PalArch's Journal of Archaeology of Egypt / Egyptology

Degrees of Freedom of the Lasso in Repeated Measurements Model

Naser Oda Jassim [1], Abdul Hussein Saber Al-Mouel [2]

[1,2] Department of Mathematics, Collage of Education for Pure Sciences, University of Basrah, Basrah, Iraq

Corresponding Author: noj1972.khalef@gmail.com

Naser Oda Jassim [1], Abdul Hussein Saber Al-Mouel [2], Degrees of Freedom of the Lasso in Repeated Measurements Model. Palarch's Journal Of Archaeology Of Egypt/Egyptology 18 (7). ISSN 1567-214x

Keywords: lasso; repeated measurements model; degrees of freedom; Stein's formula; Model selection criteria; asymptotic consistency.

ABSTRACT

First of all, we introduce the repeated measurements model and discuss the degrees of freedom of its coefficients in the structure of Stein's unbiased risk estimate (SURE). By assuming that the design matrix has full columns rank, the following results are concluded. First one, the number of non-zero coefficients is an unbiased estimate for degrees of freedom of lasso solution. Second one, the unbiased estimator of these non-zero coefficients is asymptotically consistent. In addition, It is concluded that the same above results will be obtained if there are no especial assumption on the design matrix. With all above results the optimal lasso solution can be obtained by using several model selection criteria such as C_p, AIC and BIC. Moreover, BIC-lasso shrinkage will be chosen if the variable selection is the main choice in applying lasso problem.

1. Introduction

Many scientists and researchers have been given a definition for the repeated measurements in the different periods of time. Vonesh and chinchilli [19] were defined as term used to describe the data in which observations of response variable are measured repeatedly for each experimental unit under different experimental conditions. While (Keseliman) [5] explained that the repeated measurements require two or more independent groups between the most of known experimental designs in the set of different researches type. In the other words, in the repeated measurements , the observations of experimental units are measured repeatedly in the time unit.
High dimensional data means that the number of coefficients which are to be estimated is greater than the number of observations. In other words, the number of coefficients denoted by k , are larger than the sample size which denoted by N. In this case, more dimensions will be added to a data set which leads to more difficult to predict certain quantities. By high dimensional is meant that measures and the total sample size grow together but either one would be greater than the other. In this case, the traditional methods like ordinary least squares is not unique and we must use another method to treat with this problem. These methods are called penalized least squares methods. they are common and sufficient to treat with high dimensional

data to find acceptable solution.

Tibshirani (1996) [14] proposed a penalty function for the linear regression model called 'Lasso' which is abbreviated to 'least absolute shrinkage and selected operator'. The lasso method is based on the idea that minimizing the residual sum of squares plus the sum of absolute value of coefficients. It is used to estimate coefficients and selected variable simultaneously.

In this paper, we will study the degrees of freedom of the lasso estimator in the framework of Stein's unbiased risk estimation (SURE) [13] for the high dimensional repeated measurements model and discuss their properties according to the rank of design matrix is equal to k with $k \leq N$. We will derive degrees of freedom via continuous and almost differential function and apply Stein's formula. This has an advantage, that is, lasso provides an asymptotic distribution of the degrees of freedom of lasso coefficients. Furthermore, we investigate the performance of degrees of freedom of lasso with respect to model selection according to some information criterion such as: Akaike information criteria (AIC) (Akaike (1973)), Bayes information criteria (Shwartz (1978)) and Mallows C_p (Mallows (1973)) which is very similar to AIC. The degrees of freedom is defined as the trace of the first derivative of the fitted value with respect to response variable. On the other words, it can be defined as (Efron [3], Hastie and Tibshirani [18]) a sum of covariance between each point of the response variables and its corresponding fitted values and dividing the result by the variance. It is also can be defined as the trace of the 'hat matrix' of the fitted values which is a function of response variable. The concept of degrees of freedom is connection with the complexity of the model. In the sense that, it plays an important role in determination and selection the statistical model and commonly used to quantify the actual complexity of a regression method look at e.g. Hastie and Tibshirani [7]. Generally speaking, Descriptions of degrees of freedom extremely pertinent for objectives such as model selection and model comparisons, look at, e.g. Efron (1986), Hastie and Tibshirani (1990), Tibshirani and Taylor (2012) and Tibshirani (2014). The degrees of freedom had been discussed and given the basic results for linear regression by Zou, Hastie and Tibshirani (2007). They showed that if the response variable follows a normal distribution with spherical covariance , fixed design matrix and penalty parameter such that the rank of design matrix is equal to k then the degrees of freedom is equal to expectation of active set of the unique lasso solution with respect to response variable. Moreover, the degrees of freedom is characterized by estimating the prediction accuracy of the fitted model which supports us to select the optimal model among all the candidates. In the sense that, it is selected the optimal choice of λ in the lasso.

Generally speaking, the important use of regularization is to overcome the complexity of the fitted model. It is known that the main toll of the regularization is the penalty parameter which is denoted by λ. There are two cases which are very explicitly to describe the regularization. One of these is the least regularized lasso ($\lambda = 0$) corresponds to ordinary least squares. The second one is most regularized lasso uses ($\lambda = \infty$), which leads to a constant fit. Therefore the model complexity is decreased via shrinkage.

The rest of paper is organized as follows. In section two, we present statistical model for which the degrees of freedom and other properties will be studied. In section three, we will introduce preliminary material which are considered as the basic subjects to compute the degrees of freedom of lasso by using Stein's formula. Lastly, By using degrees of freedom, we construct the adaptive model selection criteria such as C_p, AIC and BIC in section five.

2. Setting Model

Suppose Y_{it} is the value of response variable for i^{th} unit at t time, $X_{j(it)}$ is the explanatory variables, μ, β_j, are fixed parameters, V_i is the random effect with $V_i \sim N(0, \sigma_v^2)$, ε_{it} is the error term with $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$, where $i = 1, \dots, N,\ j = 1, \dots, k, t = 1, \dots, T$. Then the repeated measures model is given by

$$Y_{it} = \mu + \sum_{j=1}^{k} \beta_j X_{j(it)} + V_i + \varepsilon_{it}. \quad (1)$$

Let $\omega_{it} = V_i + \varepsilon_{it}$ with $\omega_{it} \sim N(0, \sigma_\omega^2)$, $\sigma_\omega^2 = \sigma_v^2 + \sigma_\varepsilon^2$, then the model (1) can be rewritten as

$$Y_{it} = \mu + \sum_{j=1}^{k} \beta_j X_{j(it)} + \omega_{it}. \quad (2)$$

By using matrix notation, then the model (2) becomes

$$Y = G\theta + \omega, \tag{3}$$

where $G = [e, X]$, $e = [1, 1, \dots, 1]'$ has length NT, $Y = [Y_{11}, \dots, Y_{1T}, Y_{21}, \dots, Y_{2T}, \dots, Y_{N1}, \dots, Y_{NT}]'$ has length $NT, X = [X_1, \dots, X_N]'$ is a $NT \times K$ design matrix of fixed effect, $\theta = [\mu, \beta_1, \beta_2, \dots, \beta_k]'$ has length $k+1$, and $\omega = [\omega_{11}, \dots, \omega_{1T}, \omega_{21}, \dots, \omega_{2T}, \dots, \omega_{N1}, \dots, \omega_{NT}]'$ has length NT, or equivalently

$$\begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1T} \\ Y_{21} & Y_{22} & \cdots & Y_{2T} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{N1} & Y_{N2} & \cdots & Y_{NT} \end{bmatrix} = \begin{bmatrix} 1 & X_{1(11)} & X_{2(12)} & \cdots & X_{k(1T)} \\ 1 & X_{1(21)} & X_{2(22)} & \cdots & X_{k(2T)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{1(N1)} & X_{2(N2)} & \cdots & X_{k(NT)} \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1T} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2T} \\ \vdots & \vdots & \vdots & \vdots \\ \omega_{N1} & \omega_{N2} & \cdots & \omega_{NT} \end{bmatrix}$$

From model (3) we have $Y \sim N_{NT}(G\theta, \Sigma)$, where $\Sigma = \sigma_\varepsilon^2 Q + \sigma_1^2 P$, $\sigma_1^2 = \sigma_\varepsilon^2 + T\sigma_V^2$, $Q = (I_N \otimes E_T)$, $E_T = I_T - J_T$, $P = I_N \otimes J_T$ and $\Sigma^{-1} = \frac{Q}{\sigma_\varepsilon^2} + \frac{P}{\sigma_1^2}$.

We will study the degree of freedom or the effective number of coefficients in ℓ_1 penalized repeated measurements linear model. From (3) The Lasso problem can be written as

$$\hat{\theta} \in \underset{\theta \in R^{k+1}}{\operatorname{argmin}} \frac{1}{2} \|Y - G\theta\|_2^2 + \lambda \|\theta\|_1, \tag{4}$$

where $\lambda \geq 0$ is tuning parameter.

First, we will assume that Y follows a normal distribution with spherical covariance, $Y \sim N(\mu, \sigma^2 I)$, and G, λ are considered fixed with $\operatorname{rank}(G) = k$. In this case

$$\sigma_\omega (\Sigma^{1/2})^{-1} Y \sim N_{NT}\left(\sigma_\omega (\Sigma^{1/2})^{-1} G\theta, \sigma_\omega^2 I\right), \; Y^* \sim N_{NT}(\mu^*, \sigma_\omega^2 I), \tag{5}$$

where $Y^* = \sigma_\omega (\Sigma^{1/2})^{-1} Y$, $\mu^* = \sigma_\omega (\Sigma^{1/2})^{-1} G\theta$, and $\Sigma = \sigma_\varepsilon^2 Q + \sigma_1^2 P$ with $\Sigma^{-1} = \frac{Q}{\sigma_\varepsilon^2} + \frac{P}{\sigma_1^2}$.

Therefore, (3) can be rewritten as

$$Y^* = G\theta + \omega \tag{6}$$

This assumption must be used so as to apply Stein's unbiased risk estimate for degrees of freedom in repeated measurements model.

3. Preliminary material

We will introduce the following three sections which are essential and important subjects to describe and discuss the degrees of freedom of the lasso problem in repeated measurements model.

3.1 Unbiased risk estimate of Repeated measurements model and degrees of freedom

It is known that Stein's (1981) [13] Suggested for a linear regression model a new risk estimate by using a particular unbiased estimate of degrees of freedom. Moreover, Stein's framework requires two important assumption. One of these assumption is that the response variable $Y \in R^N$ must be followed an normal distribution with spherical covariance, i.e. $Y \sim N(\mu, \sigma^2 I)$. The second assumption is that the function of response variable Y must be continuous and almost differentiable.

To apply Stein's formula for our model (3), we must have the above two assumptions. Therefore, we must use the transformation (5).

i.e. $Y^* \sim N_{NT}(\mu^*, \sigma_\omega^2 I)$, where $Y^* = \sigma_\omega (\Sigma^{1/2})^{-1} Y$, $\mu^* = \sigma_\omega (\Sigma^{1/2})^{-1} G\theta$.

Given samples $Y^* \sim N_{NT}(\mu^*, \sigma_\omega^2 I)$ and Consider a function $\hat{\mu}^* : R^N \to R^N$ such that from Y^*, provides an estimate $\hat{\mu}^*(Y^*)$ of the underlying unknown mean μ^*. In this case it can be used $\hat{\mu}^*$ to refer to this estimate and function itself. An unbiased risk estimate for repeated measurements model starts by expanding

$$E\|\mu^* - \hat{\mu}^*\|^2 = E\|\mu^* - Y^* + Y^* - \hat{\mu}^*\|_2^2$$
$$= N\sigma_\omega^2 + E\|Y^* - \hat{\mu}^*\|^2 + 2(\mu^* - Y^*)'(Y^* - \hat{\mu}^*)$$

$= N\sigma_\omega^2 + E\|Y^* - \hat{\mu}^*\| - 2(Y^* - \mu^*)'(Y^* - \hat{\mu}^*)$
$= N\sigma_\omega^2 + E\|Y^* - \hat{\mu}^*\|^2 - 2N\sigma_\omega^2 + 2\sum_{i=1}^{N} cov(Y_{it}^*, \hat{\mu}_{it}^*), \quad t = 1, \dots, T$
$= -N\sigma_\omega^2 + E\|Y^* - \hat{\mu}^*\|^2 + 2\sum_{i=1}^{N} cov(Y_{it}^*, \hat{\mu}_{it}^*),$ (7)

where $E\|Y^* - \hat{\mu}^*\|^2$ is the expected training error of $\hat{\mu}^*$.

For $\hat{\mu}^*\ (Y^*) = (\hat{\mu}_{1t}^*(Y^*), \dots, \hat{\mu}_{Nt}^*(Y^*))', t = 1, \dots, t,$ recall that the degrees of freedom is defined as,

$$df(\hat{\mu}^*) = \frac{1}{\sigma^2}\sum_{i=1}^{N} cov(Y^*_{it}, \hat{\mu}_{it}^*). \quad (8)$$

This is explained as the "effective number of coefficients" used by the function $\hat{\mu}^*$.

It can be noted that for the repeated measures linear model of Y^* onto the fix and full rank design matrix M, we have that $\hat{\mu}^*\ (Y^*) = \hat{\mu}^* = MY^*$ for some matrix M independent of Y^*, then we have that $cov(\hat{\mu}^*(Y^*), Y^*) = \sigma_\omega^2 M$ then $\frac{1}{\sigma_\omega^2} cov(\hat{\mu}^*, Y^*) = M$ and this implies that $df(\hat{\mu}^*) = tr(M) = k,$ which represents the number of nonzero coefficients. By (7) we obtain

$$E\left\{\left\|\hat{\mu}^* - Y^{*(new)}\right\|^2\right\} = E\{\|Y^* - \hat{\mu}^*\|^2 + 2\sigma_\omega^2 df(\hat{\mu})\}.$$

Thus we can define a C_k-type statistics

$$C_k(\hat{\mu}^*) = \frac{\|Y^* - \hat{\mu}^*\|^2}{N} + \frac{2\sigma_\omega^2 df(\hat{\mu})}{N} \quad (9)$$

which is unbiased estimator of the true predictor error.

Furthermore, we can denote the decomposition of $\hat{\mu}^*$ by $Risk(\hat{\mu}^*) = E\|\mu^* - \hat{\mu}^*\|^2$ as

$Risk(\hat{\mu}^*) = -N\sigma_\omega^2 + E\|Y^* - \hat{\mu}^*\|_2^2 + 2\sigma_\omega^2 df(\hat{\mu}^*)$
$= -N\,\sigma_\omega^2 + E\|Y^* - \hat{\mu}^*\|^2 + 2\sigma_\omega^2\, k.$

It is noted that the decomposition proposes an estimate of degrees of freedom $\widehat{df}(\hat{\mu}^*)$ can be used to construct an estimate of the risk,

$$\widehat{Risk}(\hat{\mu}^*) = \|Y^* - \hat{\mu}^*\|_2^2 - N\sigma_\omega^2 + 2\sigma_\omega^2\, \widehat{df}(\hat{\mu}^*). \quad (10)$$

The above estimate is an unbiased for $Risk$, i.e., $E\left[\widehat{Risk}\right] = Risk$. The above estimate $\widehat{Risk}$ is called the unbiased risk estimate. Moreover, it is easy to show that an unbiased estimate of degrees of freedom leads to an unbiased estimate of risk, this means that $df(\hat{\mu}^*) = E\left[\widehat{df}(\hat{\mu}^*)\right]$ implies $Risk(\hat{\mu}^*) = E\left[\widehat{Risk}(\hat{\mu}^*)\right]$. It is seen that the risk estimate (10) can be used for penalty parameter selection λ. If we assume that the function $\hat{\mu}^*$ depends on the penalty parameter λ denoted by $\hat{\mu}_\lambda^*(Y^*)$, then it is seemed that one can minimize the estimated risk over λ to choose a suitable value for the penalty parameter,

$$\hat{\lambda} = \underset{\lambda\in\Lambda}{\operatorname{argmin}}\ \widehat{Risk}(\hat{\mu}^*) = \underset{\lambda\in\Lambda}{\operatorname{argmin}}\ \|Y^* - \hat{\mu}^*\|^2 - N\sigma_\omega^2 + 2\sigma_\omega^2\, \widehat{df}(\hat{\mu}_\lambda^*). \quad (11)$$

This can be considered as computationally efficient alternative to choosing the penalty parameter by cross-validation for penalized linear repeated measurements problem.

It can be concluded that the main results of unbiased risk estimate is considered as alternative expression for degrees of freedom in repeated measurements model if the distribution of $Y^* \sim N_{NT}(\mu^*, \sigma_\omega^2\ I)$ and the function $\hat{\mu}^*(Y^*)$ is continuous and almost differentiable.

i.e $df(\hat{\mu}^*) = E[(\nabla.\hat{\mu}^*)(Y^*)],$ (12)

where the function $(\nabla.\hat{\mu}^*)(Y^*) = \sum_{i=1}^{N} \frac{\partial \hat{\mu}_{it}^*}{\partial Y_{it}^*} = \frac{1}{\sigma_\omega^2}\sum_{i=1}^{N} \mathrm{cov}(Y_{it}^*, \hat{\mu}_{it}^*), t = 1, \dots, T,$

is called the degree of freedom of $\hat{\mu}^*$. It is followed that the unbiased estimate of degrees of freedom,

$$\widehat{df}(\hat{\mu}^*) = (\nabla.\hat{\mu}^*)(Y^*) \quad (13)$$

3.2. Some important Notations

At the beginning, it will be defined some important notations before adopting the SURE with the lasso solution for repeated measurements model. Assume that $\hat{\mu}_\lambda^*$ represents the lasso solution by using (4). Let μ_{it}^* is the i^{th} component at time t of μ^*. For convenient, suppose $df(\lambda)$ stands for $df(\hat{\mu}_\lambda^*)$. Suppose M is the matrix with k columns. Let $h \subseteq \{1, \dots, p\}$ and denote by M_h the submatrix $\mathrm{M}_h = \left[\dots \mathrm{M}_j \dots\right]_{j\in h}$ where M_j is the j^{th} column of the matrix M. Similarly define $\theta_h = \left(\dots\ \theta_j\ \dots\right)_{j\in h}$ for any vector has length p. Assume that $Sgn(.)$ refers to

sign function such that

$$Sgn(x) = \begin{cases} 1 & if\ x > 0 \\ 0 & if\ x = 0 \\ -1 & if\ x < 0\,. \end{cases}$$

Let $S = \{j: sgn(\theta)_j \neq 0\}$ represents the active set of θ, such that $Sgn(\theta)$ be the sign vector of θ which is given by $Sgn(\theta)_j = Sgn(\theta_j)$. Let the active set $\hat{\theta}(\lambda)$ is denoted by $S(\lambda)$ and the corresponding sign vector $sgn(\hat{\theta}(\lambda))$ is denoted by $Sgn(\lambda)$. It is not distinguished between the index of coefficient and coefficient itself. Assume that For given response vector Y^*, there is a finite of $\lambda's$,

$$\lambda_0 > \lambda_1 > \cdots > \lambda_p = 0, \tag{14}$$

Such that:

1. $\hat{\theta}(\lambda) = 0$ for all $\lambda > \lambda_0$.
2. For all $\lambda \in (\lambda_{m+1}, \lambda_m)$, the active set $S(\lambda)$ and the sign vector $sgn(\lambda)_{S(\lambda)}$ are constant with respect to λ. Therefore, They are written them as S_m and sgn_m for convenience.

Definition 1. (Transition points). Are points in which the active set changes at each λ_m if λ has the following properties:

1. Some explanatory variables with zero coefficients at λ_m will have nonzero coefficients when λ decreases from $\lambda = \lambda_m - 0$, thus these coefficients attach the active set $S(\lambda)$.
2. When λ increases from $\lambda = \lambda_{m+1} + 0$ there are possibly some explanatory variables in active set $S(\lambda)$ whose coefficients reach zero, hence they do not join the active set $S(\lambda)$.

In the other hand, It will be called non-transition point for any $\lambda \in [0, \infty) \backslash \{\lambda_m\}$.

Moreover, it is important to introduce the following matrix representation of Stein's Lemma of the divergence. Suppose $\frac{\partial \hat{\mu}_{it}^*}{Y_{jt}^*}$ be a $N \times N$ matrix whose elements are.

$$\left(\frac{\partial \hat{\mu}^*}{\partial Y^*}\right)_{i,j,t} = \frac{\partial \hat{\mu}_{it}^*}{\partial Y_{jt}^*},\ i,j = 1,2,\dots,N,\ t = 1,2,\dots,T.$$

Then the trace formula can be written as

$$(15)\ \nabla.\hat{\mu}^* = tr\left(\frac{\partial \hat{\mu}^*}{\partial Y^*}\right),$$

It will be introduced some important Lemmas which are discussed the necessary and sufficient conditions to study the properties of degrees of freedom of the lasso problem. Also recall that the assumption $rank(G) = k + 1$ implies that $k + 1 \leq N$; in the other words, the result of the degrees of freedom not cover the important "high-dimensional" case $k + 1 > N$.

In the following lemma, we will discuss the properties about the uniqueness of the Lasso solution.

Lemma 1. The lasso problem has the following three properties For any Y^*, G, and $\lambda \geq 0$.

(i) The solution of the lasso estimator $\hat{G}$ in (4) is either unique or an infinitely number of solutions.

(ii) The fitted value $G\hat{\theta}$ is the same for every lasso solutions $\hat{\theta}$.

(iii) If $\lambda > 0$, then we have the same ℓ_1 penalized $\|\hat{\theta}\|_1$ for every lasso solution $\hat{\theta}$.

Proof. (i) Since the lasso problem is convex then it will be attained its minimum in R^k. Therefore lasso problem has at least one solution.

Now consider the lasso problem has two solutions $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ such that $\hat{\theta}^{(1)} \neq \hat{\theta}^{(2)}$ Since $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ are two solutions for lasso problem and convex then their addition is also convex and solution.

That is, $\delta\hat{\theta}^{(1)} + (1-\delta)\hat{\theta}^{(2)}$ is also solution for any $0 < \delta < 1$, which gives infinitely number of lasso solutions as δ varies over $(0,1)$.

(ii) It will be proved by contradiction. Assuming that we have two solutions $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$

with $G\hat{\theta}^{(1)} \neq G\hat{\theta}^{(2)}$. Suppose that m^* is the minimum value of lasso solution yielded by $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$.

for any $\delta \in (0,1)$, we have

$$\left\|Y - G(\delta\hat{\theta}^{(1)} - (1-\delta)\hat{\theta}^{(2)})\right\|_2^2 + \left\|\delta\hat{\theta}^{(1)} - (1-\delta)\hat{\theta}^{(2)}\right\|_1 < \delta\, m^* + (1-\delta)m^* = m^*,$$

Where the strictly inequality due to strictly convexity of the lasso problem. This is a contradiction because

$\delta\hat{\theta}^{(1)} - (1-\delta)\hat{\theta}^{(2)}$ has a minimum value than m^*.

(iii) If we have two lasso solutions $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$, then by (ii) must have the same fitted values

Lemma 2. The lasso coefficient estimates $\hat{\theta}(\lambda)$ is given by

$$\hat{\theta}(\lambda)_{S_m} = \left(G'_{S_m}G_{S_m}\right)^{-1}(G'_{S_m}Y^* - \frac{\lambda}{2}Sgn_m), \text{ for every } \lambda \in (\lambda_{m+1}, \lambda_m). \quad (16)$$

Proof. Since λ is in the interior of $(\lambda_{m+1}, \lambda_m)$ then the active set $S(\lambda)$ and sign vector $sgn(\lambda)_{S(\lambda)}$ are constant with respect to λ, hence they can be written as S_m and Sgn_m respectively. Let $Z(\theta, G) = \left\|Y^* - \sum_{j=1}^k G_j\theta_j\right\|^2 + \lambda\sum_{j=1}^k|\theta_j|$,

where $Y^* = \left(Y_{1t}^*, \dots, Y_{Nt}^*\right)'$ and $G_j = \left(G_{1(it)}, \dots, G_{k+1(it)}\right)', \quad t = 1, \dots, T.$

It is seen that $\hat{\theta}(\lambda)$ minimizes of $Z(\theta, G)$ for given Y^* and for every $\lambda \in (\lambda_{m+1}, \lambda_m)$. Therefore

$\frac{\partial Z(\theta,Y)}{\partial\theta_j} = 0$, for every $j \in S_m$, where S_m be active set, that is

$$-2G_j'\left(Y^* - \sum_{j=1}^k G_j\hat{\theta}(\lambda)_j\right) + \lambda Sgn\left(\hat{\theta}(\lambda)_j\right) = 0, \text{ for } j \in S_m. \quad (17)$$

We have that $\hat{\theta}(\lambda)_i = 0$ for all $i \notin S_m$, therefore $\sum_{j=1}^k G_j\hat{\theta}(\lambda)_j = \sum_{j\in S_m} G_j\hat{\theta}(\lambda)_j$.

Hence (17) becomes

$$-2G'_{S_m}(Y^* - G'_{S_m}\hat{\theta}(\lambda)_{S_m} + \lambda\, Sgn_m = 0,$$

$$2G'_{S_m}G'_{S_m}\hat{\theta}(\lambda)_{S_m} = 2G'_{S_m}Y^* - \lambda Sgn_m,$$

$$G'_{S_m}G'_{S_m}\hat{\theta}(\lambda)_{S_m} = G'_{S_m}Y^* - \frac{\lambda}{2}Sgn_m,$$

$$Hence,\ \hat{\theta}(\lambda)_{S_m}\left(= G'_{S_m}G'_{S_m}\right)^{-1}\left(G'_{S_m}Y^* - \frac{\lambda}{2}Sgn_m\right).$$

Lemma 3. The Lasso estimator of (4) is continuous function of Y^* for every λ.

Proof. Consider ordinary least squares estimator,

$\hat{\theta}_\lambda(Y^*) = (G'G)^{-1}G'Y^*$, which satisfies

$$\left|\hat{\theta}_\lambda(Y)\right|_1 \leq \left|\hat{\theta}_\lambda(Y)_{ols}\right|_1, \quad (18)$$

without loss of generality, we omit the subscript λ. Consider a sequence $\left\{Y_N^*\right\}$, $N = 1, 2, \dots$, such that Y_N^* converges to a fixed point Y_0^* as N go to infinity, that is $Y_N^* \to Y_0^*$ as $N \to \infty$, then there exists y^* such that $\left\|Y_N^*\right\| \leq y^*$ for all $N = 0, 1, 2, \dots$.

This implies that there exists an upper bound U depends on G and Y^* such that,

$\left\|\theta(Y_N^*)_{ols}\right\| \leq U$.

By using (18) and Cauchy's inequality, we have

$\left|\hat{\theta}_\lambda\left(Y_N^*\right)\right|_1 \leq \sqrt{kU}$, for $N = 0, 1, 2, \dots$.

Now to show that $\hat{\theta}(Y^*)$ is continuous function of Y^*, it must be showed that

$$\hat{\theta}\left(Y_N^*\right) \to \hat{\theta}\left(Y_0^*\right) \ as\ N \to \infty. \quad (19)$$

To prove (19) it is sufficient to show that for any convergence subsequence $\left\{\hat{\theta}\left(Y_{Np}^*\right)\right\}$ of $\left\{\hat{\theta}\left(Y_N^*\right)\right\}$ converges to $\hat{\theta}(Y_0)$ as $Np \to \infty$. To show this, assume that $\hat{\theta}\left(Y_{Np}^*\right) \to \hat{\theta}_\infty(Y^*)$ as $Np \to \infty$, and then show that $\hat{\theta}_\infty(Y^*) = \hat{\theta}\left(Y_0^*\right)$

Consider the lasso criterion $Z(\theta, Y^*)$ as mentioned in (4).

Suppose $\Delta Z\left(\theta, Y^*, Y^{*'}\right) = Z(\theta, Y^*) - Z\left(\theta, Y^{*'}\right)$. (20)

It is clear that from the definition of $\hat{\theta}\left(Y_{Np}^*\right)$, it must be had,

$$Z\left(\hat{\theta}(Y_0^*), Y_{Np}^*\right) \geq Z\left(\hat{\theta}\left(Y_{Np}^*\right), Y_{Np}^*\right). \quad (21)$$

Then from (20) and (21) we get,

$$Z(\hat{\theta}(Y_0^*, Y_0^*) = Z\big(\hat{\theta}(Y_0^*), Y_{Np}^*\big) + \Delta Z\big(\hat{\theta}(Y_0^*), Y_0^*, Y_{Np}^*\big) \geq Z\big(\hat{\theta}(Y_0^*), Y_{Np}^*\big) + \Delta Z(\hat{\theta}(Y_0^*), Y_0^*, Y_{Np}^*) \quad (22)$$
$$= Z\big(\hat{\theta}\big(Y_{Np}^*\big), Y_0^*\big) + \Delta Z\big(\hat{\theta}\big(Y_{Np}^*\big), Y_{Np}^*, Y_0^*\big) + \Delta Z\big(\hat{\theta}(Y_0^*), Y_0^*, Y_{Np}^*\big).$$

Note that, $\Delta Z\big(\hat{\theta}\big(Y_{Np}^*\big), Y_{Np}^*, Y_0^*\big) + \Delta Z\big(\hat{\theta}(Y_0^*), Y_0^*, Y_{Np}^*\big) = 2\big(Y_0^* - Y_{Np}^*\big)G'\left(\hat{\theta}\big(Y_{Np}^*\big) - \hat{\theta}(Y_0^*)\right)$.

Then (22) becomes,

$$Z\big(\hat{\theta}(Y_0^*), Y_0^*\big) \geq Z\big(\hat{\theta}\big(Y_{Np}^*\big), Y_0^*\big) + 2\big(Y_0^* - Y_{Np}^*\big)G'\left(\hat{\theta}\big(Y_{Np}^*\big) - \hat{\theta}(Y_0^*)\right). \quad (23)$$

Let $Nk \to \infty$; then we get, $2\big(Y_0^* - Y_{Np}^*\big)G'\left(\hat{\theta}\big(Y_{Np}^*\big) - \hat{\theta}(Y_0^*)\right) \to 0$.

Moreover, $Z\big(\hat{\theta}\big(Y_{Np}^*\big), Y_0^*\big) \to Z\big(\hat{\theta}_\infty\big(Y_{Np}^*\big), Y_0^*\big)$.

Therefore, (23) reduces to,

$Z\big(\hat{\theta}(Y_0^*), Y_0^*\big) \geq Z\big(\hat{\theta}_\infty\big(Y_{Np}^*\big), Y_0^*\big)$.

Since $\hat{\theta}(Y_0^*)$ is the unique minimizer of $Z(\theta, Y_0^*)$, and hence $\hat{\theta}_\infty = \hat{\theta}(Y_0^*)$, which implies that, $\hat{\theta}_\infty\big(Y_{Np}^*\big) \to \hat{\theta}(Y_0^*)$ as $Np \to \infty$, and hence $\hat{\theta}(Y_N^*) \to \hat{\theta}(Y_0^*)$ as $N \to \infty$.

Therefore, $\hat{\theta}_\lambda$ is continuous function of Y^*. ■

Lemma 4. Let S_m be the active set in the interior of $(\lambda_m, \lambda_{m+1})$ and consider λ_m, λ_{m+1} with $\lambda_{m+1} \geq 0$ are the transition points. Then

$$\lambda_m = \frac{\left(\left(G'_{S_m}G_{S_m}\right)^{-1}G'_{S_m}Y^*\right)_{i^*}}{\left(\left(G'_{S_m}G_{S_m}\right)^{-1}Sgn_m\right)_{i^*}} \quad (24)$$

if i_{add} is an index added into S_m at λ_m and its index in S_m can be written as $i_{add} = (S_m)_{i^*}$. Moreover λ_{m+1} can be written as

$$\lambda_{m+1} = \frac{\left(\left(G'_{S_m}G_{S_m}\right)^{-1}G'_{S_m}Y^*\right)_{j^*}}{\left(\left(G'_{S_m}G_{S_m}\right)^{-1}Sgn_m\right)_{j^*}}, \quad (25)$$

if j_{drop} is dropped index at λ_{m+1} and its index in S_m can be written as $j_{drop} = \big(S_m\big)_{j^*}$, where $(b)_k$ means the k^{th} item of the vector b.

Proof. Assume that $M[i,.]$ Represents i^{th} row of matrix M. Since i_{add} joins S_m at λ_m; then the lasso coefficient estimate at i_{add} equal zero, i.e. $\hat{\theta}(\lambda_m)_{i_{add}} = 0$. By using lemma 2, the lasso coefficient estimate $\hat{\theta}(\lambda)$ for $\lambda \in (\lambda_{m+1}, \lambda_m)$ is given by,

$$\hat{\theta}(\lambda)_{S_m} = \big(G'_{S_m}G_{S_m}\big)^{-1}\left(G'_{S_m}Y^* - \frac{\lambda}{2}Sgn_m\right). \quad (26)$$

Since $\hat{\theta}(\lambda)_{i_{add}}$ is continuous, then by taking the limit of the i^{th} element of (26) as $\lambda \to \lambda_m - 0$, we get,

$$2\left\{\big(G'_{S_m}G_{S_m}\big)^{-1}[i^*,.]G'_{S_m}\right\}Y^* = \lambda_m\left\{\big(G'_{S_m}G_{S_m}\big)^{-1}[i^*,.]Sgn_m\right\}.$$

$\hat{\theta}(\lambda)_{i_{add}} = 0$ for all $\lambda \in (\lambda_{m+1}, \lambda_m)$, which contradict The second $\{.\}$ is a nonzero scalar, otherwise

the assumption that i_{add} joins the active set of S_m and becomes a member of it. Therefore, we have,

$$\lambda_m = \left\{2\frac{\left(G'_{S_m}G_{S_m}\right)^{-1}[i^*,.]}{\left(G'_{S_m}G_{S_m}\right)^{-1}[i^*,.]\,Sgn_m}\right\}G'_{S_m}Y^* = f(S_m, i^*)G'_{S_m}Y^*, \quad (27)$$

Where $f(S_m, i^*) = \left\{2\frac{\left(G'_{S_m}G_{S_m}\right)^{-1}[i^*,.]}{\left(G'_{S_m}G_{S_m}\right)^{-1}[i^*,.]\,Sgn_m}\right\}$. (28)

Rearranging (27) and using $i_{add} = (S_m)_{i^*}$, we get

$$\lambda_m = \frac{\left(\left(G'_{S_m}G_{S_m}\right)^{-1}G'_{S_m}Y^*\right)_{i^*}}{\left(\left(G'_{S_m}G_{S_m}\right)^{-1}Sgn_m\right)_{i^*}}.$$

By the same way above, if j_{drop} is dropped index at λ_{m+1}. Then by continuity of $\hat{\theta}(\lambda)_{j_{drop}}$ and taking the limit of the j^{th} element of (27) as $\lambda \to \lambda_{m+1} + 0$, we have that,

$$\lambda_{m+1} = \left\{2\frac{\left(G'_{S_m}G_{S_m}\right)^{-1}[j^*,.]}{\left(G'_{S_m}G_{S_m}\right)^{-1}[j^*,.]Sgn_m}\right\}G'_{S_m}Y^* = f(S_m, j^*)G'_{S_m}Y^*, \text{ where} \quad (29)$$

$$f(S_m, j^*) = \left\{ 2 \frac{(G'_{S_m} G_{S_m})^{-1}[j^*,.]}{\left(G'_{S_m} G_{S_m}\right)^{-1}[j^*,.] Sgn_m} \right\}. \quad (30)$$

Rearranging (29) and using $j_{drop} = (S_m)_{j^*}$ we get,

$$\lambda_{m+1} = \frac{\left((G'_{S_m} G_{S_m})^{-1} G'_{S_m} Y^*\right)_{j^*}}{\left(\left(G'_{S_m} G_{S_m}\right)^{-1} Sgn_m\right)_{j^*}}. \blacksquare$$

Lemma 5. For every $\lambda > 0$ There exists a null set Ω_λ which is finite collection of hyperplanes in R^N which has the form $\{X \in R^N | a'X = b\}$ where $a \neq 0$ and $b \in R$. Assume that $\Psi_\lambda = R^N / \Omega_\lambda$. Then λ is not any of the transition points for every $Y^* \in \Psi_\lambda$. In the other words $\lambda \notin \{\lambda(Y^*)_m\}$.

Proof. First, it will be prove in conversely by assuming that for some $Y^* \notin \Psi_\lambda$ there exists a null set contains that Y^* and λ is the transition point. Assume that there exists Y^* and m such that $\lambda = \lambda(Y^*)_m$, $\lambda > 0$ and this means that m is not the last lasso step. Since λ is the transition point then by Lemma 4, we have,

$$\lambda = \lambda_m = \{f(S_m, i^*) G'_{S_m}\} Y^* = g(S_m, i^*) Y^*. \quad (31)$$

It is cleared that $g(S_m, i^*) Y^* = f(S_m, i^*) G'_{S_m} Y^*$ is nonzero vector. Now assume that g_λ be the totality of $g(S_m, i^*)$ by considering all the possible combinations of S_m, i^* and the sign vector Sgn_m. g_λ depends only on the G and is a finite set, since at most k explanatory variables are available. Hence, for every $g \in g_\lambda$, $gY^* = \lambda$ defines a hyperplane in R^N. We define

$\Omega_\lambda = \{Y: gY^* = \lambda \, for \, some \, g \in g_\lambda\}$ and $\Psi_\lambda = R^N / \Omega_\lambda$.

Hence λ in (27) can not satisfied on Ψ_λ.

Lemma 6. Consider $Y^* \in \Psi_\lambda$ as defined in Lemma (5) then the active $S(\lambda)$ and the sign vector $Sgn(\lambda)$ are locally constant with respect to Y^* at any fix $\lambda > 0$.

Proof. Consider by Ball(Y^*, r) the N-demensional ball with center Y^* and radius r. Since $Y^* \in \Psi_\lambda$ then by definition of open set, it is clear that Ψ_λ is an open set. Therefore for a fix arbitrary $Y_0^* \in \Psi_\lambda$ we can select a small enough ϵ such that $Ball(Y_0^*, \epsilon) \subseteq \Psi_\lambda$. Now, fix ϵ and let $Y_N^* \to Y^*$ as $N \to \infty$. Since Ball$(Y^*, r) \subseteq \Psi_\lambda$ and $Y_N^* \to Y^*$ Then we can assume that without loss of generality $Y_N^* \in Ball(Y_0^*, \epsilon)$ for all N. This implies that $Y_N^* \in \Psi_\lambda$ and hence λ is not a transition point for any Y_N^*.

Now to show that $S(\lambda)$ is locally constant with respect to Y^* at any fix $\lambda > 0$, it must be showed that $S(Y_0^*) = S(Y_N^*)$. By definition of active set $\hat{\theta}(Y_0^*)_j \neq 0$ for all $j \in S(Y_0^*)$.

From Lemma 3, there exists an N_1 such that $\hat{\theta}(Y_N^*)_j \neq 0$ and $Sgn\left(\hat{\theta}(Y_N^*)\right) = Sgn\left(\hat{\theta}(Y_N^*)\right)$ for all $N > N_1$, and $j \in S(Y_0^*)$. This implies that $S(Y_0^*) \subseteq S(Y_N^*)$ for all $N > N_1$.

Now to show that $S(Y_N^*) \subseteq S(Y_0^*)$, we have the equicorrelation set

$$\lambda = 2\left|G_j'\left(Y_0^* - G\hat{\theta}(Y_0^*)\right)\right| \quad \forall j \in S(Y_0^*) \quad (32)$$

$$\lambda > 2\left|G_j'\left(Y_0^* - G\hat{\theta}(Y_0^*)\right)\right| \quad \forall j \notin S(Y_0^*). \quad (33)$$

Using lemma 3 again, we see that there exists $N_2 > N_1$ such that for all $\forall j \notin S(Y_0^*)$, the strict inequalities (33) hold for Y_N^* provided $N > N_2$. This implies that,

$S^c(Y_0^*) \subseteq S^c(Y_N^*)$ for all $N > N_2$.

Hence, we have $S(Y_0^*) = S(Y_N^*)$ for all $N > N_2$.

Therefore, $S(\lambda)$ is locally constant with respect to Y^*at any fixed λ. Then the locally constant of the sign vector $Sgn(\lambda)$ follows the continuity of $\hat{\theta}(Y^*)$. ■

Theorem 1. Consider the set Ψ_λ as defined in lemma 3 and let $\Psi_0 = R^N$. The lasso solution $\hat{\mu}_\lambda^*(Y^*) = G\hat{\theta}(Y^*)$ is uniformly Lipschitz on the set Ψ_λ for any an arbitrary fix $\lambda \geq 0$. Exactly

$$\|\hat{\mu}_\lambda^*(Y^* + \Delta Y^*) - \hat{\mu}_\lambda^*(Y^*)\| \leq \|\Delta Y^*\|, \text{ for strictly small } \Delta Y^*. \quad (34)$$

In addition, we have strictly formula

$$\nabla \hat{\mu}_\lambda^*(Y^*) = |S_\lambda| \quad (35)$$

Proof. It will be discussed the two cases.

Case 1. If $\lambda = 0$ then the lasso solution is just the ordinary least squares and (16) becomes

$\hat{\theta}(Y^*)_{S_m} = \left(G'_{S_m} G_{S_m}\right)^{-1} G'_{S_m} Y^*$, and

$\hat{\mu}_\lambda^*(Y^*) = G\hat{\theta}(Y^*)_{S_m} = G_{S_m}\left(G'_{S_m}G_{S_m}\right)^{-1}G'_{S_m}Y^*,$
$\hat{\mu}_\lambda^*(Y^*) = H(Y^*)Y^*,$ (36)
Where $H(Y^*) = \left(G'_{S_m}G_{S_m}\right)^{-1}G'_{S_m}$ is the projection matrix on the space G_{S_m}, i.e. $[H(Y^*)]^2 = H(Y^*)$ $\hat{\mu}_\lambda^*(Y^* + \Delta Y^*) = H(Y^* + \Delta Y^*)(Y^* + \Delta Y^*).$ (37)
Using (34) and (35) we get
$\|\hat{\mu}_\lambda^*(Y^* + \Delta Y^*) - \hat{\mu}_\lambda^*(Y^*)\|$
$= \left\|G_{S_m}\left(G'_{S_m}G_{S_m}\right)^{-1}G'_{S_m}(Y^* + \Delta Y^*) - G_{S_m}\left(G'_{S_m}G_{S_m}\right)^{-1}G'_{S_m}Y^*\right\|$
$= \left\|G_{S_m}\left(G'_{S_m}G_{S_m}\right)^{-1}G'_{S_m}\Delta Y^*\right\| = \|H(Y^*)\Delta Y^*\| \le \|\Delta Y^*\|\|H(Y^*)\|.$
This implies that
$\|\hat{\mu}_\lambda^*(Y^* + \Delta Y^*) - \hat{\mu}_\lambda^*(Y^*)\| \le \|\Delta Y^*\|$, for sufficiently small ΔY^* and$\|H(Y^*)\| = 1.$
Hence $\hat{\mu}_\lambda(Y^*)$ is uniformly Lipschitz.
Case 2. If $\lambda > 0$ then for fix an Y^*, choose a small enough ϵ such that $Ball(Y^*, \epsilon) \subseteq \Psi_\lambda$. By definition of open set, we conclude that Ψ_λ is an open set and hence λ is not any transition point. By using (16), it can be seen that
$\hat{\mu}_\lambda^*(Y^*) = G\hat{\theta}(Y^*) = G_{S_\lambda}\left(G'_{S_m}G_{S_m}\right)^{-1}(G'_{S_m}Y^* - \frac{\lambda}{2}Sgn_m)$
$= H_\lambda(Y^*)Y^* - \lambda\psi_\lambda(Y^*),$ (38)
Where $\psi_\lambda(Y^*) = \frac{1}{2}G_{S_\lambda}\left(G'_{S_m}G_{S_m}\right)^{-1}Sgn_{S_m}.$
Assume $\|\Delta Y^*\| < \epsilon$ and by using (36) we get
$\hat{\mu}_\lambda^*(Y^* + \Delta Y^*) = H_\lambda(Y^* + \Delta Y^*)(Y^* + \Delta Y^*) - \lambda\psi_\lambda(Y^* + \Delta Y^*).$ (39)
Lemma 3 include that it can be further selected ϵ be sufficiently small such that both active set S_λ and sign vector Sgn_λ stay constant with respect to Y^* in $Ball(Y^*, \epsilon)$. Now fix sufficiently small ϵ and hence if $\|\Delta Y^*\| < \epsilon$ then we get
$H_\lambda(Y^* + \Delta Y^*) = H_\lambda(Y^*)$ and $\psi_\lambda(Y^* + \Delta Y^*) = \psi_\lambda(Y^*).$ (40)
Using (36) and (37) we get
$\|\hat{\mu}_\lambda^*(Y^* + \Delta Y^*) - \hat{\mu}_\lambda^*(Y^*)\| = \|H_\lambda(Y^* + \Delta Y^*)(Y^* + \Delta Y^*) - \lambda\psi_\lambda(Y^* + \Delta Y^*) - H_\lambda(Y^*)Y^* + \lambda\psi_\lambda(Y^*)\|$
$= \|H_\lambda(Y^*)(Y^* + \Delta Y^*) - \lambda\psi_\lambda(Y^*) - H_\lambda(Y^*)Y^* + \lambda\psi_\lambda(Y^*)\|$
$\le \|H_\lambda(Y^*)\|\|\Delta Y^*\|.$
This implies that
$\|\hat{\mu}_\lambda^*(Y^* + \Delta Y^*) - \hat{\mu}_\lambda^*(Y^*)\| \le \|\Delta Y^*\|$ for sufficiently small ΔY^* and $\|H_\lambda(Y^*)\| = 1.$
Hence $\hat{\mu}_\lambda^*(Y^*)$ is uniformly Lipschitz when $\lambda > 0.$
By the locally constant of $H_\lambda(Y^*)$ and $\psi_\lambda(Y^*)$ with respect to Y^*, we have from (36)
$\frac{\partial\hat{\mu}_\lambda^*}{\partial Y^*} = H_\lambda(Y^*).$ (41)
Then by using trace formula (15), we have that
$\nabla.\hat{\mu}_\lambda^*(Y^*) = tr\left(H_\lambda(Y^*)\right) = |S_\lambda|.$ (42)
Theorem 2. The degrees of freedom of the lasso solution $\hat{\mu}_\lambda^*(Y^*) = G\hat{\theta}(Y^*)$ which is uniformly Lipschitz on Y^* Is equal to the expectation of the active set S_λ for every $\lambda \ge 0$, that is, $df(\lambda) = E|S_\lambda|$.
Proof. If $\lambda = 0$, then get the ordinary least squares. From equation (16), we have
$\hat{\mu}_\lambda^*(Y^*) = G\hat{\theta}(Y^*) = G_{S_\lambda}\left(G'_{S_\lambda}G_{S_\lambda}\right)G'_{S_\lambda}Y^* = H_\lambda(Y^*)Y^*,$ Where $H(Y^*) = \left(G'_{S_m}G_{S_m}\right)^{-1}G'_{S_m}$ is the projection matrix on the space G_{S_m} and $\hat{\mu}_\lambda^*(Y^*)$ is the lasso solution for repeated measures linear model of Y^* on G. Then relying on the matrix form of degrees of freedom by using (8)
$df(\lambda) = \frac{1}{\sigma_\omega^2}\left(cov(H_\lambda(Y^*)Y^*, Y^*)\right)$
$= \frac{1}{\sigma_\omega^2}\, tr\left(H_\lambda(Y^*)cov(Y^*, Y^*)\right)$
$= \frac{1}{\sigma_\omega^2}\, tr\left(H_\lambda(Y^*)\right).\sigma_\omega^2$
$= tr\left(H_\lambda(Y^*)\right)$

i.e. $df(\lambda) = tr\big(H_\lambda(Y^*)\big) = k+1 = E|S_\lambda|,$ if $\lambda = 0$.

Now, if $\lambda > 0$. By theorem 1 $\hat{\mu}_\lambda^*(Y^*)$ is uniformly Lipschitz on Ψ_λ. Furthermore, $\hat{\mu}_\lambda^*(Y^*)$ is a continuous function on Y^*. Hence, $\hat{\mu}_\lambda^*(Y^*)$ is uniformly Lipschitz on R^N. This implies that $\hat{\mu}_\lambda^*(Y)$ is almost differentiable everywhere. Therefore, It can be applied divergence formula (15) and using (8) we get

$$df(\lambda) = \frac{1}{\sigma_\omega^2}\big(cov\big(H_\lambda(Y^*Y^*)\big)Y^*\big)$$
$$= E\left[\frac{\partial \hat{\mu}_\lambda(Y^*)}{\partial Y^*}(Y^*)\right]$$
$$= E[\nabla.\hat{\mu}_\lambda^*(Y^*)]$$
$$= E\big[tr\big(H_\lambda(Y^*)\big)\big] = E|S_\lambda|$$

3.4. Asymptotic consistency of the unbiased estimator $\widehat{df}(\lambda)$.

In this section, it will be showed that the unbiased estimator $\widehat{df}(\lambda)$ is consistent estimator of $df(\lambda)$. It will be adopted the similar to asymptotic analysis of Knight an Fu [9] but for repeated measurements model. Assume that the following two regularity conditions which are needed to investigate the consistency of the unbiased estimator $\widehat{df}(\lambda)$:

1. $Y_{it} = G_{it}\theta^* + \omega_{it}, \quad i = 1,\dots,N, t = 1,\dots,T,$ where the random error ω_{it} are independent and identical distributed with mean 0 and variance $\sigma_v^2 + \sigma_\varepsilon^2$, and θ^* denotes the fixed unknown coefficients of the repeated measurements linear model.

2. $A_N = \frac{1}{N}\sum_{i=1}^N G_{it}G_{it}' = A,$ (43)

where A is a positive definite matrix.

3. Define the following an objective function which is minimizing at $\theta^* = \hat{\theta}^*$

$$Z_\lambda(\theta) = (\theta - \theta^*)A(\theta - \theta^*)' + \lambda\sum_{j=1}^k|\theta_j|. \quad (44)$$

Equation (44) represents an optimizing of the lasso problem which means that minimizing a quadratic objective function with ℓ_1 norm. Moreover, there are a finite sequence of transition points $\{\lambda_m\}$ associated with optimizing (44). The following theorem explains that $\widehat{df}(\lambda_N^*)$ is a consistent estimator of $df(\lambda_N^*)$ provided that $\lambda_N^* = o(N)$.

Theorem 3. The unbiased estimator of degrees of freedom $\widehat{df}(\lambda_N^*)$ is consistent, In the sense that

$df(\lambda_N^*) - df(\lambda_N^*) \to 0$ in probability if the finite sequence of tuning parameters λ_N^* satisfies the following condition:

$\frac{\lambda_N^*}{N} \to \lambda^* > 0$, where λ^* is not transition point, that is, $\lambda^* \notin \{\lambda_m\}$ for all m.

Proof. Define the random function

$$Z_{\lambda^*}^{(N)}(\theta) = \frac{1}{N}\sum_{i=1}^N(Y_{it}^* - G_{it}'\theta)^2 + \frac{\lambda_N}{N}\sum_{j=1}^k|\theta_j| \ , \quad \lambda_N > 0, \qquad t = 1,\dots,T, \quad \text{with} \quad \lambda = \lambda_N^* \quad (45)$$

and also let $\hat{\theta}^* = \underset{\theta}{\operatorname{argmin}} Z_{\lambda^*}$. Assume the effective set of (45) denote by $S^N = \big\{j: \hat{\theta}_j^{(N)} \neq 0, 1 \le j \le k\big\}$

and let the effective set of (44) denote by $S^* = \big\{j: \hat{\theta}_j^* \neq 0, 1 \le j \le k\big\}$.

Our target is to show that $P(S^N = S^*) \to 1$ as $N \to \infty$.

First assume any $j \in S^*$ and let us show that

$\hat{\theta}^N \xrightarrow{P} \hat{\theta}^*$, where $\hat{\theta}^{(N)} = \underset{\theta}{\operatorname{argmin}} Z_{\lambda^*}^{(N)}(\theta)$ and $\hat{\theta}^* = \underset{\theta}{\operatorname{argmin}} Z_{\lambda^*}$.

To show this, we translate (45) in to matrix notation

$$Z_{\lambda^*}^{(N)}(\theta) = \frac{1}{N}\big(Y^* - G\hat{\theta}^{(N)}\big)'\big(Y^* - G\hat{\theta}^{(N)}\big) + \frac{\lambda_N}{N}\sum_{j=1}^k\left|\hat{\theta}_j^{(N)}\right|$$
$$= \frac{1}{N}\big(\omega + G\theta - G\hat{\theta}^{(N)}\big)'\big(\omega + G\theta - G\hat{\theta}^{(N)}\big) + \frac{\lambda_N}{N}\sum_{j=1}^k\left|\hat{\theta}_j^{(N)}\right|$$
$$= \frac{1}{N}\Big(\omega + G\big(\theta - \hat{\theta}^{(N)}\big)\Big)'\Big(\omega + G\big(\theta - \hat{\theta}^{(N)}\big)\Big) + \frac{\lambda_N}{N}\sum_{j=1}^k\left|\hat{\theta}_j^{(N)}\right|$$

$= \frac{1}{N}\left[\omega'\omega + 2\omega G\left(\theta - \hat{\theta}\right) + \left(\theta - \hat{\theta}\right)' G'G\left(\theta - \hat{\theta}\right)\right] + \frac{\lambda_N}{N}\sum_{j=1}^{k}\left|\hat{\theta}_j\right|$

Now letting $N \to \infty$, then we have the following results

$\frac{1}{N}G'G \to A$, $\frac{1}{N}\omega'\omega \to E(\omega'\omega) = \sigma_\omega^2$, $E\left[\frac{1}{N}\sum_{i=1}^{N}\omega_{it}\right] = \frac{1}{N}\sum_{i=1}^{N}E[\omega_{it}] = 0$,

$\frac{1}{N}\omega' G\left(\theta - \hat{\theta}^{(N)}\right) \to \sum_{i=1}^{N}E[\omega_{it}]G\left(\theta - \hat{\theta}^{(N)}\right) = 0$ and $\frac{\lambda_N^*}{N} \to \lambda^* > 0$. Therefore

$Z^{(N)}(\theta) \xrightarrow{P} \sigma_\omega^2 + (\theta - \theta^N)' A\left(\theta - \theta^{(N)}\right) + \lambda^* \sum_{j=1}^{k}\left|\hat{\theta}_j^{(N)}\right|$

$= Z\left(\hat{\theta}^*\right) + \sigma_\omega^2$.

Since $Z^{(N)}(\theta)$ pointwise convergence in probability to $Z\left(\hat{\theta}^*\right)$, we conclude that

$\hat{\theta}^{(N)} = \underset{\theta}{\operatorname{argmin}} Z^N(\theta) \xrightarrow{P} \hat{\theta}^* = \underset{\theta}{\operatorname{argmin}} Z(\theta)$.

Hence $\hat{\theta}^{(N)} \xrightarrow{P} \hat{\theta}^*$.

This implies that by using continuous mapping theorem

$Sgn\left(\hat{\theta}_j^{(N)}\right) \xrightarrow{P} Sgn\left(\hat{\theta}_j^*\right)$, since $Sgn(x)$ is continuous at all x but zero.

Therefore $P\left(S^{(N)} \supseteq S^*\right) \to 1$. (46)

Second, assume any $j' \notin S^*$, and this implies that $\theta_{j'}^* = 0$.

Then we want to show that $\hat{\theta}_j^{(N)} = 0$ and $j' \notin S^{(N)}$.

Since $\hat{\theta}^* = \underset{\theta}{\operatorname{argmin}} Z(\theta_{\lambda^*})$ and λ^* is not a transition point, then by using optimality condition of Karush-Kuhn-Tucker, it will must be have

$\lambda^* > 2\left|A_{j'}\left(\theta^* - \hat{\theta}^*\right)\right|$, (47)

where $A_{j'}$ is the $j'th$ row vector of the matrix A.

Assume $c^* = \lambda^* - 2\left|A_{j'}\left(\theta^* - \hat{\theta}^*\right)\right| > 0$.

In the same way, consider $c_N^* = \lambda_N^* - 2\left|G_{j'}'\left(Y^* - G\hat{\theta}_N^*\right)\right|$. It is seen that

$G_{j'}'\left(Y^* - G\hat{\theta}_N^*\right)G_{j'}'\left(G\theta^* + \omega - G\hat{\theta}_N^*\right) = G_{j'}'\left(G\theta^* + \omega - G\hat{\theta}_N^*\right) = G_{j'}'\left(G\left(\theta^* - \hat{\theta}_N^*\right) + \omega\right)$

$= G_{j'}'G\left(\theta^* - \hat{\theta}_N^*\right) + G_{j'}'\,\omega$. (48)

Hence $\frac{c_N^*}{N} = \frac{\lambda_N^*}{N} - 2\left|\frac{1}{N}G_{j'}'G\left(\theta^* - \hat{\theta}_N^*\right) + \frac{1}{N}G_{j'}'\,\omega\right|$.

Since $\hat{\theta}^{(N)} \xrightarrow{P} \hat{\theta}^*$ and by letting $N \to \infty$, then we obtain the following results

$\frac{1}{N}G_{j'}'\,\omega \xrightarrow{P} 0$, $\frac{1}{N}G_{j'}'G \to A_{j'}$. This implies that

$\frac{c_N^*}{N} \xrightarrow{P} c^* > 0$. Then by using optimality condition of Karush-Kuhn-Tucker, we get

$c_N^* > 0$ and this implies that $\hat{\theta}_{j'}^N = 0$. Therefore

$P(S^* \supseteq S^N) \to 1$. (49)

From (46) and (49) immediately we get

$\widehat{df}(\lambda_N^*) \xrightarrow{P} |S^*|$. Then by using convergence theorem, we have

$df(\lambda_N^*) = E\left[\widehat{df}(\lambda_N^*)\right] \xrightarrow{P} |S^*|$. (50)

Therefore $\widehat{df}(\lambda_N^*) - df(\lambda_N^*) \xrightarrow{P} 0$. ■

4. Degrees of freedom and Adaptive model selection criteria.

The first step in any penalized methods is to determine the value of penalty parameter "λ". Although the penalized least squares estimator has an oracle properties but the most important thing is the choice of penalty parameter. It controls the quantity of shrinkage for the coefficients and selects the secondary variables which include in the final model. Moreover, the penalty parameter has the most important feature for choosing the optimal model which is called model selection criteria according to some criterion formula. These criterion such as Akaike information criteria (AIC) (Akaike 1973) and Bayes information criteria (BIC) (Shwartz 1978). There is another criteria is called Mallows C_p (Mallows 1973) which is very similar to AIC.

The two criterion whether AIC or Mallows C_p criteria are provided by Stein's unbiased risk

estimate theory (SURE) (Stein 1981). In Efron (2004) C_p and SURE are suggested as covariance penalty methods for estimating the prediction error. In the previous section has been derived the degrees of freedom for the linear repeated measures model of the lasso problem for the penalty parameter λ. In spite of there is no exact value of degrees of freedom $df(\lambda)$ but it has been provided an formula to compute it. It is seen that in the spirit of SURE theory, the unbiased estimate of $df(\lambda)$ is sufficient to provide an unbiased estimate for the prediction error $\hat{\mu}_\lambda^*$. Therefore prediction error $\hat{\mu}_\lambda^*$ can be denoted by $Pe(\hat{\mu})$ and is given

$$Pe(\hat{\mu}^*) = \frac{1}{N}\|Y^* - \hat{\mu}^*\|^2 + \frac{2}{N}\widehat{df}(\hat{\mu}^*), \quad (51)$$

where $\widehat{df}$ refers to $\widehat{df}(\lambda)$. Moreover, Akaike information criteria can be defined for the linear repeated measures of the lasso problem by using (51)

$$AIC(\hat{\mu}^*) = \frac{1}{N}\|Y^* - \hat{\mu}^*\|^2 + \frac{2}{N}\widehat{df}(\hat{\mu}^*)\sigma_\omega^2. \quad (52)$$

(52) is called AIC-lasso shrinkage which is corresponding to the BIC-lasso shrinkage denoted by $BIC(\hat{\mu}^*)$ and is defined as

$$BIC(\hat{\mu}^*) = \frac{1}{N\,\sigma_\omega^2}\|Y - \hat{\mu}^*\|^2 + \frac{\log(N)}{N}\widehat{df}(\hat{\mu}^*). \quad (53)$$

Both (52) and (53) are considered as formulas for selectin lasso model but they possess different asymptotic optimality. when the true function is not included in the candidate models then the model which is chosen by AIC asymptotically provides a smallest average squared error among the candidates. In this case, the AIC estimator of the function converges at the $minmax$ optimal rate whether the true function is in the candidate models or not, as explained in shao (1997), Yang (2003) [20] and their references for linear regression. On the other hand, BIC is well Known for its consistency in choosing the true model(Shao 1997). In the sense that, if the true model in the candidate list then the probability of choosing the correct model by BIC converges to one as the sample size $N \to \infty$. Moreover, BIC-lasso shrinkage is adaptive in variable selection when the true underlying model is sparse. In addition, AIC-lasso shrinkage leads to give more non-zero coefficients than truth while BIC-lasso shrinkage is more convenient when a variable selection is the important choice in applying the lasso problem. From above, we conclude that the optimal lasso model of linear repeated measurements model can be computed either by AIC or BIC. In the sense that, we encountering an optimization problem which is

$$\lambda(optimal) = \underset{\lambda}{\operatorname{argmin}} \frac{1}{N\sigma_\omega^2}\|Y - \hat{\mu}\|^2 + \frac{\xi_N}{N}\widehat{df}(\hat{\mu}^*), \quad (54)$$

where $\xi_N = 2$ for AIC and $\xi_N = \log(N)$ for BIC. Moreover, the penalty parameter λ is considered as one of the transition points which make the searching procedure is more easier.

Theorem 5. An optimal lasso problem can be attained by solving the following optimization problem for the regularization parameter λ.

$$m^* = \min_m \frac{1}{N\,\sigma_\omega^2}\left\|Y - \hat{\mu}_{\lambda_m}^*\right\|^2 + \frac{\xi_N}{N}\widehat{df}\,(\lambda_m); \quad (55)$$

then $\lambda\,(optimal) = \lambda_{m^*}$. Where $\lambda\,(optimal)$ is one of the transition points.

Proof. Assume that $\lambda \in (\lambda_{m+1}, \lambda_m)$. By (16) and using $\hat{\mu}_\lambda = G_{S_m}\hat{\theta}_\lambda(Y)$ then we have

$$\hat{\mu}_\lambda^* = G_{S_m}\hat{\theta}_\lambda(Y) = G_{S_m}\left(G_{S_m}'G_{S_m}\right)^{-1}\left(G_{S_m}'Y - \frac{\lambda}{2}Sgn_m\right)$$
$$= \left(G_{S_m}\left(G_{S_m}'G_{S_m}\right)^{-1}G_{S_m}'\right)Y - \frac{\lambda}{2}G_{S_m}\left(G_{S_m}'G_{S_m}\right)^{-1}Sgn_m$$
$$= H_{S_m}(Y)Y - \frac{\lambda}{2}G_{S_m}\left(G_{S_m}'G_{S_m}\right)^{-1}Sgn_m,$$

where $H_{S_m} = \left(G_{S_m}\left(G_{S_m}'G_{S_m}\right)^{-1}G_{S_m}'\right)$ is the projection matrix on the space G_{S_m}, i.e. $\left[H_{S_m}(Y)\right]^2 = H_{S_m}(Y)$.

$$Y - \hat{\mu}_\lambda^* = \left(I - H_{S_m}\right)Y + \frac{\lambda}{2}G_{S_m}\left(G_{S_m}'G_{S_m}\right)^{-1}Sgn_m \quad (56)$$
$$\left\|Y - \hat{\mu}_\lambda^*\right\|^2 = Y'\left(I - H_{S_m}\right)Y + \frac{\lambda^2}{4}Sgn_m'\left(G_{S_m}'G_{S_m}\right)^{-1}Sgn_m.$$

Since the lasso estimates are continuous function on λ. Then we can conclude $\left\|Y - \hat{\mu}_\lambda^*\right\|^2$ is strictly increasing in the interval $(\lambda_{m+1}, \lambda_m)$. In the sense that for each $\lambda \in (\lambda_{m+1}, \lambda_m)$ we have

that

$$\left\|Y-\hat{\mu}^{*}_{\lambda_{m+1}}\right\|^{2} < \|Y-\hat{\mu}^{*}_{\lambda}\|^{2}\left\|Y-\hat{\mu}^{*}_{\lambda_{m}}\right\|^{2}. \quad (57)$$

On the other hand, it is noted that by Theorem 2, we have $df(\lambda) = E|S_m|$ which implies $\widehat{df}(\lambda) = |S_m|$ for every $\lambda \in (\lambda_{m+1}, \lambda_m)$ and $|S_m| \geq \left|S(\lambda_{m+1})\right|$, that is, S_m is increasing for every $\lambda \in (\lambda_{m+1}, \lambda_m)$.

Therefore the optimal choice of λ in $(\lambda_{m+1}, \lambda_m)$ is λ_{m+1}, $\lambda\,(optimal) \in \{\lambda_m\},\ m1, 2, 3, \ldots p.$

References

[1] Bühlmann P. and De Geer S. V. (2011). "Statistics for High Dimensional data". Springer Heidelberg Dordrecht London New York.

[2] Charles S. D. (2002). "Statistical Methods for the Analysis of Repeated measurements". Springer-Verlag New York, Inc.

[3] Efron, B. (1986). 'How Biased is the Apparent Error Rate of a Prediction Rule'. Journal of the American Statistical Association: Throry and Methods 81 (394),461-470.

[4] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). 'Least Angle Regression'. Annals of statistics 32, 407-499.

[5] H.J.,Keseliman, and et al. (1998). ' Statistical practices of Educational Researchers: An analysis of their, ANOVA MANOVA, and ANCOVA analysis",Rev. Educational Research, 68,350-386.

[6] Hansen, N. and Sokol, A. (2014). 'Degrees of Freedom for Nonlinear Least Squares Estimation' arXiv: 1402.2997.

[7] Hastie, T. and Tibshirani, R. (1990). 'Generalized Additive Models' . Chapman and Hall, London.

[8] Janson, L., Fithhian, W. and Hastie, T. (2013). 'Effective Degrees of Freedom: Aflawed Mataphor'. arXiv: 1312.7851.

[9] Knight K. (1998). "Limiting Distributions For L_1 Regression Estimators Under General Conditions". The Annals Of Statistics, Vol. 26, No. 2, 755-770.

[10] Knight K. and Fu W. (2000). " Asymptotic For Lasso-Type Estimator. The Annals Of Statistics, Vol. 28, No. 5, 1356-1378.

[11] Kramer, N. and Sugiyama, M. (2011). 'Degrees of Freedom of Partial Least Squares Regression'. arXiv 1002.4112.

[12] Mallows, C. (1973). 'Some Comments on C_p'. Technometrics 15(4), 661-675.

[13] Stein's, C. (1981). 'Estimation of the Mean of a Multivariate Normal Distribution'. Annals of Statistics 9(6), 1135-1151.

[14] Tibshirani, R. (1996). 'Regression Shrinkage and Selection Via the Lasso' Journal of the Royal Statistical Society: Series B 58(1), 267-288.

[15] Tibshirani, R. j. (2013). 'The Lasso Problem and Unigueness'. Electronic Journal of Statistics 7, 1456-1490.

[16] Tibshirani, R. J. (2014). 'Degrees of Freedom and Model Search' arXiv: 1402.920.

[17] Tibshirani, R. J. and Taylor, J. (2012). 'Degrees of Freedom in Lasso problem'. Annals of Statistics 40(2), 1335-1371.

[18] Tibshirani, R. and Knight, K (1999). 'The Covariance Inflation Criterion for Adaptive Model Selection'. Journal of the Royal Statistical Society: Series B 61(3), 529-546.

[19] Vonesh E. F. and Chinchilli M. (1997). "Linear and Nonlinear Models for the Analysis of Repeated Measurements". Mancel Dakker, Inc., New York.

[20] Yang,Y. (2003). 'Can the Strength of AIC and BIC be Shared? (summited. http://www.stat.iastate.edu./preprint/articles/2003-10.pdf).

[21] Yogesh Hole et al 2019 J. Phys.: Conf. Ser. 1362 012121

[22] Zou, H., Hastie, T. and Tibshirani, R. (2007). ' On the Degrees of Freedom of the Lasso'. Annals of Statistics 35(5), 2173-2192.