A Scaling and Non-Negative Garrote in Soft-Thresholding

Katsuyuki HAGIWARA\textsuperscript{(a)} Member

SUMMARY Soft-thresholding is a sparse modeling method typically applied to wavelet denoising in statistical signal processing. It is also important in machine learning since it is an essential nature of the well-known LASSO (Least Absolute Shrinkage and Selection Operator). It is known that soft-thresholding, thus, LASSO suffers from a problem of dilemma between sparsity and generalization. This is caused by excessive shrinkage at a sparse representation. There are several methods for improving this problem in the field of signal processing and machine learning. In this paper, we considered to extend and analyze a method of scaling of soft-thresholding estimators. In a setting of non-parametric orthogonal regression problem including discrete wavelet transform, we introduced component-wise and data-dependent scaling that is indeed identical to non-negative garrote. We here considered a case where a parameter value of soft-thresholding is chosen from absolute values of the least squares estimates, by which the model selection problem reduces to the determination of the number of non-zero coefficient estimates. In this case, we firstly derived a risk and construct SURE (Stein’s unbiased risk estimator) that can be used for determining the number of non-zero coefficient estimates. We also analyzed some properties of the risk curve and found that our scaling method with the derived SURE is possible to yield a model with low risk and high sparsity compared to a naive soft-thresholding method with SURE. This theoretical speculation was verified by a simple numerical experiment of wavelet denoising.

key words: soft-thresholding, SURE, non-negative garrote, scaling, wavelet denoising

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 most popular and basic method that is extensively studied [5], [6], [16], [17], [19], [20]. LASSO is an $\ell_1$ penalized least squares method and it is characterized by soft-thresholding that implements thresholding and shrinkage of coefficients. LASSO is expected to yield a sparse representation due to the thresholding. These two properties are simultaneously controlled by a single regularization parameter. If the parameter value is large then both of the threshold level and the amount of shrinkage are large. Therefore, if the parameter value is large then a sparse representation is obtained due to a high threshold level. At the same time, however, an amount of shrinkage is automatically large at the sparse representation. This leads to a possibility of excessive shrinkage that causes a high risk (prediction error) at a sparse representation. Therefore, a model selection method based on a risk estimate tends to choose a larger model even when there is a sparse representation. This drawback in model selection is actually pointed out in [13] as a dilemma between sparsity and generalization. Although SURE (Stein’s unbiased risk estimator) as a model selection criterion for LASSO has been derived by [17], [20], the work of [13] suggests that this cannot select a sparse representation even when it exists. There are several methods to relax the excessive shrinkage problem. [6] has proposed SCAD (Smoothly Clipped Absolute Deviation) penalty instead of a naive $\ell_1$ penalty. [19] has proposed adaptive LASSO that employs a kind of weighted $\ell_1$ penalty. An $\ell_1$ penalty term is modified by different ways (functions) in SCAD and adaptive LASSO while the shrinkage amounts are taken to be small for coefficient estimators whose absolute values are large. In these works, however, a model selection problem under the excessive shrinkage has not been discussed.

On the other hand, soft-thresholding is a well-established method in wavelet denoising. Especially, as a pioneering work, [3], [4] has derived SURE as a model selection criterion under a naive soft-thresholding method. This is a special case of [20] for a non-parametric orthogonal regression problem. In wavelet denoising, the problem of excessive shrinkage has been pointed out and improved in [7], [8]. [7] has proposed firm shrinkage that employs piecewise linear soft-thresholding function. [8] has introduced non-negative garrote of [1] into wavelet denoising. These works are intended to improve amounts of shrinkage for coefficient estimators whose absolute values are large; i.e. those are possibly efficient components. Especially, [8] has derived SURE for a parameter selection under the non-negative garrote. This result is important in applications since it enables us an automatic model construction procedure which is not derived for SCAD and adaptive LASSO above. Alternatively, a scaling method has been introduced by [11] to compensate excessive shrinkage and shown to be effective for improving a risk at a sparse representation. In this work, a single scaling parameter is introduced for a theoretical analysis of the excessive shrinkage problem. Therefore, it is not an application-oriented investigation.

In this paper, for an effective application of scaling, we consider to extend [11] to component-wise scaling; i.e. different scaling parameters are component-wisely introduced for every coefficients. Especially, we consider data-dependent and component-wise scaling that is indeed identical to non-negative garrote in [8]. Therefore, this paper in-
vestigates non-negative garrote under an orthogonality condition from a viewpoint of scaling. In this meaning, our work here is closely related to [8] while it is different from [8] in terms of several points. Firstly, our investigation is based on scaling of soft-thresholding estimator. This gives us clear insight into the problem of excessive shrinkage. Secondly, [8] has derived SURE at a fixed parameter value while it has recommended that the parameter value is chosen from a set of data-dependent candidates. We here directly derive SURE under the data-dependent choice of the parameter value. Thirdly, [8] has derived an asymptotic risk bound as in [3], [4]; i.e. it concerns with the risk when the number of data goes to infinity. In this paper, we mainly concern the properties of risk when the number of un-removed components changes; i.e. analysis on model selection using a risk-based criterion.

In Sect. 2, we firstly formulate a problem of non-parametric orthogonal regression that includes discrete wavelet transform. In this section, we explain scaling of soft-thresholding and also define a component-wise and data-dependent scaling that is shown to be identical to non-negative garrote. We then give SURE under the scaling method and give theoretical insights into the risk curve. In Sect. 3, we give SURE for level fixed wavelet denoising and show a numerical example of its application. Section 4 is devoted by conclusions and future works.

2. Non-Parametric Orthogonal Regression

2.1 A Framework of Orthogonal Non-Parametric Regression

\{(t_i, y_i) : i = 1, \ldots, n\} are samples in which \(y_i\) is a observed noisy sample of a target signal at \(t_i\). We then assume that \(y_i = h(t_i) + e_i, i = 1, \ldots, n\), where \(e_1, \ldots, e_n\) are i.i.d. additive noise sequence according to \(N(0, \sigma^2)\); i.e. normal distribution with mean 0 and variance \(\sigma^2\). \(h\) is a target signal or function. We define \(y = (y_1, \ldots, y_n)', \ h = (h(t_1), \ldots, h(t_n))'\) and \(e = (e_1, \ldots, e_n)', \ \text{where }'\text{ denotes a matrix transpose.} \ We then have \(y = h + e\) and \(\mathbb{E}[y] = h\), where \(\mathbb{E}\) denotes the expectation with respect to the joint probability distribution of \(y\).

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

\[ f(t) = \sum_{j=1}^{n} b_j g_j(t), \ t \in \mathbb{R}, \tag{1} \]

where \(b = (b_1, \ldots, b_n)'\) is a coefficient vector. This is a non-parametric regression problem. We call \(g_j\) a component or basis function. We assume that there exists \(n^*\) and \(\beta = (\beta_1, \ldots, \beta_n)'\) such that \(h(t) = \sum_{j=1}^{n^*} \beta_j g_j(t)\) for any \(t \in \mathbb{R}\) when \(n \geq n^*\). \(\beta_j\) can be zero for some \(j\). By assuming that \(n \geq n^*\), we define \(K^* = \{j : 1 \leq j \leq n, \beta_j \neq 0\}\) and denote the complement of \(K^*\) by \(\overline{K}\). 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. By the definition of \(K^*\), \(n^*\) is taken to be the maximum index in \(K^*\). Throughout this paper, we assume that \(n \geq n^*\) holds; i.e. true components are always included in an assumed model. We also assume that \(k^*\) is very small compared to \(n\). These two assumptions say that there exists a sparse representation of a target function in terms of a set of \(n\) components. These assumptions for sparseness is made to theoretically clarify that a scaling method can solve a problem of soft-thresholding. Although those seem to be restrictive in practices, the analysis under these assumptions is not so irrelevant in applications since we consider a nonparametric regression problem, in which a model is rich enough and an estimation bias can be negligible when \(n\) is large.

Let \(G\) be an \(n \times n\) matrix whose \((i, j)\) element is \(g_j(t_i)\). We assume that the orthogonality condition:

\[ G'G = nI_n, \tag{2} \]

where \(I_n\) denotes an \(n \times n\) identity matrix. We thus consider a non-parametric orthogonal regression problem; e.g. discrete Fourier transform and discrete wavelet transform for typical examples. Especially, wavelet denoising has been extensively studied and applied in signal processing [3], [4], [8]. In this paper, we show numerical examples of the wavelet denoising later. The least squares estimator of \(b\) under the orthogonality condition is given by

\[ \hat{c} = (\hat{c}_1, \ldots, \hat{c}_n)' = \frac{1}{n} G'y. \tag{3} \]

Note that we have \(y = G\hat{c}\) here. Since there exists a \(\beta\) such that \(h = G\beta\),

\[ \hat{c} \sim N(\beta, \sigma^2 / n) \tag{4} \]

holds by the assumption on additive noise; i.e. multivariate normal distribution with a mean vector \(\beta\) and a unit covariance matrix multiplied by \(\sigma^2 / n\). In other words, \(\hat{c}_j \sim N(\beta_j, \sigma^2 / n), j = 1, \ldots, n\) and \(\hat{c}_1, \ldots, \hat{c}_n\) are independent. We define \(s_j = \text{sign}(\hat{c}_j), j = 1, \ldots, n\), where sign is a sign function. We define \(p_1, \ldots, p_n\) as an index sequence for which \(|\hat{c}_{p_1}| \geq \cdots \geq |\hat{c}_{p_n}|\) holds. Note that we can exclude the case of ties in our probabilistic evaluations in this paper since this is guaranteed with probability one by (4).

2.2 Soft-Thresholding

Soft-thresholding estimator with a parameter \(\theta \geq 0\) is given by

\[ \hat{b}_{\theta, j} = (|\hat{c}_j| - \theta)_+, j = 1, \ldots, n \tag{5} \]

where \((a)_+ = \max(a, 0)\). Therefore, \(\theta\) determines both of a threshold level and an amount of shrinkage. We define \(\hat{b}_\theta = (\hat{b}_{\theta, 1}, \ldots, \hat{b}_{\theta, n})\). It is well known that \(\hat{b}_\theta\) is a solution to the minimization problem of
\[ L_0(b) = \frac{1}{n} \| y - Gb \|^2 + 2\theta \| b \|_1 \]  
(6)

under the orthogonality condition on \( G \), where \( \| \cdot \| \) and \( \| \cdot \|_1 \) denote the \( \ell_2 \) and \( \ell_1 \) norms. Soft-thresholding is thus a special case of LASSO [16].

We consider a case where \( \theta \) is chosen from \( \{ \hat{\theta}_1, \ldots, \hat{\theta}_n \} \), where \( \hat{\theta}_k = \| \hat{c}_p \|_1 \) for \( k < n \) and \( \hat{\theta}_n = 0 \). Therefore, un-removed components are \( g_{p_1}, \ldots, g_{p_n} \) if \( \theta = \hat{\theta}_k \); i.e. the number of un-removed components is \( k \). In applications, it is difficult to search a real line to find an optimal \( \theta \). Thus, this choice of candidates is effective and reasonable in applications since the number of un-removed components changes only at the candidates; e.g. see [8]. We here assume that \( \hat{c}_{p_i} \neq \hat{c}_{p_j} \) for any \( i \neq j \) since it holds with probability one. Indeed, soft-thresholding with this restriction is identical to Least Angle Regression (LARS) algorithm under the orthogonality condition [5], [10]. We simply denote \( \hat{b}_k = \hat{b}_{\hat{\theta}_k} \) and \( \hat{b}_n = (\hat{b}_{\hat{\theta}_1}, \ldots, \hat{b}_{\hat{\theta}_n}) \). A method using this estimator is referred as a naive soft-thresholding. We define \( \hat{K}_k = \{ j : \hat{b}_{k,j} \neq 0 \} \); i.e. a set of indexes of un-removed components.

2.3 Scaling of Soft-Thresholding Estimator

We introduce a scaling of soft-thresholding estimator. We define \( \alpha = (\alpha_1, \ldots, \alpha_n) \), \( \alpha_j \geq 0 \), \( j = 1, \ldots, n \). The soft-thresholding estimator with scaling is defined by

\[ \alpha_j \hat{b}_{k,j}, \ j = 1, \ldots, n. \]

Let \( A \) be an \( n \times n \) diagonal matrix whose diagonal vector is \( \alpha \). Then a vector of soft-thresholding estimators with scaling is written by \( A\hat{b}_k \). Since our purpose of introducing scaling is to compensate for excessive shrinkage, \( \alpha_j \) is possibly larger than one. If \( \alpha = 1 \), then we have a naive soft-thresholding estimator, where \( 1 \) is an \( n \)-dimensional vector of ones. For \( A\hat{b}_k \), we define a risk by

\[ R(n,k,\alpha) = \frac{1}{n} \mathbb{E} \left[ \| h - G\alpha\hat{b}_k \|^2 \right], \]

(8)

where \( \mathbb{E} \) denotes the expectation with respect to the joint distribution of \( y \). For a naive soft-thresholding estimator,

\[ R(n,k,1) = \frac{1}{n} \mathbb{E} \left[ \| y - G\alpha \hat{b}_k \|^2 \right] - \sigma^2 + \frac{2k\sigma^2}{n}. \]

(9)

has been derived; e.g. see [5], [10]. [11] has analyzed the case of \( \alpha = \alpha_1 1 \) in details, where \( \alpha \geq 0 \); i.e. it has introduced a single common scaling parameter. [11] has derived the risk in this case and obtained an optimal scaling value that minimizes the risk. It has showed that the optimal scaling value is larger than one and the risk is improved at the optimal scaling value. [11] has also shown that, by model selection via an unbiased estimate of the risk, the introduction of scaling tends to select a sparse representation with low risk compared to a naive soft-thresholding. However, the improvement of the risk was not satisfactory. We then need to consider the choice of component-wise scaling values for a better performance in applications.

2.4 Choice of Scaling Values and Non-Negative Garrote

The purpose of scaling is to compensate for excessive shrinkage of coefficients of un-removed components. Then, it is reasonable to choose \( \alpha_j \) so as to satisfy \( \alpha_j \hat{b}_{k,j} = \hat{c}_j \). This yields

\[ \alpha_j = 1/\left(1 - \hat{\theta}_k/|\hat{c}_j|\right) = 1 + \hat{\theta}_k/|\hat{c}_j|, \ j = 1, \ldots, n \]

(10)

when \( \hat{\theta}_k/|\hat{c}_j| \) is small. This approximation is valid since an un-removed component may have a coefficient estimate that is sufficiently larger than an appropriate threshold level. In this paper, we hence employ

\[ \hat{\alpha}_j = 1 + \hat{\theta}_k/|\hat{c}_j|, \ j = 1, \ldots, n \]

(11)

as empirical values. We may set \( \hat{\alpha}_j \) to any value if \( |\hat{c}_j| = 0 \). This do not cause any problems in the discussion below since this event does not occur with probability one. We define \( \hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_n) \). (11) gives component-wise and data-dependent scaling values. We refer this scaling method as adaptive scaling. By (11), the adaptive scaling value is always larger than one. Note also that \( \hat{\alpha}_j \) is valid only for \( j \in \hat{K}_k \) since \( \hat{b}_{k,j} = 0 \) for \( j \notin \hat{K}_k \). Let \( A \) be an \( n \times n \) diagonal matrix whose \( (j,j) \) element is \( \hat{\alpha}_j \). We define a risk of adaptive scaling by

\[ R_{\text{AS}}(n,k) = \frac{1}{n} \mathbb{E} \left[ \| h - GA\hat{b}_k \|^2 \right] = \mathbb{E} \left[ \| A\hat{b}_k - \beta \|^2 \right]. \]

(12)

The latter definition comes from the orthogonality condition.

On the other hand, as in [1], [8], non-negative garrote estimates under the orthogonality condition are given by

\[ \hat{b}_{k,j}^{\text{ng}}(\theta) = \begin{cases} \hat{c}_j(1 - \theta^2 \hat{c}_j^2) & |\hat{c}_j| > \theta \\ 0 & |\hat{c}_j| \leq \theta \end{cases} \]

(13)

for \( \theta > 0 \). Although this yields a shrinkage estimator for an un-removed component, \( \hat{b}_{k,j}^{\text{ng}} \approx \hat{c}_j \) holds if \( |\hat{c}_j| \) is sufficiently larger than \( \theta \). Therefore, the amount of shrinkage is surely relaxed for un-removed components. Indeed, it is easy to see that \( \hat{\alpha}_j \hat{b}_{k,j} = \hat{b}_{k,j}^{\text{ng}}(\hat{\theta}_k) \); i.e. adaptive scaling is identical to non-negative garrote [1], [8]. [8] has derived SURE at a fixed parameter value while it has recommended that the parameter value is chosen from \( \{ \hat{\theta}_1, \ldots, \hat{\theta}_n \} \) which are data-dependent statistics. We here directly derive the risk and SURE when the parameter value is \( \hat{\theta}_k \). Furthermore, we consider some properties of risk curve of adaptive scaling, thus non-negative garrote.

2.5 Main Results

We define \( c_j = \sqrt{n} \beta_j / \sigma \) and \( e = (e_1, \ldots, e_n) \). We also define
Theorem 3: holds. This implies that, for \( j \in K^* \),

\[
\lim_{n \to \infty} P(\hat{\alpha}_j - 1 > \epsilon) = 0
\]

holds for any \( \epsilon > 0 \). On the other hand, we assume that \( k > k^* \). Then, for \( j \in \mathcal{K} \),

\[
\lim_{n \to \infty} P(\hat{\alpha}_j < 2 - \epsilon) = 0
\]

holds for any \( \epsilon > 0 \).

Theorem 3:

\[
R(n, k^*, \mathbf{1}_n) - R_{AS}(n, k^*) \geq 2 \sigma^2 k^* \log n \frac{\log n}{n}
\]

holds for a sufficiently large \( n \).

We give some remarks.

- By Theorem 1,

\[
\tilde{R}_{AS}(n, k) = \tilde{L}(n, k) - \sigma^2 + \tilde{D} \tilde{F}_{AS}(n, k).
\]

is SURE; i.e. an unbiased estimator of risk. Therefore, this can be a model selection criterion for choosing an optimal \( k \) if we can employ an appropriate estimate of noise variance \( \sigma^2 \) in (23).

- By Lemma 4 in Appendix, the probability that all of true components are un-removed is high when \( k \geq k^* \) and \( n \) is sufficiently large; i.e. a naive soft-thresholding has a kind of consistency in selecting true components if those exist. Note that since our adaptive scaling is applied to a naive soft-thresholding estimator, this consistency result applies to adaptive scaling estimators.

- Theorem 2 says that, in a large sample situation, scaling values are larger than 2 for components that are not true. This excessive expansion of coefficient estimators for non-true components may cause a high risk for \( k > k^* \). Some of non-true components may be selected when \( k > k^* \). Therefore, \( R_{AS}(n, k) > R(n, k, \mathbf{1}_n) \) may hold for \( k > k^* \) while \( R_{AS}(n, k^*) < R(n, k^*, \mathbf{1}_n) \) holds by Theorem 3. This fact seems to be disadvantage of adaptive scaling. However, it may not be true from the viewpoint of model selection since this property allows us to identify the minimum of risk curve; i.e. risk is small at \( k = k^* \) while it is large when \( k \neq k^* \). Hence, nearly optimal \( k \) is expected to be found according to a model selection criterion given by (23). And, at such an optimal \( k \), a consistent choice of a set of true components and a low risk value are guaranteed by Lemma 4 and Theorem 3 respectively. In other words, under adaptive scaling, we can obtain a model that gives low risk at a sparse representation. This speculation is verified in numerical experiments in the next section.

- (20) in Theorem 2 implies that \( \hat{\alpha}_j - 1 = \hat{\theta}_j / \hat{c}_j, j \in K^* \) converges to zero in probability if \( k \geq k^* \). Therefore, approximations in (10) are valid for true components and naive soft-thresholding estimators for true components are nearly the least squares estimators. This fact is well understood by the definition of a non-negative garrote estimator defined in (13); i.e. amount of shrinkage for the least squares estimator is given by \( (\hat{\alpha}_j - 1)^2 \) here. Additionally, as mentioned above, true components may be consistently selected according to (23) if variance estimate is suitable. Therefore, a model estimated by our adaptive scaling scheme may be close to one estimated by a hard thresholding for which it is difficult to establish a model selection procedure.

3. Application to Wavelet Denoising

3.1 Discrete Wavelet Transform

Discrete wavelet transform is a popular tool for analysis, de-noising and compression of signals and images; e.g. see [2]. In discrete wavelet transform, we assume that \( n = 2^J \) for a natural number \( J \). Let \( a_j = (a_{j, 1}, \ldots, a_{j, n_j}) \) and \( d_j = (d_{j, 1}, \ldots, d_{j, n_j}, \ldots, d_{j-1, 1}, \ldots, d_{j-1, n_{j-1}}) \) be approximations and detail coefficients at a level \( J \) in discrete wavelet transform, where \( n_j = 2^j \). We define \( w_j = (a_j, d_j) \) in which we set \( w_j = a_j \) for \( j = J \). By setting \( a_j = y \), the decomposition algorithm with a pre-determined wavelet calculates \( w_{j-1} \) from \( a_j \) by decreasing \( j = J, J-1, \ldots, J_0 \), where \( J_0 \) is a fixed level determined by users. This procedure can be written by

\[
w_{J_0} = H_{J_0}y.
\]
Theorem 4: Under adaptive scaling defined in (27),

\[
R_{AS}(n,k) = \mathbb{E} \left[ \overline{L}(n,k) + \left( 1 - \frac{2n_0}{n} \right) \sigma^2 + \mathbb{E} \left[ \Delta \overline{F}_{AS}(n,k) \right] \right],
\]

holds, where

\[
\overline{L}(n,k) = \frac{1}{n} \| \tilde{w}_j - y \|^2
\]

\[
\Delta \overline{F}_{AS}(n,k) = \frac{2\sigma^2 k}{n} + \frac{2\sigma^2}{n} \mathbb{E} \left[ \sum_{i,j,m,k} (\tilde{a}_{i,m} - 1)^2 \right].
\]

We omit the proof since this fact can be shown by following the proof of Theorem 1 only to detail coefficients, in which we should just note that approximation coefficients are the least squares estimators. It also can be shown that risk of naive soft-thresholding is given by

\[
R(n,k,J_0) = \mathbb{E} \left[ \overline{L}(n,k) \right] + \left( 1 - \frac{2n_0}{n} \right) \sigma^2 + \frac{2\sigma^2 k}{n}.
\]

3.2 Numerical Examples

We here compare the prediction accuracy and sparsity among soft-thresholding with adaptive scaling (STAS), a naive soft-thresholding (NST) and the universal soft-thresholding (UST) in [3]. In UST, a parameter of soft-thresholding that is common for all detail coefficients is given by \( \theta_{ST} = \sqrt{2\sigma^2 \log n} \), where \( \sigma^2 \) is an estimate of noise variance. In wavelet denoising, the median absolute deviation (MAD) is a standard robust estimate of noise variance. It is given by

\[
\hat{\sigma} = \text{median}(|d_{j-1,1}|, \ldots, |d_{j-1,n_{j-1}}|)/0.6745,
\]

where \( d_{j-1,1}, j = 1, \ldots, n_{j-1} \) are the smallest scale wavelet coefficients that are empirically known to be noise dominated components. Model selection criteria for NST and STAS are SUREs of (32) and (29) respectively, in which \( \sigma^2 \) is replaced with MAD above.

We choose “heavi”sine given in [3] as a test signal and employ the orthogonal Daubechies wavelet with 8 wavelet/scaling coefficients. This choice is made because “heavi”sine is well fitted by a family of Daubechies wavelets. Therefore, we can expect that there is a nearly sparse representation and this example is suitable for verifying our theoretical speculations. Additive noise has a normal distribution with mean 0 and variance \( \sigma^2 = 1 \). Signal-to-noise ratio is set to 4. We set \( n = 2^j \) with \( J = 9 \) and \( J_0 = 2 \).

In Fig. 1, we show the averages of risks and those of SUREs for NST and STAS when \( k \) changes. The averages are calculated by 100 trials. The risk is calculated by mean squared error between output estimate and signal output. SUREs are calculated by (32) for NST and (29) for STAS respectively. In the figure, we can see that the SUREs approximate the risks well for both methods, in which slight differences are due to the MAD estimate of noise variance. The risk of STAS is smaller than that of NST at a sparse representation. This is an empirical evidence for Theorem 3. At a large \( k \), however, the risk of STAS is larger than that of NST. This is an evidence for the remark after theorems; i.e. in STAS, adaptive scaling approaches to 2 for fruitless components.
There are two important points. Firstly, risk of STAS is minimized at a sparse representation and the minimum of the risk of STAS is smaller than that of NST. Secondly, the shape of risk curve of NST is nearly flat while the minimum of risk curve of STAS is clearly identified. These suggest that a model selected by SURE in STAS tends to be smaller and have a lower risk compared to a naive soft-thresholding method with SURE. This was verified by a simple numerical experiment of an application to wavelet denoising. As a future work, in wavelet denoising, we need more application results including a choice of an optimal level and comparison to the other methods. Although we gave scaling values in a top down manner in this paper, we can test the other forms of adaptive scaling values; e.g. scaling values which are estimates of optimal values in some senses. In the context of machine learning, we need to consider a model selection criterion under a scaling method for general regression problems. Especially, we are interested in the effectiveness of scaling under a problem of large input dimensionality since it is an important topic in recent years. Since this paper considered a case where the number of data is equal to the input dimension, the above future work can be viewed as an extension of this paper to a non-orthogonal case.

4. Conclusions and Future Works

In this paper, to overcome a problem of excessive shrinkage at a sparse representation in a soft-thresholding method, we considered to introduce scaling of soft-thresholding estimators in a non-parametric orthogonal regression problem including wavelet denoising. Especially, we employed component-wise and data-dependent scaling that is identical to non-negative garrote. When candidates values of a parameter of soft-thresholding are the absolute values of the least squares estimates, we firstly derived a risk, or equivalently, prediction error. For determining an optimal parameter value that is equivalent to an optimal number of un-removed components, we then gave SURE that is an unbiased estimate of the risk. We also analyzed some properties of the risk curve and found that the SURE is possible to select a model with low risk and high sparsity compared to a naive soft-thresholding method with SURE. This was verified by a simple numerical experiment of an application to wavelet denoising. As a future work, in wavelet denoising, we need more application results including a choice of an optimal level and comparison to the other methods. Although we gave scaling values in a top down manner in this paper, we can test the other forms of adaptive scaling values; e.g. scaling values which are estimates of optimal values in some senses. In the context of machine learning, we need to consider a model selection criterion under a scaling method for general regression problems. Especially, we are interested in the effectiveness of scaling under a problem of large input dimensionality since it is an important topic in recent years. Since this paper considered a case where the number of data is equal to the input dimension, the above future work can be viewed as an extension of this paper to a non-orthogonal case.

References

Appendix A: Lemmas

We here give some lemmas that are used for proving the main theorems.

We denote $\chi^2$ distribution with one degree of freedom by $\chi^2_1$. Let $f_1$ be a probability density function of $\chi^2_1$. Let $X_1, \ldots, X_n$ be random variables. We define the $m$th largest value among $X_1, \ldots, X_n$ by $X_{(m)} = X_{(m)}(n)$.

**Lemma 1:** Let $X_1, \ldots, X_n$ be i.i.d. random variables from $\chi^2_1$. We define $t_n = 2 \log n - \log \log n - \log \pi$. Then, at each fixed $k = 1, 2, \ldots,$

$$\lim_{n \to \infty} \mathbb{E} \left[ (X_1(t_n)/2)^k \right] = (-1)^k \Gamma^k(1)$$

(A-1)

hold, where $\Gamma^k(1)$ is the $k$th derivative of the Gamma function at 1. (A-1) implies that

$$\lim_{n \to \infty} \mathbb{E} \left[ X_{(m)}^k / X_{(n)}^k \right] = 1.$$  

(A-2)

**Proof** By slightly modifying Example 3, pp.72–73 in [14], we can see that $\mathbb{P}\left( (X_1(t_n)/2 \leq x) \right)$ converges to the double exponential distribution. Then, (A-1) is a direct conclusion of Proposition 2.1 (iii) in [14].

**Lemma 2:** Let $X_1, \ldots, X_n$ be i.i.d. random variables from $\chi^2_1$. At each fixed $m$,

$$\lim_{n \to \infty} \mathbb{P} \left[ X_{(m)} \leq 2(1 - \delta) \log n \right] = 0,$$  

(A-3)

and

$$\lim_{n \to \infty} \mathbb{P} \left[ X_{(m)} > 2 \log n \right] = 0$$  

(A-4)

hold, where $\delta$ is an arbitrary positive constant.

**Proof** For a $\chi^2_1$ random variable $X$, it is easy to see that

$$\mathbb{P}[X > x] \sim 2f_1(x)$$  

(A-5)

holds for a sufficiently large $x$. By (A-5), we then obtain

$$\mathbb{P}\left[ X_{(1)} > 2 \log n \right] \leq \sum_{i=1}^{n} \mathbb{P}\left[ X_i > 2 \log n \right] \sim 2nf_1(2 \log n).$$  

(A-6)

This goes to 0 as $n \to \infty$ by the definition of $f_1$. Since $X_{(1)} \geq X_{(m)}(n)$ for any $m$, we have (A-4). On the other hand, by (A-5), we have

$$n(1 - F_1(2(1 - \delta) \log n)) \sim 2nf_1(2(1 - \delta) \log n) \sim \frac{1}{\pi(1 + \delta)} \frac{n^\delta}{\sqrt{\log n}}$$  

(A-7)

for a sufficiently large $n$. Since this goes to $\infty$, we obtain (A-3) by Theorem 2.2.1 in [12].

We define $\overline{c}_1 = \overline{c}_1 - \sqrt{n} \beta / \sigma$, by which $\overline{c}_1, \ldots, \overline{c}_n$ are i.i.d. according to $N(0, 1)$ by the definition of $\overline{c}_1, \ldots, \overline{c}_n$. For an event $E$, we denote the complement of $E$ by $\overline{E}$ and indicator function of $E$ by $I_E$. We define $E_{n,i} = \{ p_i \in K^* \}$ and $E_n = \bigcap_{i=1}^{K^*} E_{n,i}$. We also define $F_j = (\overline{c}_j, \overline{c}_j \leq \max_{i \in K} \overline{c}_i)$. We denote $F_i = (\overline{c}_i, \overline{c}_i \leq \max_{i \in K} \overline{c}_i)$.

**Lemma 3:** For any $j \in K^*$ and any $\rho > 0$,

$$\mathbb{P}\left[ F_j^c \leq \max_{i \in K} \overline{c}_i^c \right] \leq 2 \pi^{-1/2} \rho^{1/2} n^{-p}$$  

(A-8)

holds for a sufficiently large $n$.

The proof of this lemma is given in [11].

**Lemma 4:**

$$\mathbb{P}[\overline{E}_n^c] \leq k^* \pi^{-1/2} \rho^{1/2} n^{-p}$$  

(A-9)

holds for any $\rho > 0$ and a sufficiently large $n$.

**Proof** If $E_n^c$ does not occur then there exist $l \in \{1, \ldots, k^*\}$ such that $p_l \notin K^*$. This implies that there exist $j \in K^*$ and $i \in K$ that satisfy $\overline{c}_j \geq \overline{c}_i$. Therefore, we have $E_n^c \subseteq \bigcup_{j \in K^*} F_j$. By Lemma 3, we then obtain (A-9).

**Lemma 5:**

$$\lim_{n \to \infty} \mathbb{E}[\overline{c}_p^m I_{E_n}] = 0$$  

(A-10)

holds for a fixed $m = 1, 2, \ldots$.

**Proof** We define $\overline{c} = \max_{1 \leq i \leq p} \overline{c}_i$ and $\overline{\beta} = \max_{j \in K^*} |\beta_j|$. We also define an event $F = [\overline{c} > \sqrt{n} \overline{\beta} / \sigma]$. By the Cauchy-Schwarz inequality, we have

$$\mathbb{E}[\overline{c}_p^m I_{E_n}] \leq \mathbb{E}[\overline{c}_p^m + \sqrt{n} \overline{\beta}_p / \sigma]^{m} I_{E_n} \leq \mathbb{E}[(\overline{c} + \sqrt{n} \overline{\beta} / \sigma)^{2m} I_{F}]$$  

and

$$\mathbb{E}[(\overline{c} + \sqrt{n} \overline{\beta} / \sigma)^{2m} I_{F}] \leq \mathbb{E}[(\overline{c} + \sqrt{n} \overline{\beta} / \sigma)^{2m} I_{F}]$$  

(A-10)
\[
\begin{align*}
&\leq 2^{2m}\mathbb{E}[(z_{cm}I_{F}\mathcal{E}_n^c)] + 2^{2m}(\tilde{\beta}/\sigma)^{2m}n^{2m}\mathbb{E}[I_{F}\mathcal{E}_n^c] \\
&\leq 2^{2m}\mathbb{E}[(c_{cm}^2I_{F}\mathcal{E}_n^c)] + 2^{2m}(\tilde{\beta}/\sigma)^{2m}n^{2m}\mathbb{E}[I_{F}\mathcal{E}_n^c] \\
&\leq 2^{2m}\sqrt{\mathbb{E}[c_{cm}^2]} \sqrt{\mathbb{P}[^c_{cm}^2]} + 2^{2m}(\tilde{\beta}/\sigma)^{2m}n^{2m}\mathbb{P}[^c_{cm}^2]. \\
&\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \ Quad
through a simple calculation. We evaluate the three terms in the sum of (A.26). We first have
\[
\lim_{n \to \infty} \mathbb{E}[\tilde{\theta}_k^2]/(2 \log n) = 1
\]  
by Lemma 6. Hence, the proof is completed by showing that the second and third terms of (A.26) goes to zero as \( n \to \infty \). We define \( \epsilon_n = \max_{j \in K^*} \epsilon_{ja} \) and \( G_j = (\eta_{p_j}^2 > \epsilon_n^2) \), where \( \epsilon_{ja} \) is defined in (18). We have \( \eta_{p_j}^2 \leq 1 \) for \( j \leq k^* \) by the definition of \( \tilde{\theta}_k = [\tilde{\theta}_{p_{1}}] \). And, if \( E_n^* \) occurs then \( \bigcup_{j \in K^*} \{ p_j = l \} \) for any \( j \in \{ 1, \ldots, k^* \} \). We then obtain
\[
\mathbb{E} \{ \eta_{p_j}^2 \} = \mathbb{E} \{ \eta_{p_j}^2 | G_j \cap E_n^* \} + \mathbb{E} \{ \eta_{p_j}^2 | G_j \cap E_n^* \}
\leq \mathbb{E} \{ G_j \cap E_n^* \} + \epsilon_n^2 + \mathbb{P} [ E_n^* ]
\leq \sum_{j \in K^*} \mathbb{P} [ \eta_{p_j}^2 > \epsilon_n^2 ] + \epsilon_n^2 + \mathbb{P} [ E_n^* ].
\]  
(A.28)

(A.28) goes to zero as \( n \to \infty \) by (19) in Theorem 2, the definition of \( \epsilon_n \) and Lemma 4. We also have
\[
\mathbb{E} \{ \eta_{p_j}^2 \} \tilde{\theta}_k^2 = \mathbb{E} \{ \eta_{p_j}^2 \} \tilde{\theta}_k^2 G_j \cap E_n^* + \mathbb{E} \{ \eta_{p_j}^2 \} \tilde{\theta}_k^2 G_j \cap E_n^*
\leq \mathbb{E} \{ \tilde{\theta}_k^2 | G_j \cap E_n^* \} + \epsilon_n^2 \mathbb{E} \{ \tilde{\theta}_k^2 | G_j \cap E_n^* \} + \mathbb{E} \{ \tilde{\theta}_k^2 | G_j \cap E_n^* \}.
\]  
(A.29)

For the first term of (A.29), by the Cauchy-Schwarz inequality, we have
\[
\frac{\mathbb{E} \{ \tilde{\theta}_k^2 | G_j \cap E_n^* \}}{2 \log n} \leq \frac{\mathbb{E} \{ \tilde{\theta}_k^2 \}}{2 \log n} \mathbb{E} \{ G_j \cap E_n^* \}
\leq \frac{\mathbb{E} \{ \tilde{\theta}_k^2 \}}{2 \log n} \sum_{j \in K^*} \mathbb{P} [ \eta_{p_j}^2 > \epsilon_n^2 ].
\]  
(A.30)

(A.30) goes to zero as \( n \to \infty \) by Lemma 6 and (19) in Theorem 2. The second term of (A.29) goes to zero as \( n \to \infty \) by Lemma 6 and the definition of \( \epsilon_n \). By the Cauchy-Schwarz inequality, the third term of (A.29) is bounded above by \( \sqrt{\mathbb{E} \{ \tilde{\theta}_k^2 \}} \sqrt{\mathbb{P} [ E_n^* ]} \). This goes to zero as \( n \to \infty \) by Lemma 6 and Lemma 4. We thus obtain (22) as desired. \( \square \)

**Proof of Theorem 3** We write \( \eta = \tilde{\alpha}_l \) here. By (9) and (15), we have
\[
R(n, k^*, 1_a) = R_{AS}(n, k^*)
= \alpha^2 2 \log n \sum_{j = 1}^{K} \left( \frac{\mathbb{E} \{ \eta_{p_j}^2 \}}{2 \log n} - \frac{\mathbb{E} \{ \eta_{p_j}^2 \} \tilde{\theta}_k^2}{2 \log n} - \frac{2 \mathbb{E} \{ \eta_{p_j}^2 \}}{2 \log n} \right)
\]  
(A.26)