British Journal of Mathematics & Computer Science
19(4): 1-7, 2016; Article no.BJMCS.29533
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

SCIENCEDOMAIN

Prediction Consistency of Lasso Regression Does Not Need Normal Errors

Kateřina Hlaváčková-Schindler[1*]

[1] *Data Mining Group, Faculty of Computer Science, University of Vienna, Vienna, Austria.*

Author's contribution

The sole author designed, analyzed and interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/BJMCS/2016/29533
Editor(s):
(1) Morteza Seddighin, Indiana University East Richmond, USA.
Reviewers:
(1) P. Wijekoon, University of Peradeniya, Sri Lanka.
(2) Jagdish Prakash, University of Botswana, Botswana.
Complete Peer review History: http://www.sciencedomain.org/review-history/16710

Received: 15th September 2016
Accepted: 9th October 2016
Published: 28th October 2016

Short Communication

Abstract

Sourav Chatterjee in 2014 proved consistency of any estimator using orthogonal least squares (OLS) together with Lasso penalty under the conditions the observations are upper bounded, with normal errors, and being independent of observations, with a zero mean and a finite variance. Reviewing his elegant proof, we come to the conclusion that the prediction consistency of OLS with Lasso can be proven even with fewer assumptions, i.e., without assuming normality of the errors, knowing only they have a finite variance and zero mean. We give an upper bound on the convergence rate of OLS-Lasso estimator for these errors. This upper bound is not asymptotic and depends both on the number of regressors and on the size of the data set. Knowing the number of regressors in a regression problem, one can estimate how large data set is needed, to achieve a prediction error under a given value, and this in comparison to the cited work, without solving the parameter estimation problem for fitting the errors to a normal distribution. The result can encourage practitioners to use OLS Lasso as a convergent algorithm for prediction with other than normal errors satisfying these milder conditions.

Keywords: Lasso regression; orthogonal least squares; prediction approximation error; normal errors.

2010 Mathematics Subject Classification: 62J07, 60G25, 60E15.

[*] *Corresponding author: E-mail: katerina.schindler@gmail.com;*

1 Introduction

The estimation by linear ordinary least squares (OLS) has an advantage over the maximum likelihood estimation that it is a convex optimization problem. According to the Gauss-Markov theorem, in a linear regression model in which the errors are uncorrelated, have zero mean and equal variances, the best linear unbiased estimator (BLUE) of the coefficients is given by the OLS estimator. "Best" means giving the lowest variance of the estimate, as compared to other unbiased, linear estimators. The errors do not need to be normal, nor do they need to be independent and identically distributed. In other words, one can use OLS with uncorrelated errors which are not normal, only homoscedastic with a finite variance and zero mean and still can get the BLUE estimator.

OLS combined with Lasso is also a very popular regression tool. Several researchers, for example [1], [2] or [3] have recently shown, that measuring performance of the Lasso by prediction loss, only a few assumptions are required for consistency, including the normality of the errors.

[4] proved that under the conditions the observations are upper bounded, with normal errors having zero mean and being independent of observations and with constant variances is any estimator using OLS together with Lasso prediction consistent.

Similarly as for OLS, we question whether for OLS with Lasso is the assumption of normal errors necessary. In this paper we show that the normality of errors is for the convergence of prediction error not necessary. We give an upper bound on estimated prediction error of OLS-Lasso where the assumption of normality of the errors is not required.

Rates of convergence and strong consistency of the Lasso estimator with OLS and generally non-normal errors was proven in a different way in [5]. Their approach does not consider prediction consistency and their convergence rate is asymptotic, depending on the size of data set n. The approach of S. Chatterjee, on which we base and we generalize, has an advantage for practitioners that it gives an upper bound of OLS-Lasso estimator which is not asymptotic and depends both on the number of regressors p and on the size of the data set n. Knowing the number of regressors p in a regression problem one can therefore estimate how large data set should be selected, to achieve a prediction approximation error under a given value. This we extend to the case of generally non-normal errors and in comparison to [4], our upper estimate does not require to solve the parameter estimation problem fitting the errors to a normal distribution.

2 Estimated Mean Squared Prediction Error

As in [4], we adopted also his notation and definitions in this paper. Let $\mathcal{R}$ denote the set of real values. Assume that $X_1, \ldots, X_p$ are random values and M is a constant so that

$$|X_j| \leq M \text{ (almost surely) for all } j = 1, \ldots, p. \tag{2.1}$$

We rewrite the problem of OLS with Lasso by means of the original formulation of Tibshirani in his first paper on Lasso variable selection in [6] which is equivalent to the commonly used OLS with Lasso with a penalization parameter in [4].

If the parameter vector γ^* was known, the best predictor of Y based on $X_j, j = 1, \ldots, p$ would be the linear combination

$$Y = \sum_{j=1}^{p} \gamma_j^* X_j + \varepsilon. \tag{2.2}$$

where

$$\varepsilon \text{ is independent of } X_j \text{ for all } j = 1, \ldots, p. \tag{2.3}$$

Let $\mathbf{Z}$ denote the random vector $(Y, X_1, \dots, X_p)$. Let $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ be i.i.d. copies of $\mathbf{Z}$. We write $\mathbf{Z}_i = (Y_i, X_{i,1}, \dots, X_{i,p})$. The set of vectors $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ is our data set.

Denote ε_i the error of approximation by data set $(Y_i, X_{i,1}, \dots, X_{i,p})$, i.e.

$$\varepsilon_i := Y_i - \sum_{j=1}^{p} \gamma_j^* X_{i,j}. \tag{2.4}$$

Since γ^* is unknown, we estimate $\hat{Y}$ from the data. The conditions (2.1), (2.2), the independence of $\mathbf{Z}_1, \dots, \mathbf{Z}_n$, the normality of errors in (2.4) and the sparsity condition that $\sum_{j=1}^{n} |\gamma_j^*|$ is not too large, are the assumptions in the proof of [4].

Suppose that in vector $\mathbf{Z}$ the values of Y are unknown and we want to predict Y by means of $X_1, \dots, X_p$. If the parameter vector of the best estimator $\gamma^* = (\gamma_1^*, \dots, \gamma_p^*)$ was known, the best predictor of $\mathbf{Y}$ based on $X_1, \dots, X_p$ would be the linear combination

$$\hat{Y} := \sum_{j=1}^{p} \gamma_j^* X_j. \tag{2.5}$$

From data set $\mathbf{Z_1}, \dots, \mathbf{Z_p}$ we estimate ${\gamma_j}^*$. We define *mean squared prediction error* of any estimator $\tilde{\gamma}$ as expected squared error in estimating $\hat{Y}$ using $\tilde{\gamma}$

$$MSPE(\tilde{\gamma}) := E(\hat{Y} - \tilde{Y})^2, \tag{2.6}$$

where $\tilde{Y} := \sum_{j=1}^{p} \tilde{\gamma}_j X_j$. Values $\tilde{\gamma}_1, \dots, \tilde{\gamma}_p$ are computed from Z_j, $j = 1, \dots, p$, and are therefore independent of $X_1, \dots, X_p$. It can be seen in [4] that

$$MSPE(\tilde{\gamma}) := E\|\gamma^* - \tilde{\gamma}\|_\Sigma^2, \tag{2.7}$$

where Σ is a covariance matrix of $(X_1, \dots, X_p)$ and $\|.\|_\Sigma$ the norm (or seminorm) on $\mathcal{R}^p$ induced by Σ, i.e. $\|\mathbf{x}\|_\Sigma^2 = \mathbf{x}\Sigma\mathbf{x}$.

Alternatively one can define *estimated mean squared prediction error* as

$$\widehat{MSPE}(\tilde{\gamma}) := \frac{1}{n} \sum_{i=1}^{n} (\hat{Y} - \tilde{Y})^2 \tag{2.8}$$

where $\hat{Y}_i = \sum_{j=1}^{p} \gamma_j^* X_{ij}$ and $\tilde{Y}_i = \sum_{j=1}^{p} \tilde{\gamma}_j X_{ij}$. Analogously as in the previous definition it can be expressed as $\widehat{MSPE}(\tilde{\gamma}) = \|\gamma^* - \tilde{\gamma}\|_{\hat{\Sigma}}^2$ where $\|\mathbf{x}\|_{\hat{\Sigma}}^2 = \mathbf{x}\hat{\Sigma}\mathbf{x}$ and $\hat{\Sigma}$ is the sample covariance matrix of the covariates, i.e. the matrix whose (j,k)-th element is $\frac{1}{n} \sum_{i=1}^{n} X_{ij} X_{ik}$ [4].

3 Lemmas

Comparing to lemmas used in [4], our lemmas need weaker assumptions.

Lemma 3.1 (Hoeffding's Lemma)**.** *[7] Let X be any real-valued zero-mean random variable such that $a \le X \le b$ almost surely. Then for all $t \in \mathcal{R}$ holds*

$$E(e^{(tx)}) \le e^{(\frac{t^2(b-a)^2}{8})}.$$

Proof in [7].

Lemma 3.2. *(i) Let X be a bounded zero-mean random variable taking values in $[-\sigma, \sigma]$, $\sigma > 0$ and let $M_X(t) := E(e^{tX}), t \in \mathcal{R}$ be its moment generating fuction. Then for all $t \in \mathcal{R}$ holds $M_X(t) \leq e^{\frac{\sigma^2 t^2}{2}}$.*

(ii) Let X be a bounded random variable taking values in $[a, b]$ with $a \leq 0 \leq b$ and mean μ. Then for all $t \in \mathcal{R}$ holds $M_X(t) \leq e^{\frac{t^2(b-a)^2}{8}} = e^{\mu t + \frac{t^2 \sigma_{max}^2}{2}}$ where $\sigma_{max} = \frac{b-a}{2}$.

(iii) For zero mean for all $t \in \mathcal{R}$ holds $M_X(t) \leq e^{\frac{t^2(b-a)^2}{8}} = e^{\frac{t^2 \sigma_{max}^2}{2}}$.

Proof: (i) A direct consequence of the Hoeffding's lemma. (ii) Random variable X with range $[a + \mu, b + \mu]$ can be transformed into a random variable $Y = X - \mu$ which has zero mean and range $[a, b]$. $M_X(t) = e^{t\mu} M_Y(t) \leq e^{t\mu} e^{\frac{t^2 \sigma_{max}^2}{2}} = e^{\mu t + \frac{t^2 \sigma_{max}^2}{2}}$. (iii) is an obvious consequence of (ii) for zero mean.

Lemma 3.3. *Let $\xi_i, i = 1, \ldots, m$ be random variables generated from interval $[a_i, b_i]$ with $a_i < 0 < b_i$ with zero means. ξ_i need not be independent. Let $L := (\frac{b_i - a_i}{2})$. Then*

$$E(\max_i |\xi_i|) \leq L\sqrt{2logm}.$$

Proof: We follow the proof idea from [4] and at the same time apply Lemma 2 (iii). For any $t \in \mathcal{R}$ holds for the moment generating function

$$M_{\xi_i}(t) = E(e^{t\xi_i}) = e^{t^2(\frac{b_i - a_i}{2})^2/2} \leq e^{t^2 L^2/2}.$$

Then for any $t > 0$ holds

$$E(\max_i |\xi_i|) = \frac{1}{t} E(\log e^{\max_i t|\xi_i|}) \leq \frac{1}{t} E(\log \sum_{i=1}^{m} (e^{t\xi_i} + e^{-t\xi_i})).$$

The last inequality follows from the Jensen's inequality, since $M_{\xi_i}(t)$ is convex.

$$\frac{1}{t} \log \sum_{i=1}^{m} E(e^{t\xi_i} + e^{-t\xi_i}) \leq$$

$$\leq \frac{1}{t} \log \sum_{i=1}^{m} (E(e^{t\xi_i}) + E(e^{-t\xi_i})).$$

$$\leq \frac{1}{t} \log(2me^{t^2 L^2/2}).$$

The last term is equal to

$$\frac{\log 2m}{t} + \frac{t^2 L^2}{2t} = \frac{\log(2m}{t} + \frac{tL^2}{2}.$$

For $t = \frac{\sqrt{2\log(2m)}}{L}$ we get

$$E(\max_i |\xi_i|) \leq L\sqrt{2logm}.$$

4 Prediction Consistency of OLS-Lasso

Theorem 4.1. *Assume the estimator $\tilde{\gamma}^K := \{\tilde{\gamma}_j^K\}, j = 1, \ldots, p$ is the minimizer of $\sum_{i=1}^{n} (Y_i - \gamma_1 X_{i,1} - \cdots - \gamma_p X_{i,p})^2$ subject to constraint $\sum_{j=1}^{p} |\gamma_j| \leq K$ and satisfies conditions (2.1) - (2.3).*

Assume the errors ε_i in (2.4) are i.i.d. with zero mean and variance σ_i^2. Denote $\sigma_{max} = \max_i |\sigma_i|$. Then for the estimated mean squared prediction error of the OLS-Lasso estimator holds

$$E(\widehat{MSPE}(\tilde{\gamma}^K)) \leq \frac{2\sigma_{max} KM\sqrt{2\log(2p)}}{\sqrt{n}}. \tag{4.1}$$

Proof of Theorem 4.1:
We follow the proof of Chatterjee up to the point of definition of U_j where we proceed in our way.

Denote

$$\mathbf{Y} := (Y_1, \dots, Y_n), \tag{4.2}$$

and

$$\tilde{\mathbf{Y}}^K := (\tilde{Y}_1^K, \dots, \tilde{Y}_n^K), \tag{4.3}$$

where

$$\tilde{Y}_i^K := \sum_{j=1}^{p} \tilde{\gamma}_j^K X_{i,j}. \tag{4.4}$$

Similarly let

$$\tilde{Y}^K := \sum_{j=1}^{p} \tilde{\gamma}_j^K X_j. \tag{4.5}$$

Recall condition (2.1) $|X_{i,j}| \leq M$ for all $j = 1, \dots, p$. For each $j = 1, \dots, p$ let

$$\mathbf{X}_j := (X_{1,j}, \dots, X_{n,j}). \tag{4.6}$$

Denote

$$\hat{\mathbf{Y}} := (\hat{Y}_1, \dots, \hat{Y}_n)$$

where

$$\hat{Y}_i := \sum_{j=1}^{p} \gamma_j^* X_{i,j}. \tag{4.7}$$

Given Z_i, define set $C := \{\gamma_1 \mathbf{X}_1 + \dots + \gamma_p \mathbf{X}_p : |\gamma_1| + \dots + |\gamma_p| \leq K\}$. C is a compact convex subset of $\mathcal{R}^n$. By definition $\tilde{\mathbf{Y}}^K$ is a projection of $\mathbf{Y}$ on the set C. Since C is convex, it follows that for all $x \in C$ the vector $(x - \tilde{\mathbf{Y}}^K)$ must be at the obtuse angle to the vector $\mathbf{Y} - \tilde{\mathbf{Y}}^K$. In other words,

$$(x - \tilde{\mathbf{Y}}^K).(\mathbf{Y} - \tilde{\mathbf{Y}}^K) \leq 0.$$

The condition $\sum_{j=1}^{p} |\gamma_j| \leq K$ ensures that $\hat{\mathbf{Y}} \in C$. Thus

$$(\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}^K).(\mathbf{Y} - \tilde{\mathbf{Y}}^K) \leq 0.$$

This can be rewritten as

$$\|\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}^K\|^2 \leq (\mathbf{Y} - \hat{\mathbf{Y}}).(\tilde{\mathbf{Y}}^K - \hat{\mathbf{Y}}) =$$

$$= \sum_{T=1}^{n} \varepsilon_i (\sum_{j=1}^{p} (\tilde{\gamma}_{ij}^K - \gamma_{ij}^*) X_{i,j}) = \sum_{T=1}^{n} (\tilde{\gamma}_{ij}^K - \gamma_{ij}^*)(\sum_{i=1}^{n} \varepsilon_i X_{i,j}) \tag{4.8}$$

where $\varepsilon_i := \hat{Y}_i - \tilde{Y}_i$. Denote $U_j := \sum_{i=1}^{n} \varepsilon_i X_{i,j}$.

The condition $\sum_{j=1}^{p} |\gamma_j| \leq K$ gives

$$\|\hat{Y} - \tilde{Y}^K\|^2 \leq 2K \max_{j=1,\dots,p} |U_j|. \tag{4.9}$$

About ε_i with $i = 1, \dots, n$ it is assumed that $E(\varepsilon_i) = 0$ and variance $\sigma_i^2 > 0$. Let $\mathcal{F}$ be the σ-algebra generated by $A := \{X_{i,j}, j = 1, \dots, p, i = 1, \dots, n\}$ (i.e. $\mathcal{F}$ is a collection of subsets A, including the empty subset, which is closed under complement, and is closed under union or intersection of countably infinite many subsets).

Conditionally on $\mathcal{F}$, U_j is an independent identically distributed random variable having a distribution with zero mean and standard deviation $\sqrt{\sigma_{max}^2 \sum_{i=1}^n X_{i,j}^2}$.

Let $E^{\mathcal{F}}$ denote the conditional expectation given $\mathcal{F}$.

Then from Lemma 3.3 we get

$$E^{\mathcal{F}}(\max_{j=1,\dots,p}(|U_j|) \leq M\sigma_{max}\sqrt{2n\log(2p)}.$$

Using condition $\sum_{j=1}^p |\gamma_j| \leq K$ we get

$$\|\hat{Y} - \tilde{Y}^K\|^2 \leq 2KM\sigma_{max}\sqrt{2n\log(2p)}$$

and

$$E(\widehat{MSPE}(\tilde{\gamma}^K)) \leq \frac{2\sigma_{max}KM\sqrt{2\log(2p)}}{\sqrt{n}}.$$

1. Based on the central limit theorem, we are aware that every variable that can be modelled as a sum of many small independent, identically distributed variables with a finite mean and variance is approximately normal. So U_j in Theorem 4.1 is approximately normal but this does not have to be assumed a priori about errors ε_i.
2. Notice that ε_i do not have to have the same distribution for all $i = 1, \dots, n$.

Examples of non-normal distributions of errors satisfying the assumptions of Theorem 4.1 are general elliptical distributions with a finite variance, as a special case is the Student's t-distribution. Generally, errors of any distribution from the location-scale family, for reference see for example [8], having the location parameter zero and a non-negative scale parameter, satisfy Theorem 4.1.

Example 4.2. *Consider errors ε_i in (2.4) having the Student's t-distribution, $\varepsilon_i \sim t(0, \nu)$ with $\nu > 2$, where $t(\varepsilon_i) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1 + \frac{\varepsilon_i^2}{\nu})^{-\frac{\nu+1}{2}}$. The mean of this distribution is zero and variance is $\frac{\nu}{\nu-2}$. Then for the estimated mean squared prediction error of estimator $\tilde{\gamma}_t^K$ from Theorem 4.1 with the defined errors ε_i holds*

$$E(\widehat{MSPE}(\tilde{\gamma}_t^K)) \leq \frac{2\sqrt{\frac{\nu}{\nu-2}}KM\sqrt{2\log(2p)}}{\sqrt{n}}. \tag{4.10}$$

Example 4.3. *Consider errors ε_i in (2.4) having the Laplace distribution, $\varepsilon_i \sim L(0, b)$ with $b > 0$, where $f(\varepsilon_i|0, b) = \frac{1}{2b}e^{-\frac{|\varepsilon_i|}{b}}$. The mean of this distribution is zero and variance is $2b^2$. Then for the estimated mean squared prediction error of estimator $\tilde{\gamma_L^K}$ from Theorem 4.1 with the defined errors ε_i holds*

$$E(\widehat{MSPE}(\tilde{\gamma}_L^K)) \leq \frac{4bKM\sqrt{\log(2p)}}{\sqrt{n}}. \tag{4.11}$$

5 Conclusion

Having homoscedastic errors, the bound in Theorem 4.1 gives a practical hint, how for a predefined approximation error of the OLS-Lasso estimator should the size of the data be set with respect to the number of regressors p, knowing the constants K, M and the variance of errors; This in comparison to [4], is without solving the parameter estimation problem to fit the errors to a normal distribution. The result can encourage practitioners to use OLS Lasso as a convergent estimator for a prediction with other than normal errors satisfying these milder conditions.

Competing Interests

Author has declared that no competing interests exist.

References

[1] Bartlett PL, Mendelson S, Neeman J. l_1-regularized linear regression: Persistence and oracle inequalities. Probab. Theory Related Fields. 2012;154(1-2):193-224.

[2] Rigollet P, Tsybakov A. Exponential screening and optimal rates of sparse estimation. Ann. Statist. 2011;39(2):731-771.

[3] Bühlmann P, van de Geer S. Statistics for high-dimensional data. Methods, theory and applications. Springer Series in Statistics. Springer, Heidelberg; 2011.

[4] Chatterjee S. Assumptionless consistency of the Lasso. eprint arXiv:1303.5817. 2014;1-10.

[5] Chatterjee A, Lahiri SN. Strong consistency of Lasso estimators. The Indian Journal of Statistics. 2011;73-A(Part 1):55-78.

[6] Tibshirani R. Regression shrinkage and selection via the Lasso. J. Royal Statist. Soc., Ser. B. 1996;58(1):267-288.

[7] Hoeffding W. Scale-invariant correlation theory. Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik der Universität Berlin. 1940;5(3):181-233.

[8] Sinn HW. Economic Decisions under Uncertainty (Second English ed.). North-Holland; 1983.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/16710