LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Gerhard Tutz & Sebastian Petry

Shrinkage and Variable Selection by Polytopes

Technical Report Number 053, 2009
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Shrinkage and Variable Selection by Polytopes

Sebastian Petry and Gerhard Tutz

April 7, 2009

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.

Keywords: constraint regions, polytopes, lasso, elastic net, oscar

1 Introduction

We consider the linear normal regression model

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

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

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

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*

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

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

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

When c is appropriately chosen the contours of the quadratic loss

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

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

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

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

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

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 of 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 (of highly correlated predictors): Predictors that are highly correlated should have (nearly) identical coefficients. This implies a selection of groups of highly correlated variables.

A well-established shrinkage procedure that includes variable selection is the LASSO (Tibshirani, 1996). One criticism of the LASSO, which has been pointed out be Zou and Hastie (2005), is the behaviour 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 do 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 contrast to the LASSO, the EN, and the OSCAR the underlying penalty region is data driven. 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.

2.1 Some Concepts in Polytope Theory

Let in general $\mathbf{a} \leq \mathbf{b}$ denote that $a_r \leq b_r$ for all components of $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$. In the following *hyperplanes* and corresponding *halfspaces* play an important role. Definitions are given in the Appendix (see Definition A 1).

Polytopes are a class of fundamental geometric objects defined in $\mathbb{R}^p$. The dimension of a polytope is the dimension of its affine hull and a p-dimensional polytope is called *p-polytope*. There are two ways to describe polytopes: *V-polytopes* and *H-polytopes*.

Definition 1 (V-Polytope) *A* V-Polytope *is the convex hull of a finite point set* $\mathcal{V} \subset \mathbb{R}^p$*:*

$$P(\mathcal{V}) := conv(\mathcal{V}).$$

Definition 2 (H-Polytope) *A subset* $P \subset \mathbb{R}^p$ *is called an* H-Polytope *if it is the bounded intersection of a finite number of closed lower linear halfspaces. For* $\mathbf{A} \in \mathbb{R}^{m \times p}$, $\mathbf{t} \in \mathbb{R}^m$

$$P(\mathbf{A}, \mathbf{t}) := \{\mathbf{x} \in \mathbb{R}^p : \mathbf{A}\mathbf{x} \leq \mathbf{t}\}$$

describes an H-Polytope if $P(\mathbf{A}, \mathbf{t})$ *is bounded.*

The intuitive question is whether there exists a relation between H-polytopes and V-polytopes. The answer is given in Ziegler (1994) where the following theorem is shown to hold.

Theorem 1 (Main Theorem) *A subset* $P \subseteq \mathbb{R}^p$ *is the convex hull of a finite point set (a V-Polytope)*

$$P = conv(\mathcal{V}), \qquad \text{for some } \mathcal{V} \subset \mathbb{R}^{p \times n}$$

if and only if it is a bounded intersection of closed (lower linear) halfspaces (an H-Polytope)

$$P = P(\mathbf{A}, \mathbf{t}) = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{A}\mathbf{x} \leq \mathbf{t}\}, \quad \text{for some } \mathbf{A} \in \mathbb{R}^{m \times p}, \mathbf{t} \in \mathbb{R}^m.$$

However, the transformation from H- to V-representation and vice versa can be computationally expensive. The number of producing halfspaces and of vertices is an indicator for the computational costs.

Each row of the system of inequalities $\mathbf{A}\mathbf{x} \leq \mathbf{t}$ describes a linear lower closed halfspace. It represents the normal vector of a hyperplane generating a corresponding halfspace. A *vertex* of a p-polytope P is an element $\mathbf{v} \in P$ which can not be given as a convex combination of the remaining elements $P \setminus \{\mathbf{v}\}$ (see Figure 1 where the five vertices are easily identified). Although in Definition 1 a general finite set $\mathcal{V}$ is used to describe P it is sufficient to use only the vertices of P to define the same polytope P. In other words, let $P = conv(\mathcal{V})$ be a V-polytope and $E(\mathcal{V}) \subseteq \mathcal{V}$ be the set of all vertices of P then $P = conv(\mathcal{V}) = conv(E(\mathcal{V}))$ holds. We assume $\mathcal{V} = E(\mathcal{V})$ in the following. It is obvious that every point $\mathbf{x}$ of a polytope $P = conv(\mathcal{V})$ can be presented as the convex combination of all vertices,

$$\mathbf{x} = \sum_{i \in I} \lambda_i \mathbf{v}_i,\ \lambda_i \geq 0, \sum_{i \in I} \lambda_i = 1,\ \mathbf{v}_i \in \mathcal{V}, \tag{3}$$

where I is the index set of all vertices. In addition, we only consider H-polytopes whose description is not redundant. This means the leaving out of any row of $\mathbf{A}\mathbf{x} \leq \mathbf{t}$ will change the polytope.

Alternatively, a polytope can be described by its faces. The definition of faces of a polytope are based on *supporting hyperplanes* or shortly *supports* (for a definition see Def.2 in the Appendix).

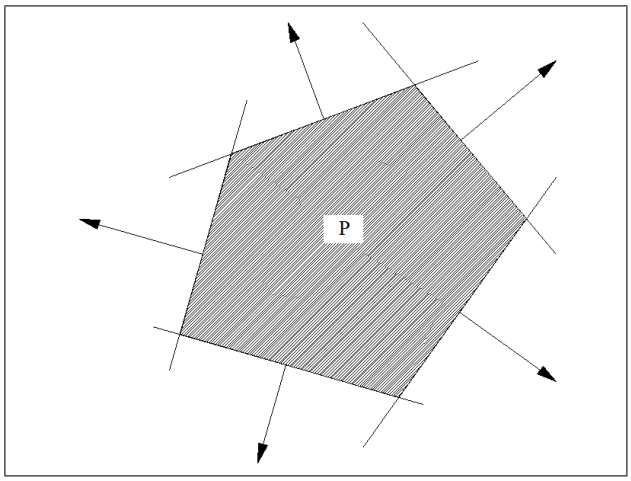

Figure 1: Illustration of the H- and V-representation of the Polytope P. The closed lower halfspaces are on the opposite side of the illustrated normal vectors. The intersection of these halfspaces is P which is shown by the shaded area. The graphed intersection of the hyperplanes are the five vertices of P. The convex hull of the five vertices produces the same polytope P.

Definition 3 (Faces of a Polytope) *Let $P \subset \mathbb{R}^p$ be a p-polytope and $H \subset \mathbb{R}^p$ be a support of P. Then the intersection $P \cap H$ is called* face of P. *A k-dimensional face is called k-*face. *A 0-face is a vertex, an 1-face is an edge, and a $(p-1)$-face is a facet.*

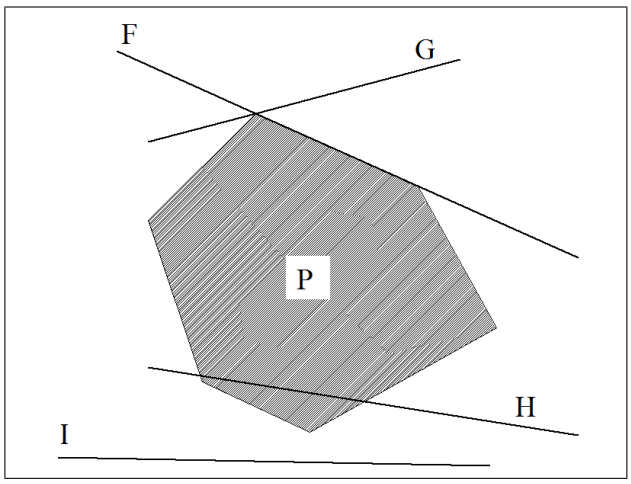

Figure 2: Illustration of Definition 3. Four hyperplanes F, G, H, I and their relationship to P: I is not a support because $I \cap P = \emptyset$. H is not a support because P is not entirely contained in one of the both closed halfspaces H^+ or H^-. F and G are supports. $P \cap G$ is a vertex of P. $P \cap F$**?** $G \cap F$ is a facet of P. (In $\mathbb{R}^2$ an edge is also a hyperplane.)

An important feature is that every face is a convex hull of vertices but not every convex hull of vertices is a face. Hence, not every convex combination of vertices lies on the surface of the polytope, but every facet is the convex hull of $q \geq p$

vertices (see Ziegler (1994)). The linear hull of these q vertices is the intersecting support which produces this facet. The intersecting support is given by one row of $\mathbf{Ax} \leq \mathbf{t}$ in Definition 2. A p-polytope P is called *simplicial* iff every facet of P contains the minimal number of p vertices.

The special class of polytopes which is of interest here is the following.

Definition 4 *A p-polytope P is called* centrosymmetric*, if*

1. *the origin is an inner point of P: $\mathbf{0} \in P$.*
2. *If $\mathbf{v} \in P$ then $-1 \cdot \mathbf{v} \in P$.*

It is intuitively clear that a centrosymmetric p-polytope can be scaled up or down in two ways

1. Multiplying the right hand side of $\mathbf{Ax} \leq \mathbf{t}$ with $s > 0$.
2. Multiplying all vertices with $s > 0$.

2.2 LASSO

The famous LASSO, proposed by Tibshirani (1996), is very popular because of its variable selection property and has been used in many fields of statistics. The LASSO constraint region is given by

$$\sum_{j=1}^{p} |\beta_j| \leq t,\ t > 0, \tag{4}$$

which corresponds to a p-polytope. The H-representation of the constraint region is obtained by solving the absolute value function $|.|$ in (4). The result is a system of inequations

$$\mathbf{L}\boldsymbol{\beta} \leq \mathbf{t}, \tag{5}$$

where $\mathbf{L}$ is a $(2^p \times p)$-matrix. Each row of $\mathbf{L}$ is one of the 2^p variations of entries -1 or $+1$ and $\mathbf{t}$ is a 2^p-dimensional vector whose entries are equal to $t > 0$. An example for the case $p = 3$ can be found in the Appendix (Example A 1). More concise, the LASSO constraint region is a p- crosspolytope, which is scaled up or down by the tuning parameter $t > 0$. (For the definition of a p-crosspolytope see Ziegler (1994), p. 8.). The underlying polytope is simplicial and this property is maintained by scaling up or down. An illustration in $\mathbb{R}^3$ is given in Figure 3.

The vertices of the LASSO penalty region are

$$\mathcal{L} = \{t \cdot \mathbf{e}_1,\ -t \cdot \mathbf{e}_1,\ ...,\ t \cdot \mathbf{e}_p,\ -t \cdot \mathbf{e}_p,\ t > 0\}, \tag{6}$$

where $\mathbf{e}_j,\ j = 1, ..., p$, denotes the j-th unit vector of $\mathbb{R}^p$. Therefore the V-representation of the LASSO penalty region is $P = conv(\mathcal{L})$.

Since the constraint (4) is determined by the 2^p constraints specified in the rows of (5), it is easy to transform the LASSO problem in constrained form,

$$\widehat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{\beta}}{\text{argmin}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2, \ s.t. \ \sum_{j=1}^{p} |\beta_j| \leq t, \ t \geq 0,$$

into a penalized regression problem,

$$\widehat{\boldsymbol{\beta}}_L = \underset{\boldsymbol{\beta}}{\text{argmin}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \ \lambda \geq 0.$$

If the OLS estimate exists and $\sum_{j=1}^{p} |\beta_{OLS_j}| = t_0$ then $\widehat{\boldsymbol{\beta}}_L$ is the contact point of the contour of the loss function $S_c(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ and the penalty region $\sum_{j=1}^{p} |\beta_j| \leq t$, $0 < t < t_0$. The variable selection property of the LASSO can be illustrated by using the V-representation. Although not all convex combinations of vertices are on the surface the solution of a polytopal constrained regression problem lies on the surface. So with respect to the simpliciality of the LASSO penalty region variable selection is performed if the solution is a convex combination of less than p vertices of its penalty region, i.e. at least one of the λ_is in (3) is zero. Thus, in $\mathbb{R}^3$ one can distinguish three cases of LASSO solutions:

1. If the LASSO solution lies on a vertex only one coefficient is nonzero, i.e. only one λ_i in (3) is 1.

2. If the LASSO solution lies on an edge that connects two axes, two λ_i's in (3) are non-zero.

3. If the LASSO solution lies on a facet, three λ_i's in (3) are non-zero.

In the first two cases variables are selected.

2.3 OSCAR

Bondell and Reich (2008) proposed a shrinkage methods called OSCAR, which stands for **O**ctagonal **S**hrinkage and **C**lustering **A**lgorithm for **R**egression. Its constraint region is

$$\sum_{j=1}^{p} \left[|\beta_j| + c \cdot \sum_{j<k} \max \{|\beta_j|, \ |\beta_k|\} \right] \leq t. \tag{7}$$

Bondell and Reich (2008) also give an alternative representation of their penalty region. Let $|\beta|_{(k)}$ denote the absolute value of the component of $\boldsymbol{\beta} \in \mathbb{R}^p$ whose

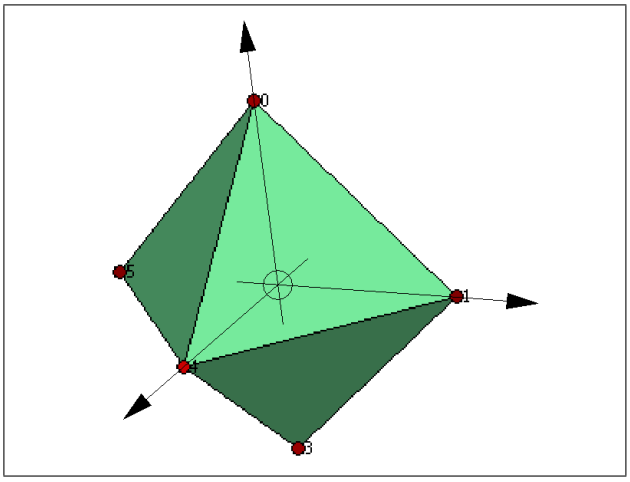

Figure 3: LASSO constraint region in $\mathbb{R}^3$.

rank is k so that $|\beta|_{(1)} \leq |\beta|_{(2)} \leq \ldots \leq |\beta|_{(p)}$ holds. With $|\beta|_{(.)}$ the OSCAR penalty region (7) is equivalent to

$$\left[\sum_{j=1}^{p} c(j-1)+1\right] \cdot |\beta|_{(j)} \leq t. \tag{8}$$

First we discuss the penalty region in the implicitly given H-representation. Then we derive the vertices as a new result. That is helpful because the V-representation allows an alternative perspective on the grouping property of OSCAR.

The analysis of the OSCAR penalty region in H-representation is based on segmentation of the p-dimensional parameter space $\mathbb{R}^p$. First we partition $\mathbb{R}^p$ in the 2^p *orthants*, which are regions for which the signs of components are fixed. Second we segment every orthant in $p!$ regions which are defined by a fixed order of ranks of $|\beta_j|$, $j = 1, ..., p$. Figure 4 illustrates the segmentation for one orthant in $\mathbb{R}^3$.

The absolute value function $|.|$ in the OSCAR penalty term corresponds to the orthants and the segmentation of each orthant is given by the sum of pairwise maximum norms. It is seen from (8) that the OSCAR penalty region is an H-polytope which depends on the order of ranks of $|\beta_j|$ and on the sign constellation with the order of ranks being linked to the weights $[c(j-1)+1]$.

For the derivation of the penalty region, $P(\mathbf{A}(c), \mathbf{t})$, we consider first the orthant with only positive signs. For this orthant we create a $(p! \times p)$-matrix $\widetilde{\mathbf{A}}(c)$ where every row represents one of the $p!$ permutation of the p weights $[c(j-1)+1]$, $j = 1, ..., p$. In a second step we form $(2^p - 1)$ matrices $\widetilde{\mathbf{A}}(c)$, which are constructed by changing the sign in one column of $\widetilde{\mathbf{A}}(c)$. Finally we combine these matrices obtaining the $(2^p \cdot p!) \times p$-matrix $\mathbf{A}(c)$. The matrices built

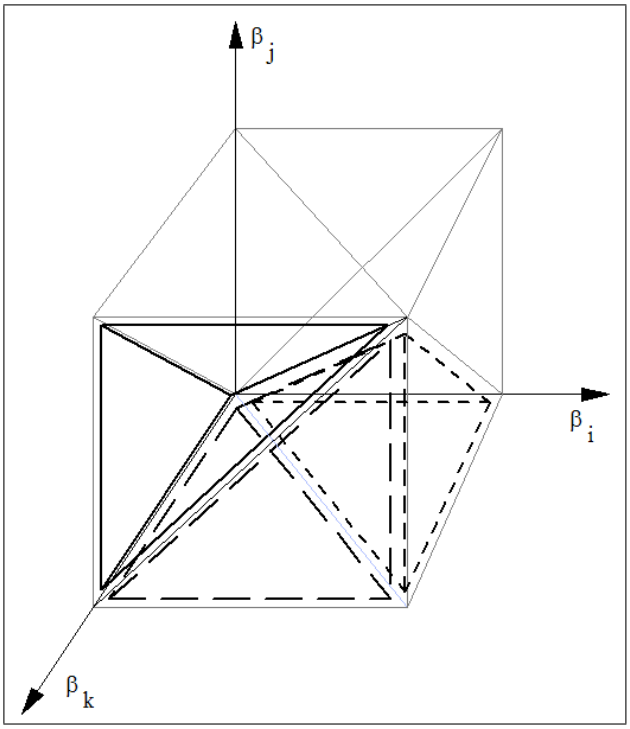

Figure 4: The region described by the shortly dashed lines corresponds to the ordering $|\beta_j| \leq |\beta_k| \leq |\beta_i|$, long dashed lines correspond to the ordering $|\beta_j| \leq |\beta_i| \leq |\beta_k|$ holds and solid lines correspond to the ordering $|\beta_i| \leq |\beta_j| \leq |\beta_k|$.

in the second step correspond to the orthants. Example A 2 in the Appendix shows the H-representation of an OSCAR penalty region.

Therefore the OSCAR penalty region with the tuning parameters $t > 0$ and $c > 0$ is represented by the intersection of $2^p \cdot p!$ hyperplanes, which shows the high complexity of the OSCAR penalty region. It is remarkable that the $2^p \cdot p!$ constraints sum up to one constraint given in (7).

On OSCAR's Vertices

Hitherto the OSCAR penalty region is considered only as a H-polytope. The Main Theorem (Theorem 1) suggests to consider the OSCAR penalty region as a V-polytope. The vertices of the OSCAR penalty region have a simple structure which is given in the following proposition.

Proposition 1 *Let an p-dimensional OSCAR penalty region with the tuning parameters $t > 0$ and $c > 0$ be given. Then the set of vertices of the OSCAR penalty region is the set of points with the following properties:*

1. *From the p components $1 \leq m \leq p$ components are nonzero and the absolute value of these components is equal. The remaining $p - m$ components are zero.*

2. *The $1 \leq m \leq p$ nonzero components of a vertex have the absolute value*

$$v(m) := \frac{t}{\sum_{j=p+1-m}^{p}[c(j-1)+1]}. \tag{9}$$

For the *proof* see Appendix (Proof A 1).

Corollary 1 *Under the conditions of Proposition 1 one obtains:*

1. *The OSCAR penalty region is the convex hull of $3^p - 1$ vertices,*

2. *The OSCAR penalty region is simplicial.*

For the *proof* see Appendix (Proof A 2). It is remarkable that (9) depends not only on the penalty level t and the tuning parameter c but also on the dimension of the problem p.

Figure 5 shows the OSCAR penalty region for different tuning parameters. For fixed tuning parameter t and p (9) becomes smaller by increasing c. So for graphical illustration we adjust t so that the axis intercepts are equal. The first

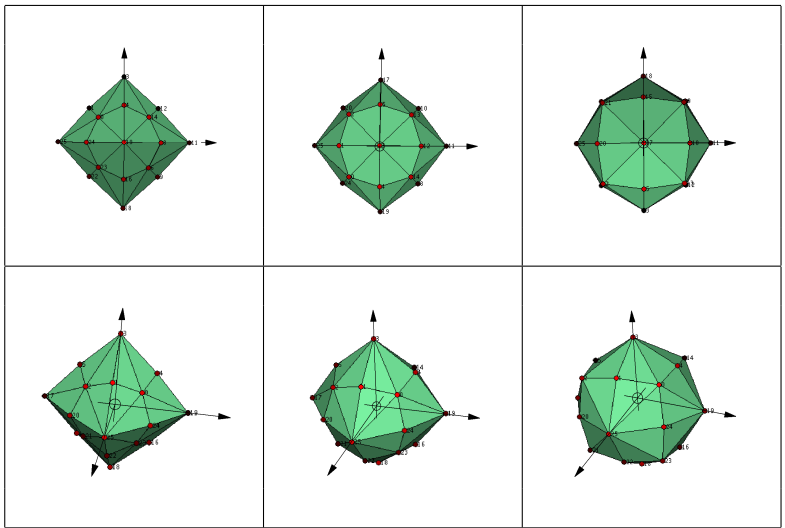

Figure 5: The OSCAR penalty region with three different tuning parameter c. In the first row the projections in to a β_i-β_j-plane is shown. In the second row a oblique view is shown.

row of Figure 5 explains the naming of OSCAR. It illustrates that orthogonal projections of an OSCAR penalty region on any β_i-β_j-plane form an octagon, which may be shown by using orthogonal projections of the vertices on any β_i-β_j-plane. Because of symmetry, in Figure 5 only one projection is shown. For further illustration the set of all vertices of an OSCAR penalty region in the case $p = 3$ are given in the Appendix (Example A 3).

In general, the parameter c controls the form of the OSCAR penalty region. For $c \to 0$ it converges to the LASSO penalty region. This can be shown by considering the limit $c \to 0$ within the system of inequations. It is noteworthy that for $c \to \infty$ and $p > 2$ the OSCAR penalty region does not converge to a p-dimensional cube (p-cube), which would enforce extreme grouping but no variable selection. A p-cube would make sense only if all predictors were very highly correlated. Rather for $c \to \infty$ the OSCAR converges to a specific polytope. This can be derived by considering the following limit: $\lim_{c\to\infty} v(m_1)/v(m_2) = (\sum_{j=p+1-m_2}^{p}(j-1))/(\sum_{j=p+1-m_1}^{p}(j-1))$, where $v(.)$ is given by (9). In the limit the ratio $v(m_1)/v(m_2)$ depends only on m_1 and m_2, the different numbers of nonzero components of vertices. Hence for $c \to \infty$ the form of the OSCAR polytope is fixed but does not converge to a p-cube.

Bondell and Reich (2008) describe the grouping (or clustering) property of OSCAR by giving a relation between correlation and grouping. Another perspective on the properties variable selection and grouping is obtained by considering vertices. From Figure 5 it is seen that grouping of three variables is forced by the vertices in the middle of the orthants. In general, for grouping of more than two predictors vertices with more than two nonzero components seem to be necessary. Grouping or variable selection is performed if less than p vertices take part in the convex combination of the OSCAR solution. Bondell and Reich (2008) give an upper bound criterion for the relationship between the tuning parameter c and the correlation of predictors but they do not use correlation directly for generating the penalty.

3 The V8 procedure

In the following a correlation driven polytope is proposed, which uses the correlation within data to define the penalty region.

3.1 The V8-polytope

The new V8-polytope is called V8 because it is a **V**-polytope for which projections on any β_i-β_j-plane are octagons. The construction focuses on the grouping property, which was advocated in particular by Zou and Hastie (2005) and is behind OSCAR (Bondell and Reich (2008)) and correlation-based penalties (Tutz and Ulbricht (2009)). It means especially that if two standardized variables are highly correlated then the estimate of their coefficients should be (nearly) equal apart from the sign. From a geometrical point of view this means: if two variables $\mathbf{x}_i$ and $\mathbf{x}_j$ are highly correlated the estimated coefficients should lie on the face of a polytopal penalty region where $|\beta_i| = |\beta_j|$ holds. This suggests to design correlation driven polytopes where the correlation between predictors determines the form of the polytope. Then no second tuning parameter is needed in contrast

to OSCAR which uses two tuning parameters.

The V8-polytope should feature the following properties:

(P1) The orthogonal projection of the polytope on every β_i-β_j-plane, $1 \leq i \leq j \leq p$, is a (convex) octagon.

(P2) The octagons are centrosymmetric.

(P3) Four vertices of each octagon lie on the axis at the values $\pm t$, two on the β_i-axis and two on the β_j-axis.

(P4) The four remaining vertices are on the bisecting line of the β_i-β_j-plane where $|\beta_i| = |\beta_j|$.

The OSCAR penalty region shares all of these properties, which may be shown by projecting the vertices of the OSCAR penalty region on any β_i-β_j-plane. For the V8-polytope in addition the penalty region is supposed to depend on the estimated correlation between two predictors, $\rho_{ij} := corr(\mathbf{x}_i, \mathbf{x}_j)$ by use of a function $c : [-1, 1] \mapsto [0, 1]$. In general, every function $c(\rho_{ij})$ with the following properties is appropriate:

(1) $c(0) = 1$.

(2) $c(1) = c(-1) = 0$.

(3) $c(\rho_{ij}) = c(-\rho_{ij})$.

(4) $c(.)$ is increasing in $[-1, 0]$ and decreasing in $(0, 1]$.

In the following we will use $c(\rho_{ij}) := 1 - |\rho_{ij}|$.

The vertices described by (P3) are defined as the same vertices as for the LASSO $\mathcal{L}$ and do not depend on the correlation. The vertices characterized by (P4) for any β_i-β_j-plane, $1 \leq i \leq j \leq p$, are specified by

$$\mathcal{B}_{ij} = \left\{ \mathbf{b} \in \mathbb{R}^p : |b_i| = \frac{t}{1 + c(\rho_{ij})},\ |b_j| = \frac{t}{1 + c(\rho_{ij})},\ b_k = 0,\ k \neq i,\ j \right\}.$$

It is obvious that $|\mathcal{B}_{ij}| = 4$. The assumptions (1)–(4) of the function $c(.)$ induce the following properties on $\mathcal{B}_{ij}$. If $\rho_{ij} \to 0$ the elements of $\mathcal{B}_{ij}$ become redundant because they are convex combinations of $\mathcal{L}$. The projection on any β_i-β_j-plane converges to a diamond with side length $\sqrt{2}t$ and so variable selection is enforced. For $|\rho_{ij}| \to 1$ the four elements $\{+t\mathbf{e}_i,\ -t\mathbf{e}_i,\ +t\mathbf{e}_j,\ -t\mathbf{e}_j\} \subset \mathcal{L}$ become redundant because they are convex combinations of $\mathcal{B}_{ij}$. In this case the octagon converges to a square with side length $2t$ and grouping of the variables $\mathbf{x}_i$ and $\mathbf{x}_j$ is enforced. This behaviour is illustrated in the first row of Figure 6. With $\mathcal{B} = \bigcup_{i<j} \mathcal{B}_{ij}$ the vertices of the V8 penalty region are $\mathcal{V} = \mathcal{L} \cup \mathcal{B}$. There are $\binom{p}{2}$ different sets $\mathcal{B}_{ij}$,

and so $|\mathcal{V}| = 2p + 4 \cdot \binom{p}{2} = 2p^2$. An example for the case $p = 4$ is given in the Appendix (Example A 4). It is obvious that $\mathcal{V}$ is convex and that for $\rho_{ij} = 0$, $\forall i \neq j$, the V8 penalty region is the same as for LASSO. Figure 7 illustrates the V8 penalty region for correlation structure given by $\rho_{12} = 0.2, \rho_{13} = 0.5, \rho_{23} = 0.8$.

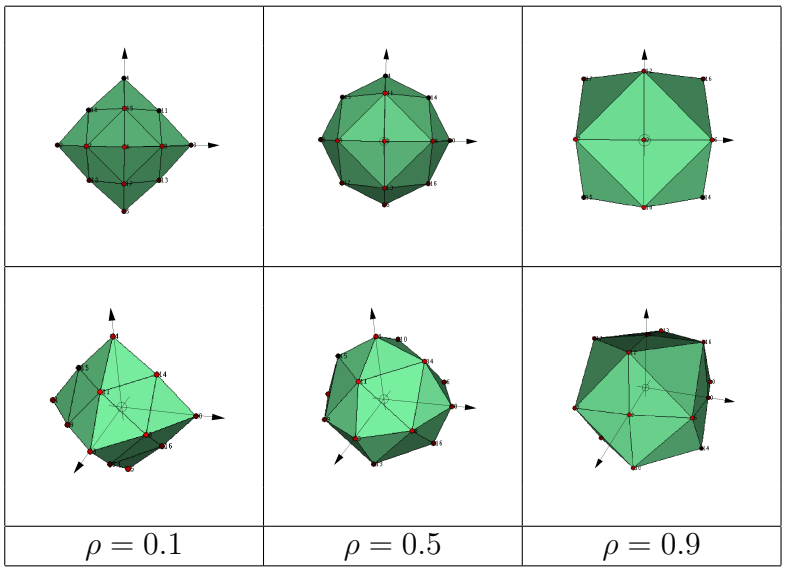

Figure 6: V8-polytopes with unique correlation ρ_{ij} between all pairs i-j.

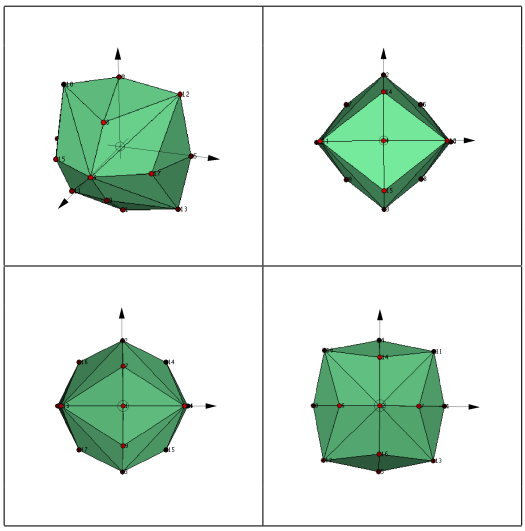

Figure 7: *Top left:* An oblique view of the V8 penalty region. *Top right:* Orthogonal projection on the β_1-β_2-plane where $\rho_{12} = 0.2$. *Bottom left:* Orthogonal projection on the β_1-β_3-plane where $\rho_{13} = 0.5$. *Bottom right:* The orthogonal projection on the β_2-β_3-plane where $\rho_{23} = 0.8$.

In summary, the V8 constraint region enforces variable selection through the LASSO vertices and enforces grouping through the vertices that are added by use of the correlation between two variables.

3.2 Solving Polytopal Constrained Regression Problems

In general, a polytopal constrained regression problem can be formulated as follows:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2, \text{ s.t. } \boldsymbol{\beta} \in P, \tag{10}$$

where P is a polytope. Based on the Main Theorem (see Theorem 1) there are two different ways to formulate (10). If P is an H-polytopes then (10) has the form

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2, \text{ s.t. } \mathbf{A}\boldsymbol{\beta} \leq \mathbf{t}. \tag{11}$$

This is a linearly constrained regression problem with the quadratic loss function which can be solved with established tools like `lsqlin` routine in `MATLAB`.

When P from (10) is a V-polytope let $\mathcal{V} = \{\mathbf{v}_1, ..., \mathbf{v}_{n_V}\}$ denote the set of vertices of P and $I := \{1, ..., n_V\}$ is the index set of $\mathcal{V}$. Every point $\boldsymbol{\beta} \in P$ is a convex combination of elements of $\mathcal{V}$. The convex combination can be written in matrix notation

$$\boldsymbol{\beta} = \mathbf{V} \cdot \boldsymbol{\lambda} \text{ with } \mathbf{V} = (\mathbf{v}_1 | \cdots | \mathbf{v}_{n_V}) \text{ and } \boldsymbol{\lambda} = (\lambda_1, ..., \lambda_{n_V})^T \tag{12}$$

with $\lambda_i \geq 0, \sum_{i \in I} \lambda_i = 1$, $\mathbf{v}_i \in \mathcal{V}$. So (10) turns into a quadratic optimization problem in $\boldsymbol{\lambda}$,

$$\widehat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \ \|\mathbf{y} - \mathbf{X}\mathbf{V} \cdot \boldsymbol{\lambda}\|^2, \text{ s.t. } \lambda_i \geq 0, \sum_{i \in I} \lambda_i = 1, \forall i \in I. \tag{13}$$

For $\widehat{\boldsymbol{\lambda}}$ the estimate $\widehat{\boldsymbol{\beta}}$ is obtained by

$$\widehat{\boldsymbol{\beta}} = \mathbf{V} \cdot \widehat{\boldsymbol{\lambda}}. \tag{14}$$

Since the transformation from H- to V-representation of a polytope can be computationally very expensive it is advisable to use the representation that is available. Thus we need an algorithm to find the optimal convex combination of vertices for solving problem (13).

The definition of centrosymmetry (cf. Definition 4) states $\mathbf{v} \in P \Leftrightarrow -1 \cdot \mathbf{v} \in P$. Thus the set $\mathcal{V}$ of all vertices of a centrosymmetric polytope includes two subsets of vertices $\mathcal{V}^+$ and $\mathcal{V}^-$ for which

$$\begin{aligned} \mathcal{V}^- &= \{-1 \cdot \mathbf{v} : \mathbf{v} \in \mathcal{V}^+\}, & \mathcal{V}^+ &= \{-1 \cdot \mathbf{v} : \mathbf{v} \in \mathcal{V}^-\}, \\ \mathcal{V}^+ \cap \mathcal{V}^- &= \emptyset, & \mathcal{V} &= \mathcal{V}^+ \cup \mathcal{V}^-, \end{aligned} \tag{15}$$

holds. The structure allows to use only one of these two subsets, because each subset is its complement multiplied by -1. The idea is graphically illustrated for $p = 2$ by Figure 8.

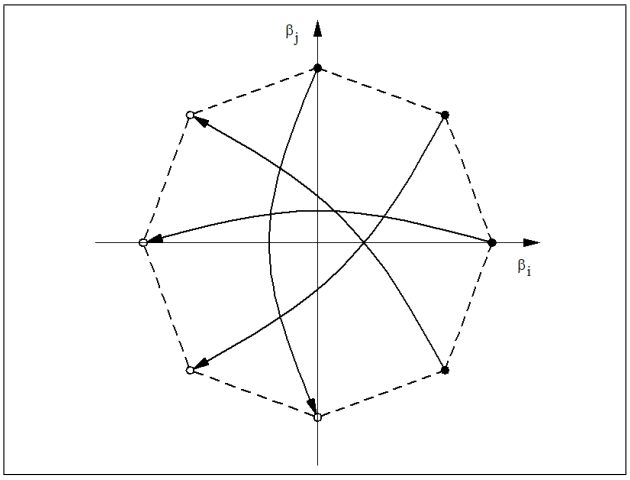

Figure 8: The solid vertices are elements of $\mathcal{V}^+$. The remaining vertices of $\mathcal{V}^-$ are produced by multiplying with -1.

It is obvious that the reduction of the set of vertices changes the constraint in (13). We take $\mathcal{V}^+$ and its index set of vertices $I^+ = \{1, ..., n_{V^+}\}$. With $\mathbf{v}_i^+$, $i \in I^+$, we denote the elements of $\mathcal{V}^+$. Now we structure $\mathcal{V}$ in the following way. The first n_{V^+} elements of $\mathcal{V}$ are equal to $\mathcal{V}^+$ and the second part of $\mathcal{V}$ is given by $\mathbf{v}_{n_{V^+}+i} = -1 \cdot \mathbf{v}_i^+$. Then, subject to the convexity constraint of $\boldsymbol{\lambda}$, for every $\boldsymbol{\beta} \in P$ holds

$$\begin{aligned}
\boldsymbol{\beta} &= \sum_{i \in I} \lambda_i \mathbf{v}_i = \sum_{i \in I^+} \lambda_i \mathbf{v}_i^+ + \sum_{i \in I^+} \lambda_{n_{V^+}+i} \mathbf{v}_{n_{V^+}+i} \\
&= \sum_{i \in I^+} \lambda_i \mathbf{v}_i^+ + \sum_{i \in I^+} \lambda_{n_{V^+}+i} \cdot (-1) \cdot \mathbf{v}_i^+ = \sum_{i \in I^+} (\lambda_i - \lambda_{n_{V^+}+i}) \mathbf{v}_i^+ \\
&= \sum_{i \in I^+} \lambda_i^+ \mathbf{v}_i^+ .
\end{aligned}$$

Due to the convexity constraint of $\boldsymbol{\lambda}$ it is easy to show that $\sum_{i \in I^+} (\lambda_i - \lambda_{n_{V^+}+i}) = \sum_{i \in I^+} \lambda_i^+ \in [-1, +1]$. Analogously to (12) we convey $\mathcal{V}^+$ into a matrix $\mathbf{V}^+ = \left(\mathbf{v}_1^+ | \cdots | \mathbf{v}_{n_{V^+}}^+ \right)$. With the reduced set of vertices (13) turns into

$$\widehat{\boldsymbol{\lambda}}^+ = \underset{\boldsymbol{\lambda}^+}{\operatorname{argmin}} \ \|\mathbf{y} - \mathbf{X}\mathbf{V}^+ \cdot \boldsymbol{\lambda}^+\|^2, \text{ s.t. } \sum_{i \in I^+} |\lambda_i^+| \leq 1 \tag{16}$$

where $\boldsymbol{\lambda}^+ = (\lambda_1^+, ..., \lambda_{n_{V^+}}^+)^T$. Analogously to (14) the estimate $\widehat{\boldsymbol{\beta}}$ is obtained by

$$\widehat{\boldsymbol{\beta}} = \mathbf{V}^+ \cdot \widehat{\boldsymbol{\lambda}}^+ . \tag{17}$$

The constraint $\sum_{i \in I^+} |\lambda_i^+| \leq 1$ in (16) is a LASSO penalty. The equal sign holds if $\widehat{\boldsymbol{\beta}}_{OLS}$ is not a inner point of the constraining polytope. We assume

that the tuning parameter t is appropriately chosen. The constrained regression problem (16) can be solved with the LARS algorithm from Efron et al. (2004) quite efficiently. So if a centrosymmetric V-polytope constrains the quadratic loss function the estimate is given by $\widehat{\boldsymbol{\beta}} = \mathbf{V}^+ \cdot \widehat{\boldsymbol{\lambda}}^+$ with $\widehat{\boldsymbol{\lambda}}^+$ given by (16)

4 Simulation study

In this section we investigate the performance of several methods. All simulations are based on the linear model

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

Each data set consists a training, a validation, and a test data set. The notation $n_{train}/n_{vali}/n_{test}$ is used to describe the number of observation of the corresponding data sets. For each simulation scenario we use 50 replications. For every method we use the following procedure to measure the performance.

We center the response and standardize the predictors of the training data set. $\overline{\mathbf{x}}_{train}^T = (\overline{\mathbf{x}}_{1,train}, ..., \overline{\mathbf{x}}_{p,train})$ denotes the vector of means in the training data set and $\overline{y}_{train}$ is the mean over the responses in the training data set. We use the transformed training data set to fit different models specified by a grid of tuning parameters. By retransformation of the coefficients we get a set of models $\mathcal{M}$. The validation data set is used to determine the model $\widehat{\boldsymbol{\beta}}^{opt} \in \mathcal{M}$ which minimizes the prediction error on the validation data set $PE_{y,vali} = \frac{1}{n_{vali}} \mathbf{r}_{vali}^T \mathbf{r}_{vali}$ with $r_{i,vali} = y_{i,vali} - (\overline{y}_{train} + (\mathbf{x}_{i,vali} - \overline{\mathbf{x}}_{train})^T \widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}} \in \mathcal{M}$. Finally we quantify the performance of $\widehat{\boldsymbol{\beta}}^{opt}$ on the test data set by computing two measures on the test data set: The prediction error on the test data set $PE_{y,test} = \frac{1}{n_{test}} \mathbf{r}_{test}^T \mathbf{r}_{test}$ with $r_{i,test} = y_{i,test} - (\overline{y}_{train} + (\mathbf{x}_{i,test} - \overline{\mathbf{x}}_{train})^T \widehat{\boldsymbol{\beta}}^{opt})$ and the mean squared error for the estimate $MSE_{\boldsymbol{\beta}} = \|\widehat{\boldsymbol{\beta}}_{opt} - \boldsymbol{\beta}_{true}\|^2$. Finally $PE_{y,test}$ and $MSE_{\boldsymbol{\beta}}$ of the 50 replications are illustrated by boxplots. The standard deviation of the medians is calculated by bootstrapping with $B = 500$ bootstrap iterations.

Because we focus on shrinkage procedures with variable selection and grouping property we compare V8, OSCAR, and Elastic Net (EN). It is remarkable that the EN penalty region is not polytopal. We add LASSO in our comparison because it is a special case of these three procedures. For the OSCAR we use the `MATLAB`-code which was available in 2007 on Bondell's homepage. The procedure tuned out to be computational very expensive. Therefore it was not possible to provide OSCAR for all settings.

The settings are described in the following:

(1) Let the underlying parameter vector be $\boldsymbol{\beta}_{true} = (3, 0, 0, 1.5, 0, 0, 0, 2)^T$ and standard error $\sigma = 3$. The correlation between the i-th and j-th

predictor follows

$$corr(i,\, j) = 0.9^{|i-j|}, \forall i,\, j \in \{1,\, ...,\, 8\}. \tag{18}$$

The numbers of observations are 20/20/200. A similar setting is used in the OSCAR paper (Bondell and Reich, 2008).

(2) This setting is the same as the first setting excepting $\boldsymbol{\beta}_{true} = (3,\, 2,\, 1.5,\, 0,\, 0,\, 0,\, 0,\, 0)^T$.

(3) In this setting the correlation is again given by (18) but the coefficient vector is $\boldsymbol{\beta}_{true} = (0.85,\, 0.85,\, 0.85,\, 0.85,\, 0.85,\, 0.85,\, 0.85,\, 0.85)^T$.

(4) In this setting there are $p = 100$ predictors. The parameter vector is structured in blocks,

$$\begin{aligned}\boldsymbol{\beta}_{true} \;=\; (&\underbrace{2,\, ...,\, 2}_{10},\, \underbrace{0,\, ...,\, 0}_{10},\, \underbrace{4,\, ...,\, 4}_{10},\, \underbrace{0,\, ...,\, 0}_{10},\, \underbrace{-2,\, ...,\, -2}_{10},\, \underbrace{0,\, ...,\, 0}_{10},\\ &\underbrace{0,\, ...,\, 0}_{15},\, \underbrace{2,\, ...,\, 2}_{5},\, \underbrace{0,\, ...,\, 0}_{20})^T\end{aligned}$$

and $\sigma = 15$. Between the first six blocks of 10 variables there is no correlation. Within these six blocks we use the correlation structure from (18). The remaining 40 variables are uncorrelated. The numbers of observations are 200/200/1000. As noted above this setting is not analyzed by OSCAR.

(5) The last setting is equal to the forth setting but numbers of observations changes to 50/50/1000.

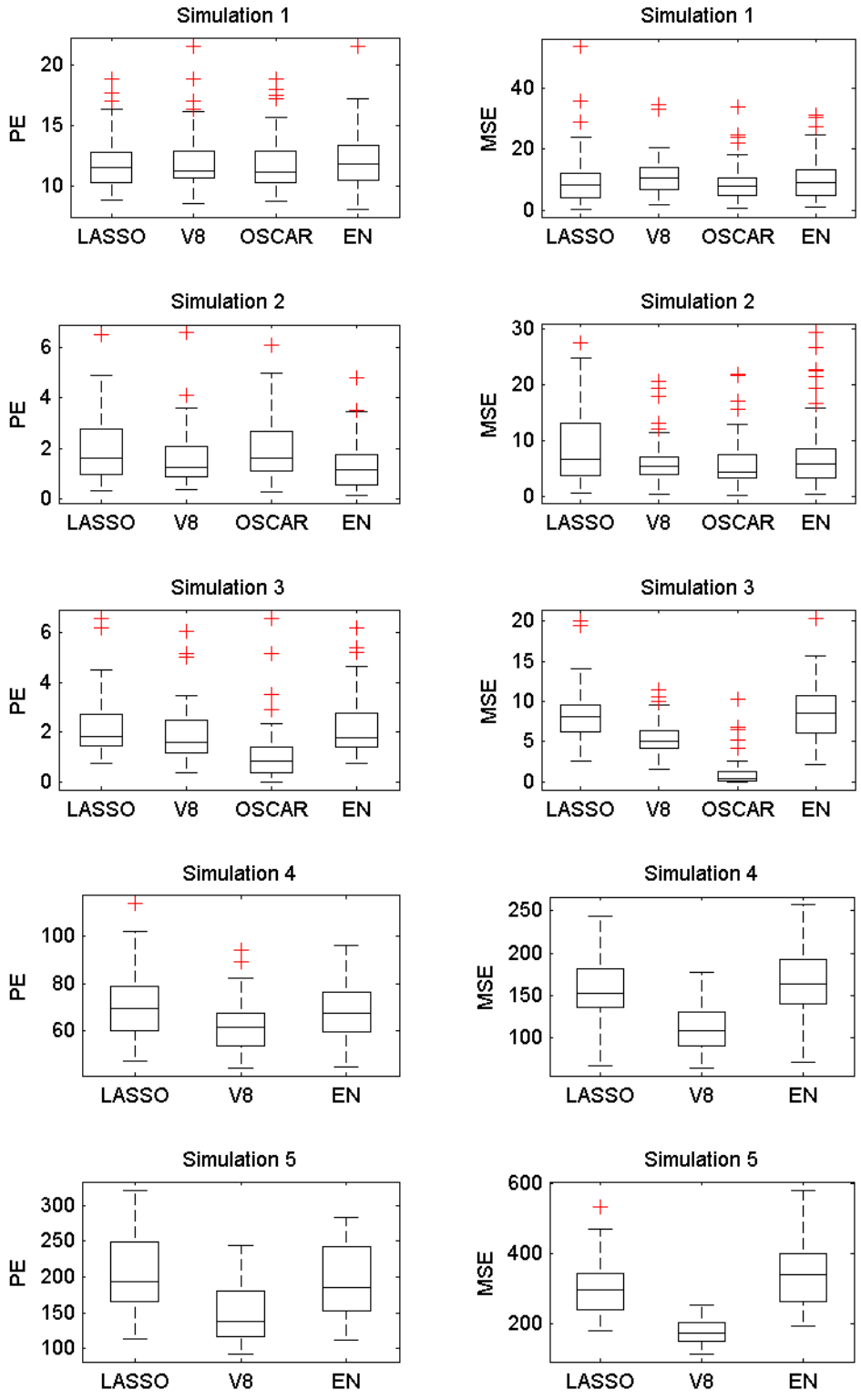

Figure 9: Boxplots of the prediction error on the test data set, $PE_{y,test}$, and MSE of $\boldsymbol{\beta}$, $MSE_{\boldsymbol{\beta}}$, for the different procedures and the five simulation settings.

	Procedure	median(PE) (Std. Err.)	median(MSE) (Std. Err.)
Simulation 1	LASSO	11.522 (0.330)	8.444 (1.081)
	V8	11.279 (0.317)	10.488 (0.877)
	OSCAR	**11.180 (0.337)**	**8.110 (0.945)**
	EN	11.809 (0.401)	9.302 (1.288)
Simulation 2	LASSO	1.630 (0.241)	6.681 (0.970)
	V8	1.256 (0.219)	5.373 (0.257)
	OSCAR	1.594 (0.230)	**4.463 (0.595)**
	EN	**1.169 (0.154)**	5.883 (0.701)
Simulation 3	LASSO	1.847 (0.199)	8.059 (0.406)
	V8	1.586 (0.143)	5.088 (0.356)
	OSCAR	**0.855 (0.158)**	**0.348 (0.067)**
	EN	1.787 (0.170)	8.463 (0.536)
Simulation 4	LASSO	69.687 (1.584)	153.182 (7.143)
	V8	**61.648 (1.958)**	**109.355 (6.946)**
	EN	67.783 (2.736)	164.480 (5.317)
Simulation 5	LASSO	192.837 (12.852)	296.555 (11.757)
	V8	**138.268 (8.502)**	**175.356 (8.342)**
	EN	185.793 (11.275)	341.093 (21.750)

Table 1: Median of prediction error on test test data set and of the MSE of $\boldsymbol{\beta}$ corresponding standard deviations estimated by bootstrapping with 500 bootstrap iterations given in brackets.

It is obvious that Simulation 1 is a challenge for the V8 procedure since V8 tries to group the influential variable with their neighbors which have no influence on the response. Nevertheless, the prediction error (Table 4) is even better than the prediction error of the LASSO and the Elastic Net. As expected, for the second setting the V8 procedure shows better performance. It is the second best in both criteria.

Although we are mainly interested in procedures with variable selection property we chose setting 3 because it was often used in the literature (cf. Bondell and Reich (2008), Tibshirani (1996), Zou and Hastie (2005)). In this somehow artificial setting the OSCAR is the best procedure because it can group all the variables. All other procedures are unable to do this. But the setting shows that adding new vertices to the LASSO penalty yields definitely better results. The performance of the LASSO is topped by both polytopes with additional vertices (OSCAR and V8). Again the V8 is the second best in both criteria.

The two last simulations show that the V8 procedure works quite well especially for the $p >> n$-case. LASSO as well as EN were outperformed by V8. The computational costs of the OSCAR were so high that it was not possible to include it in the competition.

5 Data Example

The body fat data set has been published by Penrose et al. (1985). The aim was to estimate the percentage of body fat of 252 men by use of thirteen regressors. The regressors are age (1), weight (lbs) (2), height (inches) (3), neck circumference (4), chest circumference (5), abdomen 2 circumference (6), hip circumference (7), thigh circumference (8), knee circumference (9), ankle circumference (10), biceps (extended) circumference (11), forearm circumference (12), and wrist circumference (13). All circumferences are measured in cm. Some of the predictors are highly correlated, i.e. $\rho_{ij} \approx 0.9$. The response has been calculated from the equation by Siri (Siri) using the body density determined by underwater weighting. In order to survey the performances of the different procedures we split the data at random into 25 training sets with $n_{train} = 151$ and test sets with $n_{test} = 101$. We choose the tuning parameters by tenfold cross validation on the training data set. Afterwards we estimated the model on the whole training data set. The median of prediction errors across 25 random splits were 22.03 (LASSO), 21.32 (V8), 21.99 (OSCAR) and 23.30 (Elastic Net). The corresponding boxplots are shown in Figure 5. It is seen that correlation based penalization has the best performance in terms of mean squared errors. Figure 5 shows that the V8 pro-

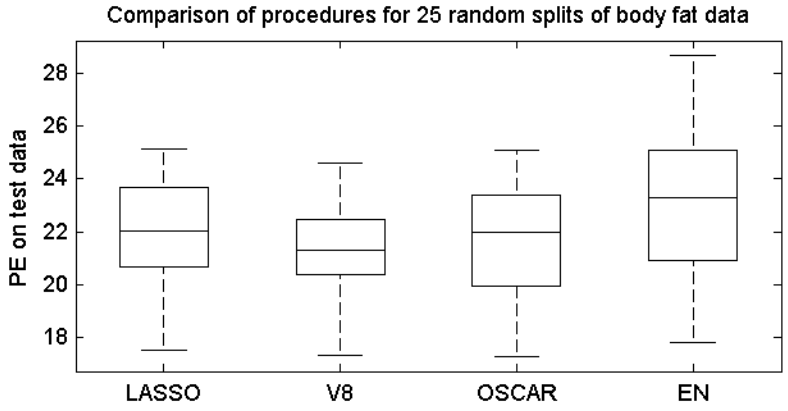

Figure 10: Boxplots of different methods for 25 random splits of the body fat data set with $n_{train} = 151$ and $n_{test} = 101$.

cedure has the lowest median and scatter. The OSCAR and the EN depend on two tuning parameters. So the computational costs are high especially for fine grids. The OSCAR procedure has the highest costs and does not preform better than the V8.

6 Concluding Remarks

It has been shown that polytopes are very flexible geometric objects which are useful for constraining regression problems. In particular their flexibility can be used to design specific polytopes that incorporate additional information contained in the data. The V8 procedure has been designed in this spirit as a correlation-based V-polytope.

For the computation of least squares problems which are constrained by centrosymmetric V-polytopes a modification of the LARS-algorithm has been proposed. V8 works quite well, in particular in the $p >> n$ case because it uses the efficient LARS-Algorithm. Therefore it can be applied where OSCAR fails because of its high computational costs. Moreover, V8 uses only one tuning parameter, which adds to reduce computational costs when searching for appropriate tuning parameters.

We restricted attention here to penalty regions which do not assume order information in the predictors. Therefore, we considered only the LASSO and OSCAR as specific polytope based procedures. If order information is available, as for example in signal regression, a successful strategy is to use the Fused Lasso (Tibshirani et al. (2005)), which is also a polytopal penalized regression problem with polytopes that reflect the order of predictors.

Acknowledgments

This work was partially supported by DFG Project TU62/4-1 (AOBJ: 548166).

A Appendix

The parameter space is the *Euclidean space* $\mathbb{R}^p$. With $(\mathbb{R}^p)^*$ we denote the *dual Euclidian space*. $\mathbb{R}^p$ represents the vector space of all column vectors of length p with real entries. $(\mathbb{R}^p)^*$ is the vector space of all linear functions $\mathbb{R}^p \to \mathbb{R}$ which are the row vectors of length p with real entries.

Definition A 1 (Hyperplane and Linear Halfspaces) *A subset $H \subset \mathbb{R}^p$ is called* hyperplane of $\mathbb{R}^p$*, if there is a linear functional $\mathbf{c} : \mathbb{R}^p \to \mathbb{R}$, $\mathbf{c} \in (\mathbb{R}^p)^* \setminus \{\mathbf{0}\}$, and a $t \in \mathbb{R}$ for which*

$$H = \{\mathbf{x} \in \mathbb{R}^p : \mathbf{cx} = t\}$$

holds.

A subset $H^- \subset \mathbb{R}^p$ is called lower linear halfspace (of $\mathbb{R}^p$)*, if there are a linear functional $\mathbf{c} : \mathbb{R}^p \to \mathbb{R}$, $\mathbf{c} \in (\mathbb{R}^p)^*$, and $t \in \mathbb{R}$ with*

$$H^-(\mathbf{c}, t) := \{\mathbf{x} \in \mathbb{R}^p : \mathbf{cx} \leq t\}.$$

Analogously a subset $H^+ \subset \mathbb{R}^p$ *is called* upper linear halfspace (of $\mathbb{R}^p$)*, if there are a linear functional* $\mathbf{c} : \mathbb{R}^p \to \mathbb{R}$, $\mathbf{c} \in (\mathbb{R}^p)^*$, *and* $t \in \mathbb{R}$ *with*

$$H^+(\mathbf{c}, t) := \{\mathbf{x} \in \mathbb{R}^p : \mathbf{c}\mathbf{x} \geq t\}.$$

Definition A 2 (Supporting Hyperplane) *Let H a hyperplane $H \subset \mathbb{R}^p$ and $K \subset \mathbb{R}^p$ a convex set. Then H is called* supporting hyperplane *or* support *for K if $H \cap K \neq \emptyset$ and K is entirely contained in one of the both closed halfspaces H^+ or H^-.*

Example A 1 *The H-representation of the LASSO penalty region in* $\mathbb{R}^3$ *is*

$$P(\mathbf{L}, \mathbf{t}) = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \leq \begin{pmatrix} t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \end{pmatrix}$$

The generalization for $p \in \mathbb{N}$ follows immediately.

Example A 2 *For $p = 3$, $t > 0$, and $c > 0$ the H-representation of the OSCAR penalty region is:*

$$\begin{pmatrix} 2c+1 & c+1 & 1 \\ 2c+1 & 1 & c+1 \\ c+1 & 2c+1 & 1 \\ c+1 & 1 & 2c+1 \\ 1 & c+1 & 2c+1 \\ 1 & 2c+1 & c+1 \\ -(2c+1) & c+1 & 1 \\ -(2c+1) & 1 & c+1 \\ -(c+1) & 2c+1 & 1 \\ -(c+1) & 1 & 2c+1 \\ -1 & c+1 & 2c+1 \\ -1 & 2c+1 & c+1 \\ 2c+1 & -(c+1) & 1 \\ 2c+1 & -1 & c+1 \\ c+1 & -(2c+1) & 1 \\ c+1 & -1 & 2c+1 \\ 1 & -(c+1) & 2c+1 \\ 1 & -(2c+1) & c+1 \\ 2c+1 & c+1 & -1 \\ 2c+1 & 1 & -(c+1) \\ c+1 & 2c+1 & -1 \\ c+1 & 1 & -(2c+1) \\ 1 & c+1 & -(2c+1) \\ 1 & 2c+1 & -(c+1) \\ -(2c+1) & -(c+1) & 1 \\ -(2c+1) & -1 & c+1 \\ -(c+1) & -(2c+1) & 1 \\ -(c+1) & -1 & 2c+1 \\ -1 & -(c+1) & 2c+1 \\ -1 & -(2c+1) & c+1 \\ -(2c+1) & c+1 & -1 \\ -(2c+1) & 1 & -(c+1) \\ -(c+1) & 2c+1 & -1 \\ -(c+1) & 1 & -(2c+1) \\ -1 & c+1 & -(2c+1) \\ -1 & 2c+1 & -(c+1) \\ 2c+1 & -(c+1) & -1 \\ 2c+1 & -1 & -(c+1) \\ c+1 & -(2c+1) & -1 \\ c+1 & -1 & -(2c+1) \\ 1 & -(c+1) & -(2c+1) \\ 1 & -(2c+1) & -(c+1) \\ -(2c+1) & -(c+1) & -1 \\ -(2c+1) & -1 & -(c+1) \\ -(c+1) & -(2c+1) & -1 \\ -(c+1) & -1 & -(2c+1) \\ -1 & -(c+1) & -(2c+1) \\ -1 & -(2c+1) & -(c+1) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \leq \begin{pmatrix} t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \\ t \end{pmatrix}$$

The generalization for $p \in \mathbb{N}$ follows immediately.

Example A 3 *Given the H-representation of an OSCAR penalty region in* $\mathbb{R}^3$ *as in Example A 2 the set of vertices of this penalty region is:*

$$\begin{aligned}\mathcal{O} \;=\; \Bigg\{ & \begin{pmatrix} \frac{\pm t}{2c+1} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \frac{\pm t}{2c+1} \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \frac{\pm t}{2c+1} \end{pmatrix}, \\ & \begin{pmatrix} \frac{\pm t}{3c+2} \\ \frac{\pm t}{3c+2} \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\pm t}{3c+2} \\ 0 \\ \frac{\pm t}{3c+2} \end{pmatrix}, \begin{pmatrix} 0 \\ \frac{\pm t}{3c+2} \\ \frac{\pm t}{3c+2} \end{pmatrix}, \begin{pmatrix} \frac{\pm t}{3c+2} \\ \frac{\pm t}{3c+2} \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\pm t}{3c+2} \\ 0 \\ \frac{\mp t}{3c+2} \end{pmatrix}, \begin{pmatrix} 0 \\ \frac{\pm t}{3c+2} \\ \frac{\mp t}{3c+2} \end{pmatrix}, \\ & \begin{pmatrix} \frac{\pm t}{3c+3} \\ \frac{\pm t}{3c+3} \\ \frac{\pm t}{3c+3} \end{pmatrix}, \begin{pmatrix} \frac{\mp t}{3c+3} \\ \frac{\pm t}{3c+3} \\ \frac{\pm t}{3c+3} \end{pmatrix}, \begin{pmatrix} \frac{\pm t}{3c+3} \\ \frac{\mp t}{3c+3} \\ \frac{\pm t}{3c+3} \end{pmatrix}, \begin{pmatrix} \frac{\pm t}{3c+3} \\ \frac{\pm t}{3c+3} \\ \frac{\mp t}{3c+3} \end{pmatrix} \Bigg\}. \end{aligned}$$

Proof A 1 (Propostion 1) *We consider a p-dimensional OSCAR penalty region for fixed tuning parameters* $t > 0$ *and* $c > 0$*. Let* $\mathcal{O}$ *denote the set of all vertices of this OSCAR penalty region. As remarked every row of the system of inequalities depends on the order of* $|\beta_i|$ *and one special orthant. For every facet determined by row of the system of inequalities one can find exactly p elements of* $\mathcal{O}$ *which confirm to the row by meanings of the order of* $|\beta_i|$ *and the signs. Consider the orthant with only positive values and the order* $|\beta_1| \geq |\beta_2| \geq \cdots \geq |\beta_p|$ *then only the following p vertices are elements of the corresponding row:*

$$\tilde{\mathcal{O}} = \left\{ \begin{pmatrix} v(1) \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} v(2) \\ v(2) \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} v(3) \\ v(3) \\ v(3) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} v(p) \\ v(p) \\ v(p) \\ v(p) \\ \vdots \\ v(p) \end{pmatrix} \right\}$$

By changing the signs and permuting the rows of the vertices of $\widetilde{\mathcal{O}}$ *we get the other orders of* $|\beta_i|$ *in every orthant.*

Hence every facet is defined by a p-elementic subset of $\mathcal{O}$ *and one row of the inequation system. The fact that no hyperplanes is ignored by a set of the kind* $\widetilde{\mathcal{O}}$ *and all elements of* $\mathcal{O}$ *are used completes the proof.*

Proof A 2 (Corollary 1) *If m of the p components of a vertex are nonzero then there are* $\binom{p}{m}$ *permutations of this m components. Further there are* 2^m *different sign combinations which are convenient. Its well known that* $\sum_{m=0}^{p} \binom{p}{m} a^{p-m} b^m = (a+b)^p$*. Now choose* $a = 1$ *and* $b = 2$*. Further* $0 < m \leq p$ *and* $\binom{p}{0} 1^p 2^0 = 1$ *holds and immediately* $\sum_{m=1}^{p} \binom{p}{m} 2^m = 3^p - 1$ *follows.*

The second statement follows directly from Proof A 1.

Example A 4 *The set of vertices the V8 penalty region is the union of the LASSO vertices* $\mathcal{L}$ *and vertices on the bisecting lines in every* β_i*-*β_j*-plane* $\mathcal{B} = \bigcup_{i<j} \mathcal{B}_{ij}$*.*

In $\mathbb{R}^4$ *the LASSO vertices are:*

$$\mathcal{L} = \left\{ \begin{pmatrix} \pm t \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \pm t \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \pm t \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ \pm t \end{pmatrix} \right\}.$$

The remaining set $\mathcal{B} = \bigcup_{i<j} \mathcal{B}_{ij}$ *is*

$$\mathcal{B} = \left\{ \begin{pmatrix} \pm\frac{t}{2-c_{12}} \\ \pm\frac{t}{2-c_{12}} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \pm\frac{t}{2-c_{12}} \\ \mp\frac{t}{2-c_{12}} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \pm\frac{t}{2-c_{13}} \\ 0 \\ \pm\frac{t}{2-c_{13}} \\ 0 \end{pmatrix}, \begin{pmatrix} \pm\frac{t}{2-c_{13}} \\ 0 \\ \mp\frac{t}{2-c_{13}} \\ 0 \end{pmatrix}, \right.$$
$$\begin{pmatrix} \pm\frac{t}{2-c_{14}} \\ 0 \\ 0 \\ \pm\frac{t}{2-c_{14}} \end{pmatrix}, \begin{pmatrix} \pm\frac{t}{2-c_{14}} \\ 0 \\ 0 \\ \mp\frac{t}{2-c_{14}} \end{pmatrix}, \begin{pmatrix} 0 \\ \pm\frac{t}{2-c_{23}} \\ \pm\frac{t}{2-c_{23}} \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ \pm\frac{t}{2-c_{23}} \\ \mp\frac{t}{2-c_{23}} \\ 0 \end{pmatrix},$$
$$\left. \begin{pmatrix} 0 \\ \pm\frac{t}{2-c_{24}} \\ 0 \\ \pm\frac{t}{2-c_{24}} \end{pmatrix}, \begin{pmatrix} 0 \\ \pm\frac{t}{2-c_{24}} \\ 0 \\ \mp\frac{t}{2-c_{24}} \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \pm\frac{t}{2-c_{34}} \\ \pm\frac{t}{2-c_{34}} \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \pm\frac{t}{2-c_{34}} \\ \mp\frac{t}{2-c_{34}} \end{pmatrix} \right\}.$$

The generalization to any finite $p \in \mathbb{N}$ *follows immediately. So The V8 penalty region is* $\mathcal{P} = conv(\mathcal{L} \cup \mathcal{B})$.

References

Bondell, H. D. and B. J. Reich (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with oscar. *Biometrics 64*, 115–123.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*, 407–499.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

James, W. and C. Stein (1961). Estimation with quadratic loss. *Proc. 4th Berkeley Symp. Math. Statist. Prob., University of California Press, Berkeley 1*, 361–380.

Luenberger, D. G. (1969). *Optimazition by Vector Space Methods*. New York: John Wiley & Sons.

Penrose, K. W., A. G. Nelson, and A. G. Fisher (1985). Generalized body composition prediction equation for men using simple measurement tech- niques. *Medicine and Science in Sports and Exercise 17*, 189.

Siri, W. B. *The gross composition of the body.*

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Kneight (2005). Sparsity and smoothness vie the fused lasso. *Journal of the Royal Statistical Society B 67*, 91–108.

Tutz, G. and J. Ulbricht (2009). Penalized regression with correlation based penalty. Statistics and Computing (to appear).

Ziegler, G. M. (1994). *Lectures on Polytopes*. New York, Berlin, Heidelberg, London, Paris Tokyo, Hong Kong, Barcelona, Budapest: Springer Verlag (in Graduates Texts in Mathematics).

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B 67*, 301–320.