Neurocomputing 194 (2016) 360–371

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

On scaling of soft-thresholding estimator

Katsuyuki Hagiwara

Faculty of Education, Mie University, 1577 Kurima-Machiya-cho, Tsu 514-8507, Japan

ARTICLE INFO

Article history:
Received 22 January 2015
Received in revised form
31 December 2015
Accepted 18 February 2016
Available online 11 March 2016

Keywords:
Soft-thresholding
LASSO
Non-parametric orthogonal regression
Shrinkage
Scaling

ABSTRACT

LASSO is known to have a problem of excessive shrinkage at a sparse representation. To analyze this problem in detail, in this paper, we consider a positive scaling for soft-thresholding estimators that are LASSO estimators in an orthogonal regression problem. We especially consider a non-parametric orthogonal regression problem which includes wavelet denosing. We first gave a risk (generalization error) of LARS (least angle regression) based soft-thresholding with a single scaling parameter. We then showed that an optimal scaling value that minimizes the risk under a sparseness condition is $1+O(\sqrt{\log n/n})$, where n is the number of samples. The important point is that the optimal value of scaling is larger than one. This implies that expanding soft-thresholding estimator shows a better generalization performance compared to a naive soft-thresholding. This also implies that a risk of LARS-based soft-thresholding with the optimal scaling is smaller than without scaling. We then showed their difference is $O(\log n/n)$. This also shows an effectiveness of the introduction of scaling. Through simple numerical experiments, we found that LARS-based soft-thresholding with scaling can improve both of sparsity and generalization performance compared to a naive soft-thresholding.

1. Introduction

In recent years, sparse modeling is an important topic in machine learning and statistics. Especially, LASSO (least absolute shrinkage and selection operator) is a popular method and has been extensively studied [19,14,21,23,7,22]. In LASSO, estimators are obtained by minimizing a cost function that is defined by squared error sum plus ℓ_1 regularizer; i.e. it is an ℓ_1 penalized least squares method. Introduction of ℓ_1 penalty yields a sparse representation under an appropriate choice of a regularization parameter. In case of orthogonal design, LASSO is known to be reduced to a soft-thresholding method. The soft-thresholding property is still kept for non-orthogonal design case; e.g. see Lemma 1 in [23].

Soft-thresholding is well established as a method of wavelet denoising in signal processing [5,6]. Since discrete wavelet transform is orthogonal transform of samples, wavelet denoising is viewed as a non-parametric orthogonal regression problem. Soft-thresholding is a combination of hard-thresholding and shrinkage in which both of threshold level and amount of shrinkage are simultaneously controlled by one common non-negative parameter. The parameter is a threshold level for removing unnecessary components. And, simultaneously, estimators of coefficients of un-removed components are shrunk toward to zero by subtracting/adding the same parameter value. If the parameter value is large then threshold level is large. Therefore, the number of un-removed components is small; i.e. it is possible to give a sparse representation. However, at the same time, the amount of shrinkage is also large. It implies that there is possible to yield a bias in representing a target function at a relatively small number of components, by which generalization error at a sparse representation may be large. Therefore, the number of un-removed components in soft-thresholding tends to be large if we choose the parameter value based on a substitution of prediction error such as cross-validation error or model selection criterion. This is an inevitable problem of soft-thresholding, which is brought about by an introduction of one parameter for controlling both of threshold level and amount of shrinkage simultaneously.

In machine learning, this dilemma between sparsity and generalization in LASSO has been discussed in [15,8,22]. Ref. [15] has showed a case where a true relation cannot be selected according to a generalization error based model selection. To solve this dilemma in LASSO, SCAD (Smoothly Clipped Absolute Deviation) penalty has been proposed instead of ℓ_1 penalty in [8] and adaptive LASSO that employs a kind of weighted ℓ_1 penalty has been proposed in [22]. An ℓ_1 penalty term is modified by different ways (functions) in SCAD and adaptive LASSO although amount of shrinkage is suppressed for large absolute values of estimators in both methods. Under their modifications, they have shown an optimality of estimators; i.e. oracle property. On the other hand, in wavelet denoising, this problem of soft-thresholding has been pointed out by [11] and [10], in which they claimed that a naive soft-thresholding imposes a large bias on components with large

http://dx.doi.org/10.1016/j.neucom.2016.02.043
0925-2312/© 2016 Elsevier B.V. All rights reserved.

absolute values of coefficients. To overcome this problem, [11] has introduced a firm shrinkage and [10] has introduced the non-negative garrote of [1]. In both of them, the amount of shrinkage is controlled to be small for large absolute values of coefficients.

Although these modifications of LASSO and soft-thresholding are shown to exhibit desirable performances in theoretical and/or practical sense, there is no direct analysis on the above problem of excessive shrinkage at a sparse representation. In this paper, to do this, we consider to introduce a simple scaling of soft-thresholding estimator under a non-parametric orthogonal regression problem in which the number of variables is consistent with the number of data. We first gave a risk. An optimal scaling value is expected to be larger than one to compensate the excessive shrinkage that is thus avoided by the introduction of scaling. Considering a non-parametric orthogonal regression problem gives us an explicit soft-thresholding estimator and enables us a detailed analysis; i.e. quantitative and qualitative evaluation of effects of scaling and a direct comparison to a naive soft-thresholding.

In Section 2, we give setting of non-parametric orthogonal regression in which we mention variations of LASSO under orthogonal designs in detail. In Section 3, we show a well known result on a risk (generalization error) of soft-thresholding and give main theorems that give a risk of soft-thresholding with scaling and an optimal scaling which minimizes the risk. Lemmas for proving theorems are given in Appendix. In Section 4, we verify our theoretical results for a toy problem of harmonic analysis and wavelet denoising. Section 5 is devoted to conclusions and future works.

2. Non-parametric orthogonal regression

2.1. Setting

Let $\boldsymbol{x}=(x_1,\dots,x_m)$ and y be input variables and an output variable, for which we have n i.i.d. samples: $\{(\boldsymbol{x}_i,y_i):i=1,\dots,n\}$, where $\boldsymbol{x}_i=(x_{i,1},\dots,x_{i,m})$. We assume that $y_i=h(\boldsymbol{x}_i)+e_i$, $i=1,\dots,n$, where $e_1,\dots,e_n$ are i.i.d additive noise sequence according to $N(0,\sigma^2)$; i.e. normal distribution with mean 0 and variance σ^2. Under this assumption, we have $\mathbb{E}[y_i|\boldsymbol{x}_i]=h(\boldsymbol{x}_i)$. Thus, h is a target function. We assume that $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ are fixed below. We define $\boldsymbol{y}=(y_1,\dots,y_n)'$, $\boldsymbol{h}=(h(\boldsymbol{x}_1),\dots,h(\boldsymbol{x}_n))'$ and $\boldsymbol{e}=(e_1,\dots,e_n)'$, where $'$ denotes a matrix transpose. We then have $\boldsymbol{y}=\boldsymbol{h}+\boldsymbol{e}$ and $\mathbb{E}_{\boldsymbol{y}}[\boldsymbol{y}]=\boldsymbol{h}$.

Let $g_1,g_2,\dots$ be a series of functions on $\mathbb{R}^m$. We consider to estimate a target function by a linear combination of n functions in this series:

$$f_{\boldsymbol{b}}(\boldsymbol{x})=\sum_{j=1}^{n}b_jg_j(\boldsymbol{x}),\quad \boldsymbol{x}\in\mathbb{R}^m, \tag{1}$$

where $\boldsymbol{b}=(b_1,\dots,b_n)'$ is a coefficient vector. This is a non-parametric regression problem. We call g_j a component or basis function. We assume that there exist n^* and $\boldsymbol{\beta}=(\beta_1,\dots,\beta_n)'$ such that $h(\boldsymbol{x})=\sum_{j=1}^{n}\beta_jg_j(\boldsymbol{x})$ for any $\boldsymbol{x}\in\mathbb{R}^m$ when $n\ge n^*$. β_j can be zero for some j. We define $K^*=\{j:1\le j\le n,\beta_j\ne 0\}$. We call g_j with $j\in K^*$ true component or non-zero component. We also define $k^*=|K^*|$ which is the number of true components or non-zero components. We assume that k^* is a relatively small constant value. This assumption says that there exists a sparse representation of a target function in terms of a set of n components if n is sufficiently large.

Let $\mathbf{G}$ be an $n\times n$ matrix whose (i,j) element is $g_j(\boldsymbol{x}_i)$. We assume that the orthogonality condition:

$$\mathbf{G}'\mathbf{G}=n\mathbf{I}_n, \tag{2}$$

where $\mathbf{I}_n$ denotes an $n\times n$ identity matrix. We thus consider an orthogonal non-parametric regression problem; e.g. discrete Fourier transform and discrete wavelet transform for typical examples. The least squares estimator under the orthogonality condition is given by

$$\widehat{\boldsymbol{c}}=(\widehat{c}_1,\dots,\widehat{c}_n)'=\frac{1}{n}\mathbf{G}'\boldsymbol{y}. \tag{3}$$

Note that we have $\boldsymbol{y}=\mathbf{G}\widehat{\boldsymbol{c}}$ here. Since there exists a $\boldsymbol{\beta}$ such that $\boldsymbol{h}=\mathbf{G}\boldsymbol{\beta}$ when $n\ge n^*$,

$$\widehat{\boldsymbol{c}}\sim N\left(\boldsymbol{\beta},\frac{\sigma^2}{n}\mathbf{I}_n\right) \tag{4}$$

holds by the assumption on additive noise; i.e. multivariate normal distribution with a mean vector $\boldsymbol{\beta}$ and a unit covariance matrix multiplied by σ^2/n. In other words, $\widehat{c}_j\sim N(\beta_j,\sigma^2/n)$, $j=1,\dots,n$ and $\widehat{c}_1,\dots,\widehat{c}_n$ are independent. We define $s_j=\mathrm{sign}(\widehat{c}_j)$, $j=1,\dots,n$, where sign is a sign function. We define $p_1,\dots,p_n$ as an index sequence for which $|\widehat{c}_{p_1}|\ge\cdots\ge|\widehat{c}_{p_n}|$ holds. Throughout this paper, we exclude the case where there are ties since this is guaranteed with probability one by (4).

2.2. LASSO, LARS, elastic net and adaptive LASSO

In this section, we assume that $\mathbf{G}'\mathbf{G}=\mathbf{I}_n$ holds; i.e. the orthonormality condition. For a fixed $\lambda_1\ge 0$, cost function of LASSO is given by

$$S_{\lambda_1}(\boldsymbol{b})=\|\boldsymbol{y}-\mathbf{G}\boldsymbol{b}\|^2+\lambda_1\|\boldsymbol{b}\|_1, \tag{5}$$

where $\|\cdot\|$ is the Euclidean norm and $\|\boldsymbol{b}\|_1=\sum_{k=1}^{n}|b_j|$. λ_1 is a regularization parameter. The second term of the right hand side of (5) is called ℓ_1 regularizer. A minimizer of (5) under the orthonormality condition is given by

$$\widehat{b}_{\mathrm{L},j}=(|\widehat{c}_j|-\lambda_1/2)_+s_j,\quad j=1,\dots,n, \tag{6}$$

where $\widehat{c}_j$ and s_j is defined above and $(u)_+=\max(u,0)$. This is a soft-thresholding estimator in which $\lambda_1/2$ is a parameter that determines threshold level and amount of shrinkage. On the other hand, for fixed $\lambda_1\ge 0$ and $\lambda_2\ge 0$, cost function of elastic net is given by

$$S_{\lambda_1,\lambda_2}(\boldsymbol{b})=\|\boldsymbol{y}-\mathbf{G}\boldsymbol{b}\|^2+\lambda_1\|\boldsymbol{b}\|_1+\lambda_2\|\boldsymbol{b}\|^2. \tag{7}$$

The third term of the right hand side of (7) is called ℓ_2 regularizer. As shown in [24], a minimizer of (7) under the orthonormality condition is given by

$$\widehat{b}_{\mathrm{E},j}=\frac{1}{1+\lambda_2}\widehat{b}_{\mathrm{L},j},\quad j=1,\dots,n. \tag{8}$$

Since $1/(1+\lambda_2)\le 1$, the solution of elastic net is an estimator that is obtained by shrinking LASSO estimator that is a soft-thresholding estimator.

On the other hand, LARS is a greedy iterative algorithm in which a component is appended to a model at each step. This can be viewed as a sparse modeling method if we can appropriately stop it. For this purpose, a C_p type criterion is derived under a mild condition in [7]. As shown in [13] and Lemma 1 in [7], LARS is also reduced to soft-thresholding under the orthonormality condition in which a threshold level is given by $|\widehat{c}_{p_{k+1}}|$ at the kth step; i.e. it is the $(k+1)$th largest absolute value among the least squares estimators. By this choice of threshold level, the number of unremoved components at the kth step is equal to k. Threshold level (amount of shrinkage) is a real number in LASSO while it is in $\{|\widehat{c}_2|,\dots,|\widehat{c}_n|\}$ in LARS.

As in [22], adaptive LASSO solution under the orthonormality condition is given by

$$\widehat{b}_{\mathrm{AL},j}=(|\widehat{c}_j|-w_j\lambda_1/2)_+s_j,\quad j=1,\dots,n, \tag{9}$$

where $w_j=1/|\widehat{c}_j|^\gamma$ for $\gamma>0$. This is a minimizer of a cost function with a weighted ℓ_1 regularizer, in which a weight for the jth