We will focus on two properties of estimates.

Variable selection: Coefficients whose corresponding predictors have vanishing or low influence on the response should be shrunk to zero.

Grouping: For a group of highly correlated variables it can be advantageous that estimated coefficients differ not too strongly.

A well-established shrinkage procedure that includes variable selection is the LASSO (Tibshirani, 1996). One criticism of the LASSO, which has been pointed out by Zou and Hastie (2005), is the behavior when predictors are highly correlated. In that case the LASSO tends to select only one or two from the group of the correlated influential predictors. Therefore, Zou and Hastie (2005) proposed the *Elastic Net* (*EN*) which tends to include the whole group of highly correlated predictors. The EN enforces the grouping effect as stated in Theorem 1 of Zou and Hastie (2005) where a relation between sample correlation and grouping was given. The EN does not use the sample correlation explicitly, the grouping effect is achieved by a second penalty term together with a second tuning parameter which does not depend on the sample correlation. In a similar way Bondell and Reich (2008) introduced the OSCAR by including an alternative penalty term that enforces grouping. OSCAR also selects variables and shows the grouping effect. Also a relation between sample correlation and grouping may be derived. An alternative penalty that explicitly uses the correlation and enforces the grouping property was proposed by Tutz and Ulbricht (2009) under the name correlation-based penalty. Variable selection was obtained by combining boosting techniques with the correlation-based penalty.

We will consider established procedures within the general framework of constraint regions based on polytopes and introduce a correlation-based penalty region called V8, which groups and selects variables. In Section 2 we give some basic concepts of polytope theory. Based on these concepts the LASSO is discussed in Section 2.2 and OSCAR in Section 2.3. The embedding into the framework of polytopes allows to derive some new results for these procedures. In Section 3 we introduce the V8 procedure and give algorithms that solve the constrained least squares problem. In Section 4 the V8 procedure is compared to established procedures on the basis of simulations.

2. Polytopes as constraint region

Polytopes provide a simple class of compact and convex regions that are useful as constraint regions. They were implicitly used in established regression procedures like LASSO (Tibshirani, 1996) or OSCAR (Bondell and Reich, 2008). In general, polytopal constrained regression problems can be reformulated as linear constrained regression problems (cf. Theorem 1). But in practice it can be hard to reformulate the polytopal constrained regression problem as a linear constrained problem. One objective of this article is to use geometrical arguments for analyzing and designing polytopal penalty regions. In the following the geometric background and the mathematical foundation of polytopes is shortly sketched.