INFONET, GIST
Journal Club (2013. 03. 21)

Scaling Up MIMO: Opportunities and challenges with very large arrays

Authors: Fredrik Rusek, Thomas L. Marzetta, et al.
Publication: IEEE Signal Processing Magazine
Speaker: Woongbi Lee

Short summary: Very large MIMO systems is an emerging research area in antenna systems, electronics, and wireless communication systems. A base station with an antenna array serves a multiplicity of single-antenna terminals. In this presentation, the fundamental principle of massive MIMO technology and several issues are introduced.

I. INTRODUCTION

Multiple-Input, Multiple-Output (MIMO) technology is becoming mature, and incorporated into emerging wireless broadband standard like LTE. Basically, the more antennas the transmitter/receiver is equipped with, and the more degrees of freedom that the propagation channel can provide, the better performance in terms of data rate or link reliability. However, MIMO technology requires increased complexity of the hardware, the complexity and energy consumption of the signal processing, and the physical space for accommodating antennas including rents of real estate.

Today, as mobile data traffic exponentially increases, further capacity enhancement is needed. As a solution for the high capacity demand, Massive MIMO (very large MIMO, Large-Scale Antenna System, Full Dimension MIMO) technology has been widely studied for last few years. Massive MIMO adopts hundreds of antennas at base station (BS) serving a much smaller number of terminals. The number of terminals that can be simultaneously served is limited, not by the number of antennas, but rather by inability to acquire channel-state information for an unlimited number of terminals. With an unlimited number of antennas, the transmit power can be made arbitrarily small and the uncorrelated interference and noise can be vanished. But, the performance is limited by pilot contamination.

This paper approaches to Massive MIMO according to three directions: Information-theoretic performance limit, and antennas and propagation aspects of large MIMO, and transmit and receive schemes.

II. Information Theory for Very Large MIMO Arrays

According to the noisy-channel coding theorem in information theory, for any communication link, there is a capacity or achievable rate, such that for any transmission rate less than the capacity, there exists a coding scheme that makes the error-rate arbitrarily small.

A. Point-to-point MIMO

1) Channel model

Transmitter has an array of n_t antennas and a receiver has an array of n_r antennas. The simplest narrowband memoryless channel has the following mathematical description,

$$\mathbf{x} = \sqrt{\rho}\mathbf{G}\mathbf{s} + \mathbf{w}$$

where $\boldsymbol{s}$ is the n_t component vector of transmitted signals, $\boldsymbol{x}$ is the n_r component vector of received signals, $\boldsymbol{G}$ is the $n_r \times n_t$ propagation matrix of complex-valued channel coefficients, and $\boldsymbol{w}$ is the n_r component vector of receiver noise. The components of the additive noise vector are i.i.d. zero mean and unit-variance circular-symmetric complex-Gaussian random variables ($CN(0,1)$). The scalar ρ is a measure of the Signal-to-Noise Ratio (SNR) of the link.

2) Achievable rate

With the assumption that the receiver has perfect knowledge of the channel matrix, $\boldsymbol{G}$, the mutual information between the input and the output of the point-to-point MIMO channel is

$$C = I(x;s) = \log_2 \det\left(\mathbf{I}_{n_r} + \frac{\rho}{n_t}\mathbf{G}\mathbf{G}^H\right)$$

where $\mathbf{I}_{n_r}$ denotes the $n_r \times n_r$ identity matrix. The propagation matrix can be decomposed by

$$\mathbf{G} = \mathbf{\Phi}\mathbf{D}_v\mathbf{\Psi}^H,$$

where $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are unitary matrices of dimension $n_r \times n_r$ and $n_t \times n_t$ respectively, and $\mathbf{D}_\nu$ is a $n_r \times n_t$ diagonal matrix whose diagonal elements are the singular values, $\left\{\nu_1, \nu_2, \cdots, \nu_{\min(n_t, n_r)}\right\}$. The achievable rate can be written as

$$C = \sum_{l=1}^{\min(n_t, n_r)} \log_2\left(1 + \frac{\rho \nu_l^2}{n_t}\right)$$

With the decomposed propagation matrix,

$$\sum_{l=1}^{\min(n_t, n_r)} \nu_l^2 = \mathrm{Tr}\left(\mathbf{G}\mathbf{G}^H\right)$$

where "Tr" denotes "trace". There can exist two extreme cases: the worst case when all except one of the singular values are equal to zero and the best case when all of the $\min\left(n_t, n_r\right)$ singular values are equal. The two cases bound the achievable rate as follows,

$$\log_2\left(1 + \frac{\rho \cdot \mathrm{Tr}\left(\mathbf{G}\mathbf{G}^H\right)}{n_t}\right) \le C \le \min\left(n_t, n_r\right) \cdot \log_2\left(1 + \frac{\rho \cdot \mathrm{Tr}\left(\mathbf{G}\mathbf{G}^H\right)}{n_t \min\left(n_t, n_r\right)}\right)$$

The rank-1 (worst) case occurs either for compact arrays under Line-of-Sight (LOS) propagation conditions such that the transmit array cannot resolve individual elements of the receive array and vice-versa, or under extreme keyhole propagation conditions. The equal singular value (best) case is approached when the entries of the propagation matrix are IID random variables. Under favorable propagation conditions and a high SNR, the achievable rate is proportional to the smaller of the number of transmit and receive antennas.

3) Limiting cases

a) Low SNRs

Low SNRs can be experience by terminals at the edge of a cell. For low SNRs, only beamforming gains are important and the achievable rate becomes

$$C_{\rho \to 0} \approx \frac{\rho \cdot \mathrm{Tr}\left(\mathbf{G}\mathbf{G}^H\right)}{n_t \ln 2}$$

$$\approx \frac{\rho n_r}{\ln 2}$$

which is independent of n_t, and thus, even under the most favorable propagation conditions the multiplexing gains are lost, and multiple transmit antennas are of no value under low SNRs.

b) Number of transmit antennas grow large

It is assumed that the row-vectors of the propagation matrix are asymptotically orthogonal. Then,

$$\left(\frac{\mathbf{G}\mathbf{G}^H}{n_t}\right)_{n_t>>n_r} \approx \mathbf{I}_{n_r}$$

and the achievable rate becomes

$$\begin{aligned} C_{n_t>>n_r} &\approx \log_2 \det\left(\mathbf{I}_{n_r} + \rho\cdot\mathbf{I}_{n_r}\right) \\ &= n_r\cdot\log_2\left(1+\rho\right) \end{aligned}$$

c) Number of receive antennas grow large

It is also assumed that the column-vectors of the propagation matrix are asymptotically orthogonal,

$$\left(\frac{\mathbf{G}^H\mathbf{G}}{n_r}\right)_{n_r>>n_t} \approx \mathbf{I}_{n_t}$$

and the achievable rate becomes

$$\begin{aligned} C_{n_r>>n_t} &= \log_2 \det\left(\mathbf{I}_{n_t} + \frac{\rho}{n_t}\cdot\mathbf{G}^H\mathbf{G}\right) \\ &\approx n_t\cdot\log_2\left(1+\frac{\rho n_r}{n_t}\right) \end{aligned}$$

Thus, a large number of transmit or receive antennas, combined with asymptotic orthogonality of the propagation vectors (i.i.d. complex Gaussian), can increase the achievable rate. Extra receive antennas can compensate for a low SNR and restore multiplexing gains.

B. Multi-user MIMO

Multi-user MIMO consists of an array of *M* antennas and *K* autonomous terminals. We assume that each terminal has only one antenna. Multi-user MIMO differs from point-to-point MIMO in two respects: first, the terminals are typically separated by many wavelengths, and second, the terminals cannot collaborate among themselves.

1) Propagation

We assume Time Division Duplex (TDD), so the reverse link propagation matrix is merely the transpose of the forward link propagation matrix. Assumption on the TDD comes from the need to acquire channel state-information between extreme numbers of service antennas and much smaller numbers of terminals. The propagation matrix, $\mathbf{G} \in \mathbb{R}^{M \times K}$, can be decomposed by

$$\mathbf{G} = \mathbf{H}\mathbf{D}_\beta^{1/2}$$

where $\mathbf{H} \in \mathbb{R}^{M \times K}$ represents small scale fading and $\mathbf{D}_\beta^{1/2} \in \mathbb{R}^{K \times K}$ whose diagonal elements constitute a $K \times 1$ vector, and β is large scale fading coefficients. By assumption, the antenna array is sufficiently compact that all of the propagation paths for a particular terminal are subject to the same large scale fading.

For multi-user MIMO with large arrays, the number of antennas greatly exceeds the number of terminals. Under the most favorable propagation conditions the column-vectors of the propagation matrix are asymptotically orthogonal,

$$\begin{aligned}\left(\frac{\mathbf{G}^H\mathbf{G}}{M}\right)_{M>>K} &= \mathbf{D}_\beta^{1/2}\left(\frac{\mathbf{H}^H\mathbf{H}}{M}\right)_{M>>K}\mathbf{D}_\beta^{1/2} \\ &\approx \mathbf{D}_\beta\end{aligned}$$

2) Reverse link

On the reverse link, for each channel use, the K terminals collectively transmit a $K \times 1$ vector of QAM symbols, $\mathbf{q}_r$, and the antenna array receives a $M \times 1$ vector, $\mathbf{x}_r$,

$$\mathbf{x}_r = \sqrt{\rho_r}\mathbf{G}\mathbf{q}_r + \mathbf{w}_r$$

Under the assumption that the columns of the propagation matrix are nearly orthogonal, i.e., $\mathbf{G}^H\mathbf{G} = M \cdot \mathbf{D}_\beta$, the base station could process its received signal by a matched-filter (MF),

$$\begin{aligned}\mathbf{G}^H\mathbf{x}_r &= \sqrt{\rho_r}\mathbf{G}^H\mathbf{G}\mathbf{q}_r + \mathbf{G}^H\mathbf{w}_r \\ &\approx M\sqrt{\rho_r}\mathbf{D}_\beta\mathbf{q}_r + \mathbf{G}^H\mathbf{w}_r\end{aligned}$$

3) Forward link

For each use of the channel the base station transmits a $M \times 1$ vector, $\mathbf{s}_f$, through its M antennas, and the K terminals collectively receive a $K \times 1$, $\mathbf{x}_f$,

$$\mathbf{x}_f = \sqrt{\rho_f}\mathbf{G}^T\mathbf{s}_f + \mathbf{w}_f$$

III. ANTENNA AND PROPAGATION ASPECTS OF VERY LARGE MIMO

The performance of all types of MIMO systems strongly depends on properties of the antenna arrays and the propagation environment in which the system is operating. With well separated ideal antenna elements, in a sufficiently complex propagation environment and without directivity and mutual coupling, each additional antenna element in the array adds another degree of freedom that can be used by the system. But, in reality, the antenna elements are not ideal, they are not always well separated, and the propagation environment may not be complex enough to offer the large number of degrees of freedom that a large antenna array could exploit. These practical issues are presented in this section.

A. Spatial focus with more antennas

The field strength is not necessarily focused in the direction of the intended receiver, but rather to a geographical point where the incoming multipath components add up constructively. As a technique for focusing transmitted energy to a specific location, Time Reversal (TR) has drawn attention, where the transmitted signal is a time-reversed replica of the channel impulse response. In this paper, the Time-Reversal Beam Forming (TRBF) is considered.

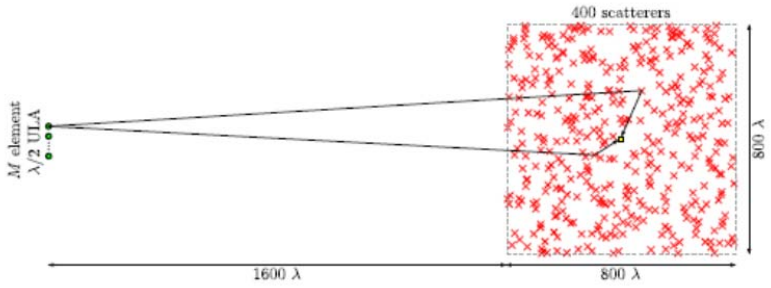


Figure 1. Geometry of the simulated dense scattering environment.

Figure 1 shows a simple geometrical channel model. The channel is composed of 400 uniformly distributed scatterers in a square of dimension $800\lambda \times 800\lambda$, where λ is the signal wavelength. The broadside direction of the *M*-element Uniform Linear Array (ULA) with adjacent element spacing $d = \lambda / 2$ is pointing towards the center of the scatterer area. This model creates a field strength that varies rapidly over the geographical area, typical of

small-scale fading. With a complex enough scattering environment and a sufficiently large element spacing in the transmit array, the filed strength resulting from different elements in the transmit array can be seen as independent.

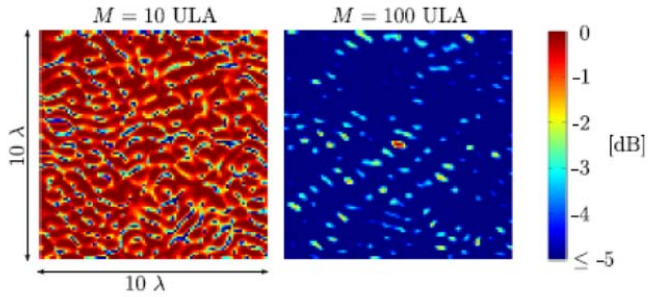
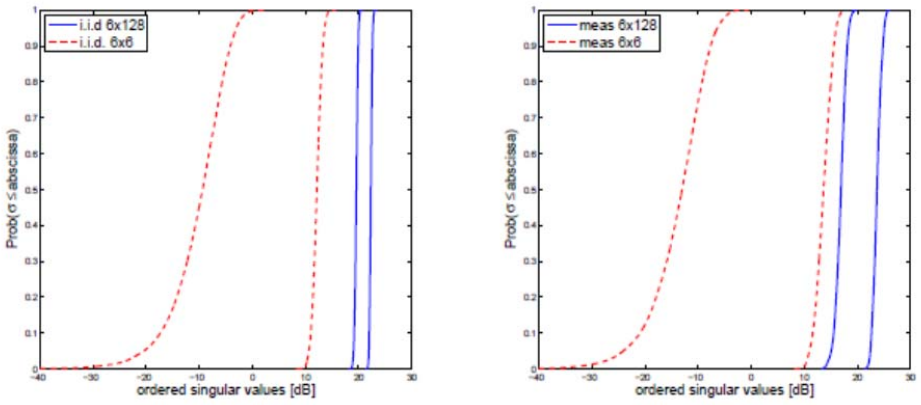


Figure 2. Normalized filed strength in a $10\lambda \times 10\lambda$ area

Figure 2 shows the resulting normalized field strength in a small $10\lambda \times 10\lambda$ environment around the receiver to which we focus the transmitted signal (using MF precoding), for ULAs with $d = \lambda/2$ of size $M = 10$ and $M = 100$ elements. Figure 2 illustrates two important properties of the spatial MF precoding: (i) that the field strength can be focused to a point rather than in a certain direction and (ii) that more antennas improve the ability to focus energy to a certain point, which leads to less interference between spatially separated users.

B. Antenna aspects

Massive MIMO relies to a large extent on a property of the radio environment called *favorable propagation*. Favorable propagation means that propagation channel responses from the base station to different terminals are sufficiently different. One way of quantifying how different the channel responses to different terminals are, is to look at the spread between the smallest and largest singular value of the channel. Figure 3(a) shows this for a computer simulated "i.i.d." channel. The figure shows the cumulative density function for the smallest respectively the largest singular value for two cases: A conventional array of 6 elements serving 6 terminals (red curves), and a massive array of 128 elements serving 6 terminals (blue curves)



(a) Theory. (b) Measured results.

Figure 3. Singular value spread of massive MIMO channels

For the 6-element array, the singular value spread is about 30 dB, meaning that 1000 times more power would be required to serve all six terminals, as compared to the power required to serve just one of them. With the massive array, the gap is less than 3 dB.

In real channel implementation, measurements were conducted using an indoor 128-antenna base station consisting of four stacked double polarized 16 element circular patch arrays as shown in Fig. 4. Three of the terminals are indoors at various positions and 3 users are outdoors. The measurements were performed at 2.6 GHz with a bandwidth of 50 MHz, and the results were averaged over this bandwidth and over a physical displacement of 10 meters.

Figure 4. Massive MIMO antenna used in measurements

The blue curves in Fig 3. (b) show the corresponding singular value distributions. It is striking how well reality resembles the ideal case in Fig. 3. (a). The spread between the smallest and the largest singular value is a bit larger than for the ideal case, but the probability that the spread

exceeds 10 dB is negligible. As a reference, Fig. 3(b) also shows the result when only 6 of the 128 elements are activates (red curves). Overall, there is compelling evidence that the assumption on favorable propagation that underpin massive MIMO are substantially valid in practice.

IV. DISCUSSION

After meeting, please write discussion in the meeting and update your presentation file.

In the antenna implementation, they compared system of 6-antenna array and 6 terminals with single antenna and system of 128-antenna array and 6 terminals with single antenna. They used $f_c[\text{cycle}/\text{sec}] = 2.6\times10^9 \approx 3.0\times10^9$, $c\left[\text{m}/\text{sec}\right] = 3.0\times10^8$. Thus, the wavelength $\lambda\left[\text{m}/\text{cycle}\right] = \frac{c}{f_c} = 0.1$. According to the wavelength, if they used $d = \lambda/2$, then the distance between patched antennas is $d = \lambda/2 = 50\text{cm}$.

Appendix

Reference

[1]

Shrinkage Methods to Sparse Linear Regression

Main ref.:T. Hastie, R. Tibshirani, and J. Friedman,

The Elements of Statistical Learning 2nd Edition, Springer 2008 [1]

Presenter: Jaewook Kang

CS Journal Club in GIST, Apr. 2013

This report will be appeared in Ph.D. dissertation of Jaewook Kang

Abstract

In this report, we introduce linear regression approaches using shrinkage methods. The shrinkage method have got attention to solve the problem of linear systems $\mathbf{y} = \mathbf{A}\mathbf{x}$ because the method enables us to obtain the solution with lower variance than the conventional least square estimator having the minimum variance unbiasedness. First we will introduce basic concept of two shrinkage methods in the linear regression, *ridge and lasso*. Then, we move our focus to problems of the Lasso variants such as *Fused lasso* and *Elastic-net*. For the discussion in this report, we have partially referred to the chapter 3 of the book [1].

I. Introduction

A linear regression problem starts from an assumption that the corresponding regression function $\underline{Y} = f(\underline{X})$ is linear where $\underline{Y} \in \mathbb{R}^M$ is a measurement vector generated by the function $f(\cdot)$ given a vector $\underline{X} \in \mathbb{R}^N$. This assumption allows us to describe the function $f(\cdot)$ using a linear projection, given by

$$\underline{Y} = \mathbf{A}\underline{X} \in \mathbb{R}^M, \tag{1}$$

where a measurement matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ specifies the linear relation between $\underline{Y}$ and $\underline{X}$. In such a regression problem, a typical aim is to estimate the unknown vector $\underline{X}$ from a set of known inputs or training data $(Y_1, \underline{a}_{\text{1st-row}})$ ... $(Y_M, \underline{a}_{\text{Mst-row}})$ where $\underline{a}_{\text{jth-row}} = [a_{j1}, a_{j2}, ..., a_{jN}]$ denotes the j-th row vector of the matrix $\mathbf{A}$. In addition, as before, we confine our focus to the linear regression problems which is underdetermined $(M < N)$ such that there exists infinitely many solutions for $\underline{X}$.

The most standard approach to the linear regression problems is *least square estimation* (LSE). LSE obtains its estimate by solving the following optimization problem, given by

$$\begin{aligned}(P_{\text{LSE}}) : \underline{\widehat{X}}_{\text{LSE}} &= \arg\min_{\underline{X}} \|\underline{Y} - \mathbf{A}\underline{X}\|_2^2 \\ &= \arg\min_{\underline{X}} \sum_{j=1}^{M} \left(Y_j - \sum_{i=1}^{N} a_{ji} X_i \right)^2 . \end{aligned} \tag{2}$$

It is well known that the LSE solver obtains an estimate $\underline{\widehat{X}}_{\text{LSE}}$ by projecting the measurement vector $\underline{Y}$ to a subspace of $\mathbb{R}^M$ spanned by the column vectors of the matrix $\mathbf{A}$. Namely, the minimization task in (2) chooses $\underline{\widehat{X}}_{\text{LSE}}$ which makes the vector difference $\underline{Y} - \mathbf{A}\underline{\widehat{X}}_{\text{LSE}}$ to be orthogonal to the subspace. Such a LSE solution can be represented as a linear function of the measurement vector, *i.e.*,

$$\underline{\widehat{X}}_{LSE} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\underline{Y}. \tag{3}$$

The popularity of LSE is originated from the Gauss-Markov theorem, one of the famous results in statistics. The Gauss-Markov theorem states that the LSE solver provides the smallest variance among all linear unbiased estimators. Let $\underline{\widetilde{X}}$ denote an unbiased linear estimate, *i.e.*, $E\left[\underline{\widetilde{X}}\right] = \underline{X}$. The mean squared error (MSE) of $\underline{\widetilde{X}}$ is calculated as

$$\begin{aligned}\text{MSE}\left(\underline{\widetilde{X}}\right) :=& E\left[\left(\underline{\widehat{X}} - \underline{X}\right)^2\right] = \text{Var}\left(\underline{\widetilde{X}}\right) + \left(E\left[\underline{\widehat{X}}\right] - \underline{X}\right)^2 \\ =& \text{Var}\left(\underline{\widetilde{X}}\right). \end{aligned} \tag{4}$$

Then, the Gauss-Markov theorem shows that

$$\text{Var}\left(\underline{\widehat{X}}_{LSE}\right) \leq \text{Var}\left(\underline{\widetilde{X}}\right) \tag{5}$$

for any other unbiased linear estimate $\underline{\widetilde{X}}$ (We omit the proof here. please refer to [1]).

However, there may exist a biased estimator which can offer smaller MSE then the LSE solver. Such an estimator would provide a significant reduction in MSE at the expense of losing the unbiasedness [1],[3]. This is one motivation to use the shrinkage method to linear regression problem. The shrinkage methods is a biased estimation approach to impose a penalty to the optimization setting of (2). If the imposed penalty can properly catch the characteristic of the target unknown $\underline{X}$, the shrinkage methods greatly improve the estimation accuracy. In this report, we first introduce two types of the most well-known shrinkage methods, *ridge* and *lasso*, by partially referring to [1],[3]. Then, we extend our discussion to shrinkage methods which estimates the vector $\underline{X}$ having piecewise smooth or approximately sparse property.

II. SHRINKAGE WITH RIDGE PENALTY

In ridge regression, the elements of $\underline{X}$ are shrunk by imposing a penalty on the energy of $\underline{X}$ [2]. Therefore, the ridge penalty takes a quadratic form of $\underline{X}$, leading to the following optimization setup

$$\begin{aligned}(P_{\text{Ridge}}) : \underline{\widehat{X}}_{\text{Ridge}} =& \arg\min_{\underline{X}} \|\underline{Y} - \mathbf{A}\underline{X}\|_2^2 + \lambda \|\underline{X}\|_2^2 \\ =& \sum_{j=1}^{M} \left(y_j - \sum_{i=1}^{N} a_{ji} X_i \right)^2 + \lambda \sum_{i=1}^{N} X_i^2, \end{aligned} \tag{6}$$

where $\lambda \geq 0$ denote a parameter to control the amount of ridge shrinkage. Note that by applying the quadratic penalty $\|\underline{X}\|_2^2 = \underline{X}^T \underline{X}$, the ridge estimation can be represented as a closed form function of $\underline{Y}$ if $M \geq N$, given by

$$\underline{\widehat{X}}_{\text{Ridge}} = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}_N)^{-1} \mathbf{A}^T \underline{Y}. \tag{7}$$

This imposition of the ridge penalty adds a positive constant to the diagonal $\mathbf{A}^T\mathbf{A}$ in (7) before inversion. It is noteworthy that this addition makes the regression problem nonsingular even when $\mathbf{A}^T\mathbf{A}$ does not have full rank. Namely, the ridge solution is necessarily unique regardless of the condition of the matrix $\mathbf{A}$. This is a strong motivation to use ridge regression.

Ridge regression shrinks the coordinate of $\underline{\widehat{X}}_{\text{Ridge}}$ according to the singular value of the matrix $\mathbf{A}$. The singular value decomposition (SVD) of $\mathbf{A}$ has the form

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{8}$$

where $\mathbf{U} \in \mathbb{R}^{M \times N}$ and $\mathbf{V} \in \mathbb{R}^{N \times N}$ are orthogonal matrices, and $\mathbf{D} \in \mathbb{R}^{N \times N}$ is a diagonal matrix with singular values $d_1 \geq d_2 \geq ... \geq d_N \geq 0$ of $\mathbf{A}$. By applying SVD to the ridge solution, we can efficiently compute a ridge estimate $\underline{\widehat{X}}_{\text{Ridge}}$ associated with the orthonormal basis $\mathbf{U}$ and $\mathbf{V}$, as LSE does using the QR decomposition

$$\begin{aligned}\underline{\widehat{X}}_{\text{Ridge}} =& (\mathbf{A}^T\mathbf{A} + \lambda \mathbf{I}_N)^{-1} \mathbf{A}^T \underline{Y} \\ =& \mathbf{V}(\mathbf{D}^2 + \lambda \mathbf{I}_N)^{-1} \mathbf{D}\mathbf{U}^T \underline{Y} \\ =& \sum_{i=1}^{N} \underline{v}_i \frac{d_i}{d_i^2 + \lambda} \underline{u}_i^T \underline{Y}, \end{aligned} \tag{9}$$

where the $\underline{u}_i \in \mathbb{R}^M$ and $\underline{v}_i \in \mathbb{R}^N$ are the column vectors of $\mathbf{U}$ and $\mathbf{V}$ respectively. In (9), ridge regression shrinks the elements of $\underline{\widehat{X}}_{\text{Ridge}}$ by the factors $d_i/(d_i^2 + \lambda)$. This means that a greater amount of shrinkage is applied to the elements of $\underline{\widehat{X}}_{\text{Ridge}}$ associated with $\underline{v}_i$ having smaller singular values d_i. Namely, ridge

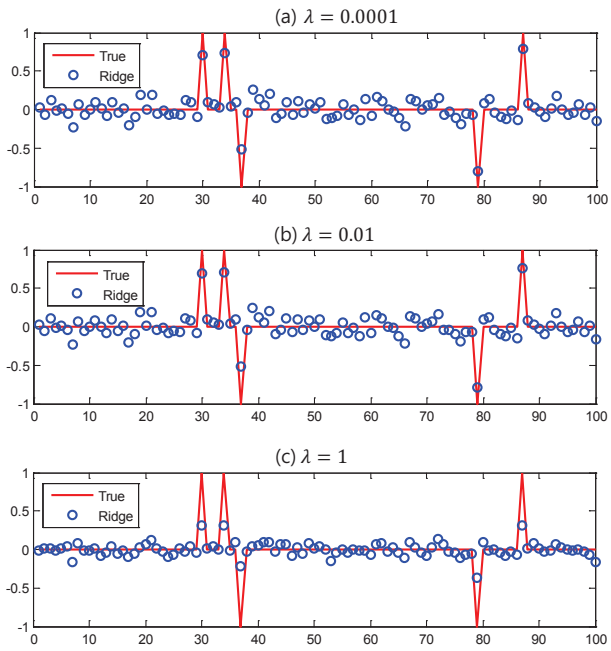


Fig. 1. Sparse estimation via ridge estimator with different λ where $N = 100, M = 70, K = 5$

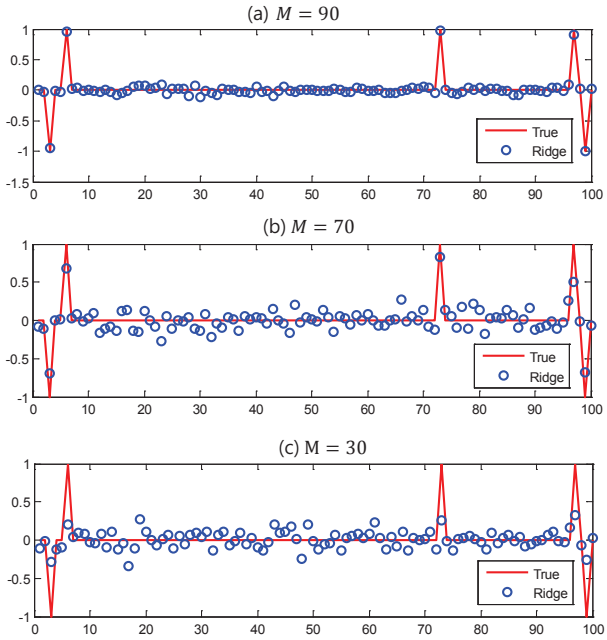


Fig. 2. Sparse estimation via ridge estimator with different M where $N = 100, K = 5, \lambda = 0.001$

regression shrinks together the correlated elements of $\underline{X}$ with respect to $\underline{v}_i$ if the direction of $\underline{v}_i$ has small energy in the column space of $\mathbf{A}$.

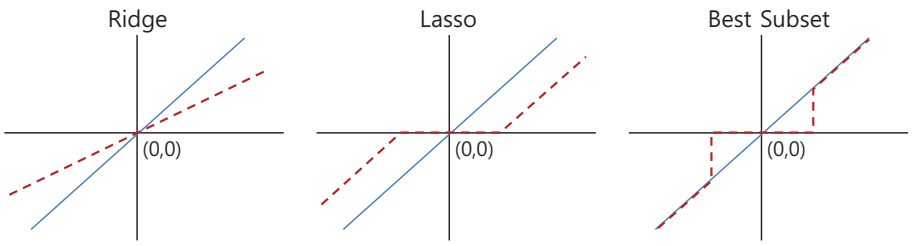


Fig. 3. Shrinkage characteristic of Ridge, Lasso and Best subset selection where the orthonomal matrix $\mathbf{A}$ is assumed. In addition, the blue solid line in the figure is the 45° line to show the LSE solution as a reference (The figure is borrowed from Table 3.4 of [1]).

The ridge penalty can be used to estimate sparse vectors $\underline{X} \in \mathbb{R}^N$ in undetermined systems $\underline{Y} = \mathbf{A}\underline{X} \in \mathbb{R}^M$. In order to apply the expression of (7), we need an augmented matrix $\mathbf{A}' \in \mathbb{R}^{N\times N}$ which additionally includes $N - M$ zero rows from $\mathbf{A} \in \mathbb{R}^{M\times N}$. Let us consider sparse vectors $\underline{X}$ which contains K nonzero signed elements having unit magnitude. The ridge penalty shrinks the elements of $\underline{X}$ with respect to non-principal basis of $\mathbf{A}'$. Hence, the ridge regression enables the K largest elements, which are most related to the principal basis of $\mathbf{A}'$, to have exceptionally large magnitude.

We examine the ridge regression on the parameter $N = 100, K = 5$ with standard Gaussian matrix $a_{ji} \in \mathbf{A} \sim \mathcal{N}(0, 1/M)$. Fig.1 shows that the ridge estimation can finds the K largest elements of $\underline{X}$ with appropriately chosen λ. Another example is shown in Fig.2 where we show the behavior of ridge regression according to the number of M. We note in Fig.2 that the magnitude of the K largest elements of $\widehat{\underline{X}}$ becomes smaller as M decreases. This means that for clear distinction of the K largest elements, the ridge method requires M close to N. In addition, we know from Fig.1 and Fig.2 that the ridge solver cannot exactly fit the nonzero elements of $\underline{X}$.

III. Shrinkage with Lasso Penalty

The main characteristic of lasso is that the elements of $\underline{X}$ are shrunk by imposing a L_1-norm penalty of $\underline{X}$ [3]. Namely, the penalty in the lasso setup takes an absolute sum of $\underline{X}$, *i.e.*, $\|\underline{X}\|_1 = \sum_{i=1}^{N} |X_i|$. Following that, its optimization setup is represented as

$$\begin{aligned}(P_{\text{Lasso}}) : \widehat{\underline{X}}_{\text{Lasso}} &= \arg\min_{\underline{X}} \|\underline{Y} - \mathbf{A}\underline{X}\|_2^2 + \lambda \|\underline{X}\|_1 \\ &= \sum_{j=1}^{M} \left(y_j - \sum_{i=1}^{N} a_{ji} X_i \right)^2 + \lambda \sum_{i=1}^{N} |X_i|, \end{aligned} \tag{10}$$

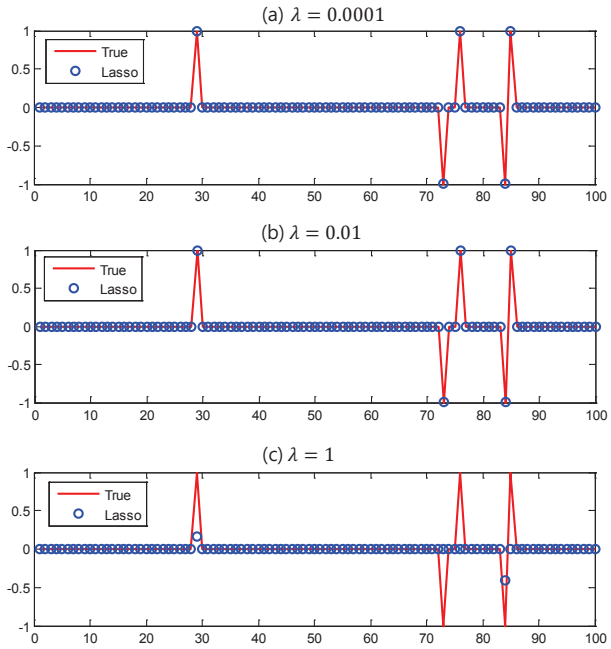


Fig. 4. Sparse estimation via lasso estimator with different λ where $N = 100, M = 70, K = 5$

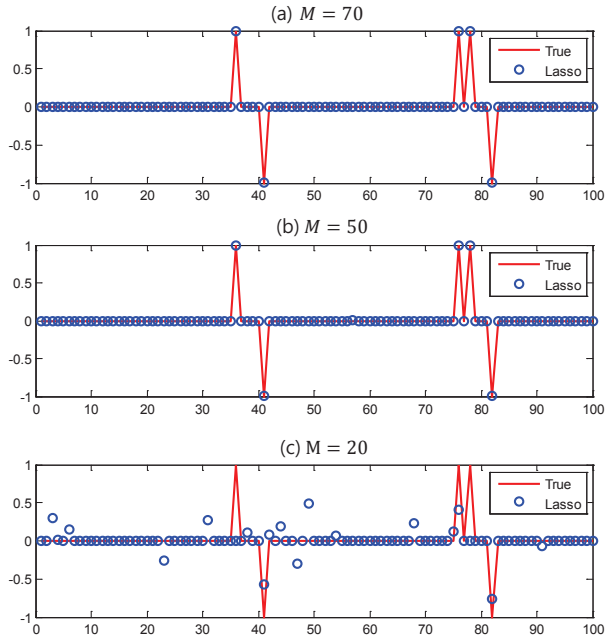


Fig. 5. Sparse estimation via lasso estimator with different M where $N = 100, K = 5, \lambda = 0.001$

where λ is a parameter to control the amount of lasso shrinkage. The larger λ leads to the stronger shrinkage against the elements of $\underline{X}$. When $\lambda = 0$ the solution is equivalent to the LSE solution. This L_1 penalty generates the solutions of (10) nonlinear with respect to the measurement vector $\underline{Y}$; therefore, there is no closed form solver as in ridge regression. The lasso solver can be implemented via a quadratic programming. In addition, the *LARs* algorithm is well known as a computationally efficient algorithm for the lasso solution [4].

To see the detail of the lasso behavior, we investigate the *Karush-Kuhn-Tucker* (KKT) condition with the Lagrangian $\mathcal{L}(\underline{X}, \lambda)$ of the setup in (10).

$$\begin{aligned}
&1) \text{ Stationarity: } \nabla_{\underline{X}} \mathcal{L}(\underline{X}, \lambda) = \mathbf{G}\underline{X} - \mathbf{A}^T\underline{Y} + \lambda \underline{B} = 0, \\
&2) \text{ Dual feasibility: } \lambda \geq 0, \\
&3) \text{ Prime feasibility: } \nabla_{\lambda} \mathcal{L}(\underline{X}, \lambda) = \|\underline{X}\|_1 \leq 0, \\
&4) \text{ Complementary slackness for strong duality: } \lambda \|\underline{X}\|_1 = 0,
\end{aligned} \tag{11}$$

where we define a Gram matrix $\mathbf{G} := \mathbf{A}^T\mathbf{A}$ and

$$\underline{B} := \nabla_{\underline{X}} \|\underline{X}\|_1 = \left[\frac{\partial \sum |X_i|}{\partial X_1}, \frac{\partial \sum |X_i|}{\partial X_2}, ..., \frac{\partial \sum |X_i|}{\partial X_N} \right]. \tag{12}$$

Since $\sum |X_i|$ is not differentiable, we apply the concept of sub-differential to $\frac{\partial \sum |X_i|}{\partial X_1}$. Then, each element of $\underline{B}$ is given by

$$B_i = \frac{\partial \sum |X_i|}{\partial X_1} := \begin{cases} \text{sign}(X_i) & \text{if } |X_i| \geq \lambda \\ B_i \in [-1, 1] & \text{if } |X_i| < \lambda \end{cases}. \tag{13}$$

We note the stationarity condition in (11), which can be rewritten as

$$\mathbf{A}^T\underline{Y} - \lambda \underline{B} = \mathbf{G}\underline{X}. \tag{14}$$

Insight about the lasso shrinkage can be obtained by assuming that the matrix $\mathbf{A}$ is orthonomal, *i.e.*, $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. By applying the orthonomal assumption to (14), we have

$$\underline{\widehat{X}}_{\text{Lasso}} = \mathbf{A}^T\underline{Y} - \lambda \underline{B}. \tag{15}$$

Then, the expression in (15) can be represented by a soft thresholding function with the parameter λ [5], *i.e.*,

$$\underline{\widehat{X}}_{\text{Lasso}} = \eta\left(\mathbf{A}^T\underline{Y}; \lambda\right), \tag{16}$$

where the thresholding function $\eta(\tau;\lambda)$ is elementwisely defined as

$$\eta\left(\tau_i;\lambda\right) = \begin{cases} \tau_i - \lambda & \text{if } \tau \geq \lambda_i, \\ \tau_i + \lambda & \text{if } \tau \leq -\lambda_i, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

We know from (17) that lasso shrinks the elements of $\underline{X}$ according to their magnitude. For the comparison purpose, we also consider ridge estimate with the orthonomal matrix $\mathbf{A}$, given by

$$\underline{\widehat{X}}_{\text{Ridge}} = \frac{1}{1+\lambda}\mathbf{A}^T\underline{Y}. \tag{18}$$

Differently from the lasso case, the ridge estimate is obtained with a proportional shrinkage $\frac{1}{1+\lambda}$ (radically the proportional shrinkage of ridge is determined by singular values of $\mathbf{A}$). We borrow Fig.3 from the reference book (the figure in Table 3.4 of [1]) to depict the shrinkage characteristic of ridge and lasso, compared to the best subset selection which is an optimal estimator to find the $K(\leq M)$ largest elements of $\underline{X} \in \mathbb{R}^N$. Fig.3 explicitly shows the difference among those three estimators.

We examine the lasso solver to estimate the signed K-sparse vectors $\underline{X} \in \mathbb{R}^N$ from the undetermined system $\underline{Y} = \mathbf{A}\underline{X} \in \mathbb{R}^M$, as in the ridge regression. Fig.4 shows that the lasso solver perfectly finds the K largest elements with appropriate λ. We note in Fig.4 that the lasso estimate of the case $\lambda = 1$ does not fit to the true of $\underline{X}$ because in this case, the lasso penalty shrinks the elements too much. Fig.5 shows the lasso recovery of $\underline{X}$ for a variety of the number of measurements M. In the figure, we see that the lasso solver finds an accurate solution when $M = 50, 70$, but fails in the estimation when $M = 20$.

IV. Variants of Lasso

A. Elastic-Net for Approximately Sparse Signal

We can generalize the ridge and the lasso penalty by using the concept of L_p-norm, *i.e.*,

$$(P_{L_p}) : \underline{\widehat{X}}_{L_p} = \arg\min_{\underline{X}} \sum_{j=1}^{M}\left(y_j - \sum_{i=1}^{N} a_{ji}X_i\right)^2 + \lambda\sum_{i=1}^{N}|X_i|^p, \tag{19}$$

for $p \geq 0$, where the case $p = 0$ corresponds to the best subset selection which is non-convex; $p = 2$ corresponds to ridge regression which is convex; $p = 1$ is the lasso case which has the smallest p such that the problem is convex. Value of $p \in (1, 2)$ suggests a compromise between the lasso and ridge regression. If p is closer to 1, the solver has the ability to put small elements close to zero which is the nature of the lasso solver, If p is closer to 2, the solver more tends to shrink signal elements associated with the singular values of $\mathbf{A}$ which is the nature of ridge regression.

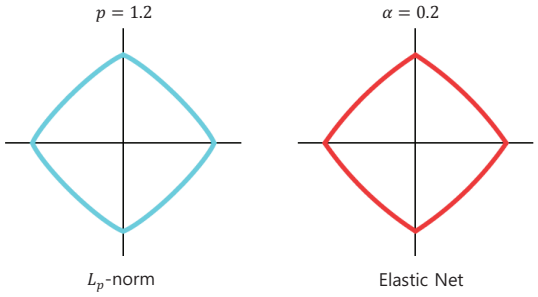


Fig. 6. Contours of the L_p penalty for $p = 1.2$ (left plot) and the elastic-net penalty ($\alpha = 0.2$) (right plot) (The figure is borrowed from Figure 3.13 of [1].)

Elastic-net, proposed by Zou and Hastie, introduced a different compromise between ridge and lasso [7]. The elastic-net selects the largest elements like lasso, and shrinks the remaining small elements like ridge, using a mixture penalty. Therefore, the elastic-net solver is useful for approximately sparse signals whose small elements are originally not exactly zero. The optimization setting of the elastic-net solver is given by

$$\begin{aligned}(\mathrm{P_{EN}}): \ \underline{\widehat{X}}_{EN} = \arg\min_{\underline{X}} \|\underline{Y} - \mathbf{A}\underline{X}\|_2^2 \\ + \lambda \left(\alpha \|\underline{X}\|_2^2 + (1-\alpha)\|\underline{X}\|_1\right),\end{aligned} \tag{20}$$

where α is a mixing rate of the mixture penalty. We borrow Fig.6 from the book (Figure 3.13 of [1]). This figure compares contours of the L_p norm penalty with $p = 1.2$ and the mixture penalty with $\alpha = 0.2$. It is very difficult to distinguish those two penalties by eyes. Although those two are visually very similar, there exists a fundamental difference. The elastic-net has sharp (non-differentiable) corners such that it can put the elements exactly zero, whereas the L_p penalty does not [7]. Likewise with lasso, the elastic-net can be solved via quadratic programming, and the *LARs-EN* algorithm was introduced as a LARs type algorithm to solve the elastic-net problem by Zou and Hastie [7]

We compare the elastic-net solver to the lasso solver in Fig.7 where the problem size is $N = 100, M = 70$. For the comparison, we test an approximately sparse signal generated from *i.i.d* two-state Gaussian mixture density, *i.e.*,

$$f_{\underline{X}}(\underline{x}) = \prod_{i=1}^{N} q\mathcal{N}(x_i; 0, \sigma_{X_1}^2) + (1-q)\mathcal{N}(x_i; 0, \sigma_{X_0}^2), \tag{21}$$

with $q = 0.07, \sigma_{X_1} = 0.05, \sigma_{X_1} = 1$. We set the elastic-net parameters $\alpha = 0.4$, and we use $\lambda = 0.001$ for the lasso and elastic-net both. In Fig.7, the elastic-net with appropriately calibrated parameters α, λ

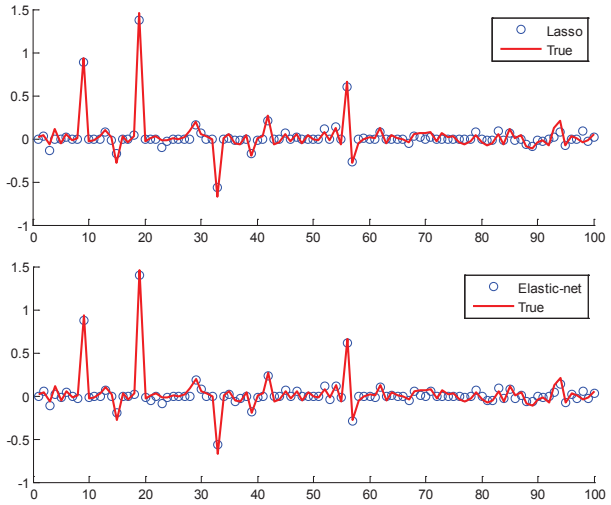


Fig. 7. Approximately sparse signal estimation ($N = 100, M = 70, \lambda = 0.001$) via lasso (Upper plot), and via the elastic-net (bottom plot) where the MSE of lasso estimate is 0.0509, while that of the elastin-net is 0.0359 in this example.

surely improves the estimation accuracy from lasso although it is very hard to be distinguished by eyes. Indeed, the MSE of lasso estimate is 0.0509, while that of the elastin-net is 0.0359 in this example.

B. Fused Lasso for Piecewise Smooth Signals

The use of various types of penalties enables us to solve the $\underline{Y} = \mathbf{A}\underline{X}$ problem adaptively to the characteristic of the signal $\underline{X}$. The *fused lasso* is one of such solvers to find piecewise smooth signals. The fused lasso solves the problem given by

$$\begin{aligned}(\mathrm{P_{FL}}): \ \underline{\widehat{X}}_{FL} = \arg\min_{\underline{X}} &\|\underline{Y} - \mathbf{A}\underline{X}\|_2^2 \\ &+ \lambda\left(\alpha\sum_{i=2}^{N}|X_i - X_{i-1}| + (1-\alpha)\|\underline{X}\|_1\right),\end{aligned} \tag{22}$$

where the difference penalty, $\sum_{i=2}^{N}|X_i - X_{i-1}|$, enforces the estimate $\underline{\widehat{X}}_{FL}$ to be piecewise smooth by considering the order of the features. Namely, the fuse lasso encourages both sparsity of the signal values and sparsity of difference between adjacent elements. Fig.8 shows contour plot of the fused lasso penalty compared to that of the lasso penalty. As shown in Fig.8, the fused lasso penalty has asymmetric contour owing the difference penalty, and it becomes severe as α increases. This asymmetricity of the fused lasso encourages the smoothness of the signal.

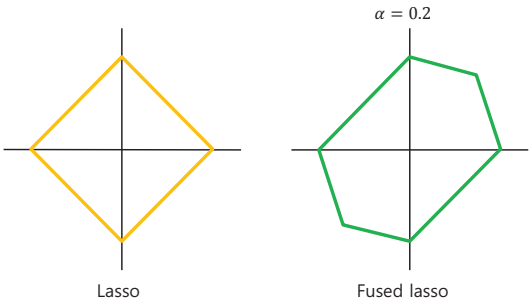


Fig. 8. Contours of the Lasso penalty (left plot) and the fuse lasso penalty ($\alpha = 0.2$) (right plot)

We show an example of the piecewise smooth signal recovery using a fused lasso solver in Fig.9, where measurements $\underline{Y}$ is generated from a piecewise smooth signal $\underline{X}$ with $N = 100, M = 50$ using a standard Gaussian matrix $\mathbf{A}$. This example shows that the piecewise smooth signal can be recovered via the fused lasso as α increases although the signal $\underline{X}$ itself is not sparse. We also checked that the signal can be recovered even from $M = 25$ measurements when $\alpha = 0.9$. In the figure, the case of $\alpha = 0$ is noteworthy because the case is equivalent to the conventional lasso case. This case informs us that such a piecewise smooth recovery is not successful via the normal lasso solver.

V. Conclusive Remarks

We have discussed about shrinkage method to solve the linear system $\underline{Y} = \mathbf{A}\underline{X}$. Estimation through such a method has smaller MSE than LSE at the expense of losing the unbiasedness. Ridge regression is one of the shrinkage methods applying a penalty on the energy of $\underline{X}$. This ridge penalty makes the solver to shrinks together the correlated elements of $\underline{X}$ with respect to the matrix $\mathbf{A}$. Estimation accuracy of ridge is not satisfied for the K sparse signal estimation because the ridge solver cannot exactly fit the nonzero elements of $\underline{X}$. We also have introduced the lasso solver which imposes L_1-norm penalty of $\underline{X}$. The lasso solver shrinks the elements of $\underline{X}$ according to their magnitude, performing the shrinkage as a soft thresholding function. The estimation accuracy of lasso is very good for K sparse signals by putting the nonzero elements of $\underline{X}$ exactly to zero. Elastic-net solver is a compromise of ridge and lasso using a mixture penalty. This solver is useful for approximately sparse signals whose small elements are not exactly zero. The fused lasso solver was devised to find piecewise smooth signals. Imposing of difference penalty, which reflects the order of signal features, enables us to estimate the piecewise smooth signal effectively.

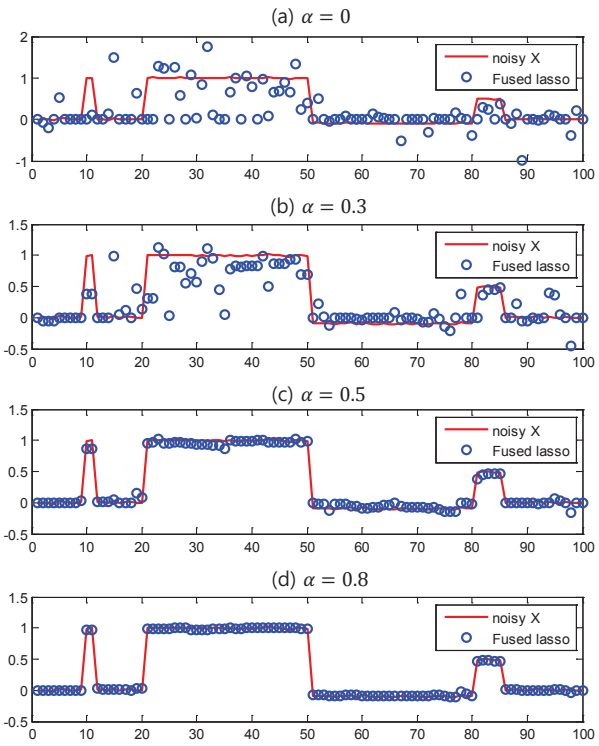


Fig. 9. Piecewise smooth signal estimation via fused lasso for a variety of α when $N = 100, M = 50, \lambda = 0.01$

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning 2nd Edition,* Springer 2008

[2] E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, issue 1, 1970.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statisti. Soc., Ser. B*, vol. 58, no. 1, pp. 267-288, 1996.

[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407-499, 2004.

[5] D. L. Donoho, "De-nosing by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 613-627, May. 1995.

[6] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso, *J. R. Statist. Soc. Ser. B*, vol.67, pp. 91-108, 2005.

[7] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Statist. Soc. Ser. B*, vol. 67 pp. 301-320, 2005.