Statistics and Probability Letters 83 (2013) 2108–2112

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

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

On grouping effect of elastic net☆

Ding-Xuan Zhou

Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China

ARTICLE INFO

Article history:
Received 8 May 2012
Received in revised form 15 May 2013
Accepted 15 May 2013
Available online 23 May 2013

MSC:
68T05
42B10

Keywords:
Elastic net
Grouping effect
Approximation error
Variable selection
Reproducing kernel Hilbert space

ABSTRACT

Grouping effect of the elastic net asserts that coefficients corresponding to highly correlated predictors in a linear regression setting have small differences. A quantitative estimate for such small differences was given in Zou and Hastie (2005) when the coefficients have the same sign. We show that the same estimate holds true even when the coefficients have different signs. The estimate is also improved by means of an empirical approximation error when the model fits the data well.

1. Introduction and main results

Elastic net is a method for variable selection introduced in Zou and Hastie (2005). It can be stated as a coefficient-based regularization scheme. In a linear regression setting with p predictors $\mathbf{x}_1, \ldots, \mathbf{x}_p \in \mathbb{R}^n$ and a response $\mathbf{y} \in \mathbb{R}^n$, the elastic net produces a coefficient vector $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^T \in \mathbb{R}^p$ given by

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \left| \mathbf{y} - \left[\mathbf{x}_1, \ldots, \mathbf{x}_p \right] \beta \right|^2 + \lambda_1 |\beta|_1 + \lambda_2 |\beta|^2 \right\}, \tag{1.1}$$

where λ_1, λ_2 are positive regularization parameters and $|\beta|_1 = \sum_{j=1}^p |\beta_j|$, $|\beta| = \left(\sum_{j=1}^p |\beta_j|^2 \right)^{1/2}$ are norms on $\mathbb{R}^p$ defined for $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$.

The classical ordinary least squares estimator takes the form (1.1) with $\beta_1 = \beta_2 = 0$ while ridge regression (Hoerl and Kennard, 1988) corresponds to the case $\beta_1 = 0$. With a motivation of sparse representations, a regularization technique called Lasso was introduced by Tibshirani (Tibshirani, 1996). It can be expressed as (1.1) with $\beta_2 = 0$.

While Lasso yields sparse solutions in many cases, it may have limitations in some situations. One situation is when $p > n$ in which case Lasso chooses at most n variables. The other is when a group of variables are highly correlated in which case Lasso often chooses only one variable from the group. This may not be desirable in some applications. For example, in gene sequence analysis of microarray data with thousands of genes (predictors) and $n \ll p$, it is quite common that a group of highly correlated genes responding to the same biological change are equally important to be included in studying the biological mechanism of the change, which is missed by Lasso. The elastic net has the advantage of including automatically

☆ The work described in this paper is supported by a grant from the Research Grants Council of Hong Kong [Project No. CityU 104710].
E-mail address: mazhou@cityu.edu.hk.

0167-7152/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spl.2013.05.014

all the highly correlated variables in the group. This is called the *grouping effect*. A rigorous mathematical theorem about the grouping effect of the elastic net was proved in Zou and Hastie (2005) as follows. Denote $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$.

Theorem 1. *Assume that the response is centered and the predictors are standardized as*

$$\sum_{i=1}^{n} y_i = 0, \qquad \sum_{i=1}^{n} x_{ij} = 0, \quad \text{and} \quad \sum_{i=1}^{n} x_{ij}^2 = 1, \quad \text{for } j = 1, \ldots, p. \tag{1.2}$$

If

$$\widehat{\beta}_i \widehat{\beta}_j > 0, \tag{1.3}$$

then

$$\frac{|\widehat{\beta}_i - \widehat{\beta}_j|}{|\mathbf{y}|} \le \frac{\sqrt{2(1 - \mathbf{x}_i^T \mathbf{x}_j)}}{\lambda_2}. \tag{1.4}$$

Note that $\mathbf{x}_i^T \mathbf{x}_j$ is the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$. It is close to 1 when $\mathbf{x}_i$ and $\mathbf{x}_j$ are highly correlated. In this case, Theorem 1 asserts that the difference between the coefficient paths of predictors i and j is small. Thus (1.5) gives a quantitative description of the grouping effect of the elastic net.

While assumption (1.2) is not essential, condition (1.3) in Theorem 1 requires that the coefficient paths of predictors i and j have the same sign. This is a data dependent condition and Theorem 1 does not estimate $|\widehat{\beta}_i - \widehat{\beta}_j|$ when this condition fails. The first purpose of this paper is to prove that the conclusion of Theorem 1 still holds without condition (1.3).

Theorem 2. *Under the assumption (1.2), we have*

$$|\widehat{\beta}_i - \widehat{\beta}_j| \le \frac{\sqrt{2(1 - \mathbf{x}_i^T \mathbf{x}_j)}\,|\mathbf{y}|}{\lambda_2}. \tag{1.5}$$

The above estimate for the grouping effect depends on the correlation, not on the data fitting. The second purpose of this paper is to improve the estimate in Theorem 2 when the model fits the data well. We measure the data fitting by the *empirical approximation error* defined similarly to approximation error (Smale and Zhou, 2003) as

$$\widehat{\mathcal{D}}(\lambda_1, \lambda_2) = \min_{\beta \in \mathbb{R}^p} \left\{ \left| \mathbf{y} - \left[\mathbf{x}_1, \ldots, \mathbf{x}_p\right] \beta \right|^2 + \lambda_1 |\beta|_1 + \lambda_2 |\beta|^2 \right\}. \tag{1.6}$$

Since $\widehat{\beta}$ is a minimizer of (1.6), by taking $\beta = 0$, we see that

$$\left| \mathbf{y} - \left[\mathbf{x}_1, \ldots, \mathbf{x}_p\right] \widehat{\beta} \right| \le \sqrt{\widehat{\mathcal{D}}(\lambda_1, \lambda_2)} \le |\mathbf{y}|. \tag{1.7}$$

Thus a bound becomes tighter when $|\mathbf{y}|/\sqrt{n} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2}$ is replaced by $\sqrt{\frac{1}{n}\widehat{\mathcal{D}}(\lambda_1, \lambda_2)}$, a quantity which might tend to 0 as the sample size increases. We state our bound in a general setting without the standardizing assumption (1.2) under which $\sqrt{2(1 - \mathbf{x}_i^T \mathbf{x}_j)} = |\mathbf{x}_i - \mathbf{x}_j|$. Theorem 3, to be proved in Section 2, implies Theorem 2.

Theorem 3. *Denote the $n \times p$ matrix $\left[\mathbf{x}_1, \ldots, \mathbf{x}_p\right]$ as $\mathbf{X}$. If $\lambda_1, \lambda_2 > 0$ and $\widehat{\beta}$ is the solution to (1.1), then we have*

$$|\widehat{\beta}_i - \widehat{\beta}_j| \le \frac{\left|(\mathbf{x}_i - \mathbf{x}_j) \cdot (\mathbf{y} - \mathbf{X}\widehat{\beta})\right|}{\lambda_2} \le \frac{|\mathbf{x}_i - \mathbf{x}_j| \sqrt{\widehat{\mathcal{D}}(\lambda_1, \lambda_2)}}{\lambda_2}, \quad \forall i, j \in \{1, \ldots, p\}. \tag{1.8}$$

2. Proof of main results

In this section we prove our main results on grouping effect which might be used for discussing other properties such as sparsity. To this end, we need some explicit formulas for the elastic net which were derived by the Karush–Kuhn–Tucker Theorem in Yuan and Lin (2007) and for a weighted case in Hong and Zhang (2010). For completeness, we give an elementary proof here. From these formulas we shall see that $\widehat{\beta}_i = 0$ if and only if $\left|\mathbf{x}_i^T \left(\mathbf{y} - \mathbf{X}\widehat{\beta}\right)\right| \le \frac{\lambda_1}{2}$.

Theorem 4. *Let $i \in \{1, \ldots, p\}$. Then $\widehat{\beta}_i \ne 0$ if and only if*

$$\left|\mathbf{x}_i^T \left(\mathbf{y} - \mathbf{X}\widehat{\beta}\right)\right| > \frac{\lambda_1}{2}. \tag{2.1}$$