LASSO Variable Selection Techniques in Data Envelopment Analysis

Jia-Ying Cai
Institute of Manufacturing Information and Systems
National Cheng Kung University, Tainan City 701, Taiwan
Tel: (+886) 6-275-7575 ext.34223, Email: interact236@gmail.com

Chia-Yen Lee
Institute of Manufacturing Information and Systems
National Cheng Kung University, Tainan City 701, Taiwan
Tel: (+886) 6-275-7575 ext.34223, Email: cylee@mail.ncku.edu.tw

Abstract. The number of inputs and outputs factors has significant impacts on the production function estimation by data envelopment analysis (DEA). That is, "curse of dimensionality" is an issue when using a small number of observations for high-dimensional frontier estimation. The study conducts a data generating process (DGP) to argue the typical "rule of thumbs", e.g. the number of observations should be at least larger than twice of the number of inputs and outputs, used in DEA is ambiguous and may lead to large deviations in technical efficiency. The paper proposes a LASSO variable selection technique usually used in data mining for extracting significant factors in the formulation of sign-constrained convex nonparametric least squares (CNLS) as DEA, and the results show that the proposed LASSO-CNLS method is useful for providing guidelines of dimension reduction in DEA.

Keywords: data envelopment analysis, LASSO variable selection, efficiency estimation, convex nonparametric least squares, dimension reduction

1. INTRODUCTION

Data envelopment analysis (DEA) is a nonparametric approach for estimating frontiers by decision making units (DMUs). As we all known, when performing DEA analysis, there are several advantages to having larger data sets. However, in a general situation, it is time-consuming and costly to collect more data sets. In this case, if we use a smaller size of data set, we will face the problem of the curse of dimensionality which is a phenomena happens in high-dimensional spaces. That is, using a small size sample (i.e., DMUs) for high-dimensional frontier estimation. If doing that, there is a problem of imprecise estimation of production frontiers because most of the efficiency score of each DMU is going to be one. Therefore, variable selection raises an issue.

Variable selection aims to remove the less important variables and keep the significant inputs and outputs. By removing the unimportant variables, we can estimate our production frontiers more precisely and release the problem of the curse of dimensionality as well. Nataraja and Johnson (2011) also promoted variable selection methods in DEA and showed that DEA loses explanatory power as the dimensional space increases. This study proposes a LASSO variable selection technique for extracting significant factors.

The paper is organized as follows. Section 2 presents the rule of thumb of minimal requirements in number of observations and a validation of the insufficient minimal requirements. Section 3 develops the proposed LASSO-convex nonparametric least square (CNLS) method and estimates the accuracy of production frontiers by calculating Mean Square Error (MSE). And then we'll give the conclusion and future study in section 4.

2. LITERATURE REVIEW

In DEA literatures, it's better to have larger data sets, but there are some criteria proposed for the minimal requirements. There are some "rule of thumbs" proposed in literatures (e.g., the number of observations should be at least larger than twice of the number of inputs and outputs). Table 1 shows the promotion about minimal size of data sets from different scholars, where m means the total number of outputs and n means the total number of inputs. Boussofiane et al. (1991) said that for effective discrimination and the flexibility in the choice of weights, the number of inputs and outputs which is selected has to be smaller than the total number of DMUs. A specific ratio of an input to an output is large enough can own all its weight and become efficiency. The total number of such ratios will be the product of the number of efficient units. Hence the minimal number of DMUs should multiply the number of inputs and outputs.

Golany and Roll (1989) expounded that a larger set of units enables a sharper identification of typical relations between inputs and outputs in the set. A rule of thumb established in the paper is that the number of units should be at least twice the number of inputs and outputs considered. Bowlin (1998) elaborated that a general rule of thumb is that three decision making units are needed for each input and output variable used in the model for the purpose of insuring sufficient degree of freedom for a meaningful analysis. Dyson (2001) suggested that to achieve a reasonable level of discrimination, the DMUs need the number of units to be at least two times of the number of inputs multiply the number of outputs. For instance with a three inputs and four outputs model, Boussofiane et al. (1991) recommended 12 DMUs, Golany and Roll (1989) suggested 14 DMUs, Bowlin (1998) recommended 21 DMUs, and Dyson (2001) suggested 24 DMUs.

Table 1: Minimal Size of Data Set Promoted from different scholars

Scholar	Promotion
Boussofiane et al. (1991)	m*n
Golany and Roll (1989)	2(m*n)
Bowlin (1998)	3(m*n)
Dyson et al. (2001)	2m*n

Nataraja and Johnson (2011) used four methods to testify the selection process in DEA, including efficiency contribution measure (ECM), principle component analysis (PCA-DEA), a regression-based test (RB), and bootstrapping. The result showed that in higher correlated inputs (larger than 0.8) with smaller data set (less than 300 DMUs) PCA-DEA performs well. Otherwise, RB and ECM is a good choice under lower correlation (smaller than 0.2) and larger data set (at least 300 DMUs). However, the implement of these methods is time-consuming and PCA transforms multiple variables into PCs for dimension reduction rather than removing the original variable.

This study proposes a LASSO variable selection technique for variable selection in DEA when dimension is high or data set is limited. The reason to propose LASSO method is because LASSO is a regression-based method and also DEA can be regarded as a sign-constraint CNLS formulation (Kuosmanen and Johnson, 2010) which builds efficient frontier via regression hyperplane. Thus, LASSO fits the variable selection in DEA. Furthermore, LASSO provides the slope and coefficients of regression hyperplane and shrink the coefficients to zero when penalty parameter becomes larger. This investigates the effect of input factors on the production function and implies the production function can be rationally estimated with fewer variables.

3. Curse of Dimensionality in DEA

This study is separated into two parts. The first part provides a proof of an insufficient number of DMUs (i.e., observations) causing the issue of curse of dimensionality. The second part proposes the LASSO-CNLS method for dimension reduction.

3.1 A Validation of Insufficient DMUs

In this section, we will illustrate our data generating process (DGP) and then argue the insufficient of typical rule of thumbs by comparing the MSE in different dimensions.

3.1.1 Data Generating Process (DGP)

In data generation, we follow Nataraja and Johnson (2011) and randomly generate inputs (x_i) and inefficiency term (μ). As the paper suggested, the values for the inputs are independently and identically distributed (i.i.d.) and generated from a uniform distribution on the interval (10, 20), and the inefficiency term (μ) is half-normal with mean zero and variance 0.7. This study uses Cobb-Douglas production function VRS model to calculate our output, y and y^{true}. Parameter y^{true} is the true frontier which is not affected by noise while y represents the output that is affected by noise. An example can be seen in Figure 1. True production function (y^{true}) should be a "smooth" production function passing through the "origin". Frontier y will be lower than y^{true} because of the noise. And we try to get closer to the true frontier by adding our number of observations. Eq. 1 shows Cobb-Douglas production function y. Eq. 2 is the true production function that we assumed. Here, i is the index of input, k is the index of observation, n is the total number of input, x_i means i^{th} input, μ is the inefficiency term and e is Euler's number.

$$y = \prod_{i=1}^{n} x_{ki}^{(\frac{1}{n+1})} * e^{-\mu} \quad ,\forall k \tag{1}$$

$$y^{true} = \prod_{i=1}^{n} x_{ki}^{(\frac{1}{n+1})} \quad , \quad \forall k \tag{2}$$

Table 2: MSE value in different observations of each dimension

Dimension = 4		**Dimension = 5**		**Dimension = 6**		**Dimension = 7**	
observation	**MSE**	**observation**	**MSE**	**observation**	**MSE**	**observation**	**MSE**
12	5.302	12	9.014	12	12.5	12	16.013
25	3.481	25	6.807	25	9.676	25	13.291
50	2.434	50	4.79	50	6.988	50	10.407
100	1.442	100	3.104	100	5.454	100	7.927
200	0.896	200	2.121	200	3.855	200	6.119
300	0.65	300	1.632	300	2.937	300	4.865
500	0.455	500	1.166	500	2.355	500	3.866
1000	0.278	1000	0.717	1000	1.538	1000	2.682
Dimension = 8		**Dimension = 9**		**Dimension = 10**		**Dimension = 20**	
observation	**MSE**	**observation**	**MSE**	**observation**	**MSE**	**observation**	**MSE**
12	19.1	12	21.767	12	23.163	12	32.315
25	15.814	25	18.745	25	21.507	25	32.156
50	13.404	50	16.037	50	18.489	50	32.054
100	10.878	100	13.954	100	16.727	100	31.615
200	8.359	200	11.065	200	13.799	200	30.675
300	6.904	300	9.321	300	12.284	300	30.25
500	5.676	500	7.994	500	10.253	500	29.352
1000	4.115	1000	5.893	1000	8.0374	1000	27.7476

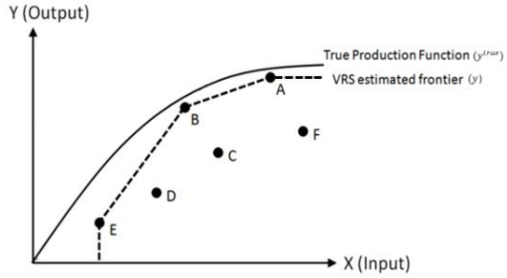
Y (Output)
True Production Function (y^{true})
VRS estimated frontier (y)
A
B
C
D
E
F
X (Input)

Figure 1: Difference between y and y^{true}
(Lee and Johnson, 2015)

After finishing generating process, we calculate the efficiency by DEA output-oriented dual VRS model which is shown in Eq. 3.

Max θ_r

Subject to

$$\begin{array}{l} \sum_{k=1}^{p}\lambda_k * x_{ki} \le x_{ri},\ \forall i \\ \sum_{k=1}^{p}\lambda_k * y_{kj} \ge \theta_r * y_{rj},\ \forall j \\ \sum_{k=1}^{p}\lambda_k = 1 \\ \lambda_k \ge 0 \end{array} \qquad (3)$$

Where r indicates one specific observation, and r is alias of index k. θ is a decision variable for efficiency estimation. If $\theta = 1$, then the DMU is efficient; otherwise it is inefficient if $\theta > 1$. λ_k is a decision variable representing the intensity multiplier for linear combination of DMUs. Index p is the total number of observations and j is the index of output. The first constraint and the second constraint represent input and output constraints. The third constraint is used for convex combination of DMUs representing the variable returns of scale (VRS) DEA. The final constraint is non-negativity constraint.

3.1.2 MSE Calculation to Prove Inadequate DMUs

We calculate the efficiency when dimension is 4, 5, 6, 7, 8, 9, 10 and 20 separately. Dimension is 4 means that we have 3 inputs and 1 output, and dimension equals 5 means we have 4 inputs and 1 output and so on. In each dimension, we give different number of observations (12, 25, 50, 100, 200, 300, 500 and 1000) and run 100 times of replication and then take the average MSE. Eq. 4 shows the formula of MSE.

$$MSE = \frac{\sum_{r=1}^{p}[y_r^{true}-(\theta_r * y_r)]^2}{n} \qquad (4)$$

The MSE in different observations of each dimension is shown in Figure 2 and Table 2. In Figure 2, we can find out that MSE decreases exponentially in each dimension while observations increase. Therefore, the minimal requirements scholars promoted are not sufficient and the data set should be increased exponentially while the dimension become higher. Table 2 shows the MSE value in detail.

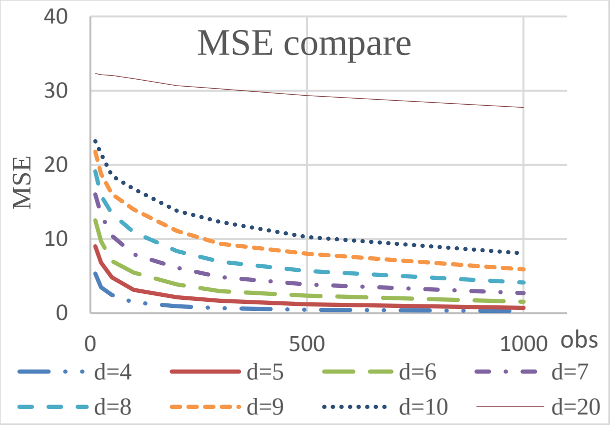

Figure 2: MSE compare in chart

3.2 A New Model for Dimension Reduction

In this section we will first introduce the LASSO variable selection technique and prove the formulation of sign-constrained convex nonparametric least squares (CNLS) as DEA. In the last of this section, we will show how the new model selects the significant variables.

3.2.1 Introduction of LASSO

LASSO (Tibshirani, 1996) is a common technique for variable selection because of its maintaining prediction accuracy and discovering relevant variables. A good selection procedure should have some oracle properties and also continuous shrinkage. However, LASSO has been shown inconsistent results in some scenarios. Besides, LASSO is sensitive to outliers (Zou, 2006).

LASSO is a variant of ridge regression which shrinks coefficients by imposing penalty on their size (Hastie, Tibshirani and Friedman, 2008). The formulation about ridge regression can be seen in Eq. 5.

$$\beta^{ridge} = \operatorname{argmin}\left\{\frac{1}{2}\left(\sum_{i=1}^{N} y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{P} {\beta_j}^2\right\} \qquad (5)$$

The concept of LASSO is that it uses penalty term to shrink coefficients of each variable. The formula is shown in Eq. 6. By increasing the value of lambda, we can shrink coefficients of less important variables to 0, so as to get significant variables.

$$\beta^{lasso} = \operatorname{argmin}\left\{\frac{1}{2}\left(\sum_{i=1}^{N} y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{P} |\beta_j|\right\} \qquad (6)$$

An example about the difference between ridge regression and LASSO is shown in Figure 3 and 4 (Hastie, Tibshirani and Friedman, 2008). Ridge regression is a model with proportional shrinkage while LASSO penalizes β and truncates at zero. Figure 3 describes an example of ridge regression in prostate cancer. The words on the right hand side of the figure are variables (e.g., lcavol, svi lweight, and so on). It is important to say that $\mathrm{df}(\lambda) = \mathrm{p}$ when $\lambda = 0$ and $\mathrm{df}(\lambda) \to 0$ when $\lambda \to \infty$. Therefore, while λ value becomes higher (i.e., $\mathrm{df}(\lambda)$ gets lower), the coefficient of each variable shrinks. Figure 4 illustrates an example of LASSO in prostate cancer. When λ value becomes higher (i.e., shrinkage factor gets lower), the coefficient of each variable shrinks and truncates at zero.

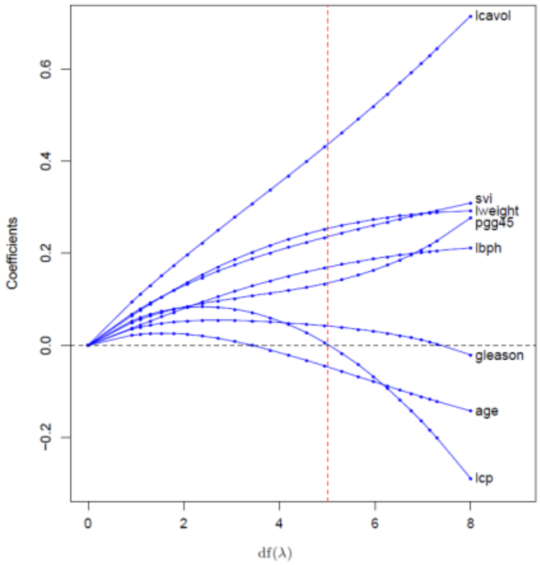

Figure 3: An example of ridge regression (Hastie et al., 2008)

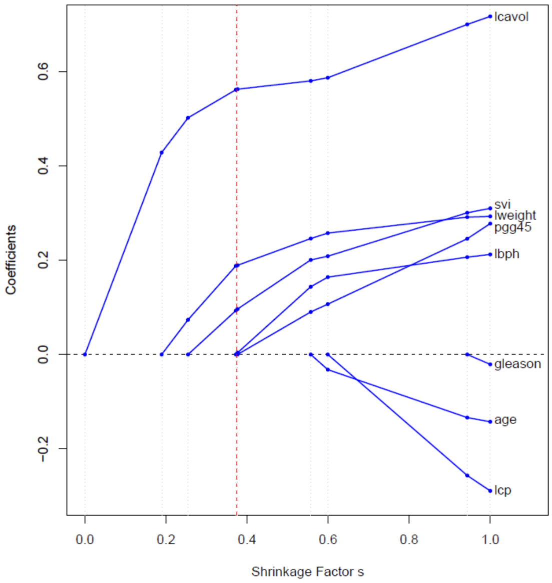

Figure 4: An example of lasso (Hastie et al., 2008)

3.2.2 DEA As a Sign-constrained CNLS

Kuosmanen and Johnson (2010) proved that DEA is a "sign-constrained" CNLS when we add $\varepsilon_k < 0$ in CNLS model. Before the combination of Lasso and sign-constrained CNLS, we use numerical experiment to verify it. When dimension equals to 10, we calculate the MSE of original DEA model and sign-constrained model in different observations (12, 25, 50 and 100). Repeating for 100 times and taking the average MSE in different observations. The result can be seen in Table 3.

Table 3: Comparison between DEA model and Sign-constrained CNLS model

DEA model		Sign-constrained CNLS	
observations	MSE	observations	MSE
12	23.558	12	23.558
25	21.6	25	21.6
50	18.159	50	18.163
100	16.095	100	16.081

3.2.3 LASSO-CNLS model

In this section we apply the concept of LASSO technique into "sign-constrained" CNLS model and try to reduce our dimensions without losing the explanatory power of our data set. The formulas can be seen in Eq. 7.

$$Min\ {\varepsilon_k}^2 + \lambda \sum_{k=1}^{p} \sum_{i=1}^{n} \beta_{ki}$$

Subject to

$$y_k = \alpha_k + \sum_{i=1}^{n} \beta_{ki} * x_{ki} + \varepsilon_k \ \ , \forall k \qquad (7)$$

$$\alpha_\mathrm{k} + \sum_{i=1}^{n} \beta_{ki} * x_{ki} \le \alpha_h + \sum_{i=1}^{n} \beta_{hi} * x_{ki},\ \forall \mathrm{k}, \mathrm{h} \text{ and } \mathrm{k} \ne \mathrm{h}$$

$$\varepsilon_k < 0 \ \ , \forall \mathrm{k}$$

$$\beta_{ki} \ge 0 \ \ , \forall \mathrm{k}, \mathrm{i}$$

Where ε_k is the composite error that represents the deviation of observation k from the estimated function. Decision variables α_k and β_{ki} characterize the intercept and slope parameters regarding the marginal products of the inputs for each observation. The objective function minimizes the sum of the square with respect to the disturbance terms. The first equality constraint represents a basic linear regression for each observation k, that is, there are m different regression lines estimated rather than one specific line as in OLS. The second inequality constraint imposes concavity using Afriat inequalities which are the key points in modeling concavity constraints in multiple regression setting. The third inequality imposes the negative sign on error for formulating DEA frontier. The last constraint imposes monotonicity of the inputs on the underlying unknown function.

Due to multiple solutions in this model, we propose a variant of the model suggested by Kuosmanen and Johnson (2010) to address the issue. The formulation is shown as Eq. 8 for one specific observation r.

$$Min\ \alpha + \sum_{i=1}^{n} \beta_i * x_{ri}$$

Subject to

$$\alpha + \sum_{i=1}^{n} \beta_i * x_{ki} \ge \hat{\alpha}_k + \sum_{i=1}^{n} \hat{\beta}_{ki} * x_{ki} \ \ , \forall k \qquad (8)$$

$$\beta_i \le \sum_{k=1}^{p} \hat{\beta}_{ki} \ \ , \forall \mathrm{i}$$

Where $\hat{\alpha}_k$ and $\hat{\beta}_{ki}$ are the optimal solution obtained from equation (7). The objective function minimizes the linear regression line for each specific observation because only the minimum bound satisfies monotonicity and concavity properties. The first constraint is used for showing that the new regression line must larger than the regression line we obtained in equation (7). The second constraint ensures the new value of β_i smaller than the value of $\hat{\beta}_{ki}$ that we obtained before.

We calculate these two models sequentially in GAMS solver by increasing lambda (i.e., penalty) value each time so as to decrease the dimensions for ten times. In the following process, we calculate the MSE in each dimension of each replication. The result can be seen in Figure 5. In Figure 5, we let number of observations equal 25 and dimensions equal 10 (i.e., 9 inputs and 1 output) and random generate data from Eq. 1 and Eq. 2 for ten times (i.e., data1, data2 and so forth). Each time we increase lambda value in order to decrease the dimensions. We calculate MSE in each dimension by Eq. 4 and give this chart. In Figure 6, we take the average of MSE calculated by these 10 replications in each dimension. In Figure 6, we can find that when dimensions reduced, the MSE decreased as well. It means that we eliminate the less important variables and get closer to the true frontier.

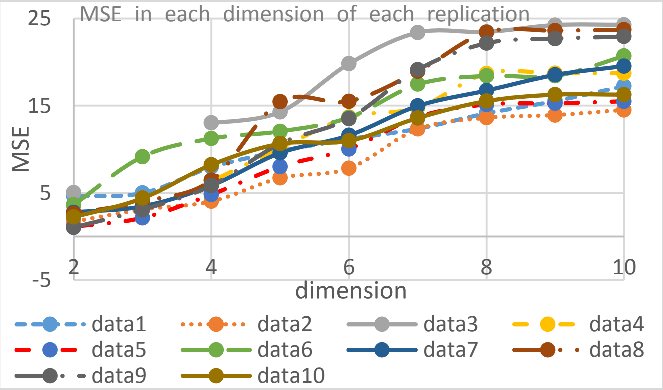

Figure 5: MSE in each dimension of each replication

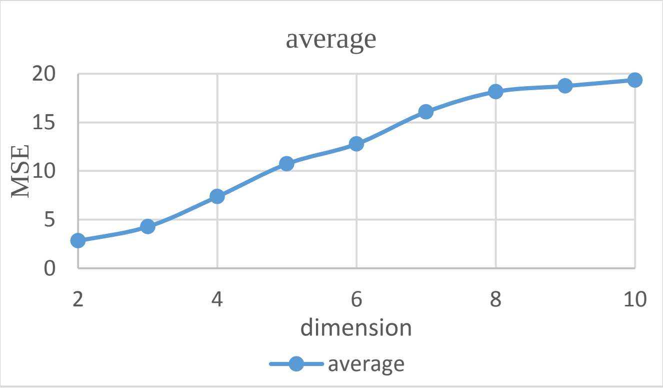

Figure 6: Average MSE of each dimension

4. Conclusion

By combing LASSO and sign-constrained CNLS, we successfully remove some unimportant variables for addressing curse of dimensionality. That is, we keep more critical variables. Therefore, if we compare MSE between the variable LASSO chosen and another variable which was not picked, MSE of previous one should be lower.

In Table 4, there is a comparison about what we mentioned above. The variable Lasso chose in the final is in bold while others are not.

We can find out that though Lasso may not do well in choosing the most significant variable since it is a bias estimator, it provide an effective way for dimension reduction efficiently.

In the future work, due to a bias estimate chosen by LASSO estimator, our suggestion is to use adaptive LASSO which tries to give larger penalty to zero coefficients and smaller penalty to nonzero coefficients and tries to decrease the bias of estimation and increase selection accuracy (Zou, 2006).

REFERENCES

Boussofiane, A., Dyson, R.G., and Thanassoulis, E. (1991) *Applied Data Envelopment Analysis,* European Journal of Operational Research 52, 1-15.

Bowlin, W.F. (1998) *Measuring Performance: An Introduction to Data Envelopment Analysis (DEA),* Journal of Cost Analysis 7, 3-27.

Dyson, R.G., Allen, R., Camanho, A.S., Podinovski, V.V., Sarrico, C.S., and Shale, E.A. (2001) *Pitfalls and Protocols in DEA*, European Journal of Operational Research, 132, 245-259.

Golany, B. and Roll, Y. (1989) *An Application Procedure for DEA,* Omega 17, 237-250.

Hastie, T., Tibshirani, R., and Friedman, J. (2008) *The Elements of Statistical Learning,* Springer Series in Statistics, Second Edition.

Kuosmanen, T., Johnson, A. L. (2010) *Data Envelopment Analysis as nonparametric Least-Square Regression,* Operations Research Vol.58, No. 1, January-February 2010, pp. 149-160.

Lee, C-Y, Johnson, A. L. (2015) *Measuring Efficiency in Imperfectly Competitive Markets: An Example of Rational Inefficiency*, Journal of Optimization Theory and Applications, 164 (2), 702–722.

Nataraja, N. R., Johnson, A. L. (2011) *Guidelines for using variable selection techniques in data envelopment analysis,* European Journal of Operational Research 215 (2011) 622-669.

Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso,* J. Roy. Statist. Soc.Ser. B 58 267-288. MR1379242

Zou, H. (2006) *The adaptive Lasso and its oracle properties,* J. Amer. Statist. Assoc. 101, 1418-1.

Table 4: MSE comparison in different variables

Data1	*MSE*	*Data2*	*MSE*	*Data3*	*MSE*	*Data4*	*MSE*	*Data5*	*MSE*
x8	0.261	x9	0.385	x9	1.228	x9	0.277	x7	0.4
x6	0.363	x6	0.953	x7	1.396	x5	0.377	x6	0.494
x7	0.634	x5	0.989	x1	2.516	x2	0.416	x4	0.814
x9	0.846	x1	1.184	x8	3.952	x8	1.433	x3	1.038
x1	0.874	x7	1.401	**x3**	**5.011**	x6	1.891	**x2**	**1.089**
x2	2.042	x2	1.517	x2	5.128	x4	2.478	x9	1.365
x3	3.297	x8	1.619	x5	5.324	x1	2.479	x1	1.958
x4	**4.644**	**x4**	**1.704**	x4	6.386	x3	3.017	x8	2.841
x5	5.403	x3	1.747	x6	7.399	**x7**	**3.496**	x5	3.87

Data6	*MSE*	*Data7*	*MSE*	*Data8*	*MSE*	*Data9*	*MSE*	*Data10*	*MSE*
x2	0.483	x2	1.104	x4	0.422	**x8**	**1.016**	x5	0.25
x7	0.715	x6	1.121	x3	0.743	x4	1.387	x2	0.274
x6	1.017	x1	1.519	x6	1.222	x9	1.752	x4	0.289
x1	1.562	x7	1.705	x2	1.317	x5	1.838	x1	1.187
x5	1.638	x3	1.946	x7	2.386	x3	2.57	x8	1.258
x9	3.412	x8	2.666	**x9**	**2.616**	x6	2.86	**x6**	**2.216**
x4	3.549	**x4**	**2.755**	x8	2.819	x2	2.883	x7	2.228
x8	3.593	x9	2.811	x1	3.889	x7	4.094	x9	2.272
x3	**3.628**	x5	3.81	x5	7.358	x1	5.111	x3	2.424