Regularized sliced inverse regression through the elastic net penalty

Ali Alkenani[1] and Mohamed Abdulkadhim[2]

[1,2]Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, Al Diwaniyah, Iraq.
[1]Corresponding Author: Ali Alkenani, E-mail: ali.alkenani@qu.edu.iq
[1]http://orcid.org/0000-0001-5067-2321

Abstract

A model-free variable selection method (SSIR-EN) was proposed in this article. The elastic net (EN) penalty was employed with sliced inverse regression (SIR) method to introduce SSIR-EN. Without assuming a parametric model, the SSIR-EN provides better prediction accuracy and easier interpretations. Under sufficient dimension reduction (SDR) settings, the SSIR-EN produces a shrinkage estimation when the predictors are highly correlated and perform groups. The SSIR-EN extended EN to nonlinear and multi-dimensional regression under SDR settings. Also, the SSIR-EN enables SIR method to work with problems were the predictors are highly correlated and perform groups. In addition, SSIR-EN can exhaustively estimate dimensions, while selecting the important covariates simultaneously. The effectiveness of SSIR-EN was checked by both simulation and real data analysis.

Keywords: Sufficient dimension reduction, Variable selection, sliced inverse regression, Elastic Net.

1. Introduction

In high-dimensional regressions, the SDR received great attention. Let y is a response variable and $\mathrm{X} = (\mathrm{x}_1, \ldots, \mathrm{x}_p)^T$ is a p-dimensional predictors vector. Without assuming any parametric model, the main idea of SDR is to replace X with a lower-dimensional orthogonal projection $P_s\mathrm{X}$ on to S with keeping the information about the distribution of $Y|\mathrm{X}$. The aim of SDR is to find the central subspace ($\mathrm{S}_{Y|\mathrm{X}}$). The $\mathrm{S}_{Y|\mathrm{X}}$ is the intersection of all subspaces S such that $Y \perp\!\!\!\perp \mathrm{X}|P_s\mathrm{X}$, where $\perp\!\!\!\perp$ indicates independence. Consequently, $P_\beta\mathrm{X}$ extracts all of the information from X about Y, where β is a basis of $\mathrm{S}_{Y|\mathrm{X}}$ (Cook, 1998).

For finding $\mathrm{S}_{Y|\mathrm{X}}$, many methods was proposed. The SIR (Li, 1991) is one of the well-known methods for estimating $\mathrm{S}_{Y|\mathrm{X}}$. The SIR applied in vary areas such as marketing, economics and bioinformatics. However, SIR suffers from that the interpretation of the SIR resulting estimates could be difficult because that SIR produces linear combinations of all the original predictors. In SIR analysis and for better interpretability, there is need to reduce the number of unimportant nonzero coefficients in the SIR directions.

For achieving better interpretability under ordinary least squares settings, many of regularisation methods were proposed. See, for example, the Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), Elastic Net (Zou and Hastie, 2005), adaptive Lasso (Zou, 2006), group Lasso (Yuan and Lin, 2006), OSCAR (Bondell and Reich, 2008), MCP (Zhang, 2010).

Under SDR framework, Naik and Tsai (2001) proposed model-selection method for single-index models. For assessing the contribution of predictors, Cook (2004) proposed a model-free variable selection method. Also, the shrinkage SIR (SSIR) estimator through adding Lasso penalty to least squares formulation of SIR was proposed by Ni et al. (2005). Li and Nachtsheim (2006) combined the lasso and LARS with SIR to produce sparse SIR (SPSIR). Li (2007) combined a number of SDR methods with the concept of regularisation estimation. This strategy was applied to SIR and many of SDR methods. A regularised SIR (RSIR) method was proposed by Li and Yin (2008) to enable SIR to work when $p > n$ and the predictors are highly correlated. The p and n are the number of predictors and the sample size,

respectively. The sliced inverse quantile regression SIQR method was proposed by Alkenani and Dikheel (2016). Moreover, the authors combined the ideas of Lasso and Adaptive Lasso with SIQR to propose sparse SIQR. For the multiple index model, the Lasso-SIR method was proposed by Lin et.al (2018). The Lasso-SIR is consistent and achieve the optimal convergence rate under $p > n$ settings (Lin et.al, 2018).

In this article, we proposed SSIR-EN method. It is a shrinkage estimation method under SDR framework. SSIR-EN was proposed to work when there is a group of predictors among which the predictors are highly pairwise correlated. SSIR-EN has merits over the existing sparse SIR methods. SSIR-EN benefited from the advantages of EN. The first advantage of EN is that the parameters estimation and variable selection are carried out simultaneously. The second advantage is that EN has the ability to select groups of highly correlated variables. The second advantage does not hold for Lasso, adaptive lasso, SCAD, MCP and bridge penalties which are employed in the existing methods.

The rest of this article is as follows. In Section 2, we presented a short review of SIR and shrinkage SIR. In Section 3, the SSIR-EN is proposed. Simulation studies are carried out in Section 4. In Section 5, we applied the studied methods to real data. In Section 6, the conclusions are given.

2. SIR and shrinkage SIR

For estimating the basis of $S_{y|\mathbf{x}}$, the SIR method was proposed by Li (1991). The SIR requires $Z = \Sigma^{-\frac{1}{2}}\left(X - E(X)\right)$, satisfy the condition $E(\mathrm{Z}|P_c\mathrm{Z}) = P_c\mathrm{Z}$, where $\Sigma_x = Cov(X)$ is the population covariance matrix of X and c is a basis for $S_{Y|\mathrm{Z}}$. This condition connects $S_{Y|\mathrm{Z}}$ with the inverse regression of Z on Y. The symmetric kernel matrix of SIR is $M = cov\,[E(Z|Y)]$ and $Span(M) \subseteq S_{Y|\mathrm{Z}}$.

Let a random sample of size n of (X, Y), which has a joint distribution. Let $\bar{X}$ is the sample mean of X. Also, assume that $\hat{Z} = \hat{\Sigma}^{-\frac{1}{2}}(X - \bar{X})$ is the sample version of Z, where $\hat{\Sigma}$ is the sample covariance matrix of X. Let h is the number of slices and n_y is the number of observations in the yth slice. Thus, $\widehat{M} = \sum_{y=1}^{h} \hat{f}_y \hat{Z}_y \hat{Z}_y^T$ is the sample version of M, where $\hat{f}_y = n_y/n$ and $\hat{Z}_y$ is the average of Z in the slice y. Let $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \lambda_p \geq 0$ are the eigenvalues corresponding to the

eigenvectors $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_p$ of $\widehat{M}$. If the dimension d of $S_{Y|Z}$ is known, $span(\hat{\beta}) = span(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_d)$ is a consistent estimator of $S_{Y|X}$, where $\hat{\beta}_i = \hat{\Sigma}^{-\frac{1}{2}} \hat{v}_i$.

The SIR provides an estimator $span(\hat{\beta})$ of $S_{Y|X}$. Usually, the elements of $\hat{\beta} \in \mathbb{R}^{p \times d}$ are nonzero. In the construction of 'sufficient predictors', only the important predictors are needed if the number of predictors is large or the predictors are highly-correlated. To this end, a number of regularizations methods were employed with SIR by many researchers to compress some rows of $\hat{\beta}$ to 0's.

To improve interpretability, the SIR was formulated as a regression type optimisation problem by Cook (2004) through minimising

$$F(A, C) = \sum_{y=1}^{h} \left\| \hat{f}_y^{1/2} \hat{Z}_y - A C_y \right\|^2, \qquad (1)$$

over $A \in \mathbb{R}^{p \times d}$ and $C_y \in \mathbb{R}^d$, with $C = (C_1, \ldots, C_h)$. Let $\hat{A}$ and $\hat{C}$ are the values of A and C that minimise F. Then $span(\hat{A})$ equals the space spanned by the d largest eigenvectors of M. By focusing on the coefficients of the X variables, Ni et al. (2005) rewrite (1) as

$$G(B, C) = \sum_{y=1}^{h} \left(\hat{f}_y^{1/2} \hat{\Sigma}^{-\frac{1}{2}} \hat{Z}_y - B C_y \right)^T \hat{\Sigma} \left(\hat{f}_y^{1/2} \hat{\Sigma}^{-\frac{1}{2}} \hat{Z}_y - B C_y \right), \qquad (2)$$

where $B \in \mathbb{R}^{p \times d}$. The value of B which minimises (2) is exactly $\hat{\beta}$ and $span(\hat{\beta}) = span\left(\hat{\Sigma}^{-\frac{1}{2}} \hat{A}\right)$ is the estimator of $S_{Y|X}$. After that, Ni et al. (2005) proposed a shrinkage SIR estimator (SSIR) of $S_{Y|X}$ is $span(diag(\tilde{\alpha})\hat{\beta})$, where the shrinkage indices $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_p)^T \in \mathbb{R}^p$ are determined by minimising

$$\sum_{y=1}^{h} \left\| \hat{f}_y^{1/2} \hat{Z}_y - \hat{\Sigma}^{\frac{1}{2}} diag(\hat{B} \hat{C}_y) \alpha \right\|^2 + \lambda \sum_{i=1}^{p} |\alpha_i| , \qquad (3)$$

where, $\hat{B}$ and $\hat{C} = \left(\hat{C}_1, \ldots, \hat{C}_h\right)$ minimise (2).

A standard Lasso algorithm can be employed to carry out (3). To be specific, let $\tilde{Y} = vec(\hat{f}_1^{1/2} \hat{Z}_1, \ldots, \hat{f}_h^{1/2} \hat{Z}_h) \in \mathbb{R}^{ph}$ and $\tilde{X} = \left(diag(\hat{B}\hat{C}_1)\hat{\Sigma}^{\frac{1}{2}}, \ldots, diag(\hat{B}\hat{C}_a)\hat{\Sigma}^{\frac{1}{2}}\right)^T \in \mathbb{R}^{ph \times p}$, where $vec(.)$ is a matrix operator that stacks the matrix's columns to a single vector. Then the vector α, is exactly the estimator of Lasso for the regression $\tilde{Y}$ on $\tilde{X}$.

3. SSIR-EN

Ni et al. (2005) introduced SSIR through adding Lasso penalty to least squares formulation of SIR. Li and Nachtsheim (2006) combined the lasso and LARS with SIR to produce SPSIR. Li and Yin (2008) proposed RSIR when $p > n$ and the predictors were highly correlated. Under $p > n$ settings, Lasso-SIR method for the multiple index model was introduced by Lin et.al (2018).

The SSIR, SPSIR, RSIR and Lasso-SIR methods employed penalties that fail to work with grouped predictors situation and do not have the ability to select groups of highly correlated predictors. The limitations of the mentioned methods motivate us to propose SSIR-EN method. In this article, we propose SSIR-EN to minimise

$$\sum_{y=1}^{h} \left\| \hat{f}_y^{1/2} \hat{Z}_y - \hat{\Sigma}^{\frac{1}{2}} diag(\hat{B}\hat{C}_y)\alpha \right\|^2 + \lambda_1 \sum_{i=1}^{p} \alpha_i^2 + \lambda_2 \sum_{i=1}^{p} |\alpha_i| \,, \ldots\ldots (4)$$

The minimisation in (4) consists of three parts. The first part is the loss function of SIR. The second part is the ridge penalty function and the third part is the Lasso penalty function. The EN penalty consists of the second and the third parts. Also, $\lambda_1$and λ_2 are the tuning parameters of Elastic Net.

By employing a standard EN algorithm, the constrained optimisation of (4) can be done and can be conveniently carried out in standard software. Then, the vector α is exactly the estimator of EN for the regression of $\tilde{Y}$ on $\tilde{X}$. To select $\lambda_1$and λ_2, Cross-validation or an information criterion like AIC or BIC could be used.

In summary, SSIR-EN is a two-step procedure: first, apply SIR to obtain d, $\tilde{Y}$ and $\tilde{X}$; secondly, compute α via EN algorithm by choosing $\lambda_1$and λ_2 through Cross-validation or AIC or BIC.

SSIR-EN combines EN into the "OLS formulation" of SIR. Thus, under the same conditions as those for SIR and EN, the minimisation algorithm for solving (4) is guaranteed to converge to the global minimum. Based on our extensive simulations, the algorithm usually converges fast. The *R* code for SSIR-EN is available from the author.

4. Simulation study

In terms of prediction accuracy and variable selection, the performance of SSIR-EN was compared with SSIR, SPSIR, RSIR and Lasso-SIR methods under different settings. Also, the ability of SSIR-EN to achieve groups selection was checked.

The performance of the SSIR-EN was examined through a number of examples as are reported below. In each example, the simulated data were divided into three sets. They are the training set, the independent validation set and the independent test set. We fitted the models through the training data, and we employed the validation data to select the tuning parameters. By using the test data, the mean-squared error (MSE) and the average number of zero coefficients (Ave 0's) were computed to check the prediction accuracy and the ability of variable selection for the considered methods, respectively. We used the notation ././. to represent the number of observations in the training set, the independent validation set and the independent test set, respectively.

In all examples, SSIR-EN was computed as described in section3. The R code made by Liqiang Ni was employed to carry out SSIR method. Using the R codes made by Lexin Li, the SPSIR and RSIR methods were implemented. The function *LassoSIR* from the R package (LassoSIR) was employed to compute Lasso-SIR estimates. The *R* code for SSIR-EN is available from the author. The tuning parameters were selected by tenfold cross-validation (C.V) for each competitor.

The examples as follow:

Model 1. (Single-index model, d =1)

We simulated the data from the following single index model

$$y = 1 + 2(\mathbf{x}^T\boldsymbol{\beta} + 3)log(3|\mathbf{x}^T\boldsymbol{\beta}| + 1) + \varepsilon,$$

where ε is from $N(0,1)$ and x_i is normally distributed according to the below described examples.

Example1.

We generated 100 data sets with 20/20/200 observations and $p = 8$. The $\boldsymbol{\beta}$ is selected as $\boldsymbol{\beta} = (3,1.5,0,0,2,0,0,0)$. Let $\boldsymbol{\rho}_{ij} = 0.5^{|i-j|}$ is the pairwise correlation between x_i and x_j.

Example2.

The settings in this example are similar to the settings in example 1, except that $\beta_j = 0.85$ for all j.

Example3.

We generated 100 data sets with 100/100/400 observations and $p = 40$. The $\boldsymbol{\beta}$ is selected as follows

$$\boldsymbol{\beta} = (\underbrace{0,\dots,0}_{10},\underbrace{2,\dots,2}_{10},\underbrace{0,\dots,0}_{10},\underbrace{2,\dots,2}_{10}),$$

where $\boldsymbol{\rho}_{ij} = 0.5$ for all i and j.

Example4.

We generated 100 data sets with 50/50/400 observations and $p = 40$. The $\boldsymbol{\beta}$ is selected as follows:

$$\boldsymbol{\beta} = (\underbrace{3,\dots,3}_{15},\underbrace{0,\dots,0}_{25})$$

and the predictors were

$\mathrm{x}_i = Z_1 + \varepsilon^*, Z_1 \sim N(0,1), i = 1,\dots,5,$

$\mathrm{x}_i = Z_2 + \varepsilon^*, Z_2 \sim N(0,1), i = 6,\dots,10,$

$\mathrm{x}_i = Z_3 + \varepsilon^*, Z_3 \sim N(0,1), i = 11,\dots,15,$

x_i is i.i.d from $N(0,1), i = 16,\dots,40.$

where ε^* is i.i.d from $N(0,0.01), i = 1,\dots,5$. We can notice that there are 3 groups in this model and there are 5 predictors in each group. Also, there are 25 predictors with zero coefficients.

Model 2. (Multiple-index model, d =2)

Example5.

We generated 100 data sets with 20/20/200 observations and $p = 8$ from the following model:

$$y = \frac{\mathbf{x}^T\boldsymbol{\beta}_1}{0.5+(1.5+\mathbf{x}^T\boldsymbol{\beta}_2)} + \sigma\,\varepsilon,$$

where x_i is normally distributed and ε is i.i.d from an $N(0,1)$. Also, $\boldsymbol{\beta}_1 = (3,1.5,2,0,0,0,0,0)^T$, $\boldsymbol{\beta}_2 = (0,0,0,0,0,3,1.5,2)^T$ and $\sigma = 3$ .For $\boldsymbol{\beta}_1$, the first three predictors were highly correlated with pairwise correlation $r = o.7$, while the rest

were uncorrelated. For $\boldsymbol{\beta}_2$, the first five predictors were uncorrelated, while last three predictors were highly correlated with pairwise correlation $r = o.7$.

Table1: Median MSE results (MMSE) for the considered methods according to the five examples. Standard errors are in parentheses.

Methods	Examples					
	Example1	Example2	Example3	Example4	Example5	
					$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$
SSIR	3.02 (0.53)	3.84 (0.62)	67.64 (3.13)	46.40 (4.20)	5.03 (0.75)	5.09 (0.76)
RSIR	2.74 (0.50)	3.58 (0.57)	65.25 (2.96)	44.24 (4.00)	4.70 (0.71)	4.68 (0.72)
SPSIR	2.80 (0.50)	3.65 (0.58)	65.84 (3.02)	45.03 (4.11)	4.82 (0.72)	4.86 (0.74)
Lasso-SIR	2.73 (0.46)	3.50 (0.51)	63.34 (2.88)	42.40 (3.92)	4.70 (0.64)	4.65 (0.66)
SSIR-EN	2.59 (0.42)	3.19 (0.36)	57.98 (1.82)	38.50 (1.74)	4.40 (0.50)	4.36 (0.41)

From Table 1., we can summarise the prediction accuracy results according to MMSE as follows. For all the examples, it is clear that the worst performance is for the SSIR method. From another side, the performance of SSIR-EN method is better and more accurate than all the competitors. In general, the performance of Lasso-SIR was better than RSIR, SPSIR and SSIR methods for all the examples. The Lasso-SIR was a good competitor for SSIR-EN in all the examples. The results of simulation indicate that the SSIR-EN dominates all the competitors under collinearity.

Table2: The Ave 0's results for the considered methods according to the five examples.

Methods	Examples					
	Example1	Example2	Example3	Example4	Example5	
					$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$
SSIR	3.01	0	10.13	12.03	2.88	2.73
RSIR	3.60	0	13.29	15.77	3.51	3.40
SPSIR	3.50	0	12.18	14.38	3.36	3.24
Lasso-SIR	3.75	0	15.35	18.03	3.55	3.48
SSIR-EN	3.90	0	15.67	18.25	3.79	3.70

From Table 2, we can see that the SSIR-EN produces sparse models. Compared to the competitors, the SSIR-EN tends to select the true important predictors accurately. The performance of SSIR-EN was very well especially when grouped selection is required. The ability of 'grouped selection' of EN makes the performance of SSIR-EN better than the performance of the Lasso-SIR, RSIR, SPSIR and SSIR methods in term of variable selection. In all the examples, the performance of Lasso-SIR for variable selection was better than the the performance of RSIR, SPSIR and SSIR methods. In terms of selection accuracy, the performance of SSIR was the worst for all the examples.

5. Prostate cancer (P.C) data

In this section, we analysed the P.C data via the SSIR-EN, Lasso-SIR, RSIR, SPSIR and SSIR methods. The data related with prostate cancer study (Stamey et al., 1989). The data are public and available from "lasso2" package in R. The predictors are 8 of clinical indexes and $n = 97$. The predictors are: log (volume of cancer) (lcavol), log(weight of prostate) (lweight), age, log (prost. hyperplas.) (lbph), semi. ves. inv. (svi), log(caps. penetr.) (lcp), Gleas. score (gleas.) and percent. gleas. 4 or 5 (pgg45). The y is log prost.- antig. (lpsa).

The P.C data were randomly split into a training and test sets with $n = 67\ and\ 30$, respectively. Selection of penalty parameters by tenfold CV and model fitting were implemented on the training data. The performance of SSIR-EN, Lasso-SIR, RSIR, SPSIR and SSIR methods was compared via computing their prediction MSE on the test data.

Table 3. The results of Test MSE and the selected variables according to SSIR-EN, Lasso-SIR, RSIR, SPSIR and SSIR methods based on P.C. data.

Method	Test MSE	Variables selected
SSIR	0.478 (0.143)	(1,2,4,5,8)
RSIR	0.447(0.134)	(1,2,3,6)
SPSIR	0.465 (0.138)	(1,2,4,5,8)

Lasso-SIR	0.443 (0.128)	(1,2,3,6)
SSIR-EN	0.428 (0.120)	(1,2,5,6,8)

From Table 3, it is clear that the good performance of SSIR-EN was confirmed based on the analysis the P.C. data. In terms of sparsity and prediction precision, the performance of SSIR-EN is better than the performance of the competitors. In general, the performance of SSIR was the worst among the competitors. The performance of Lasso-SIR is better than the performance of RSIR, SPSIR and SSIR methods. The Lasso-SIR estimator was a good competitor for SSIR-EN. The SSIR-EN selects lcavol, lweight, svi, lcp and pgg45 as the most important predictors. Also, the prediction error of the SSIR-EN was lower than that of all the competitors.

6. Conclusion

In this article, we proposed SSIR-EN method. SSIR-EN combined the EN penalty within SIR regression type formulation. SIR can estimate $S_{y|\mathbf{x}}$ while EN does continuous shrinkage and variable selection simultaneously and it encourages groups selection of highly correlated predictors. The SSIR-EN benefits from the strength of SIR and Elastic Net. The SSIR-EN extends EN to nonlinear and multi-dimensional regression under SDR settings. Computationally, the SSIR-EN is shown to be ease implemented with an effective algorithm. The results of simulation and real data analysis showed that SSIR-EN can yield promising predictive accuracy, as well as encourages groups variable selection for the highly pairwise correlated predictors under SDR settings.

The idea of SSIR-EN can be extended to other SDR methods, such as SAVE (Cook and Weisberg, 1991) and PHD (Li, 1992). Also, the SSIR-EN can be extended to binary response models. Moreover, robust SSIR-EN is another possible extension of the proposed method.

References

Alkenani, A. and Reisan, T. (2017). Robust Group Identification and Variable Selection in Regression. Journal of Probability and Statistics. Volume 2017 (2017), Article ID 2170816, 8 pages.

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

Cook, R. (1998). Regression graphics: ideas for studying the regression through graphics. New York, Wily.

CooK, R. (2004). Testing predictor contributions in sufficient dimension reduction. Annals of Statistics 32, 1061–92.

Cook, R. and Weisberg, S. (1991). Discussion of Li (1991). Journal of the American Statistical Association 86, 328–32.

Fan, J. and Li, R. Z. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96,1348–1360.

Li, K. (1991). Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association 86, 316–342.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. Journal of the American Statistical Association 87, 1025–1039.

Li, L. (2007). Sparse sufficient dimension reduction. Biometrika 94, 603–613.

Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. Technometrics 48, 503–510.

Li, L. and Yin, X. (2008). Sliced Inverse Regression with regularizations. Biometrics 64, 124–131.

Lin, Q., Zhao, Z., and Liu J. (2019). Sparse sliced inverse regression via lasso. Journal of the American Statistical Association, pages 1–33.

Naik, P. A. and Tsai, C.-L. (2001). Single-index model selections. Biometrika 88, 821–32.

Ni, L., Cook, R. D. and Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. Biometrika 92, 242–247.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989) .Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii: radical prostatectomy treated patients. *J. Urol.*, **16**, 1076–1083.

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

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society B, 68, 49-67.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics 38, 894–942.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society, Series B 67, 301–320.
Zou, H. (2006). The adaptive Lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–142.