Computational Statistics and Data Analysis 94 (2016) 221–237

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

On stepwise pattern recovery of the fused Lasso

Junyang Qian, Jinzhu Jia *

School of Mathematical Sciences and Center for Statistical Science, LMAM, LMEQF, Peking University, China

HIGHLIGHTS

- We provided necessary and sufficient conditions such that fused Lasso consistently recovers the piecewise constant pattern.
- We found that in general the fused Lasso is not consistent.
- We proposed a preconditioned fused Lasso to overcome the non-consistent issue.
- Simulation studies support our findings.

ARTICLE INFO

Article history:
Received 8 October 2014
Received in revised form 29 June 2015
Accepted 20 August 2015
Available online 31 August 2015

Keywords:
Fused Lasso
Non-asymptotic
Pattern recovery
Preconditioning

ABSTRACT

We study the property of the Fused Lasso Signal Approximator (FLSA) for estimating a blocky signal sequence with additive noise. We transform the FLSA to an ordinary Lasso problem, and find that in general the resulting design matrix does not satisfy the irrepresentable condition that is known as an almost necessary and sufficient condition for exact pattern recovery. We give necessary and sufficient conditions on the expected signal pattern such that the irrepresentable condition holds in the transformed Lasso problem. However, these conditions turn out to be very restrictive. We apply the newly developed preconditioning method — Puffer Transformation (Jia and Rohe, 2015) to the transformed Lasso and call the new procedure the *preconditioned fused Lasso*. We give non-asymptotic results for this method, showing that as long as the signal-to-noise ratio is not too small, our preconditioned fused Lasso estimator always recovers the correct pattern with high probability. Theoretical results give insight into what controls the ability of recovering the pattern — it is the noise level instead of the length of the signal sequence. Simulations further confirm our theorems and visualize the significant improvement of the preconditioned fused Lasso estimator over the vanilla FLSA in exact pattern recovery.

1. Introduction

Assume we have a sequence of signals $(y_1, y_2, \ldots, y_n)$ and it follows the additive model

$$y_i = \mu_i^* + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is the observed signal vector, $\mu^* = (\mu_1, \ldots, \mu_n)^T \in \mathbb{R}^n$ the expected signal vector, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ the white noise such that $\epsilon_1, \ldots, \epsilon_n$ are assumed to be i.i.d. Gaussian random variables with mean 0 and variance σ^2. The model is assumed to be blocky in the sense that the signals come in blocks and have only a few change-points. To be exact, there exists a partition of $\{1, 2, \ldots, n\} = \cup_{j=1}^{J}\{L_j, L_j + 1, \ldots, U_j\}$ with $L_1 = 1, U_J = n, U_j \geq L_j, L_{j+1}$

* Correspondence to: 5 Summer Palace Road, School of Mathematical Sciences, Peking University, Beijing, 100871, China.
E-mail address: jzjia@math.pku.edu.cn (J. Jia).

http://dx.doi.org/10.1016/j.csda.2015.08.013
0167-9473/

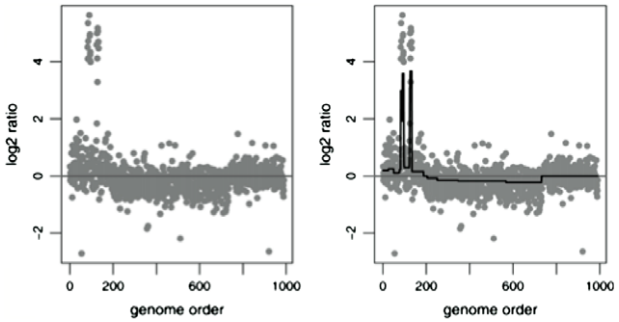

Fig. 1. This figure is from Tibshirani and Wang (2008). The fused Lasso is applied to some CGH data. The data are shown in the left panel, and the solid line in the right panel represents the estimated signals by the fused Lasso. The horizontal line is for $y = 0$.

$= U_j + 1$, and the following stepwise function holds:

$$\mu_i^* = \sum_{j=1}^{J} \nu_j^* 1_{L_j \le i \le U_j},$$

with ν_j^*, L_j, U_j fixed but unknown. We also assume that the vector $\nu = (\nu_1, \nu_2, \ldots, \nu_J)$ is sparse, meaning that only a few of ν_j's are nonzeros. We point out that the Gaussian noise is not necessary, but we still use it to get insight of the fused Lasso. The variance σ^2 of ϵ_i is the measure of noise level and does not have to be a constant here. In many cases, each observation of y_i can be an average of multiple measurements and so σ^2 decreases when the number of measurements increases. Rinaldo (2009) considers the model when $\sigma^2 = \sigma_0^2/n$, where σ_0 is a constant. We do not make this specific assumption in the development of our theory.

Featured by blockiness and sparseness, this model has many applications. For example, in tumor studies, based on the Comparative Genomic Hybridization (CGH) data, it can be used to automatically detect the gains and losses in DNA copies by taking the "signal" above as the log-ratio between the number of DNA copies in tumor cells and that in reference cells (Tibshirani and Wang, 2008). For more applications, see Tibshirani and Taylor (2011), Friedman et al. (2007) and Hoefling (2010).

One way to estimate the unknown parameters is via the Fused Lasso Signal Approximator (FLSA) defined as follows (Tibshirani et al., 2004; Friedman et al., 2007):

$$\hat{\mu}(\lambda_1, \lambda_2) = \underset{\mu}{\operatorname{argmin}} \frac{1}{2}\|Y - \mu\|_2^2 + \lambda_1\|\mu\|_1 + \lambda_2\|\mu\|_{TV}, \tag{2}$$

where $\|\mu\|_1 = \sum_{i=1}^n |\mu_i|$, $\|\mu\|_2^2 = \sum_{i=1}^n \mu_i^2$ and $\|\mu\|_{TV} = \sum_{i=1}^{n-1} |\mu_{i+1} - \mu_i|$. The L_1-norm regularization controls the sparsity (number of zeros) and the total variation seminorm ($\|\mu\|_{TV}$) regularization controls the blockiness (number of blocks or partitions).

Fig. 1 gives some CGH data, a typical example of signals with such features and a proper FLSA estimate on the data. More details and examples can be seen in Tibshirani and Wang (2008).

One important question for the FLSA is how good the estimator defined in Eq. (2) is. We analyze in this paper if the FLSA can recover the "stepwise pattern" or not. We also try to answer the following question: what do we do if the FLSA does not recover the "stepwise pattern"? To measure how good an estimator is, we introduce the following definition of Pattern Recovery.

Definition 1 (*Pattern Recovery*)**.** An FLSA solution $\hat{\mu}(\lambda_{1n}, \lambda_{2n})$ recovers the signal pattern if and only if there exist λ_{1n} and λ_{2n}, such that

$$sign(\hat{\mu}_{i+1}(\lambda_{1n}, \lambda_{2n}) - \hat{\mu}_i(\lambda_{1n}, \lambda_{2n})) = sign(\mu_{i+1}^* - \mu_i^*), \quad i = 1, \ldots, n-1. \tag{3}$$

We use $\hat{\mu} =_{js} \mu^*$ to shortly denote (3) (*js* is the acronym for *jump sign*). The FLSA with the property of pattern recovery means that it can be used to identify both the groups and the jump directions (up or down) between groups.

The concept of pattern recovery of the FLSA is very similar to the sign recovery of the Lasso (Zhao and Yu, 2006). In fact, we will see in Section 2 that the pattern recovery property of the FLSA is equivalent to the sign recovery property of the Lasso after transformation.

For observation pairs (x_i, y_i), $i = 1, 2, \ldots, n$ with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, the Lasso estimator is defined as follows (Tibshirani, 1996):

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|_1,$$

or equivalently, in the matrix form,

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\|Y - X\beta\|^2 + \lambda\|\beta\|_1, \tag{4}$$

where $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ with x_i^T as its ith row. We use X_j to denote the jth column of X. Lasso is a regularized method that gives sparse solutions. The L_1 term in the objective function makes the solution sparse. For different purposes, other loss functions could be used. Xu et al. (2015) used Cauchy loss in the objective function to get robust estimators. For matrix parameters, nuclear norm and its variations could be used to get low rank estimators (Cai et al., 2010; Xu et al., 2014).

Sign recovery of the Lasso is defined as follows.

Definition 2 (*Sign Recovery (Zhao and Yu, 2006)*)**.** Suppose that data (X, Y) follow a linear model: $Y = X\beta^* + \epsilon$, where $Y = (y_1, \ldots, y_n)^T, X \in \mathbb{R}^{n\times p}$ with x_i^T as its ith row, $\beta^* \in \mathbb{R}^{p\times 1}$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^{n\times 1}$ with $E(\epsilon_i) = 0$. A Lasso estimator $\hat{\beta}(\lambda_n)$ has the sign recovery property if and only if there exists λ_n such that

$$sign(\hat{\beta}_j(\lambda_n)) = sign(\beta_j^*), \quad j = 1, \ldots, p. \tag{5}$$

We will use $\hat{\beta} =_s \beta^*$ to shortly denote this property. In particular, this property, if satisfied, means that the Lasso selects the correct set of predictors. Asymptotically, we say the Lasso estimator $\hat{\beta}(\lambda_n)$ is sign consistent if there exists a sequence of λ_n such that $P(\hat{\beta}(\lambda_n) =_s \beta) \to 1$, as the sample size $n \to \infty$.

A rich theoretical literature has studied the consistency of the Lasso, highlighting several potential pitfalls (Knight and Fu, 2000; Fan and Li, 2001; Greenshtein and Ritov, 2004; Donoho et al., 2006; Meinshausen and Bühlmann, 2006; Tropp, 2006; Zhao and Yu, 2006; Zhang and Huang, 2008; Wainwright, 2009). The sign consistency of the Lasso requires the irrepresentable condition, a stringent assumption on the design matrix (Zhao and Yu, 2006). Now it is well understood that if the design matrix violates the irrepresentable condition, the Lasso will perform poorly in sign recovery, even with increased sample size.

The analyses of the FLSA are fairly recent. Tibshirani and Taylor (2011) consider the FLSA as a special case of the generalized Lasso problem and give the boundary lemma, a remarkable property of the FLSA saying that along the solution path, the coordinates on the boundary will always stay on the boundary as λ decreases. Rinaldo (2009) considers the sign consistency of the FLSA and proposes the adaptive fused Lasso. It shows that under some conditions, the FLSA can be consistent in both block reconstruction and model selection. The fused Lasso applied to general graphs, called the edge Lasso, is considered in Sharpnack et al. (2012). The ability of exact pattern recovery depends heavily on the structural properties of the graph. In particular, the author mentions the sign inconsistency of the 1D fused Lasso, the FLSA here, but does not attempt at addressing this issue.

There is a very close connection between the fused Lasso and change point estimation. In Harchaoui and Lévy-Leduc (2008) and Harchaoui and Lévy-Leduc (2010), the authors study the change point estimation problem using the total variation penalty, which is the fused Lasso when λ_1 is taken as 0. Harchaoui and Lévy-Leduc (2010) have a very nice result for consistency on the estimation of the signal mean. They also point out that the total variation penalized least squares cannot consistently estimate the locations of change points. Instead, they show that under a few regularity conditions, the estimated locations of the change points are close to the true locations. In this paper, we give deeper analysis of the total variation penalized least squares. We give a necessary and sufficient condition on consistently recovering the piecewise constant pattern, which is also called consistent recovery of change points.

We borrow analytical tools used to study the LASSO in the paper. There are several differences between the fused Lasso and Lasso that should be highlighted, though the former can be transformed to the latter by reparametrization. First, though could be seen as a specific kind of Lasso problem under reparametrization, the recovery of piecewise constant signals, or detection of multiple change points itself is an important class of problems and has very wide applications. So it is worthwhile to gain more insight into this problem beyond general Lasso's and to consider fast and consistent algorithms as we do here. Second, the analysis of sign consistency is different than classic Lasso problems in that the attempt to transform to Lasso introduces an intercept term that is not penalized as other coefficients. The intercept can be replaced by the mean of observations, however, extra care should be taken before directly applying the Lasso theories. We adapt the argument for standard Lasso to general noise structures other than independent noise.

In this paper, we not only study the sign recovery of the FLSA using the irrepresentable condition for the Lasso, but more importantly, by preconditioning the design matrix, our newly proposed method significantly improves the performance of pattern recovery. For preparation, we first prove that even for the linear model with correlated noise, the irrepresentable condition is still necessary for sign consistency. We then analyze the design matrix in the transformed Lasso problem. We give necessary and sufficient condition such that the design matrix in the transformed Lasso problem complies with the irrepresentable condition. We show that, only for a special class of models (with special designed stepwise function on μ_i^*), the irrepresentable condition holds. For other signal patterns, the irrepresentable condition does not hold and thus the FLSA may fail to keep consistent. A recent paper "Preconditioning to comply with the irrepresentable condition" by Jia and

Rohe (2015) suggests a Puffer Transformation that will improve the Lasso and make the Lasso estimator sign consistent under some mild conditions. We apply this technique, propose the *preconditioned fused Lasso* and show that it significantly improves the FLSA and recovers the signal pattern with high probability. We also point out that we did not fully solve the change point estimation problem. Our results show that if the signal is strong and the noise is small, then no matter what the pattern is, the preconditioned fused Lasso gives consistent estimation of the change point locations. In change point literature, people also study some special patterns — say when the block size is big enough. With more information on the signal pattern, one may have stronger results. This is now out of our research scope and will be our further study. In this paper, we try to understand the fused Lasso (equivalently the total variation penalized least squares) for one dimensional signals.

The rest of the paper is organized as follows. In Section 2, we establish connection between the FLSA and a Lasso problem via proper transformation. Section 3 discusses when the FLSA can recover the signal pattern and when it cannot. In Section 4, we propose a new algorithm called the preconditioned fused Lasso that improves the FLSA by using the preconditioning technique. We show that for a wide range of designs of the stepwise function on μ^*, this algorithm can recover the signal pattern with high probability. In Section 5, simulations are implemented to compare the performances between the preconditioned fused Lasso and the vanilla FLSA. Section 6 concludes the paper. The proofs are given in the Appendix.

2. The FLSA and the Lasso

In this section, we transform the FLSA problem into a Lasso problem by change of variables. Define the soft thresholding function $SH_\lambda(x)$ as

$$SH_\lambda(x) = \begin{cases} x+\lambda & x < -\lambda \\ 0 & -\lambda \le x \le \lambda \\ x-\lambda & x > \lambda. \end{cases}$$

Let $\hat{\mu}(\lambda_1, \lambda_2)$ be the fused Lasso estimator defined in (2). We have the following result.

Lemma 1 (*Friedman et al., 2007*)**.**

$$\hat{\mu}(\lambda_1, \lambda_2) = SH_{\lambda_1}(\hat{\mu}(0, \lambda_2)).$$

From Lemma 1, to study the properties of $\hat{\mu}(\lambda_1, \lambda_2)$, we can set $\lambda_1 = 0$ first. Since pattern recovery is our main concern here, we only consider the case $\lambda_1 = 0$ in this paper. When $\lambda_1 = 0$, we can solve the FLSA by change of variables. Let $\theta_1 = \mu_1, \theta_i = \mu_i - \mu_{i-1}, i = 2, \ldots, n$, or in the matrix form, $\mu = A\theta$, where $A \in \mathbb{R}^{n\times n}$ is the lower triangular matrix with nonzero elements equal to one. So by using θ instead of μ, we have an equivalent solution of $\hat{\mu}(0, \lambda_2)$ via the following $\hat{\theta}(\lambda_2)$:

$$\hat{\theta}(\lambda_2) = \underset{\theta}{\operatorname{argmin}} \frac{1}{2}\|Y - A\theta\|_2^2 + \lambda_2\|\tilde{\theta}\|_1, \tag{6}$$

where $\tilde{\theta} = (\theta_2, \theta_3, \ldots, \theta_n)^T \in \mathbb{R}^{n-1}$. Once we obtain $\hat{\theta}(\lambda_2)$, we have $\hat{\mu}(0, \lambda_2) = A\hat{\theta}(\lambda_2)$. With the special form of the design matrix A, (6) is a Lasso problem with intercept. In fact, (6) can be rewritten as

$$\hat{\theta}(\lambda_2) = \underset{(\theta_1, \tilde{\theta})}{\operatorname{argmin}} \frac{1}{2}\|Y - \mathbf{1}\cdot\theta_1 - X\tilde{\theta}\|_2^2 + \lambda_2\|\tilde{\theta}\|_1 \tag{7}$$

where $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$, $\tilde{\theta} = (\theta_2, \ldots, \theta_n)^T$ and $X = (x_{ij}) \in \mathbb{R}^{n\times(n-1)}$:

$$x_{ij} = \begin{cases} 1 & i > j \\ 0 & i \le j. \end{cases}$$

Define the centered version of $X \in \mathbb{R}^{n\times(n-1)}$ and $Y \in \mathbb{R}^n$ as follows:

$$\tilde{X} = [X_1 - \bar{X}_1, \ldots, X_{n-1} - \bar{X}_{n-1}] \quad \text{and} \quad \tilde{Y} = Y - \bar{Y}, \tag{8}$$

where $\bar{u}$ is the vector with all elements equal to the average of u. It is easy to see that (7) is equivalent to the following standard Lasso problem without intercept:

$$\hat{\tilde{\theta}}(\lambda_2) = \underset{\tilde{\theta}}{\operatorname{argmin}} \frac{1}{2}\|\tilde{Y} - \tilde{X}\tilde{\theta}\|_2^2 + \lambda_2\|\tilde{\theta}\|_1, \quad \text{and} \quad \hat{\theta}_1(\lambda_2) = \bar{Y} - \bar{X}\hat{\tilde{\theta}}(\lambda_2). \tag{9}$$

Define $\theta^* = A^{-1}\mu^*$; that is, $\theta_1^* = \mu_1^*, \theta_i^* = \mu_i^* - \mu_{i-1}^*, i = 2, \ldots, n$. Let $\tilde{\theta}^* = (\theta_2^*, \theta_3^*, \ldots, \theta_n^*)^T \in \mathbb{R}^{n-1}$. Since the observation $Y = (y_1, \ldots, y_n)$ follows the model defined in (1), we have that (X, Y) satisfies the linear model:

$$Y = A\theta^* + \epsilon = \theta_1^* + X\tilde{\theta}^* + \epsilon,$$

where X is defined at (2). Thus the centered version of (X, Y) satisfies the following linear model:

$$\tilde{Y} = \tilde{X}\tilde{\theta}^* + \tilde{\epsilon}, \tag{10}$$

where $\tilde{\epsilon} = \epsilon - \bar{\epsilon}$, the centered version. Now we see that $\hat{\tilde{\theta}}(\lambda_2)$ defined in (9) has the sign recovery property if and only if $\hat{\tilde{\theta}}(\lambda_2) =_s \tilde{\theta}^*$. By the relationship between μ and θ ($\mu = A\theta$), $\hat{\tilde{\theta}}(\lambda_2) =_s \tilde{\theta}^*$ is equivalent to $\hat{\mu}(0, \lambda_2) =_{js} \mu^*$. In other words, the pattern recovery property of the FLSA is equivalent to the sign recovery of the corresponding Lasso estimator.

Property 1. *The pattern recovery of the FLSA $\hat{\mu}(0, \lambda_2)$ defined in (2) is equivalent to the sign consistency of the Lasso estimator $\hat{\tilde{\theta}}(\lambda_2)$ defined in (9).*

Note that the main purpose of the change of variables here is for theoretical analysis rather than computational considerations. Although there are many efficient algorithms for the Lasso such as LARS and coordinate descent, it is not recommended in practice to solve the FLSA by transforming to the Lasso and then applying these algorithms. The transformation makes the design matrix in (9) dense and is computationally unfavorable. Instead, Friedman et al. (2007) develop specialized algorithm for the FLSA based on the coordinate-wise descent. Hoefling (2010) proposes the path algorithm and extends that to more general fused Lasso problems. In our theoretical analysis, however, such transformation helps since we can use well understood techniques for the Lasso to analyze the properties of the FLSA.

3. The transformed Lasso

From the above argument, the study of the FLSA is reduced to analyzing the Lasso problem defined in (9). It is now well understood that in a standard linear regression problem the Lasso is sign consistent when the design matrix satisfies some stringent conditions. One such condition is the irrepresentable condition (Zhao and Yu, 2006) defined as follows:

Definition 3 (*Irrepresentable Condition*)**.** Suppose that data (X, Y) follows a linear model: $Y = X\beta^* + \epsilon$, where $Y = (y_1, \ldots, y_n)^T, X \in \mathbb{R}^{n\times p}, \beta^* \in \mathbb{R}^{p\times 1}$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^{n\times 1}$ with $E(\epsilon_i) = 0$. The design matrix X satisfies the *Irrepresentable Condition* for β^* with support $S = \{j : \beta_j^* \neq 0\}$ if, for some $\eta \in (0, 1]$,

$$\left\| X_{S^c}^T X_S \left(X_S^T X_S\right)^{-1} sign(\beta_S^*) \right\|_\infty \leq 1 - \eta, \tag{11}$$

where for a vector x, $\|x\|_\infty = \max_i |x_i|$, and for $T \subset \{1, \ldots, p\}$ with $|T| = t$, $X_T \in \mathbb{R}^{n\times t}$ is a matrix which contains the columns of X indexed by T.

The irrepresentable condition is a key condition for the Lasso's sign consistency. A lot of researchers noticed that the irrepresentable condition is a necessary condition for the Lasso's sign consistency (Zhao and Yu, 2006; Wainwright, 2009; Jia et al., 2013). We also state this conclusion under a more general linear model with correlated noise terms.

Theorem 1. *Suppose that data (X, Y) follows a linear model $Y = X\beta^* + \epsilon$, with Gaussian noise $\epsilon \sim N(0, \Sigma_\epsilon)$. The irrepresentable condition (11) is necessary for the sign consistency of the Lasso. In other words, if*

$$\left\| X_{S^c}^T X_S \left(X_S^T X_S\right)^{-1} sign(\beta_S^*) \right\|_\infty \geq 1, \tag{12}$$

we have

$$P(\hat{\beta}(\lambda) =_s \beta^*) \leq \frac{1}{2}.$$

The proof of Theorem 1 is postponed to Appendix. This theorem says if the irrepresentable condition does not hold, it is very likely that the Lasso cannot correctly recover the signs of the coefficients.

With the above theorem, we come back to the transformed Lasso problem defined in (9) and examine if the irrepresentable condition holds or not in this case. Recall that for the Lasso problem induced from the FLSA, we have the design matrix

$$\tilde{X} = [X_1 - \bar{X}_1, \ldots, X_{n-1} - \bar{X}_{n-1}].$$

Let $S = \{j : \tilde{\theta}_j^* \neq 0\}$ denote the index set of the relevant variables in the true model. Then (11) can be written as

$$\left| \tilde{X}_j^T \tilde{X}_S (\tilde{X}_S^T \tilde{X}_S)^{-1} sign(\tilde{\theta}^*) \right| < 1, \quad \forall j \notin S.$$

This is equivalent to

$$|\hat{b}_j^T sign(\tilde{\theta}^*)| < 1, \quad \forall j \notin S,$$

where $\hat{b}_j \in \mathbb{R}^{|S|}$ is the OLS estimate of b_j in the following linear regression equation:

$$\tilde{X}_j = b_j{}^T \tilde{X}_S + \epsilon. \tag{13}$$

Since $\tilde{X}$ is the centered version of X, it can be easily shown that $\hat{b}_j$ is also the OLS estimate of b_j in the following linear regression equation:

$$X_j = b_0 + b_j{}^T X_S + \epsilon, \tag{14}$$

where $b_0 \in \mathbb{R}$ is the intercept term. Recall the irrepresentable condition defined in (11):

$$\left\| X_{S^c}^T X_S \left(X_S^T X_S \right)^{-1} sign(\beta_S^*) \right\|_\infty \leq 1 - \eta.$$

We can thus find all the signal patterns in (1) such that the irrepresentable condition holds for the corresponding transformed Lasso.

Theorem 2. *Assume $y = (y_1, \ldots, y_n)$ satisfies model (1), the collection of the indices of jump points are $S = \{j_1, j_2, \ldots, j_s\}$ with $j_k (1 \leq k \leq s)$ increasing. Formally, $S = \{j : \mu_j^* \neq \mu_{j-1}^*, j = 2, \ldots, n\}$. Then the irrepresentable condition (11) holds for the transformed Lasso if and only if one of the following two conditions holds.*

(1) *The jump points are consecutive. That is, $s = 1$ or*

$$\max_{1 \leq k < s} (j_{k+1} - j_k) = 1.$$

(2) *If there exists one group of data points (with more than 1 point) between some two consecutive jump points and these data points have the same expected signal strength, then the two jumps are of different directions (up or down). Formally, let j_k and j_{k+1} be two jump points and $\mu_{j_k}^* = \cdots = \mu_{j_{k+1}-1}^*$, then $(\mu_{j_k}^* - \mu_{j_k-1}^*)(\mu_{j_{k+1}}^* - \mu_{j_{k+1}-1}^*) < 0$.*

The proof is given in the Appendix. Theorem 2 implies that only a few configurations of μ^* can make the transformed Lasso comply with the irrepresentable condition. In applications, most signal patterns do not satisfy either of the two conditions in Theorem 2. Harchaoui and Lévy-Leduc (2010) also provide a result considering the irrepresentable condition. They show that the irrepresentable condition never holds for a slightly modified total variation penalty. Note that the penalty term for original total variation penalty term is $\sum_{i=2}^p |\beta_{i+1} - \beta_i|$, while Harchaoui and Lévy-Leduc (2010) give a result for a slightly different penalty $\sum_{i=2}^p [|\beta_{i+1} - \beta_i|] + \beta_1$. With the small modification, the result is also different.

We noticed that Rinaldo (2009) gave a result (Theorem 2.3 on page 2930) saying that under some regularity conditions, the fused Lasso consistently recovers the signal pattern. This is a contradiction with Theorem 2, since those regularity conditions only depend on the signal strengths, the number of blocks and the minimal size of the blocks. Because these conditions do not guarantee the irrepresentable condition, that result missed this irrepresentable condition.

For sign recovery defined in (5), Jia and Rohe (2015) proposed a Puffer Transformation that preconditions the design matrix in order to comply with the irrepresentable condition. The connection between sign recovery and pattern recovery defined in (3) enables us to apply the same technique and thus improve the performance over the vanilla FLSA in pattern recovery.

4. Preconditioned fused Lasso with puffer transformation

Jia and Rohe (2015) introduce the Puffer Transformation to the Lasso when the design matrix does not satisfy the irrepresentable condition. They showed that when $n \geq p$, even if the Lasso is not sign consistent, after the Puffer Transformation, the Lasso is sign consistent under some mild conditions.

We assume that the design matrix $X \in \mathbb{R}^{n \times p}$ has rank $d = \min\{n, p\}$. By the singular value decomposition, there exist matrices $U \in \mathbb{R}^{n \times d}$ and $V \in \mathbb{R}^{p \times d}$ with $U^T U = V^T V = I_d$ and a diagonal matrix $D \in \mathbb{R}^{d \times d}$ such that $X = UDV'$. Define the *Puffer Transformation* (Jia and Rohe, 2015),

$$F_{n \times n} = UD^{-1}U^T. \tag{15}$$

The preconditioned design matrix FX has the same singular vectors as X. However, all of the nonzero singular values of FX are set to unity: $FX = UV'$. When $n \geq p$, the columns of FX are orthonormal. When $n \leq p$, the rows of FX are orthonormal. Jia and Rohe (2015) have a non-asymptotic result for the Lasso on (FX, FY) stated in Theorem 4 in the Appendix. We see that with the Puffer Transformation, the Lasso does not need the irrepresentable condition any more.

As shown previously, the FLSA can be transformed to a standard Lasso problem. We have already shown that for most configurations of μ^*, the design matrix $\tilde{X}$ does not satisfy the irrepresentable condition. Now we turn to the Puffer Transformation and obtain a concrete non-asymptotic result for the preconditioned fused Lasso.

Theorem 3. *Assume* $y = (y_1, \ldots, y_n)$ *satisfies model (1).* $\tilde{X}$ *and* $\tilde{Y}$ *are defined in (8). Let* $\theta^* = A^{-1}\mu^*$ *(equivalently,* $\theta_1^* = \mu_1^*, \theta_i^* = \mu_i^* - \mu_{i-1}^*, i = 2, \ldots, n$*), where A is defined to be the lower triangular matrix with nonzero elements equal to one. Let* $\tilde{\theta}^* \in \mathbb{R}^{n-1} = (\theta_2^*, \theta_3^*, \ldots, \theta_n^*)^T$*. Define the singular value decomposition of* $\tilde{X}$ *as* $\tilde{X} = UDV^T$*. Denote the Puffer Transformation by* $F = UD^{-1}U^T$*. Let* $Z = F\tilde{X}$ *and* $a = F\tilde{Y}$*. Define*

$$\tilde{\beta}(\lambda) = \underset{b}{\operatorname{argmin}} \ \frac{1}{2}\|a - Zb\|_2^2 + \lambda\|b\|_1. \tag{16}$$

If $\min_{j\geq 2, \theta_j^* \neq 0} |\theta_j^*| \geq 2\lambda$*, then*

$$P\left(\tilde{\beta}(\lambda) =_s \tilde{\theta}^*\right) \geq 1 - 2n \exp\left\{-\frac{\lambda^2}{8\sigma^2}\right\}.$$

The proof is given in the Appendix. By the relationship between θ^* and μ^*, if $\tilde{\beta}(\lambda)$ – the estimate of $\tilde{\theta}^*$ has the sign recovery property, then the estimate of μ^* defined as follows has the property of pattern recovery.

$$\hat{\mu}^* = A\hat{\theta}^* \tag{17}$$

with

$$\hat{\theta}^* = [\hat{\theta}_1, \tilde{\beta}(\lambda)] \quad \text{and} \quad \hat{\theta}_1 = \bar{Y} - \bar{X}\tilde{\beta}(\lambda).$$

Theorem 3 shows that the pattern recovery of the preconditioned fused Lasso depends on the signal jump strength ($\min_{j\geq 2, \theta_j^* \neq 0} |\theta_j^*|$) and the noise level σ^2. To get a pattern-consistent estimate, we need σ small enough and $\min_{j\geq 2, \theta_j^* \neq 0} |\theta_j^*|$ big enough. To think about the small σ^2 issue, we can treat each y_i as an average of multiple Gaussian measurements. If the number of measurements is m, then $\sigma^2 = \frac{\sigma_0^2}{m}$ with some constant σ_0^2. If $m \gg \log(n)$, we can find a very small λ to make the estimator defined in (17) have the pattern recovery property. One choice of λ is such that $\lambda^2 = \frac{\log(n+1)}{\sqrt{m}}$. For this choice of λ, the probability of $\hat{\mu}^* =_{js} \mu^*$ is greater than $1 - 2\exp\left(-[\frac{\sqrt{m}}{8\sigma_0^2} - 1]\log(n+1)\right)$, which goes to 1 as n goes to ∞.

5. Numerical examples

We use several examples to illustrate our theory. Recall the model is set to be

$$y_i = \mu_i^* + \epsilon_i,$$

where the errors ϵ_i's are i.i.d. Gaussian variables with mean 0 and standard deviation σ. In the following simulations, the length of the signal is set to be 430, not for others, but is just the same as the signal length in the example in Rinaldo (2009) and more convenient for comparison. μ_i^* will be specified case by case, reflecting the characteristics of the signal pattern. σ will generally vary between 0.05 and 0.5 to illustrate the recovery ability as a function of the noise level. Here are some implementation details of the two procedures.

FLSA When calculating the FLSA solution, we use a path algorithm proposed by Hoefling (2010) which is very efficient to give the entire solution path of the FLSA. An R package ("flsa") for this algorithm is available in http://cran.r-project.org/web/packages/flsa/index.html. In fact, the entire solution path is piecewise linear in λ. "flsa" only stores the λ's corresponding to the breakpoints at which the directions of the linear function change.

Preconditioned fused Lasso We calculate the solution defined in (16). After the SVD and the Puffer Transformation, the task becomes much easier. It suffices to do soft-thresholding to obtain the entire solution path. This is because

$$Z^TZ = X^TF^TFX = (VDU^T)(UD^{-1}U^T)(UD^{-1}U^T)(UDV^T) = I_n$$

and the property of the Lasso allows us to solve it directly by soft-thresholding (Tibshirani, 1996):

$$\hat{b}(\lambda) = SH_\lambda(Z^Ta).$$

Obviously, $\hat{b}(\lambda)$ is piecewise linear in λ and the breakpoints are $\lambda_i = |Z^Ta|_{(i)}, i = 1, 2, \ldots, n$, where $x_{(i)}$ denotes the ith largest value in vector x and n is the dimension of the vector Z^Ta.

By a little further algebra analysis, we see that the Preconditioned Fused Lasso estimator can be calculated via soft thresholding of the successive differences of the observed signals. This is because

$$Z^Ta = \tilde{X}^TF^TF\tilde{Y} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y} = \arg\min_\theta \|\tilde{Y} - \tilde{X}\theta\|_2$$

$$= \arg\min_\theta \|Y - X\theta - \theta_1\|_2 = \left\{\arg\min_\theta \|Y - A\theta\|_2\right\}_{[2:n]} = \left\{A^{-1}Y\right\}_{[2:n]},$$

where $A \in \mathbb{R}^{n\times n}$ is the lower triangular matrix with nonzero elements equal to one, and $x_{[2:n]}$ stands for the vector consisting of the 2^{nd} to n^{th} elements of x. So $\hat{b}_i(\lambda) = SH_\lambda(Y_{i+1} - Y_i)$, $i = 1, 2, \ldots, n-1$.

Despite equivalence to this simplified thresholding algorithm, we point out that the preconditioned fused Lasso can be extended easily to more general settings. For example, suppose we would like to recover the blocks in an unknown coefficient sequence by solving

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda \sum_{i=1}^{n-1} |\beta_{i+1} - \beta_i|. \tag{18}$$

We can do the transformation $\beta = A\gamma$, where A is the same as in (6). The objective can again be transformed to a Lasso-type, only no penalty on γ_1. Then we can apply the preconditioned technique, in particular the Puffer Transformation as in Theorem 3, to the response y and the new matrix XA after reparametrization, and solve resulting problem. We do a few experiments for this problem and report one simulated result at the end of this section.

There are many criteria for comparison. In the context of exact pattern recovery, our principle is to check the solution path and see if there is a solution that has exactly the same jump points as the true signals.

For each σ selected, we draw 1000 sample sequences and define P_{FLSA} and $P_{PCD\ FL}$ to be the proportion of samples that the recovered blocks (jump points) exactly match the true blocks (jump points) by using the FLSA and using the preconditioned fused Lasso, respectively.

We will demonstrate that when irrepresentable condition holds, FLSA can recover the true blocks with high probability; when irrepresentable condition does not hold, FLSA performs poorly, and for this case Puffer Transformation helps a lot.

5.1. The irrepresentable condition holds

We give specific examples of the signal patterns such that the irrepresentable condition of the transformed Lasso holds and thus the FLSA can recover the pattern under mild conditions. Theorem 2 provides the necessary and sufficient conditions for irrepresentable condition. Since the second condition in Theorem 2 is more commonly met, we focus on the signals that satisfy this condition in the following.

We used two examples to show that when signal pattern follows the second condition in Theorem 2, both FLSA and preconditioned FLSA can recover the signal pattern when noise is small. The sample data and results are shown in Fig. 2. The top row gives the expected signal (plotted in lines) and the sampled noisy data (plotted in dots). The two signal patterns are slightly different — the left one is more regular in the sense that the block size is almost the same. It is not clear when transformed FLSA is better than vanilla FLSA when irrepresentable condition holds. But when irrepresentable condition does not hold, which happens more likely than it holds, preconditioned FLSA definitely outperforms the vanilla FLSA. We demonstrate this in the next subsection.

5.2. The irrepresentable condition fails

When the irrepresentable condition does not hold, the FLSA cannot reliably recover the exact pattern. We analyze in more detail the numerical performances of the two procedures.

We use the same example as in Rinaldo (2009) except for larger noise ($\sigma = 0.25$ here). Recall that the signal pattern is set to be

$$\mu_i^* = \begin{cases} 0, & 1 \le i \le 100 \\ -2, & 101 \le i \le 110 \\ -0.1, & 111 \le i \le 210 \\ 2, & 211 \le i \le 220 \\ 0.1, & 221 \le i \le 320 \\ -2, & 321 \le i \le 330 \\ 0, & 331 \le i \le 430. \end{cases}$$

Fig. 3 shows the sample data and true signals.

We apply the FLSA and the preconditioned fused Lasso to this sample data and compare the recovery performances. Fig. 4 plots two solutions by different selection of tuning parameter. The solution in the left plot is the one with tuning parameter selected by recovering the same number of blocks as that of μ^*, the true signals; Right panel plots the FLSA solution (in red lines) with tuning parameter selected by minimizing the ℓ_2 error between $\hat{\mu}^*$ and μ^*. Fig. 4 shows that the FLSA cannot locate the jump points correctly and the right subfigure illustrates that a good estimate under the Euclidean norm is not reliable in exact pattern recovery. In sharp contrast, the preconditioned fused Lasso applied shown in Fig. 5 precisely locates all the jump points and recovers the pattern.

Note that the reported preconditioned FLSA estimate in Fig. 5 is very biased from the expected value. There is a tradeoff between the unbiasedness and the quality of pattern recovery. One possible solution for the unbiasedness is via a two-stage estimator — for the first stage the signal patten is recovered and for the second stage an unbiased estimate is obtained.

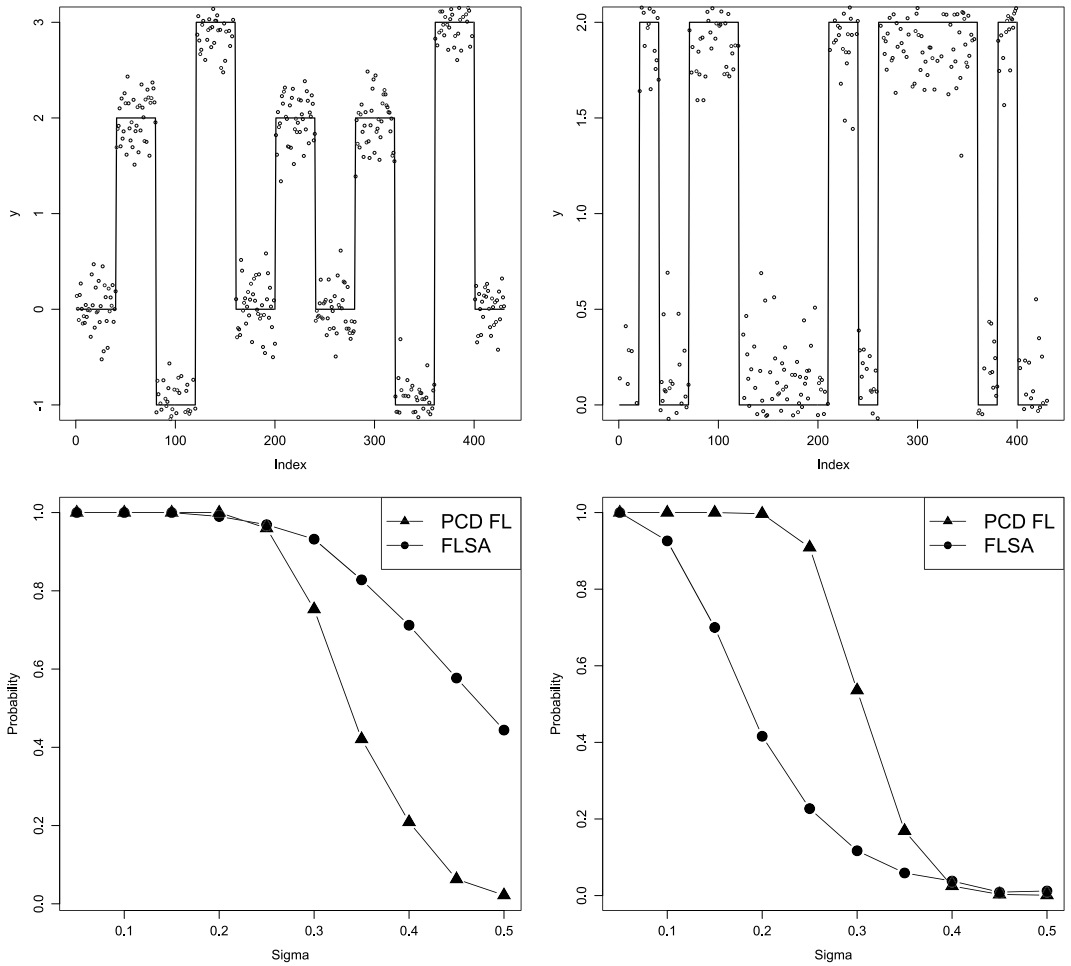

Fig. 2. Signal patterns and recovery performances. The first row presents the true signals (lines) and the sample data (points) under noise level $\sigma = 0.25$. The second row shows the recovery performances measured by the approximate probability of exact pattern recovery when the variance of noise increases. Each point is estimated with 1000 randomly generated datasets.

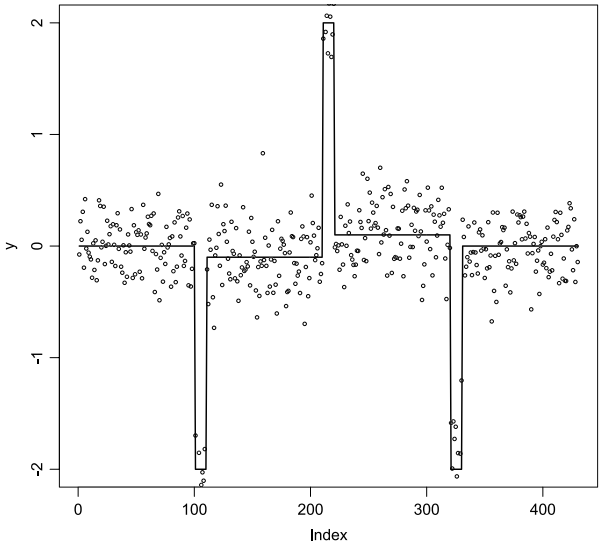

Fig. 3. Sample data (points) and the expected signals (lines) at $\sigma = 0.25$.

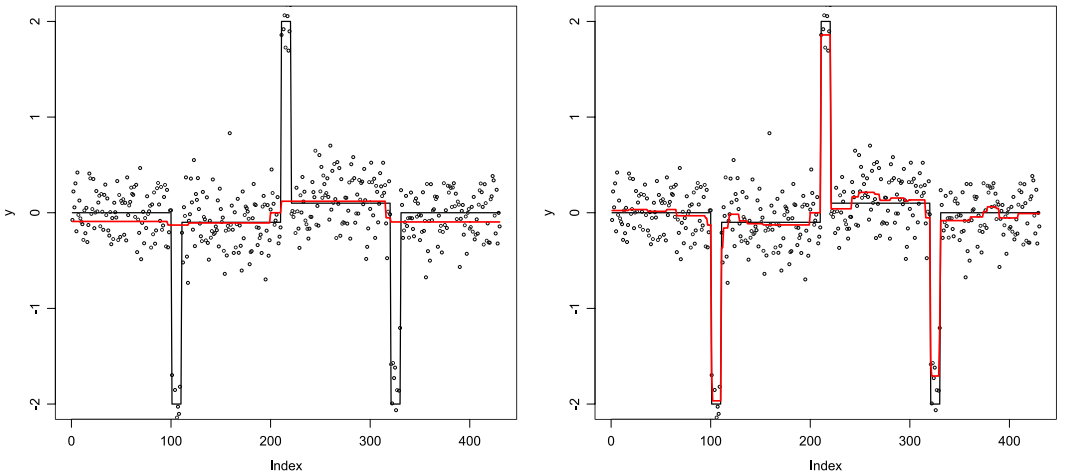

Fig. 4. FLSA solutions. Left panel: the FLSA solution(in red lines) with tuning parameter selected by recovering the same number of blocks as that of μ^*, the true signals; Right panel: the FLSA solution (in red lines) with tuning parameter selected by minimizing the ℓ_2 error between $\hat{\mu}^*$ and μ^*. The black lines are the expected signal sequences.

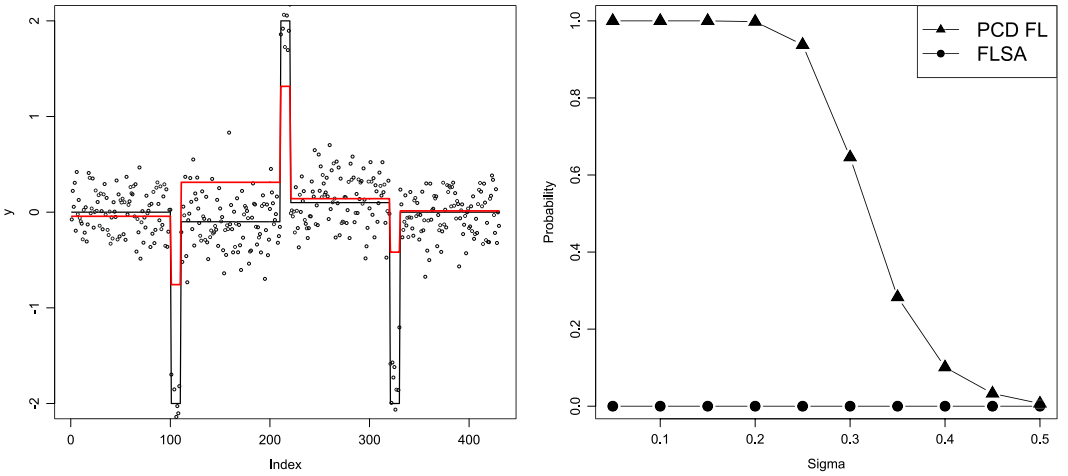

Fig. 5. The preconditioned fused Lasso solutions. Left: a sample solution of the preconditioned fused Lasso. Right: the estimated probability of pattern recovery under different noise levels for the preconditioned fused Lasso. Each point is estimated with 1000 randomly generated datasets.

We further compare the recovery performances under different noise levels. We draw 1000 random sample sequences and compare the approximate exact recovery probability. Fig. 5 visualizes the result.

The FLSA can hardly recover the signal pattern exactly even when the noise level is as small as $\sigma = 0.05$. This is supported by the theory above. In contrast, the preconditioned fused Lasso has fairly satisfactory recovery performance till $\sigma = 0.25$.

5.3. Extensions

Finally, we provide one simulation result for the problem of

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_{TV}.$$

In this simulation study, we choose each row of X i.i.d. from $N(0, \Sigma)$, where $\Sigma_{ij} = 0.4$, for $i \neq j$ and $\Sigma_{ii} = 1$. We have a design matrix $X \in \mathbb{R}^{n\times p}$ with $n = 200$ and $p = 100$. We take β to have the pattern of piecewise constant. The true coefficients are plotted in Fig. 6 (left). Y is designed as $X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. We vary σ from 0.25 to 5. We apply both preconditioned method (denoted as PCFL) and the one without preconditioning (denoted as FLSA). For each noise level σ, we do 500 repetitions and then we calculate the proportion of successful cases when one method could recover the piecewise pattern of the coefficients. The results are plotted in Fig. 6 (right), from which we see the necessity of preconditioning.

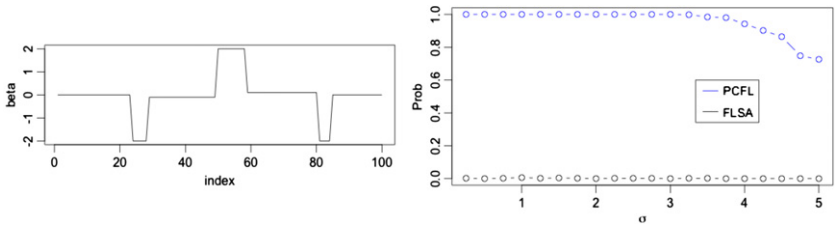

Fig. 6. True coefficients (left) and recovery performances (right). Recovery performance is measured by the approximate probability of exact pattern recovery. Each point is estimated with 500 randomly generated datasets.

6. Conclusions and discussions

In this paper we provided more understanding of the FLSA and shed some light on the insight into the FLSA. The FLSA can be transformed to a standard Lasso problem. The sign recovery of the transformed Lasso problem is equivalent to the pattern recovery of the FLSA problem. Theoretical analysis showed that the transformed Lasso problem is not sign consistent in most situations. So the FLSA might also meet this consistency problem when it is used to recover signal patterns. To overcome such problem, we introduced the preconditioned fused Lasso. We gave non-asymptotic results on the preconditioned fused Lasso. The result implies that when the signal-to-noise ratio is not so small, the preconditioned fused Lasso can recover the signal pattern with high probability. Through simulation studies, we also found that the FLSA, if the irrepresentable condition holds, is only more apt at recovering regular signals while our preconditioned fused Lasso is more robust to various kinds of signals. Some attempts also imply that the preconditioned fused Lasso does not work so well if all the signals are equally weak.

A good pattern recovery will facilitate many things afterwards. The preconditioned fused Lasso is reliable for pattern recovery, and so it can be incorporated into other processes — such as the recovery of sparsity.

Acknowledgments

This work is partially supported by the National Basic Research Program of China (973 Program 2011CB809105), NSFC-61121002, NSFC-11101005, DPHEC-20110001120113, and MSRA. Junyang Qian is also affiliated with Department of Statistics, Stanford University.

Appendix

We prove the theorems in the appendix.

A.1. Proof of Theorem 1

We first give a well-known result that makes sure the Lasso exactly recovers the sparse pattern of β^*, that is $\hat{\beta}(\lambda) =_s \beta^*$. The following Lemma gives necessary and sufficient conditions for $\text{sign}(\hat{\beta}(\lambda)) = \text{sign}(\beta^*)$, which follows from the KKT conditions. The proof of this lemma can be found in Wainwright (2009).

Lemma 2. *For the linear model $Y = X\beta^* + \epsilon$, assume that the matrix $X_S^T X_S$ is invertible. Then for any given $\lambda > 0$ and any noise term $\epsilon \in \mathbb{R}^n$, there exists a Lasso estimate $\hat{\beta}(\lambda)$ described in (4) which satisfies $\hat{\beta}(\lambda) =_s \beta^*$, if and only if the following two conditions hold*

$$\left|X_{S^c}^T X_S (X_S^T X_S)^{-1} \left[X_S^T \epsilon - \lambda sign(\beta_S^*)\right] - X_{S^c}^T \epsilon\right| \leq \lambda, \tag{19}$$

$$sign\left(\beta_S^* + (X_S^T X_S)^{-1} \left[X_S^T \epsilon - \lambda sign(\beta_S^*)\right]\right) = sign(\beta_S^*), \tag{20}$$

where the vector inequality and equality are taken elementwise. Moreover, if the inequality (19) holds strictly, then

$$\hat{\beta} = (\hat{\beta}^{(1)}, 0)$$

is the unique optimal solution to the Lasso problem in Eq. (4), where

$$\hat{\beta}^{(1)} = \beta_S^* + (X_S^T X_S)^{-1} \left[X_S^T \epsilon - \lambda sign(\beta^*)\right]. \tag{21}$$

Remark. As in Wainwright (2009), we state an equivalent condition for (19). Define

$$\overrightarrow{b} = sign(\beta_S^*),$$

and define

$$V_j = X_j^T \left\{ X_S (X_S^T X_S)^{-1} \lambda \overrightarrow{b} - \left[X_S (X_S^T X_S)^{-1} X_S^T - I \right] \epsilon \right\}.$$

By rearranging terms, it is easy to see that (19) holds if and only if

$$\mathcal{M}(V) = \left\{ \max_{j \in S^c} |V_j| \leq \lambda \right\} \tag{22}$$

holds.

With Lemma 2 and the above comments, now we prove Theorem 1. Without loss of generality, assume for some $j \in S^c$ and $\zeta \geq 0$,

$$X_j^T X_S \left(X_S{}^T X_S \right)^{-1} \overrightarrow{b} = 1 + \zeta.$$

Then

$$V_j = \lambda(1 + \zeta) + \tilde{V}_j,$$

where $\tilde{V}_j = -X_j^T [X_S \left(X_S{}^T X_S \right)^{-1} X_S{}^T - I] \frac{\epsilon}{n}$ is a Gaussian random variable with mean 0, so $P(\tilde{V}_j > 0) = \frac{1}{2}$. Therefore,

$$P(V_j > \lambda) \geq \frac{1}{2}$$

and the equality holds when $\zeta = 0$. This implies that for any λ, Condition (19) (a necessary condition) is violated with probability greater than 1/2.

In the proof of Theorem 1, we need an algebra result as follows.

Lemma 3. *For $k \geq 3$, $a_1, \ldots, a_k \in \mathbb{R}$ and are not equal to each other. $A = (a_{ij})_{k \times k}$, with $a_{ij} = a_\ell$ where $\ell = \max\{i, j\}$. That is,*

$$A = \begin{pmatrix} a_1 & a_2 & \vdots & a_k \\ a_2 & a_2 & \vdots & \vdots \\ \cdots & \cdots & \cdot & \vdots \\ a_k & \cdots & \cdots & a_k \end{pmatrix}.$$

Then the inverse of A

$$(A)^{-1} = \begin{bmatrix} r_{11} & r_{12} & & & & & \\ r_{21} & r_{22} & r_{23} & & & & \\ & r_{32} & r_{33} & r_{34} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & r_{k-1,k-2} & r_{k-1,k-1} & r_{k-1,k} \\ & & & & r_{k,k-1} & r_{k,k} \end{bmatrix}$$

where

$$r_{ij} = \begin{cases} \dfrac{1}{a_1 - a_2} & i = j = 1 \\ -\dfrac{1}{a_{j-1} - a_j} & i = j - 1 \\ -\dfrac{1}{a_j - a_{j+1}} & i = j + 1 \\ \dfrac{a_{j-1} - a_{j+1}}{(a_{j-1} - a_j)(a_j - a_{j+1})} & 1 < i = j < k \\ \dfrac{a_{k-1}}{(a_{k-1} - a_k)(a_k)} & i = j = k \\ 0 & \text{otherwise}. \end{cases}$$

Proof. This lemma can be directly verified via the following equations:

$$\sum_i a_{ji} r_{ij} = 1 \quad \text{and} \quad \sum_i a_{\ell i} r_{ij} = 0,\ \ell \neq j.$$

We first verify $\sum_i a_{ji} r_{ij} = 1$, for all j.

When $j = 1$,

$$\sum_i a_{1i} r_{i1} = a_1 r_{11} + a_2 r_{21} = a1 \cdot \frac{1}{a_1 - a_2} + a_2 \cdot \frac{-1}{a_1 - a_2} = 1.$$

When $1 < j < k$,

$$\begin{aligned}\sum_i a_{ji} r_{ij} &= a_{j,j-1} r_{j-1,j} + a_{j,j} r_{j,j} + a_{j,j+1} r_{j+1,j} \\ &= a_j \cdot \frac{-1}{a_{j-1} - a_j} + a_j \cdot \frac{a_{j-1} - a_{j+1}}{(a_{j-1} - a_j)(a_j - a_{j+1})} + a_{j+1} \cdot \frac{-1}{a_j - a_{j+1}} \\ &= 1.\end{aligned}$$

When $j = k$,

$$\begin{aligned}\sum_i a_{ji} r_{ij} &= a_{k,k-1} r_{k-1,k} + a_{k,k} r_{k,k} \\ &= a_k \cdot \frac{-1}{a_{k-1} - a_k} + a_k \cdot \frac{a_{k-1}}{(a_{k-1} - a_k) a_k} \\ &= 1.\end{aligned}$$

We next verify $\sum_i a_{\ell i} r_{ij} = 0$ for all $\ell \neq j$. We only verify the general case when there are three elements in one column of A^{-1}. The other verifications are the same. $\sum_i a_{\ell i} r_{ij} = a_{\ell,j-1} r_{j-1,j} + a_{\ell,j} r_{j,j} + a_{\ell,j+1} r_{j+1,j}$. Since $\ell \neq j$, there are only two situations we need to consider. (1) $\ell \leq j - 1$ and (2) $\ell \geq j + 1$.

When $\ell \leq j - 1$,

$$\begin{aligned}\sum_i a_{\ell i} r_{ij} &= a_{\ell,j-1} r_{j-1,j} + a_{\ell,j} r_{j,j} + a_{\ell,j+1} r_{j+1,j} \\ &= a_{j-1} r_{j-1,j} + a_j r_{j,j} + a_{j+1} r_{j+1,j} \\ &= 0.\end{aligned}$$

When $\ell \geq j + 1$,

$$\begin{aligned}\sum_i a_{\ell i} r_{ij} &= a_{\ell,j-1} r_{j-1,j} + a_{\ell,j} r_{j,j} + a_{\ell,j+1} r_{j+1,j} \\ &= a_\ell r_{j-1,j} + a_\ell r_{j,j} + a_\ell r_{j+1,j} \\ &= a_\ell \cdot \left[\frac{-1}{a_{j-1} - a_j} + \frac{a_{j-1} - a_{j+1}}{(a_{j-1} - a_j)(a_j - a_{j+1})} + \frac{-1}{a_j - a_{j+1}} \right] \\ &= 0. \quad \square\end{aligned}$$

A.2. Proof of Theorem 2

Proof. Note that the OLS estimate of the coefficients in the linear regression equation (14) is

$$\begin{pmatrix} \hat{b}_0 \\ \hat{b}_j \end{pmatrix} = (Z_S{}^T Z_S)^{-1} Z_S{}^T X_j, \tag{23}$$

where $Z_S = \begin{pmatrix} 1_n & X_S \end{pmatrix}$ and $\hat{b}_j \in \mathbb{R}^s$. We first attempt to find the necessary and sufficient condition such that $\|\hat{b}_j\|_1 < 1$ for all j, which is a stronger condition for (11) to hold.

We know that $Z_S^T Z_S = (t_{k\ell}) \in \mathbb{R}^{(s+1)\times(s+1)}$ with

$$t_{k\ell} = n - \max\{j_{k-1}, j_{\ell-1}\}.$$

where we assume $j_0 = 0$. According to a linear algebra result stated in Lemma 3 in the Appendix, the inverse of this matrix is a tridiagonal matrix:

$$(Z_S^T Z_S)^{-1} = \begin{bmatrix} r_{11} & r_{12} & & & & & \\ r_{21} & r_{22} & r_{23} & & & & \\ & r_{32} & r_{33} & r_{34} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & r_{s,s-1} & r_{s,s} & r_{s,s+1} \\ & & & & r_{s+1,s} & r_{s+1,s+1} \end{bmatrix}$$

where

$$r_{k\ell} = \begin{cases} \dfrac{1}{j_1} & k = \ell = 1 \\ -\dfrac{1}{j_{\ell-1} - j_{\ell-2}} & k = \ell - 1 \\ -\dfrac{1}{j_\ell - j_{\ell-1}} & k = \ell + 1 \\ \dfrac{j_\ell - j_{\ell-2}}{(j_{\ell-1} - j_{\ell-2})(j_\ell - j_{\ell-1})} & 1 < k = \ell < s+1 \\ \dfrac{n - j_{s-1}}{(j_s - j_{s-1})(n - j_s)} & k = \ell = s+1 \\ 0 & \text{otherwise.} \end{cases}$$

Denote $v = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_j \end{pmatrix} = (Z_S{}^T Z_S)^{-1} Z_S{}^T X_j$. There are three pattern types that we need to consider.

(i) If there exists $1 \le k < s$ such that $j_{k+1} - j_k \ge 2$, then for any j with $j_k < j < j_{k+1}$,

$$Z_S{}^T X_j = (\underbrace{n-j, n-j, \ldots, n-j}_{k+1}, n - j_{k+1}, n - j_{k+2}, \ldots, n - j_s)^T.$$

We have

$$v = \left(0, \ldots, 0, \frac{j_{k+1} - j}{j_{k+1} - j_k}, -\frac{j_{k+1} - j}{j_{k+1} - j_k} + 1, 0, \ldots, 0\right)^T.$$

Hence,

$$\|\hat{b}_j\|_1 = \left|\frac{j_{k+1} - j}{j_{k+1} - j_k}\right| + \left| -\frac{j_{k+1} - j}{j_{k+1} - j_k} + 1\right| = 1, \quad \text{since } j_k < j < j_{k+1}. \tag{24}$$

(ii) If $j < j_1$,

$$Z_S{}^T X_j = (n - j, n - j_1, n - j_2, \ldots, n - j_s)^T.$$

We have

$$v = \left(1 - \frac{j}{j_1}, \frac{j}{j_1}, 0, \ldots, 0\right)^T.$$

Hence, $\|\hat{b}_j\|_1 = |\frac{j}{j_1}| < 1$.

(iii) If $j > j_s$,

$$Z_S{}^T X_j = (\underbrace{n-j, \ldots, n-j}_{s+1})^T.$$

We have

$$v = \left(0, \ldots, 0, \frac{n-j}{n-j_s}\right)^T.$$

Hence, $\|\hat{b}_j\|_1 = |\frac{n-j}{n-j_s}| < 1$.

These three cases for the position of $j \in S^c$ show that as long as j is not between two jump points, $\|\hat{b}_j\|_1 < 1$. Otherwise $\|\hat{b}_j\|_1 = 1$. So

$$s = 1 \quad \text{or} \quad \max_{1 \le k < s} (j_{k+1} - j_k) = 1$$

is necessary and sufficient for all $\|\hat{b}_j\|_1 < 1, j \in S^c$.

If the above condition does not hold, that is, there are at least two jump points and all the jump points are not consecutive in indices, the irrepresentable condition (11) holds if and only if for each k such that $j_{k+1} - j_k \ge 2$, $\tilde{\theta}^*_k$ and $\tilde{\theta}^*_{k+1}$ have different signs. By the definition of $\tilde{\theta}$, we see that $(\mu^*_{j_k} - \mu^*_{j_k - 1})(\mu^*_{j_{k+1}} - \mu^*_{j_{k+1}-1}) < 0$ is equivalent to $\tilde{\theta}^*_k$ and $\tilde{\theta}^*_{k+1}$ having different signs. □

A.3. Proof of Theorem 3

To prove the non-asymptotic result for the preconditioned fused Lasso, we first state a general result for the Lasso that will be the main ingredient of our proof.

Theorem 4 (*Jia and Rohe, 2015*)**.** *Suppose that data (X, Y) follows a linear model $Y = X\beta^* + \epsilon$, where $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^{n\times 1}$, $X \in \mathbb{R}^{n\times p}$ with x_i^T as its ith row, $\beta^* \in \mathbb{R}^{p\times 1}$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^{n\times 1}$ with $\epsilon \sim N(0, \sigma^2 I_n)$. Define the singular value decomposition of X as $X = UDV'$. Suppose that $n \ge p$ and X has rank p. Let $\Lambda_{\min}(X)$ be the minimal eigenvalue of the matrix X and $C_{\min} = \Lambda_{\min}(X_S^T X_S)$, where S is the same as that in Definition 3. We further assume that the minimal eigenvalue $\Lambda_{\min}(\frac{1}{n}X'X) \ge \tilde{C}_{\min} > 0$. Define the Puffer Transformation, $F = UD^{-1}U^T$. Let $Z = FX$ and $a = FY$. Define*

$$\tilde{\beta}(\lambda) = \underset{b}{\operatorname{argmin}} \frac{1}{2}\|a - Zb\|_2^2 + \lambda\|b\|_1.$$

If $\min_{j\in S} |\beta^*_j| \ge 2\lambda$, then with probability greater than

$$1 - 2p \exp\left\{-\frac{n\lambda^2 \tilde{C}_{\min}}{2\sigma^2}\right\} \tag{25}$$

$\tilde{\beta}(\lambda) =_s \beta^*$.

The proof can be found in Jia and Rohe (2015). From the proof we see that the assumption that $\epsilon \sim N(0, \sigma^2 I_n)$ can be relaxed to $\epsilon \sim N(0, \Sigma)$ with $\max_i \Sigma_{ii} \le \sigma^2$. Note that in Theorem 4, the minimum singular value of the design matrix plays a critical role. The following lemma addresses this issue for the special design matrix of the preconditioned fused Lasso.

Lemma 4. *$\tilde{X} \in \mathbb{R}^{n\times(n-1)}$ is defined in (8). Let $\sigma_j(\cdot)$ denote the jth largest singular value of a matrix. Then*

$$\sigma_1(\tilde{X}) \ge \sigma_2(\tilde{X}) \ge \cdots \ge \sigma_{n-1}(\tilde{X}) \ge 0.5.$$

To prove Lemma 4, we need the following two results.

Lemma 5. *Let $X \in \mathbb{R}^{n\times n}$ be a lower triangular matrix with elements 1 on and below the diagonals and 0 in other places.*

$$X_{ij} = \begin{cases} 1 & i \ge j \\ 0 & i < j. \end{cases}$$

The minimal singular value is greater or equal to 0.5.

Proof. Let $X = (a_{ij}) \in \mathbb{R}^{n\times n}$ be the matrix satisfying the condition of the lemma. Note that the singular values of this matrix X are the non-negative square roots of the eigenvalues of X^TX. Hence it suffices to show that all the eigenvalues of X^TX are greater or equal to 0.25.

The explicit expression of $C_n = X^TX = (c_{ij}) \in \mathbb{R}^{n\times n}$ is

$$c_{ij} = n + 1 - \max\{i, j\}.$$

By Lemma 3, we have

$$C_n^{-1} = \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}.$$

Then for any vector $u \in \mathbb{R}^{n\times 1}$,

$$\begin{aligned} u^T C_n^{-1} u &= u_1^2 + \sum_{i=2}^{n}(2u_i^2) - 2\sum_{i=1}^{n-1} u_i u_{i+1} \\ &\le 2\sum_{i=1}^{n} u_i^2 - 2\sum_{i=1}^{n-1} u_i u_{i+1} \\ &\le 2\sum_{i=1}^{n} u_i^2 + 2\sum_{i=1}^{n-1} |u_i u_{i+1}|. \end{aligned}$$

By the fact that $\sum_{i=1}^{n-1} |u_i u_{i+1}| \le \frac{1}{2}\sum_{i=1}^{n-1}(u_i^2 + u_{i+1}^2) \le \sum_{i=1}^{n} u_i^2$, we have

$$u^T C_n^{-1} u \le 4\sum_{i=1}^{n} u_i^2,$$

which implies that the eigenvalues of C_n^{-1} are less or equal to 4 and thus the eigenvalues of C_n are all greater or equal to 0.25. □

The following lemma states the relationship between eigenvalues of second moments for centered and non-centered data. Let $X \in \mathbb{R}^{n\times p}$ be a data matrix. Define the (empirical) covariance matrix of X to be

$$S = \frac{X_c' X_c}{n},$$

where X_c is the centered version of X with the jth column of X_c be $X_j - \bar{X}_j$. Let the second moments of the dataset X be

$$T = \frac{X'X}{n}.$$

Then the eigenvalues of S and T have the following property.

Lemma 6 (*Cadima and Jolliffe, 2009*)**.** *Let S be the covariance matrix for a given dataset, and T its corresponding matrix of non-central second moments. Let $\lambda_j(\cdot)$ be the jth largest eigenvalue of a matrix. Then*

$$\lambda_{j+1}(T) \le \lambda_j(S) \le \lambda_j(T).$$

Lemma 6 can be found on page 5 in Cadima and Jolliffe (2009).
With the results from Lemmas 5 and 6, we now prove Lemma 4.

Proof. Let $X \in \mathbb{R}^{n\times n}$ be a lower triangular matrix with elements 1 on and below the diagonals and 0 in other places.

$$X_{ij} = \begin{cases} 1 & i \ge j \\ 0 & i < j. \end{cases}$$

Let $\sigma_j(\cdot)$ denote the jth largest singular value of a matrix. By Lemma 5, the smallest singular value is not less than 0.5, that is $\sigma_j(X) \ge 0.5, \forall j = 1, \ldots, n$. Now let X_c be the centered version of X, then $X_c = [\mathbf{0}, \tilde{X}]$, where $\mathbf{0}$ is a column vector with all elements 0, and $\tilde{X} \in \mathbb{R}^{n\times(n-1)}$ as defined in Eq. (8). Let $\sigma_j(\cdot)$ denote the jth largest singular value of a matrix. By Lemma 6, we have

$$\sigma_{j+1}(X) \le \sigma_j(X_c) \le \sigma_j(X), \quad \forall j = 1, 2, \ldots, n-1.$$

In particular, take $j = n-1$ in the above inequalities and we have $\sigma_{n-1}(X_c) \ge \sigma_n(X) \ge 0.5$. Since X_c is singular, the minimal singular value $\sigma_n(X_c) = 0$. Therefore,

$$\sigma_{n-1}(\tilde{X}) = \sigma_{n-1}(X_c) \ge 0.5. \quad \square$$

With the above tools, we can now prove Theorem 3.

Proof. By (10),

$$\tilde{Y} = \tilde{X}\tilde{\theta}^* + \tilde{\epsilon},$$

where $\tilde{\epsilon} = \epsilon - \bar{\epsilon}$ with $E(\tilde{\epsilon}) = 0$ and

$$var(\tilde{\epsilon}_i) = var(\epsilon_i - \bar{\epsilon}) = \frac{n-1}{n}\sigma^2 \le \sigma^2.$$

According to the comments below Theorem 4, we can apply Theorem 4 to have a lower bound on $P(\tilde{\beta}(\lambda) =_s \tilde{\theta}^*)$. Let $s_1 \leq s_2 \leq \cdots \leq s_n$ be the singular values of $\tilde{X}$. From Lemma 4 in the Appendix, $s_1 \geq 0.5$. So $\Lambda_{min}(\frac{1}{n}\tilde{X}'\tilde{X}) = \frac{s_1^2}{n} \geq \frac{1}{4n}$. Put $\tilde{C}_{\min} = \frac{1}{4n}$ in expression (25) and note that $\tilde{X}$ has $n-1$ columns, we have

$$P(\tilde{\beta}(\lambda) =_s \tilde{\theta}^*) \geq 1 - 2(n-1)\exp\left\{-\frac{\lambda^2}{8\sigma^2}\right\} \geq 1 - 2n\exp\left\{-\frac{\lambda^2}{8\sigma^2}\right\}. \quad \square$$

References

Cadima, J., Jolliffe, I., 2009. On relationships between uncentred and column-centred principal component analysis. Pak. J. Stat. 25 (4), 473–503.
Cai, J., Candes, E., Shen, Z., 2010. A singular value thresholding algorithm for matrix completion. SIAM J. Optim. 20 (4), 1956–1982.
Donoho, D.L., Elad, M., Temlyakov, V.N., 2006. Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inf. Theory 52 (1), 6–18.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.
Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., 2007. Pathwise coordinate optimization. Ann. Appl. Stat. 1 (2), 302–332.
Greenshtein, E., Ritov, Y.A., 2004. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. Bernoulli 10 (6), 971–988.
Harchaoui, Zaıd, Lévy-Leduc, Céline, 2008. Catching change-points with Lasso. Adv. Neural Inf. Process. Syst. 20 (161–168), 18.
Harchaoui, Zaıd, Lévy-Leduc, Céline, 2010. Multiple change-point estimation with a total variation penalty. J. Amer. Statist. Assoc. 105 (492), 1480–1493.
Hoefling, H., 2010. A path algorithm for the fused lasso signal approximator. J. Comput. Graph. Statist. 19 (4), 984–1006.
Jia, J., Rohe, K., 2015. Preconditioning the lasso for sign consistency. Electron. J. Stat. 9, 1150–1172.
Jia, J., Rohe, K., Yu, B., 2013. The lasso under heteroscedasticity. Statist. Sinica 23, 99–118.
Knight, K., Fu, W., 2000. Asymptotics for lasso-type estimators. Ann. Statist. 1356–1378.
Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. Ann. Statist. 34 (3), 1436–1462.
Rinaldo, A., 2009. Properties and refinements of the fused lasso. Ann. Statist. 37 (5B), 2922–2952.
Sharpnack, James, Rinaldo, Alessandro, Singh, Aarti, 2012. Sparsistency of the edge lasso over graphs. In: International Conference on Artificial Intelligence and Statistics (AISTATS, JMLR: W&CP), volume 22, pp. 1028–1036.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 267–288.
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2004. Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (1), 91–108.
Tibshirani, Ryan J., Taylor, Jonathan, 2011. The solution path of the generalized lasso. Ann. Statist. 39 (3), 1335–1371.
Tibshirani, R., Wang, P., 2008. Spatial smoothing and hot spot detection for cgh data using the fused lasso. Biostatistics 9 (1), 18–29.
Tropp, J.A., 2006. Just relax: Convex programming methods for identifying sparse signals in noise. IEEE Trans. Inf. Theory 52 (3), 1030–1051.
Wainwright, M.J., 2009. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. IEEE Trans. Inf. Theory 55 (12), 5728–5741.
Xu, C., Liu, T., Tao, D., Xu, C., 2014. Local rademacher complexity for multi-label learning. arXiv preprint arXiv:1410.6990.
Xu, C., Tao, D., Xu, C., 2015. Multi-view Intact Space Learning. IEEE Trans. Pattern Anal. Mach. Intell. 1 (1), 1.
Zhang, C.H., Huang, J., 2008. The sparsity and bias of the lasso selection in high-dimensional linear regression. Ann. Statist. 36 (4), 1567–1594.
Zhao, P., Yu, B., 2006. On model selection consistency of lasso. J. Mach. Learn. Res. 7 (2), 2541.