High-dimensional linear model selection via multiple testing

Konrad Furmańczyk[a], $Wojciech Rejchel^b$

Warsaw University of Life Sciences, Institute of Information Technology[a], $Warsaw, Poland, email : konfur@wp.pl$

Nicolaus Copernicus University[b], $Torun, Poland, email : wrejchel@gmail.com$

Mathematical Methods of Modern Statistics 2, Luminy 15-19/6/2020

Introduction We propose a new model selection algorithm (called „LassoSD"), which combines a stepdown multiple testing approach with penalized minimization. First, we use the Lasso to discard a substantial part of irrelevant predictors. In the next step of the algorithm we consider only the support of the Lasso and select the final model in a way that is motivated by multiple testing (analogously to the Bonferroni correction or the Holm method). We state nonasymptotic probabilistic inequalities, that upper bounds the model selection error of LassoSD in the high-dimensional linear model, i.e. the number of predictors can be much larger than the sample size.

Linear model W consider the high-dimensional linear model with sub-Gaussian errors, that is

$$Y = X\beta^\star + \varepsilon, \tag{1}$$

where $Y = (Y_1, ..., Y_n)'$ is a response vector, $X = (x_1, \ldots, x_p)$ is a fixed $n \times p$ design matrix and x_j is its j-th column, $\beta^\star$ is the true $p \times 1$ parameter and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$ is an $n \times 1$ stochastic error. We assume that coordinates of ε are i.i.d. zero-mean random variables having a sub-Gaussian distribution with the parameter $\sigma > 0$, that is $E\exp(u\varepsilon_1) \leq \exp\left(\sigma^2 u^2/2\right)$ for $u \in \mathbb{R}$. The true model is denoted by $I_0 = \{1 \leq j \leq p : \beta_j^\star \neq 0\}$ and its number of elements by $|I_0| = p_0$. We work in the high-dimensional scenario, so the number of predictors p can depend on the sample size n and can be much larger than n.

LassoSD

Our algorithm, called LassoSD, is a three-step procedure:

(1) We perform the Lasso, that is we minimize the expression

$$\frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{2}$$

for some parameter $\lambda > 0$. We denote the minimizer of (2) by $\hat\beta^L$ and its support by $S^L = \left\{1 \leq j \leq p : \hat\beta_j^L \neq 0\right\}$.

(2) We discard from the set S^L these predictors, that are smaller in absolute values than the prespecified level $\delta > 0$, namely we obtain the set

$$S = \left\{j \in S^L : |\hat\beta_j^L| \geq \delta\right\} \tag{3}$$

and $s := |S|$.

(3) We calculate

$$\hat\beta^{LS} = \left(X_S' X_S\right)^{-1} X_S' Y, \tag{4}$$

which is the ordinary least squares estimator in the linear model "restricted" to predictors from the set S. Obviously, to calculate (4) we need that the number of predictors in S is not greater than the sample size. Next, we compute statistics $t_j = \frac{\hat\beta_j^{LS}}{\sigma\sqrt{(X_S'X_S)^{-1}_{jj}}}$ for $j \in S$. We sort statistics $\{t_j : j \in S\}$ in decreasing order with respect to their squares, i.e.

$$t_{[1]}^2 \geq t_{[2]}^2 \geq \ldots \geq t_{[s]}^2,$$

in particular $|t_{[1]}| = \max(|t_1|, \ldots, |t_s|)$. Let $\gamma_1 \geq \ldots \geq \gamma_s > 0$ be given thresholds. The estimator $\hat I$ of the true set I_0 is constructed as follows: if $t_{[1]}^2 < \gamma_1^2$, then $\hat I$ is the empty set. Otherwise, we look for the largest r satisfying $t_{[1]}^2 \geq \gamma_1^2, \ldots, t_{[r]}^2 \geq \gamma_r^2$ and $\hat I$ consists of r-predictors, which correspond to ordered statistics $t_{[1]}, t_{[2]}, \ldots, t_{[r]}$.

Motivated by Multiple Testing

The third step of the algorithm is motivated by the stepdown multiple testing procedure. We propose to choose thresholds $\gamma_1 \geq \ldots \geq \gamma_s > 0$ in the following two ways: for a (small) number $\alpha > 0$ we consider: the procedure with equal thresholds

$$\gamma = \sqrt{2\log\binom{p}{s} - 2\log\alpha}, \tag{5}$$

the procedure with decreasing thresholds

$$\gamma_j = \sqrt{2\log\binom{p+1-j}{s+1-j} - 2\log\alpha}, \tag{6}$$

where $j = 1, ..., s$.

SCIF(The Sign-Restricted Cone Invertibility Factor)

For $\xi > 1$ and the set I_0 we define two cones

$$\mathcal{C}(I_0, \xi) = \left\{\beta \in \mathbb{R}^p : ||\beta_{I_0'}||_1 \leq \xi||\beta_{I_0}||_1, \quad \beta_j x_j' X\beta/n \leq 0 \text{ for } j \in I_0'\right\}$$

and

$$\tilde{\mathcal{C}}(I_0, \xi) = \{\beta \in \mathbb{R}^p : |\beta_{I_0'}|_1 \leq \xi|\beta_{I_0}|_1\} \ ,$$

where I_0' is the complement of I_0. In the case when $p >> n$ three different conditions measuring the degree of correlations between predictors are often used (SCIF),

$$F_q(I_0, \xi) = \inf_{\beta \in \mathcal{C}(I_0,\xi)} \frac{p_0^{1/q}\|X'X\beta/n\|_\infty}{\|\beta\|_q}$$

with $q \geq 1$. Note that $m_j(J) = [(X_J'X_J)^{-1}]_{jj}$, where $j \in J$ and $\beta^\star_{\min} = \min_{j \in I_0}|\beta_j^\star|$.

Main Result-Theorem Let $a \in (0,1), b > 0, \xi > 1$ be arbitrary. Consider the LassoSD algorithm with the parameter λ satisfying

$$\frac{\sigma(\xi+1)}{\sqrt{a}(\xi-1)}\sqrt{\frac{2\log(p)}{n}} \leq \lambda \leq \frac{(\xi+1)\beta^\star_{\min}F_\infty(I_0,\xi)}{2\xi + (\xi+1)bF_\infty(I_0,\xi)}, \tag{7}$$

$\delta = b\lambda$ and thresholds (5) or (6) in the third step with $\alpha \geq \frac{1}{p}$. Moreover, suppose that $K := p_0\left(1 + \frac{2\xi}{b(\xi+1)F_1(I_0,\xi)}\right) \leq n$ and

$$\frac{(\beta^\star_{\min})^2}{\sigma^2 m} \geq 16\log(p^K/K!), \tag{8}$$

where

$$m = \max_{k=0,\ldots,K-p_0}\ \max_{J \subset I_0' : |J| = k}\ \max_{j \in I_0} m_j(I_0 \cup J). \tag{9}$$

Then we obtain that

$$P(\hat I \neq I_0) \leq 2\exp\left(-\frac{(1-a)(\xi-1)^2\lambda^2 n}{2\sigma^2(\xi+1)^2}\right) \tag{10}$$

$$+\ \alpha(K - p_0 + 1)\left[p_0\left(\frac{K}{p - p_0 + 1}\right)^{p_0} + K - p_0\right]. \tag{11}$$

Conclusions In Furmańczyk, K and Rejchel, W. (2020) we propose the new algorithm for high-dimensional linear models. It adapts multiple hypothesis testing to the model selection problem. We prove that this procedure is able to find the true model under mild conditions. The experiments confirm that LassoSD performs well in model selection. In fact, it works similarly or better than competitive algorithms based on multi-splitting or minimization of information criteria.

References

Bunea, F., Wegkamp, M.H., Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *J. Statist. Plann. Infer.* **136**, 12, 4349-4364. with sparsity. The Lasso and generalizations. A Chapmann & Hall Book.

Furmańczyk, K and Rejchel, W. (2020). High-dimensional linear model selection motivated by multiple testing. *Statistics* **54**, 1, 152-166.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p-values for high dimensional regression. *Journal of the American Statistical Association* **104**, 1671–1681.