Optimization Writing Assignment

L1-constrained fitting for statistics and data mining (LASSO)

1st Dajin Han
Hanyang University
Seoul, Republic of Korea
dajinstory@hanyang.ac.kr

***Abstract*—This report has a content about L1-constrained fitting methods and researches like research paper. There are summary of each paper.**

***Index Terms*—LASSO, L1-Constrained Fitting, Data Mining, Bridge Regression**

I. Researches

A. Regression Shrinkage and Selection via the Lasso

B. The Lasso Method for Variable Selection in the Cox Model

C. Penalized Regressions: The Bridge Versus the Lasso

D. Asymptotics for Lasso-type Estimators

II. Regression Shrinkage and Selection via the Lasso

A. Prior Studies

There are ordinary least squares(OLS) which is an loss function obtained by minimizing the residual squared error. But this loss function is not used alone because there are obvious drawback in this method. First, prediction accuracy is not good because of the ”large variance”. Some researchers judged that the cause of large variance is small bias ironically so they sacrifice a little bias to reduce the variance and this improved the overall prediction performance. Second drawback is ”Interpretation”. It is hard to interpret the model trained with very large number of predictors. Traditional techniques to solve above two drawback are ”Subset selection” and Ridge regression” each. But subset selection often ruin the performance because artificial small subset selection is lack of general attribute so affect to the final prediction performance bad. Ridge regression cannot set coefficients to 0 and this does not make model simpler and easier to interpret. So Author suggest the Lasso.

B. What is Lasso

$(x^i, y_i), i = 1, 2, ..., N$, where $x^i = (x_{i1}, ..., x_{ip})^T$ are the dataset with the predictor variables x and the responses y. We assume either that the observation are independent or that the y_is are conditionally independent about given x_{ij}s. So generalize these variable by average value and the standard deviation. So $\Sigma_i x_{ij} = 0, \Sigma_i x_{ij}^2/N = 1$.

$$(\hat{\alpha}, \hat{\beta}) = \arg\min \Sigma_i y_i - \alpha - \Sigma_j \beta_j x_{ij} \text{ subject to } \Sigma_j |\beta_j| \leq t. \tag{1}$$

Then we omit α to generalize the loss. Equation (1) is solved by quadratic programming problem with Linear inequality constraints. parameter t constrols the amount of shrinkage for model performance. By doing so, some coefficients become 0, and this is called shrinkage.

$$\Sigma_i(y_i - \alpha - \Sigma_j c_j \hat{\beta}_j^0 x_{ij})^2 \text{ subject to } c_j \geq 0, \Sigma c_j \leq t. \tag{2}$$

Lasso is motivated from the proposal of non-negative garotte minimizes by Breiman(1993). The Garotte run OLS first, and then next add some constraints on sum of c_j to t. The garotte get lower prediction error than the subset selection and is competitive against to ridge regression. Only the drawback is the condition when the model hass too many small non-zero coefficients. The garotte advanced the OLS estimation but the thing that garotte advanced from OLS is the drawback of garotte. The OLS's own limitation is not solved. So Lasso avoids the explicit use of the OLS estimation to overcome the drawback of OLS.

$$\hat{\beta}_j = sign(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+ \tag{3}$$

Let X be the nxm dimension matrix with (i,j)th value is x_{ij} and $X^T X = I$.This equation show that the form of soft shrinkage and a minimum L_1-norm penalty. The connection between soft shrinkage and a minimum L_1-norm penalty is now pointed out by Donoho et al. With the idea of orthonormal design, selecting best subset of size k is just to pick k largest coefficient and set the rest to 0. For some case, this is equivalent to setting $\hat{\beta}_j = \hat{\beta}_j^0$ if $|\hat{\beta}_j^0| > \lambda$. Ridge regression minimizes

$$\Sigma_i(y_i - \Sigma_j \beta_j x_{ij})^2 + \lambda \Sigma_j \beta_j^2 \tag{4}$$

or, equivalently, minimizes

$$\Sigma_i(y_i - \Sigma_j \beta_j x_{ij})^2 \text{ subject to } \Sigma \beta_j^2 \leq t \tag{5}$$

The ridge solutions are

$$(1 - \frac{\gamma}{\hat{\beta}_j^{0\,2}})^+ \hat{\beta}_j^0 \tag{6}$$

This shows that the ridge regression scales the coefficients by a fixed, constant value however the Lasso adjust factors then translate the constants and then truncating to zero constant to specific constraints which judged to be less important. Lasso function is similar to the garotte function advanced mechanism based on OLS. The lasso is more shrinkage for larger coefficient. Strengths of lasso over garotte become larger when the function design is not orthogonal.

Lasso often adjust its own coefficients that would be multiplied to constraints to 0. This is general situation in lasso algorithm. It is confused. Because it is natural to think that while ridge regression progressing this situation is occured. But This is not occur with ridge regression and also lasso use the constraint $\Sigma\beta_j^2 \leq t$ rather than $\Sigma|\beta_j| \leq t$.

$$(\beta - \hat{\beta}_0)^T X^T X(\beta - \hat{\beta}^0) \tag{7}$$

In equation (7), this quadratic function is equals to the criterion $\Sigma_i^N y_i - \Sigma_j \beta_j x_{ij}{}^2$.

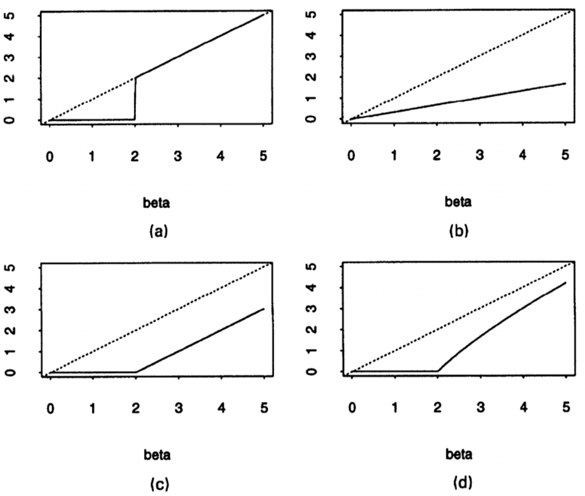


Fig. 1. (a) Subset regression, (b) ridge regression, (c) the lasso and (d) the garotte

It is clear that the lasso is overperform other method or algorithms like subset regression, ridge regression and garotte methods. lasso performs good on small subset with the value almost closed to zero and also for the larger subset size, lasso showed more stable performance against garotte on Figure 1. The garotte algorithm tends to closed to reference line with the larger subset size.

In Figure 2,Estimation pictures for lasso and ridge regression each is appeared. These full curves are centered at the OLS estimation which is the constraint region colored black. Let the point p_{opt} where the ellipse meet the constraint region first. For the lasso, constraint region is rotated square with 45^o angle. When the optimized point p_{opt} is on axis, it means that the coefficient becomes zero. The ellipse function and the colored region meet on the vertex or edge of colored region. For the lasso, optimized point is at the vertex of colored region with high probability, which is on the axis, so the coefficient becomes zero with high probability. However the standard ridge regression method is hard to make coefficient to zero because there is no specific vertex on axis with the shape to be easily met. In conclusion normal standard ridge regression could get the small coefficient but is hard to get the zero coefficient like lasso.

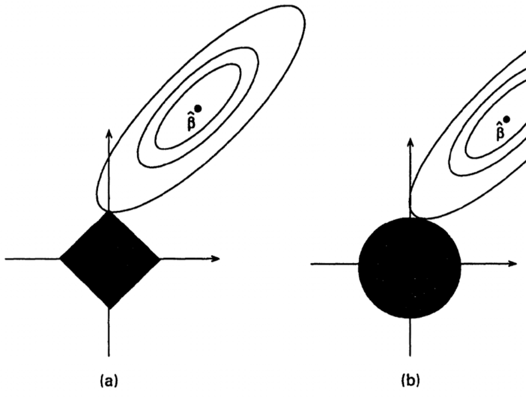


Fig. 2. (Estimation picture for (a) the lasso and (b) ridge regression

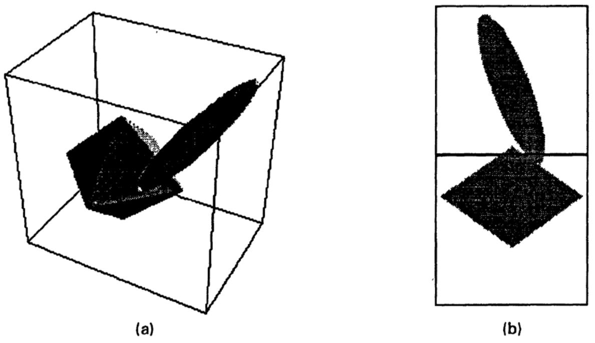


Fig. 3. (a) Example in which the lasso estimate falls in an octant different from the overall least squares estimate; (b) overhead view

Figure 3 shows an three dimension example. Garotte keep the sign of $\hat{\beta}_j^0$ and the lasso can change the sign of $\hat{\beta}_j^0$. Because the value keeps sign, the model can be re-defined. The model $Sigmac_j\hat{\beta}_j^0 x_{ij}$ with constraint $\Sigma c_j \leq t$ can be redefined as the new model $\Sigma\beta_j x_{ij}$ with constraint $\Sigma\beta_j/\hat{\beta}_j^0 \leq t$. With these new model and constraints, larger values of β_1 and smaller values of β_2 are gained.

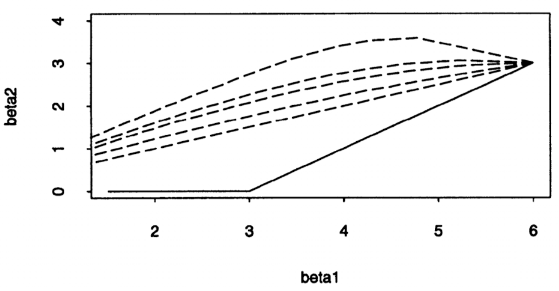


Fig. 4. Lasso and ridge regression for the two-predictor example: the curves show the (β_1, β_2) pairs as the bound on the lasso or ridge parameters is varied; starting with the bottom broken curve and moving upwards, the correlation ρ is 0, 0.23, 0.45, 0.68 and 0.90

$$\hat{\beta} = (\hat{\beta}_j^0 - \gamma)^+ \tag{8}$$

$$\hat{\beta}_1 = (\frac{t}{2} + \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{)^+} \tag{9}$$

$$\hat{\beta}_2 = (\frac{t}{2} - \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{)^+} \tag{10}$$

In figure 4, 100 generated data points from the model $y = 6x_1 + 3x_2$ with no noise. x_1 and x_2 are standard normal variations with correlation ρ. In Figure 4, the curve shows that the normal line representing lasso is at the most bottom position. Curves like dotted line are representing the ridge regression with different ρ. From bottom to upwards, ρ increase. For the bigger ρ the line sometimes has tendancy to get the decreased bound and this is because of the strong tendancy to minimize the squared norm regardless of the overall performance.

Lasso estimation is a non-linear and non-differential function. So it is hard to get the powerful performance, higher accuracy with its standard error. One suggested approach is bootstrap. Progressively update t by analogous method to select the best subset sample and then use the least square standard error for the selected subset. By approximating the penalty $\Sigma|\beta_j|$ like $\Sigma\beta_j^2/|\beta_j|$. By approximating the solution using ridge regression with the form $\beta^* = (X^TX + \lambda W^-)^{-1}X^Ty$, a diagonal matrix W with diagonal elements $|\hat{\beta}_j$. λ and W^- is for generalized inverse W. So the $\Sigma|\beta_j|^* = t$. Keep calculating with the below equation.

$$(X^TX +^-)^{-1}X^TX(X^TX + X^-)^{-1}\hat{\sigma}^2 \tag{11}$$

$\hat{\sigma}^2$ is an error variance variable. By converting the equation easier to approach with linear programming, iterated ridge regression algorithm is covered. There is one drawback that this mechanism is not efficient but that is resolved with useful selection of lasso parameter t.

Figure 5 shows that the absolute value of coefficient tends to be 0 when s goes to 0, but not general. This is the lack of lasso. This is caused by the ridge regression and subset regression. On the broken line with $s = 0.44$, the optimal value is selected through cross validation calculation.

Predictor	Least squares results			Subset selection results			Lasso results		
	Coefficient	*Standard error*	*Z-score*	*Coefficient*	*Standard error*	*Z-score*	*Coefficient*	*Standard error*	*Z-score*
1 intcpt	2.48	0.07	34.46	2.48	0.07	34.05	2.48	0.07	35.43
2 lcavol	0.69	0.10	6.68	0.65	0.09	7.39	0.56	0.09	6.22
3 lweight	0.23	0.08	2.67	0.25	0.07	3.39	0.10	0.07	1.43
4 age	−0.15	0.08	−1.76	0.00	0.00	—	0.00	0.01	0.00
5 lbph	0.16	0.08	1.83	0.00	0.00	—	0.00	0.04	0.00
6 svi	0.32	0.10	3.14	0.28	0.09	3.18	0.16	0.09	1.78
7 lcp	−0.15	0.13	−1.16	0.00	0.00	—	0.00	0.03	0.00
8 gleason	0.03	0.11	0.29	0.00	0.00	—	0.00	0.02	0.00
9 pgg45	0.13	0.12	1.02	0.00	0.00	—	0.00	0.03	0.00

TABLE I
RESULTS FOR THE PROSTATE CANCER EXAMPLE

In table 1, the results of the prostate cancer example is shown. The lasso results is balanced with both coefficient, standard error and Z-score. Both coefficient and Z-score is better than rest methods, and standard error is also competitive against the rest methods. The result of lasso has relatively worse standard error (of course it is competitive enough) because lasso method pointed out that the prior methods minimize standard error too much and it occured the degradation of overall model accuracy. Lasso adjust the standard error adequately and get the accurate accuracy on coefficient, standard error and Z-score.

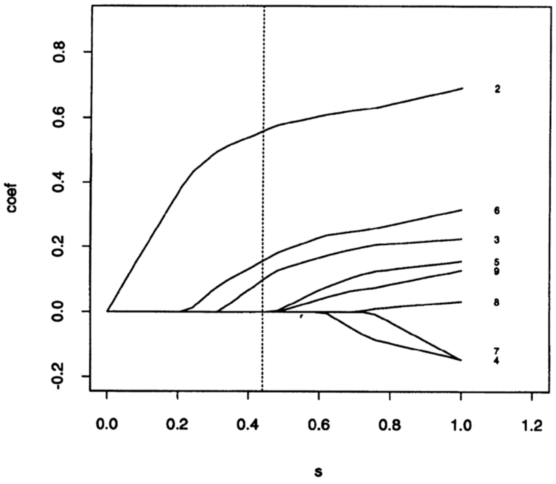


Fig. 5. Lasso shrinkage of coefficients in the prostate cancer example: each curve represents a coefficient (labelled on the right) as a function of the (scaled) lasso parameter $s = t/\Sigma|\beta_j^0|$, and the broken line represents the model for $\hat{s} = 0.44$, selected by generalized cross-validation

Predictor	Coefficient	Bootstrap standard error		Standard error approximation (7)
		Fixed t	*Varying t*	
1 intcpt	2.48	0.07	0.07	0.07
2 lcavol	0.56	0.08	0.10	0.09
3 lweight	0.10	0.06	0.08	0.06
4 age	0.00	0.04	0.05	0.00
5 lbph	0.00	0.04	0.07	0.00
6 svi	0.16	0.09	0.09	0.07
7 lcp	0.00	0.03	0.07	0.00
8 gleason	0.00	0.02	0.05	0.00
9 pgg45	0.00	0.03	0.06	0.00

TABLE II
STANDARD ERROR ESTIMATES FOR THE PROSTATE CANCER EXAMPLE

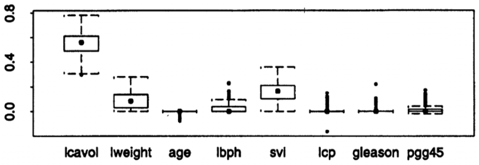


Fig. 6. Box plots of 200 bootstrap with the values of lasso coefficient

In table 2, standard error on various predictor is shown. There are various ridge approximation about equation(11) with fixed bootstrap t. The results show that the approximations works stable and specific result on the case with the coefficient being 0. The ridge regression model works bad on the case with 0 coefficient but lasso covered that cases. So in Table

1,2 and Figure 6 we can see the stability of lasso model on various case.

C. Methods for estimating the lasso parameter

The lasso parameter t could be approached by cross-validation, generalized cross-validation and the analytical unbiased estimation. First two methods are advanced version of 'X-random' case that assumed the (X, Y) is unknown distribution. The last method is based on "X-fixed" case. But in real-life case, those methods are not distinguished clearly so in this paper convenient method is introduced.

1) Mean Squared Error: There are function

$$Y = \eta(X) + \epsilon$$

Define ME as the mean squared error with zero average and the standard derivation value.

$$ME = E(hat\eta(X) - \eta(X))^2$$

The expected value is calculated via joint distribution of X and Y. $\hat{\eta}(x)$ is fixed.

$$PE = E(Y - \hat{\eta}(X))^2 = ME + \sigma^2 \tag{12}$$

To trace the simulation result, ME is often used rather than PE. The form of mean squared error becomes simple. with population covariance matrix X.

$$ME = (\hat{\beta} - \beta)^T V(\hat{\beta} - \beta)$$

2) Generalized Cross Validation: We can get GCV via linear approximation to the lasso estimation. Lagrangian penalty and the reidual sum of squares becomes the constraint of β. constrained solution $\hat{\beta}$ on the ridge regression is

$$\hat{\beta} = (X^T X + \lambda W^-)^{-1} X^T y \tag{13}$$

with $W = diag(|\hat{\beta}_j|)$

$$p(t) = tr(X(X^T X + \lambda W^-)^{-1} X^T))$$

Use $rss(t)$ be the residual sum of squares of the constrained fit with constraint t. Then can get the Generalized cross validation value.

$$GCV(t) = \frac{1}{N} \frac{rss(t)}{(1 - p(t)/N)^2} \tag{14}$$

3) Lasso parameter: To get the lasso parameter, we need to use method based on Stein's unbiased estimation. z is normal random vector with mean μ. Via the differential formula we can get

$$E_\mu ||\hat{\mu} - \mu||^2 = p + E_\mu(||g(z)||^2 + 2\Sigma_p dg_i/dz_i) \tag{15}$$

From the Equation (15), replace the original variable to different expression from other constraints. then we can get the γ by minimizer $R(\hat{\beta}(\gamma))$

$$\hat{\gamma} = \arg\min_{\gamma \geq 0} R(\hat{\beta}(\gamma))$$

$$\hat{t} = \Sigma(|\hat{\beta}_j^0| - \hat{\gamma})^+$$

The derivation of $\hat{t}$ assumed that the orthogonal design would be adjust on function. With these three methods, computational statistical calculation can be used on this problem.

D. Bayes Estimation

We can derive the lasso estimation as the Bayes posterior mode under independent double exponential states about β_j and this function is appeared.

$$f(\beta_j) = \frac{1}{2\tau} exp(-\frac{|\beta_j|}{\tau}), \tau = \frac{1}{\lambda}$$

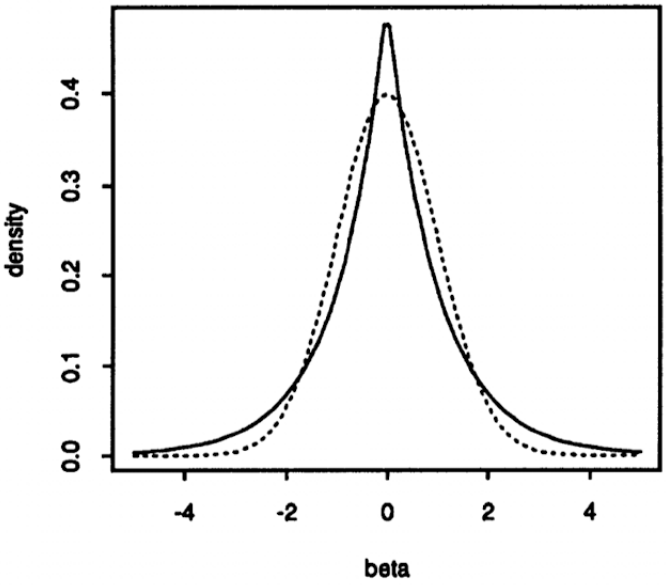


Fig. 7. line : double-exponential density(lasso), plotted line : normal density(ridge regression)

E. Algorithms for Lasso Solution

Let $t \geq 0$. Formula (1) can be expressed as a Least Square problem. Figure 7 shows the result of density of lasso method result and the ridge regression method result. Both shows the gausian-like shape. There are $m = 2^p$ different possible signs because of m linear inequality constraints about p rank vector. This is too much. So use the Kuhn-Tucker Condition. This mechanism must converge in a finite number of steps. Step by Step optimizing via computational method.

1) Start with $E = i_0$ where $\delta_{i_0} = sign(\hat{\beta}^0)$:

2) Find $\hat{\beta}$ to minimize $g(\beta)$ subject to $G_E \beta \leq t1$:

3) While $\Sigma|\hat{\beta}_j| > t$,:

4) add i to the set E where $\delta = sign(\hat{\beta})$. Find $\hat{\beta}$ to minimize $g(\beta)$ subject to $G_E \beta \leq t1$:

F. Experiments

Experiments on various condition to get the optimal solution for given problem.

III. The Lasso Method for Variable Selection in the Cox Model

A. Goal

Advanced model based on the lasso model in first paper. Use shrinkage in Cox's proportional hazards model. Same author

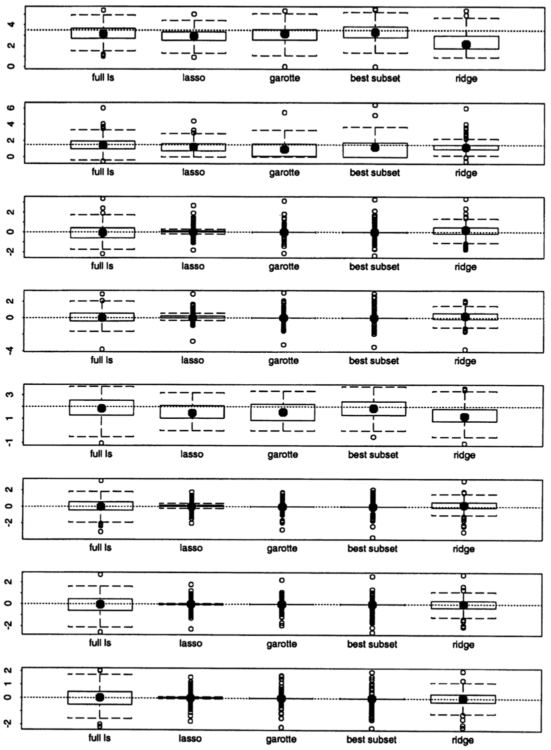


Fig. 8. Estimates for the eight coefficients in example 1, excluding the intercept: ·········, true coefficients

Fig. 8. Experiments. eight coefficients in Problem (1)

on the first paper "Regression shrinkage and selection via the lasso". Maybe in 1996, the new model lasso is borned and the next year they optimize the optimizing model. They minimize the log partial likelihood and use the absolute sum of bounded constant. This maximize the effect of using zero coefficient. This reduces the variance error while providing interpretable final model. "Interpretable final model" is valuable result. optimization method is Linear regression version based on the lasso method.

B. Introduction

This paper starts with the assumption maybe to make more limitation for their needs and reduce limitation for generality. The Proportional hazards model for "real data" is called Cox model and it assumed the condition that

$$\lambda(t|x) = \lambda_0(t)exp(\Sigma_j x_j \beta_j) \tag{16}$$

$\lambda(t|x)$ is hazard when t, $x = (x_1, x_2, ..., x_p)$, and $\lambda_0(t)$ is artificially set on baseline hazard function. $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ is the parameter in the proportional hazards model.

$$L(\beta) = \Pi \frac{exp(\beta^T x^{j_r})}{\Sigma_j exp(\beta^T x^j)} \tag{17}$$

In Equation(17), D is the set of indices of the failures. R_r is the set of indices of the individuals. Simplicity that there are no tied faiulre times, then it means suitable modification of the partial likelihood for the case of ties which really exists. Assuming the censoring not informable, So now we can use the partial likelihood on the lasso model. by using partial likelihood we use the connection between objective function and the surviving function. Those are Linear-Log relation. Cox model use this rule and make optimizing more cool. For this paper, They estimate β with the criterion like below.

$$\hat{\beta} = \arg\min \ log(L(\beta)) \text{ subject to } \Sigma|\beta_j| \leq s \tag{18}$$

In Linear regression flow, minimization of residual sum of squares with the new constraint above. The ridge regression approach shrinks their coefficients and decrease it almost closely 0 but not exactly 0. This is the difference. In actuality exact 0 coefficient rarely occur. So lasso is meaningful to make a coefficient exact zero with high probability via make the constraint region sharp for the edge that make it easier for objective function to meet constraint region on the axis which means current condition has the zero coefficient.

C. Algorithms

1) *Fix s and initialize $\hat{\beta} = 0$:*

2) *Compute η, u, A, and z based on the current value of $\hat{\beta}$:*

3) *Minimize $(z - X\beta)^T A(z - X\beta)$ subjet to $\Sigma|\beta_i| \leq s$:*

4) *Repeat steps 2 and 3 until $\hat{\beta}$ does not change:* Step 3 is doing minimization and it is done when through quadratic programming procedure meet the stop condition. This mechanism is similar to the Newton-Raphson algorithm and when use the unconstrained minimization, it is equivalent to the usual Newton-Raphson algorithm.

D. Experiments

Figure 9 shows the estimated coefficients from lasso with the constraint of the standardized parameter $u = s/\Sigma|\hat{\beta}_j^0|$. β_j^0 is unconstrained partial likelihood estimates. Vertical broken line is drawn at $u = 0.45$ which is the generalized cross validation result about u. Karnofsky score is used to get the coefficient to get the best optimized parameters easily.

Variables	Full			Stepwise			Lasso		
	Coefficient	SE	*Z*-score	Coefficient	SE	*Z*-score	Coefficient	SE	*Z*-score
1	– 0·06	0·11	– 0·58	–	–	–	0·00	0·00	0·00
2	0·30	0·12	2·49	0·33	0·11	3·08	0·17	0·09	1·89
3	– 0·12	0·10	– 1·17	–	–	–	– 0·01	0·03	– 0·31
4	0·02	0·10	0·23	–	–	–	0·04	0·07	0·63
5	0·01	0·13	0·10	–	–	–	0·00	0·00	0·00
6	0·05	0·11	0·42	–	–	–	0·02	0·05	0·40
7	0·27	0·11	2·56	0·22	0·09	2·37	0·18	0·11	1·71
8	0·37	0·12	3·14	0·39	0·09	4·39	0·35	0·12	2·97
9	0·12	0·10	1·11	–	–	–	0·00	0·01	0·28
10	– 0·30	0·12	– 2·40	– 0·29	0·11	– 2·63	– 0·22	0·10	– 2·27
11	0·22	0·10	2·13	0·25	0·09	2·90	0·21	0·11	1·98
12	0·00	0·08	0·03	–	–	–	0·00	0·00	0·00
13	0·23	0·11	2·08	0·25	0·10	2·42	0·09	0·08	1·04
14	– 0·06	0·09	– 0·75	–	–	–	0·00	0·00	0·00
15	0·08	0·11	0·76	–	–	–	0·00	0·00	0·00
16	0·23	0·11	2·19	0·23	0·10	2·25	0·09	0·09	0·97
17	0·39	0·15	2·59	0·37	0·12	2·97	0·21	0·09	2·28

TABLE III
RESULTS FOR LIVER DATA EXAMPLE

For Table 3, Lasso method modified from the original lasso with help of partial probability skills make the best performance on every metrics, coefficient, SE and Z-score for

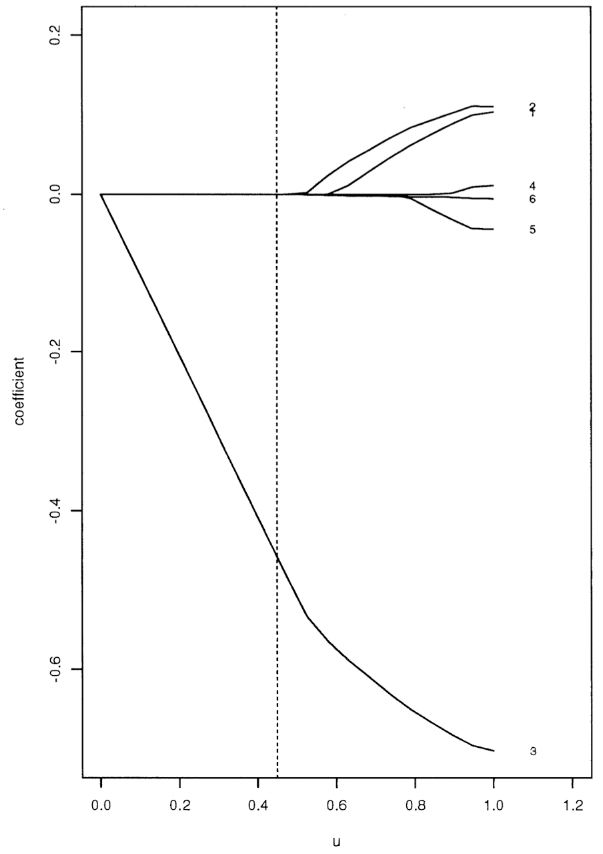


Fig. 9. Coefficient estimates for lung cancer

almost variables. Also it cover every case's variable with less constraints. It means generality is also guaranteed.

In this paper, author minimize the approximate generalized cross validation statistics and used the partial likelihood so computational flow also become more simple. Using standard matrix manimpuation, Constrained solution $\hat{\beta}$ is

$$\hat{\beta} = (X^T DX + \lambda W)^{-1} X^T Dz \tag{19}$$

Of course, W = diag(W_j), $W_i = 1/|\hat{\beta}_j|$ if $|\hat{\beta}_j| > 0$ and 0 otherwise. So we use the Equation(19), we approximate the effective parameter number with the constrained region.

$$p(s) = tr(X(X^T DX + \lambda W^-)^{-1} X^T D)$$

and it becomes the GCV-style equation

$$GCV(s) = \frac{1}{N} \frac{-logL(\beta)_s}{N(1 - p(s)/N)^2} \tag{20}$$

With the standardized constraint parameter $u = s/\Sigma|\beta^T x|$ we now can get the optimized model easier.First, find the u that make GCV minimal then, with the result value we can conduct the model.

E. Additional Results

1) Few Large Effects: General effect of this method on the statistical analysis.

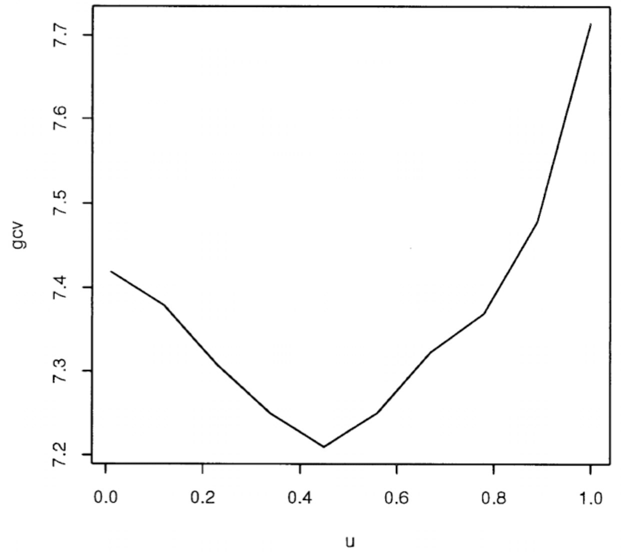


Fig. 10. GCV plot for lung cancer example. with standardized constraint parameter u

Method	Median MSE (standard errors)	Average numbers of zero coefficients
Null	0·44 (—)	9·0
Full model	0·82 (0·13)	0·0
Stepwise	0·63 (0·12)	5·6
Lasso	0·26 (0·07)	6·7

TABLE IV
BIG GENERAL EFFECT

In Table 4, median of the Mean Squared Error is shown. Lasso keeps smallest median mean squared errors and also keep a lot of coefficients with zero value. Stable for various case, Generalized on real data. and the Figure 11 plotted the result of Table 4. Lasso has the smallest standard deviation on each cases, and also median value of each cases are also similar, so very stable.

Method	Median MSE (standard errors)	Average numbers of zero coefficients
Null	0·15 (—)	9·0
Full model	0·57 (0·04)	0·0
Stepwise	0·53 (0·04)	5·5
Lasso	0·15 (0·00)	7·8

TABLE V
SMALL EFFECTS

In table 5, compared with table 4, Average number of zero coefficients increased and the median mean squared error decreased. The lasso model outperforms the full and stepwise models by shrinking with partial likelihood skills. In this experiment they show that the shrinking some constraints by multiplying zero is important for real dataset. There are many non-informative columns so filtering those conditions is one of the most important step in optimizing study. Table 6 shows

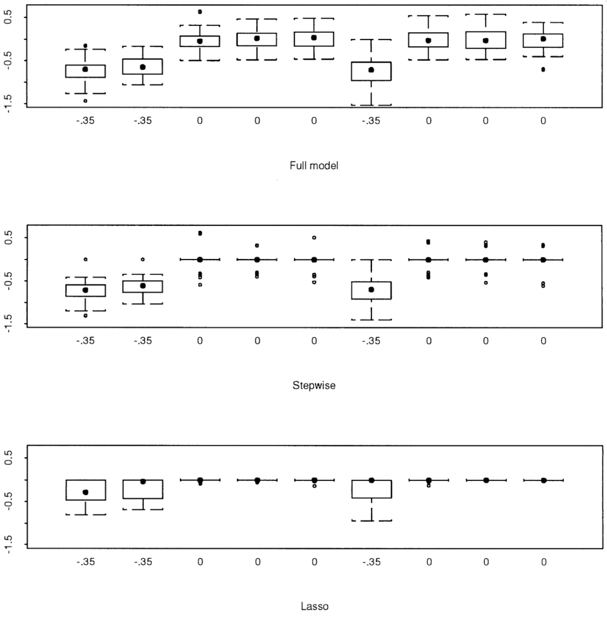
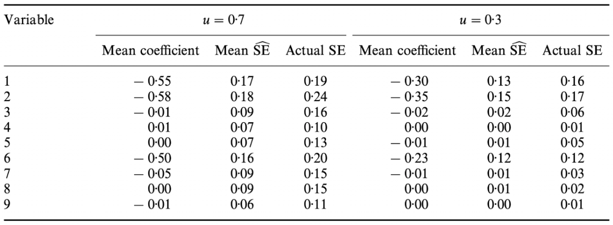


Fig. 11. Big General Effect Plotting



Variable	$u = 0{\cdot}7$			$u = 0{\cdot}3$		
	Mean coefficient	Mean $\widehat{\text{SE}}$	Actual SE	Mean coefficient	Mean $\widehat{\text{SE}}$	Actual SE
1	– 0·55	0·17	0·19	– 0·30	0·13	0·16
2	– 0·58	0·18	0·24	– 0·35	0·15	0·17
3	– 0·01	0·09	0·16	– 0·02	0·02	0·06
4	0·01	0·07	0·10	0·00	0·00	0·01
5	0·00	0·07	0·13	– 0·01	0·01	0·05
6	– 0·50	0·16	0·20	– 0·23	0·12	0·12
7	– 0·05	0·09	0·15	– 0·01	0·01	0·03
8	0·00	0·09	0·15	0·00	0·01	0·02
9	– 0·01	0·06	0·11	0·00	0·00	0·01

TABLE VI
OPTIMIZING PARAMETER U

that the optimal point of optimizing parameter u exists. for every variable, $u = 0.3$ outperforms. except a little metric on a little number of variable. This strengthen the method to find the optimal optimizing point to optimize the model.

IV. Penalized Regressions: The Bridge Versus the Lasso

A. Regression models

Logistic regression model is one of the generalized linear model. It can be used when the variable y is a Categorical variable. So this type of models are optimize the problem to calculate the posterior probability. This posterior probability is about the probability that the input x in on the certain class and is calculated through linear function about x. So many problems in real life trying to convert the regression model to minimizing problem. They make objective function with a format like minimizing problem. So with this purpose using negative log-likelihood function is considered. There are various Lasso-like model. Ridge, Lasso, etc. are those. But in this paper, they study the bridge estimation and tried to expand to the lasso so they can use the regression approach on the bridge estimation. Simply link lasso techniques on the bridge estimation and it works. Structure of the bridge estimators, algorithm for the bridge, parameter-finding method for shrinkage parameter γ and the tuning parameter λ for bridge regression, and the bridge penalty are introduced.

B. Structure

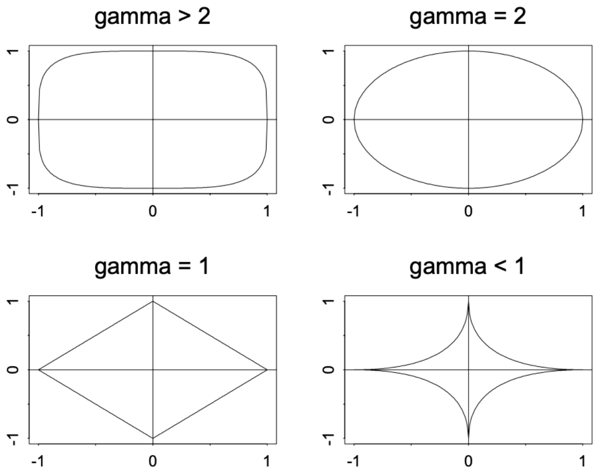


Fig. 12. constrained region with t=1

Figure 12 show the various constrained region with each γ and fixed $t = 1$. Previous Lasso experiments take $\gamma = 1$ so get the sharp edge shape.

$$\text{Given } \gamma \geq 1 \text{ and } \lambda \geq 0, \ \min_{\beta}(RSS + \lambda \Sigma |\beta_j|^{\gamma}) \tag{21}$$

$$S_1(\beta, X, y) + d(\beta_1, \lambda, \gamma) = 0$$

$$\ldots$$

$$S_p(\beta, X, y) + d(\beta_p, \lambda, \gamma) = 0 \tag{22}$$

1) Theorem 1: Given $\gamma > 1$, $\lambda > 0$. If function S is continuously differentiable with respect to β and the Jacobian $(dS/d\beta)$ is positive-semi-definite, then Equation(21) has a unique solution and the limit of the unique solution exists as $\gamma \to 1+$

2) Theorem 2: Given $\gamma > 1$, $\lambda > 0$. if functions S_j's are -2 multiples of the score functions of a joint likelihood function for Gaussian distribution, and the Jacobian $(dS/d\beta)$ is positive definite then unique solution of Equation(22) is equal to the unique estimator of the penalized regression (21) and the limit of the unique solution of Equation(22) is equal to the lasso estimator of the penalized regression (21).

V. On the Application of the Global Matched Filter to DOA Estimation with Uniform Circular Arrays

A. Modified Newton-Raphson (M-N-R) Algorithm for the Bridge $\gamma > 1$

1. Initial $\hat{\beta}_0 = \hat{\beta}_{ols}$

2. At step m, for each $j = 1, ..., p$, Let $S_0 = S_j(0, \hat{\beta}^{-j}, X, y)$ and Set $\hat{\beta}_j = 0$ if $S_0 = 0$. Apply Newton-Raphson if $gamma \geq 2$ to get unique solution $\hat{\beta}_j$. if $\gamma < 2$, modify function $-d$ using tangent line at interior point between the solution and the current. Keep approaching to the converged point and keep updating all $\hat{\beta}_j$

3. Repeat second step until $\hat{\beta}_m$ converges

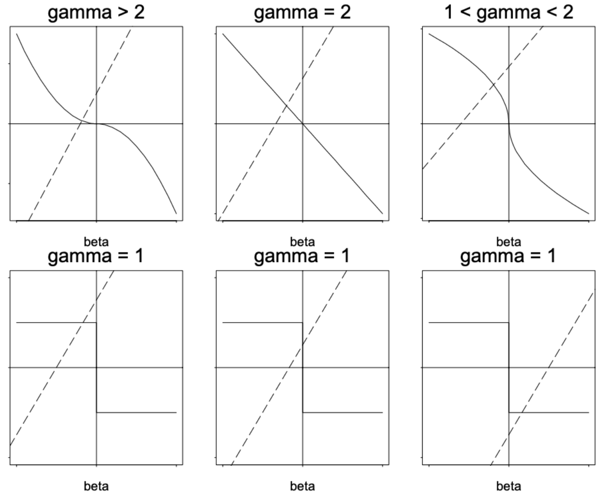


Fig. 13. Function in Equation (21)

The dotted line indicates the solution of Equation (21). While γ updating the solution also converges to the optimal solution. Solution converged on the $\gamma \to 1+$ and this looks strictly optimized.

B. Shooting Algorithm for the Lasso

1. Start with $\hat{\beta}_0 = \hat{\beta}_{ols} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T)$

2. At step m, for each $j = 1, ..., p$ let $S_0 = S_j(0, \hat{\beta}^{-j}, Xmy)$ and

$$\hat{\beta}_j = \begin{cases} \frac{\lambda - S_0}{2x_j^T x_j}, & \text{if } S_0 > \lambda \\ \frac{-\lambda - S_0}{2x_j^T x_j}, & \text{if } S_0 < -\lambda \\ 0, & \text{if } |S_0| \leq \lambda \end{cases}$$

where x_j is the j_{th} column vector of X and then get new estimator $\hat{\beta}_m$ after updating all $\hat{\beta}_j$

3. Repeat second step until $\hat{\beta}_m$ converges.

C. The Variance of the Bridge Estimator and Select parameter

Using Delta method, we can derive the variance of Equation(22). Since the Lasso make some coefficients to zero, delta method cannot be applied well. However bootstrap method can be used for calculating the variance. So we use the parameter making good variance estimator announced on the first paper, with a small difference on equation.

$$GCV = \frac{RSS}{n(1 - p(\lambda)/n)^2} \tag{23}$$

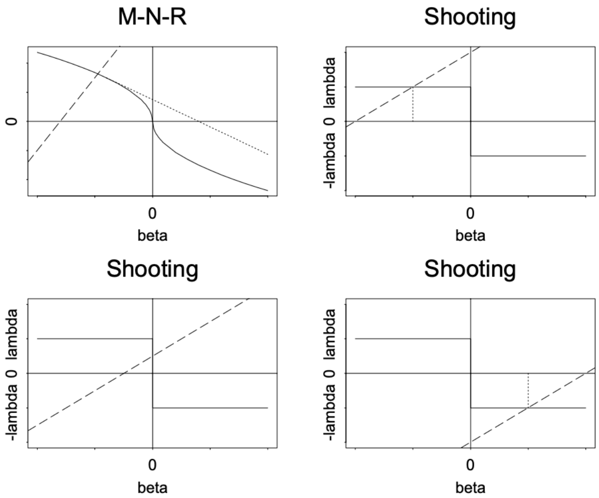


Fig. 14. Algorithm for Lasso

D. Bridge Regression of Orthonormal Matrix X

The bridge regression of orthonormal regression matrix is available. It means, bridge regression cover the special case about shrinkage effect about $gamma$. The Orthonormal matrix $X = (x_{ij}, \Sigma_i x_{ij} x_{il} = 1$ if $j = 1$, pr 0 otherwise. So lets simplify the Equation (21) to the independent parameter p.

$$2\beta_j - 2\Sigma_i x_{ij} y_i + \lambda\gamma|\beta_j|^{\gamma-1} sign(\beta_j) = 0 \tag{24}$$

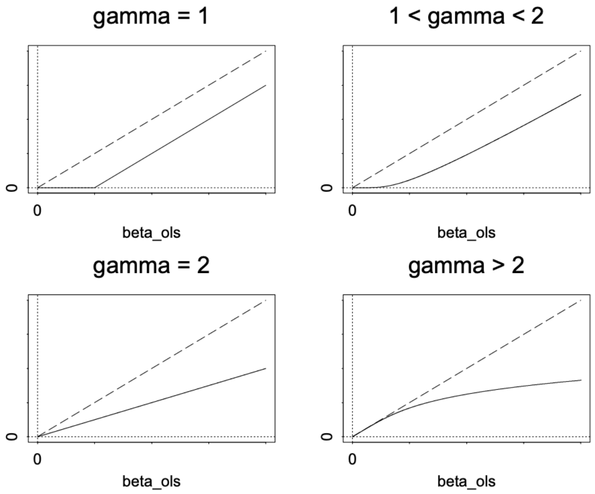


Fig. 15. Shrinkage Effect of Bridge Regression

In figure 15, solid line is the bridge estimator and the dashed line is OLS estimator. solid line becomes similar to OLS estimator when $\gamma \to 1$ and the low variables tends to be zero. Small coefficients becomes zero. Bridge regression of large value of γ tends to retain small parameters, while small value of γ tends to shrink small parameters to zero.

E. Bridge Penalty as Bayesian Prior

In Figure 16, Penalty looks like a gaussian distribution shape. with $\gamma \to 1$ the sharp point appeared which has power to make a zero coefficient well. with these penalty Lasso mechanism also can be understanded as specific Bridge regression model.

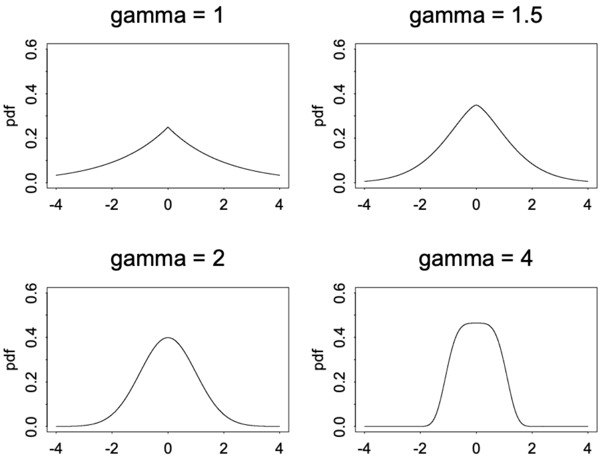


Fig. 16. Bridge penalty as a bayesian prior

VI. Asymptotics for LASSO-type Estimators

A. Introduction

This paper study the asymptotic approach for the regression estimation that minimize the residual sum of squares and the proportional to $\Sigma|\beta_j|^\gamma$ for some $\gamma > 0$.Also advanced study from the third paper. In this paper, by limiting the distribution, we can get positive probability mass at zero if the true value of the parameter is zero. This used the conceptually nearly singular matrix design. Below equation is the linear regression model.

$$Y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \epsilon = \beta_0 + x_i^T \beta + \epsilon \quad (25)$$

with mean zero and variance ρ^2. We estimate β by minimizing the penalized least squares criterion.

$$\Sigma_i (Y_i - x_i^T \theta)^2 + \lambda_n \Sigma_j |\theta_j|^\gamma \quad (26)$$

This is the Bridge estimators. And when γ is 1, it becomes lasso, special case of bridge regression.

B. Limiting distribution

Let's assume that the matrix C is non-singular.

$$C_n = \frac{1}{n} \Sigma_i x_i x_i^T \to C \quad (27)$$

1) Theorem 1: If C in equation (27) is non-singular, and $\lambda/n \to \lambda \geq 0$ then $\hat{\beta}_n \to_p \arg\min(Z)$ where

$$Z(\theta) = (\theta - \beta)^T C(\theta - \beta) + \lambda_0 \Sigma_j |\theta_j|^\gamma$$

So if $\lambda_n = o(n)$, $\arg\min(Z) = \beta$ and so $\hat{\beta}_n$ is consistent.

2) Theorem 2: If $\gamma \geq 1$, $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ and C is non-singular, then

$$\sqrt{n}(\hat{\beta}_n - \beta) \to_d \arg\min(V)$$

3) Theorem 3: If $\gamma < 1$m $\lambda_n/n^{\gamma/2} \to \lambda_0 \geq 0$ then

$$\sqrt{n}(\hat{\beta}_n - \beta) \to_d \arg\min(V) \quad (28)$$

Using these three theorems, put the value on the original equations. C get some constraints.

$$C_{11}u_1 - W_1 = -\frac{\lambda_0}{2} sgn(\beta) \quad (29)$$

$$-\frac{\lambda_0}{2} 1 \leq C_{21}u_1 - W_2 \leq \frac{\lambda_0}{2} 1 \quad (30)$$

Figures(17, 18 and 19) show scattered plots of random 500 samples limiting distributions for the Least Squares estimator with three condition repectively. Each are $(\lambda_0 = 0), (\lambda_0 = 1, \gamma = 1)$, and $(\lambda_0 = 0.5, \gamma = 0.5)$ with same distribution $\sigma^2 = 1$

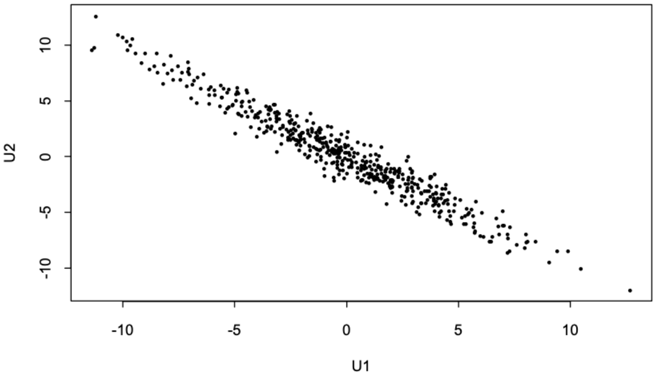


Fig. 17. $(\lambda_0 = 0)$

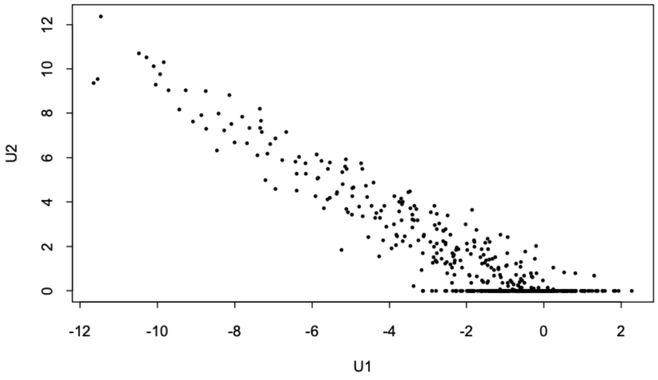


Fig. 18. $(\lambda_0 = 1, \gamma = 1)$

In figure 17, it is LS estimation, and the strong correlation between two variables is found. This means the overestimation of β_1 and is generally accompanied with underestimation of

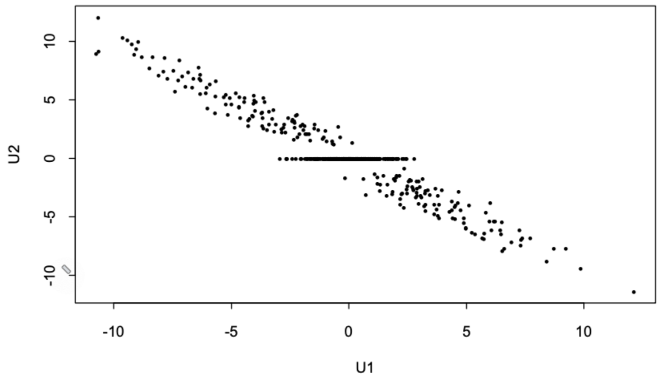


Fig. 19. $(\lambda_0 = 0.5, \gamma = 0.5))$

β_2. In figure 18, Lasso estimation also has relation between two parameters complementary. when β_1 increase, then β_2 decrease. But LASSO effectively sets the estimation of β_2 to zero if β_1 is fully overestimated. In figure 19, the shrinkage of coefficients to zero by estimation of β_2 is more active. And Big, overall, general shape does not changed. overall estimates does note changed a lot. So we can see that the asymptotic distributions approximate finite sample distribution very well. Limiting the distributions have positive mass at zero when the true parameter value is zero, and keep continuous on other range.

C. Asymptotics for nearly singular designs

Suppose that C_n is non-singular but almost singular matrix. For the consistency attribute and the limiting distribution, function need to have unique minimizer. So convert u to satisfy a conditions

$$0 \in Cu - \lambda_0 \tau(W, \beta) \tag{31}$$

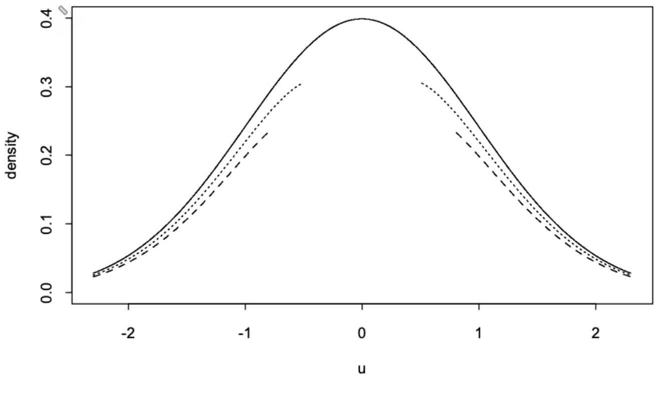


Fig. 20. Density function in figure 17, 18, 19 up to down

With the singular condition, parameters closed to zero, then the density disappeared. but in other cases, density function is continuous. In this study they find for the new logic to cover the center blank space and keep using singular structure to advance the calculating flow.

References

[1] Tibshirani R. "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society, Series B (Methodological), Volume 58, 1996, pp.267-288.

[2] Tibshirani R. "The lasso method for variable selection in the cox model," Statistics in Medicine, Volume 16, 1997, pp.385–395.

[3] Wenjiang J. Fu. "Penalized Regressions: The Bridge Versus the Lasso," Journal of Computational and Graphical Statistics, Volume 7, 1997, pp.397-416.

[4] Keith K, Wenjiang J. Fu. "Asymptotics for LASSO-type Estimators," The Annals of Statistics, Volume 28, 2000, pp.1356-1378