ON PATTERN RECOVERY OF THE FUSED LASSO [*]

By Junyang Qian [†,‡] and Jinzhu Jia[†]

Peking University [†] *and University of Pennsylvania* [‡]

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. By studying the property of the design matrix in the transformed Lasso problem, we find that the irrepresentable condition might not hold, in which case we show that the FLSA might bot be able to recover the signal pattern. We then apply the newly developed preconditioning method – Puffer Transformation [Jia and Rohe, 2012] on the transformed Lasso problem. We call the new method the *preconditioned fused Lasso* and we give non-asymptotic results for this method. Results show that when the signal jump strength (signal difference between two neighboring groups) is big and the noise level is small, our preconditioned fused Lasso estimator gives the correct pattern with high probability. Theoretical results give insight on what controls the signal pattern recovery ability – it is the noise level instead of the length of the sequence. Simulations confirm our theorems and show significant improvement of the preconditioned fused Lasso estimator over the vanilla FLSA.

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

$$(1) \qquad y_i = \mu_i^* + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

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, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is the white noise that is assumed to be i.i.d. and each has a normal distribution with mean 0 and variance σ^2. The model is assumed to be sparse in the sense that the signals come in blocks and only a few of the blocks are nonzero. To be exact, there exists a partition of $\{1, 2, \ldots, n\} = \cup_{j=1}^J [L_j, U_j]$ with $L_1 = 1, U_J = n, U_j \geq L_j, L_{j+1} = U_j + 1$, and the following stepwise function holds:

$$\mu_i^* = \nu_j^* \mathbf{1}_{L_j \leq \mu_i \leq U_j},$$

with ν_j^*, L_j, U_j fixed but unknown. We also assume that the vector $\nu = (\nu_1, \nu_2, \ldots, \nu_n)$ is sparse, meaning that only a few of ν_j's are nonzeros. We point out that the Gaussian noise

1

is not necessary. But we still use it to study the 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. For a lot of real data problem, each observation of y_i can be an average of multiple measurements and so σ^2 decreases when the number of measurements increases.

This model featured by blockiness and sparseness 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 "signals" as the log-ratio between the number of DNA copies in the tumor cells and that in the reference cells [Tibshirani and Wang, 2008].

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]:

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

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 of the signal and the total variation seminorm ($\|\mu\|_{TV}$) regularization controls the number of blocks (partitions or groups).

Figure 1 gives one example of the signal sequence and the FLSA estimate on CGH data. More details and examples can be seen in Tibshirani and Wang [2008].
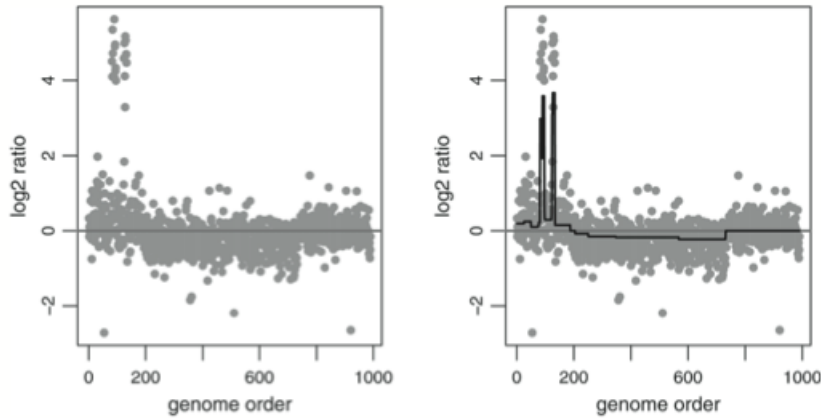

FIG 1. *This figure is from Tibshirani and Wang [2008]. The fused Lasso is applied to some CGH data. The data are shown in the left panel, and the solid line in the right panel represents the estimated signals by the fused Lasso. The horizontal line is for $y = 0$.*

One important question for the FLSA is how good the estimator defined in Equation (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 exists λ_{1n} and λ_{2n}, such that*

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

We use $\hat{\mu} =_{js} \mu$ to shortly denote Equation (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 jump directions (up or down) between groups.

The concept of pattern recovery is very similar to sign recovery of the Lasso. In fact, some simple calculations in Section 2 tell us that the pattern recovery property of the FLSA can be transformed to the sign recovery property of the Lasso estimator.

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 is defined as follows [Tibshirani, 1996].

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

Equivalently, in matrix form,

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

where $Y = (y_1, \ldots, y_n)^T$ 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.

Sign Recovery of the Lasso estimator is defined as follows.

DEFINITION 2 (Sign Recovery). *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*

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

We will use $\hat{\beta} =_s \beta^*$ to shortly denote $sign(\hat{\beta}_j(\lambda_n)) = sign(\beta_j^*)$, $j = 1, \ldots, p$. The Lasso estimator with the sign recovery property implies that it selects the correct set of predictors. If $P(\hat{\beta}(\lambda_n) =_s \beta) \to 1$, as the sample size $n \to \infty$, we say that $\hat{\beta}(\lambda_n)$ is sign consistent.

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 and the estimation performance will not be improved by increasing the sample size.

Our study of the pattern recovery of the FLSA begins with a transformation that changes the FLSA to a special Lasso problem. The data defined in the transformed Lasso problem has correlated noise terms instead of independent ones. We prove that even for the linear model with correlated noise, the irrepresentable condition is still necessary for sign consistency. We then analyze the property of the design matrix in the transformed Lasso problem. We give necessary and sufficient condition such that the design matrix in the transformed Lasso problem satisfies the irrepresentable condition. We show that, 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 [2012] shows that a Puffer Transformation 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 improves the FLSA and recovers the signal pattern with high probability.

In Rinaldo [2009], the author also considers the consistency conditions for the FLSA. They showed that under some conditions, the FLSA can be consistent both in block reconstruction and model selection. The author says in Rinaldo [2009] that the asymptotic results may have little guidance to the practical performance when n is finite. However, our method, as we will see, can not only provide mild conditions for the estimator to be consistent in block recovery but also give an explicit non-asymptotic lower bound on the probability that the true blocks are recovered. Numerical simulations also illustrate that in many cases our method turns out to be more effective in block recovery.

The rest of the paper is organized as follows. In Section 2, we transform the FLSA problem into a Lasso problem and analyze the property of the design matrix in the transformed Lasso problem. Section 3 illustrates 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 the technique of Puffer Transformation (defined in Equation (20)). We show that for a wide range 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. Some proofs are given in the appendix.

2. FLSA and the Lasso.   We turn 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 Equation (2). We have the following result.

LEMMA 1.
$$\hat{\mu}(\lambda_1, \lambda_2) = SH_{\lambda_1}(\hat{\mu}(0, \lambda_2))$$

The proof of Lemma 1 can be found in Friedman et al. [2007]. From Lemma 1, to study the property of $\hat{\mu}(\lambda_1, \lambda_2)$, we can set $\lambda_1 = 0$ first. In the whole paper, since pattern recovery is our main concern, so we only consider the case when $\lambda_1 = 0$. When $\lambda_1 = 0$, we can solve the FLSA by change of variables. Let $\theta_1 = \mu_1, \theta_i = \mu_i - \mu_{i-1}, i = 2, \ldots, n$. In matrix form, we have $\mu = A\theta$, with

$$(6) \qquad A_{n \times n} = \begin{pmatrix} 1 & 0 & \ldots & \ldots & \ldots & 0 \\ 1 & 1 & 0 & \ldots & \ldots & 0 \\ 1 & 1 & 1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 1 & 1 & \ldots & \ldots & 1 \end{pmatrix}$$

So by using θ instead of μ, we have an equivalent solution of $\hat{\mu}(0, \lambda_2)$ via the following $\hat{\theta}(\lambda_2)$.

$$(7) \qquad \hat{\theta}(\lambda_2) = \underset{\theta}{\operatorname{argmin}} \frac{1}{2}\|Y - A\theta\|_2^2 + \lambda_2\|\tilde{\theta}\|_1,$$

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 $\hat{\mu}(0, \lambda_2) = A\hat{\theta}(\lambda_2)$. Notice the special form of the design matrix A, Expression (7) is a Lasso problem with interception. In fact, Expression (7) can be rewritten as

$$(8) \qquad \hat{\theta}(\lambda_2) = \underset{(\theta_1, \tilde{\theta})}{\operatorname{argmin}} \frac{1}{2}\|Y - \theta_1 - X\tilde{\theta}\|_2^2 + \lambda_2\|\tilde{\theta}\|_1.$$

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

$$(9) \qquad x_{ij} = \begin{cases} 1 & i > j \\ 0 & i \le j. \end{cases}$$

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

$$(10) \qquad \tilde{X} = [X_1 - \bar{X}_1, \ldots, X_{n-1} - \bar{X}_{n-1}] \text{ and } \tilde{Y} = Y - \bar{Y}$$

with $\bar{u}$ being the average of the vector u. It is easy to see that Expression (8) is equivalent to the following standard Lasso problem without interception.

$$(11) \qquad \hat{\tilde{\theta}}(\lambda_2) = \underset{\tilde{\theta}}{\operatorname{argmin}} \frac{1}{2} \|\tilde{Y} - \tilde{X}\tilde{\theta}\|_2^2 + \lambda_2 \|\tilde{\theta}\|_1, \text{ and } \hat{\theta}_1(\lambda_2) = \bar{Y} - \bar{X}\hat{\tilde{\theta}}(\lambda_2).$$

Since the observation $Y = (y_1, \ldots, y_n)$ follows the model defined in Equation (1). Define $\theta^* = A^{-1}\mu^*$, (equivalently, $\theta_1^* = \mu_1^*, \theta_i^* = \mu_i^* - \mu_{i-1}^*, i = 2, \ldots, n$), where A is defined in Equation (6). Let $\tilde{\theta}^* \in \mathbb{R}^{n-1} = (\theta_2^*, \theta_3^*, \ldots, \theta_n^*)^T$. We have that (X, Y) satisfy the following linear model:

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

where X is defined at Equation (9). Consequently the centered version of (X, Y) satisfy the following linear model:

$$(12) \qquad \tilde{Y} = \tilde{X}\tilde{\theta}^* + \tilde{\epsilon},$$

where $\tilde{\epsilon} = \epsilon - \bar{\epsilon}$ with $E(\tilde{\epsilon}) = 0$. Now we see that $\hat{\tilde{\theta}}(\lambda_2)$ defined at (11) has the sign recovery property if and only if $\hat{\tilde{\theta}}(\lambda_2) =_s \tilde{\theta}^*$. By the relationship between μ and θ, $\hat{\tilde{\theta}}(\lambda_2) =_s \tilde{\theta}^*$ is equivalent to $\hat{\mu}(0, \lambda_2) =_{js} \mu^*$. In other words, the pattern recovery property of an FLSA can be viewed as sign recovery of a Lasso estimator.

PROPERTY 1. *The pattern recovery of the FLSA $\hat{\mu}(0, \lambda_2)$ defined in Equation (2) is equivalent to the sign consistency of the the Lasso estimator $\hat{\tilde{\theta}}(\lambda_2)$ defined in Equation (11).*

Note that this change of variables serves mainly for theoretical analysis rather than computational facilitation. Although there are many mature algorithms for the Lasso, transforming the FLSA to the Lasso is not recommended in practice because it makes the design matrix in (11) much more dense, which is unfavorable to the efficiency of computation. Instead, Friedman et al. [2007] develops specialized algorithm for the FLSA based on the coordinate-wise descent. Hoefling [2010] generalizes the path algorithm and extends it to the general fused Lasso problem. However, in our consistency analysis, this transformation works since we can use the well understood techniques on the Lasso to analyze the theoretical properties of the FLSA.

We now turn to analyze the Lasso problem defined in Equation (11).

3. The Transformed Lasso. 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 defined as follows.

DEFINITION 3 (Irrepresentable Condition). *The design matrix X satisfies the **Irrepresentable Condition** for β^* with support $S = \{j : \beta_j^* \neq 0\}$ if, for some $\eta \in (0, 1]$,*

$$(13) \qquad \left\| X_{S^c}^T X_S \left(X_S^T X_S \right)^{-1} sign(\beta_S^*) \right\|_\infty \leq 1 - \eta,$$

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

Let $\Lambda_{\min}(X)$ be the minimal eigenvalue of the matrix X and

$$C_{\min} = \Lambda_{\min}\left(X_S^T X_S \right) > 0.$$

Define

$$\Psi(\mathbf{X}, \beta^*, \lambda) = \lambda \left[\frac{\eta}{\sqrt{C_{\min}} \max_{j \in S^c} \|X_j\|_2} + \left\| \left(X_S^T X_S \right)^{-1} sign(\beta_S^*) \right\|_\infty \right].$$

With the above notation, we have a general non-asymptotic result for the sign recovery of the Lasso when data (X, Y) follow a linear model.

THEOREM 1. *Suppose that data $(\mathbf{X}, Y)$ follow a linear model $Y = \mathbf{X}\beta^* + \epsilon$, 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 $\epsilon \sim N(0, \Sigma_\epsilon)$. Assume that the irrepresentable condition (13) holds. If λ satisfies*

$$M(\beta^*) > \Psi(\mathbf{X}, \beta^*, \lambda),$$

then with probability greater than

$$1 - 2p \exp\left\{ -\frac{\lambda^2 \eta^2}{2[\Lambda_{\max}(\Sigma_\epsilon) \max_{j \in S^c} \|X_j\|_2^2]} \right\},$$

the Lasso has a unique solution $\hat{\beta}(\lambda)$ with $\hat{\beta}(\lambda) =_s \beta^*$.

The proof of Theorem 1 is very similar to that of Lemma 3 in Jia and Rohe [2012] (pp. 24). The only difference is that in Jia and Rohe [2012], they scale each column of X to be bounded with $\|X_j\|_2 \leq 1$. Here we do not have any assumption for the ℓ_2 norm of X_j. If we further have the assumption that $\|X_j\|_2 \leq 1$ for each j, then we have exactly the same result as in Jia and Rohe [2012]. So we omit the proof for Theorem 1 .

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., 2010]. We also state this conclusion under a more general linear model with correlated noise terms.

THEOREM 2. *Suppose that data* $(\mathbf{X}, Y)$ *follow a linear model* $Y = \mathbf{X}\beta^* + \epsilon$, *with Gaussian noise* $\epsilon \sim N(0, \Sigma_\epsilon)$. *The irrepresentable condition* (13) *is necessary for the sign consistency of the Lasso. In other words, if*

$$(14) \qquad \left\| X_{S^c}^T X_S \left(X_S^T X_S \right)^{-1} sign(\beta_S^*) \right\|_\infty \geq 1,$$

we have

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

A proof of Theorem 2 can be seen in the appendix. Theorem 2 says that if the irrepresentable condition does not hold, it is very likely that the Lasso does not recover signs of the coefficients.

With the above theorem, we now come back to the transformed Lasso problem defined in Equation (11) and examine if the irrepresentable condition holds or not in this case. Recall that for the Lasso problem transformed from the FLSA, we have the design matrix

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

Denote $S = \{j : \tilde{\theta}_j^* \neq 0\}$ as the index set of the relevant variables in the true model. Let j be the index of any of the irrelevant variables. Then (13) can be written as

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

which is equivalent to

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

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

$$(15) \qquad \tilde{X}_j = b_j^T \tilde{X}_S + \epsilon.$$

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:

$$(16) \qquad X_j = b_0 + b_j^T X_S + \epsilon,$$

where $b_0 \in \mathbb{R}$ is the intercept term.

A stronger version of irrepresentable condition is as follows

$$(17) \qquad \left\| \tilde{X}_j^T \tilde{X}_S (\tilde{X}_S^T \tilde{X}_S)^{-1} \right\|_1 < 1, \forall j \notin S.$$

If (17) holds, then for any μ^* (equivalently, for any $\tilde{\theta}^*$) the irrepresentable condition always holds. Otherwiese, if (17) does not hold, then there exists some $\tilde{\theta}^*$ such that the irrepresentable condition fails to hold. We have a necessary and sufficient condition on μ^* such that the stronger version of the irrepresentale condition (17) holds.

THEOREM 3. *Assume $y = (y_1, \ldots, y_n)$ satisfies model* (1), *the collection of the indexes of jump points is $S = \{j_1, j_2, \ldots, j_s\}$ with $j_k (1 \leq k \leq s)$ increasing. Formally, $S = \{j : \mu_j^* \neq \mu_{j-1}^*, j = 2, \ldots, n\}$. Then the stronger version of the irrepresentable condition* (17) *holds if and only if the jump points are consecutive. That is, $s = 1$ or*

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

PROOF. Note that the OLS estimate of the coefficients in the linear regression equation (16) is

$$(18) \qquad \begin{pmatrix} \hat{b}_0 \\ \hat{b}_j \end{pmatrix} = (Z_S^T Z_S)^{-1} Z_S^T X_j,$$

where $Z_S = \begin{pmatrix} 1_n & X_S \end{pmatrix}$. 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 4 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

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

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 $1 \leq k < s$ such that $j_{k+1} - j_k \geq 2$, 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
$$v = \big(0, \ldots, 0, \frac{j_{k+1} - j}{j_{k+1} - j_k}, -\frac{j_{k+1} - j}{j_{k+1} - j_k} + 1, 0, \ldots, 0\big)^T.$$

Hence,

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

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

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

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 $\|\hat{b}_j\|_1 < 1$, $j \in S^c$. $\square$

The above theorem shows that only a few special structures on μ^* make the stronger version of the irrepresentable condition hold. From the proof, we can propose a necessary and sufficient condition for the irrepresentable condition.

THEOREM 4.   *Assume* $y = (y_1, \ldots, y_n)$ *satisfies model* (1), *the collection of the indexes of jump points are* $S = \{j_1, j_2, \ldots, j_s\}$ *with* $j_k(1 \le k \le s)$ *increasing. Formally,* $S = \{j : \mu_j^* \ne \mu_{j-1}^*, j = 2, \ldots, n\}$. *Then the irrepresentable condition* (13) *holds if and only if one of the following two conditions holds.*

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

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

(2) If there exists one group of data points (with more than 1 point) between some two jump points and these data point 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 $\mu^*_{j_k} = \ldots = \mu^*_{j_{k+1}-1}$, then $(\mu^*_{j_k} - \mu^*_{j_k-1})(\mu^*_{j_{k+1}} - \mu^*_{j_{k+1}-1}) < 0$.

PROOF. From Theorem 3, if condition (1) in Theorem 4 holds, a stronger version of the irrepresentable condition holds and thus the irrepresentable condition (13) holds. If condition (1) does not hold, then there exists two jump points j_k and j_{k+1} such that $j_{k+1} \geq j_k + 2$ and $\mu^*_{j_k} = \ldots = \mu^*_{j_{k+1}-1}$. From Equation (19) in the proof of Theorem 3, we see that the irrepresentable condition (13) holds if and only if $\tilde{\theta}^*_k$ 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. $\square$

Theorem 4 says that only a few configurations of μ^* make the irrepresentable condition hold. In practice, a lot of signal patterns do not satisfy either of the two conditions listed in Theorem 4. For the Lasso problem, to comply with the irrepresentable condition, Jia and Rohe [2012] proposed a Puffer Transformation. We now introduce the Puffer Transformation and apply it to solve the fused Lasso problem, which we call the preconditioned fused Lasso.

4. Preconditioned Fused Lasso. Jia and Rohe [2012] introduces 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 $\mathbf{X} \in \mathbb{R}^{n \times p}$ 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 $\mathbf{X} = UDV'$. Define the **Puffer Transformation** [Jia and Rohe, 2012],

$$(20) \qquad F_{n \times n} = UD^{-1}U^T.$$

The preconditioned design matrix $F\mathbf{X}$ has the same singular vectors as $\mathbf{X}$. However, all of the nonzero singular values of $F\mathbf{X}$ are set to unity: $F\mathbf{X} = UV'$. When $n \geq p$, the columns of $F\mathbf{X}$ are orthonormal. When $n \leq p$, the rows of $F\mathbf{X}$ are orthonormal. Jia and Rohe [2012] has a non-asymptotic result for the Lasso on $(F\mathbf{X}, FY)$ stated as follows.

THEOREM 5 (Jia and Rohe [2012]). *Suppose that data $(\mathbf{X}, Y)$ follow a linear model $Y = \mathbf{X}\beta^* + \epsilon$, 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 $\mathbf{X}$ as $\mathbf{X} = UDV'$. Suppose that $n \geq p$ and $\mathbf{X}$ has rank p. We further assume that the minimal eigenvalue $\Lambda_{\min}(\frac{1}{n}X'X) \geq \tilde{C}_{\min} > 0$. Define the Puffer Transformation, $F = UD^{-1}U^T$. Let $Z = F\mathbf{X}$ and $a = FY$. Define*

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

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

(21) $$1 - 2p \exp\left\{-\frac{n\lambda^2 \tilde{C}_{\min}}{2\sigma^2}\right\}$$

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

The proof of Theorem 5 can be found in Jia and Rohe [2012]. 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 $\max_i \Sigma_{ii} \leq \sigma^2$. Compare Theorem 5 to Theorem 1, we see that with the Puffer Transformation, the Lasso does not need the irrepresentable condtion any more.

The FLSA problem 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. First we have the following result on the singular values of $\tilde{X}$.

LEMMA 2. $\tilde{X} \in \mathbb{R}^{n \times (n-1)}$ is defined in Equation (10). Let $\sigma_j(\cdot)$ denote the j-th largest singular value of a matrix. Then

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

A proof of Lemma 2 can be found in the appendix. With the lower bound on singular values of $\tilde{X}$ and applying Theorem 5, we have the following result for our preconditioned fused Lasso.

THEOREM 6. Assume $y = (y_1, \ldots, y_n)$ satisfies model (1). $\tilde{X}$ and $\tilde{Y}$ are defined in Equation (10). 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 in Equation (6). 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 $\tilde{X} = UDV'$. Denote the Puffer Transformation, $F = UD^{-1}U^T$. Let $Z = F\tilde{X}$ and $a = F\tilde{Y}$. Define

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

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

$$1 - 2n \exp\left\{-\frac{\lambda^2}{8\sigma^2}\right\}$$

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

PROOF. By Equation (12)

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

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

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

According to the comments below Theorem 5, we can apply Theorem 5 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 2, $s_1 \geq 0.5$. 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 (21) and note that $\tilde{X}$ has $n-1$ columns, we have

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

$\square$

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.

(23) $$\hat{\mu}^* = A\hat{\theta}^*$$

with

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

Theorem 6 shows that the ability of pattern recovery 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 Equation (23) 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 m goes to ∞.

In the next section, we use simulations to illustrate that for general signal patterns, the FLSA does not have the pattern recovery property while the preconditioned fused Lasso has, which enhances our findings.

5. Simulations. We use simulation examples to confirm our theorems. We first set the model to be

(24) $$y_i = \mu_i^* + \epsilon_i,$$

where

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

and the errors are i.i.d. Gaussian variables with mean 0 and standard deviation $\sigma = 0.25$. This one is similar to the example in Rinaldo [2009] except that the noise here is larger ($\sigma = 0.2$ in Rinaldo [2009]). Figure 2 shows one sequence of sample data (points) along with the true expected signal (lines).
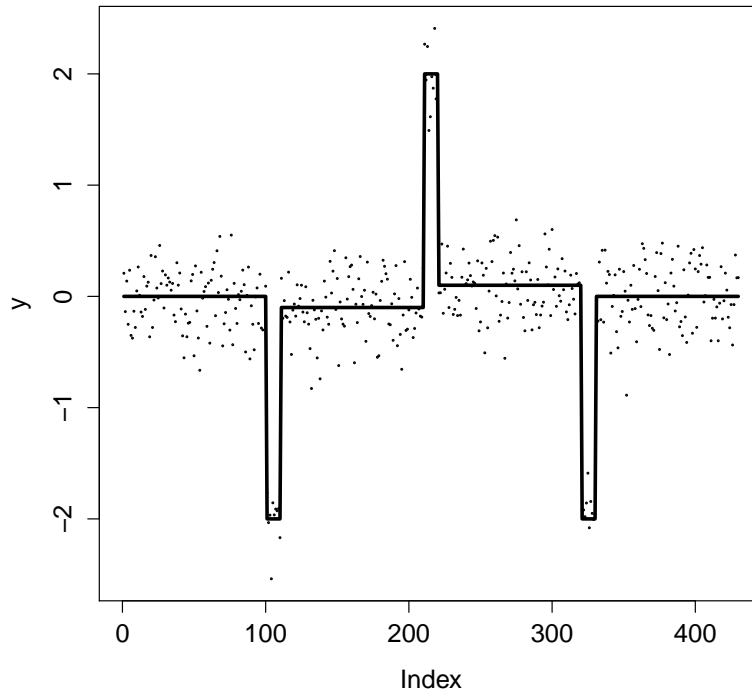

FIG 2. *Sample data (points) and the expected signals (lines).*

From Figure 2 we see that the data points are grouped into seven clusters and featured by three spikes. The points can be well separated due to small noise. We will use this

typical example to compare the performances of the two methods, the FLSA and the preconditioned fused Lasso, in recovering the signal patterns. There are many criteria that can be used in comparison. In the context of pattern recovery, it is natural to define a loss function, which we call the *pattern loss* (ℓ_{PA}) of the recovered sequence of signals $\hat{\mu}^* = (\hat{\mu}_1^*, \hat{\mu}_2^*, \ldots, \hat{\mu}_n^*) \in \mathbb{R}^n$ as follows:

$$\ell_{PA}(\hat{\mu}^*) = |\{i : \text{sign}(\hat{\mu}_{i+1}^* - \hat{\mu}_i^*) \neq \text{sign}(\mu_{i+1}^* - \mu_i^*), i = 1, \ldots, n-1\}|$$

where $\mu^* = (\mu_1^*, \mu_2^*, \ldots, \mu_n^*)$ is the expected signals and $|\cdot|$ the cardinality of a set. Note that the pattern loss achieves 0 if and only if the pattern of the signals is recovered exactly. We compare the solutions under the two methods (FLSA and preconditioned fused lasso). For each method, the solution chosen is the one that minimizes the pattern loss on the solution path.

We first apply the FLSA to estimate $\mu_i^*, i = 1, 2, \ldots, 430$. When calculating the FLSA solution, we use a path algorithm proposed by Hoefling [2010] which is very efficient to give the whole 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 whole FLSA solution path is piecewise linear in λ. "flsa" only stores the sequence of λ's when the direction of the linear function changes. Note that the pattern loss does not change with λ on every linear piece of the solution path. By comparing the signal pattern of all the estimated signals on the solution path with the true signals μ^*, we see that there is no one solution that recovers the original signal pattern. That is, all the FLSA solutions have a positive pattern loss. We present in Figure 3 (left panel) the solution that minimizes such loss. We see that this estimate is just the trivial estimate that averages all the signals, which obviously does not give satisfactory recovery of signal patterns.

For each λ in the sequence, we also calculate the common ℓ_2 distance between the estimated signals μ^* and the true ones. The estimate with the smallest ℓ_2 distance is reported in Figure 3 (right panel). We see that for this estimate, it does not recover the original signal pattern either.

To compare, we calculate the solution of the Lasso defined in Equation (22). After the SVD and the Puffer Transformation, this becomes much easier. We only need to do a soft-thresholding with the given λ. This is because

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

and the property of the Lasso allows us to solve it directly by soft-thresholding

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

Obviously, $\hat{b}(\lambda)$ is also a piecewise linear function in λ and the break points are $\lambda_i = |Z^T a|_{(i)}, i = 1, 2, \ldots, n$, where $x_{(i)}$ denotes the ith largest value in vector x and n is the dimension of vector $Z^T a$. On the solution path, for each λ_i we have an estimate $\hat{\mu}$ for
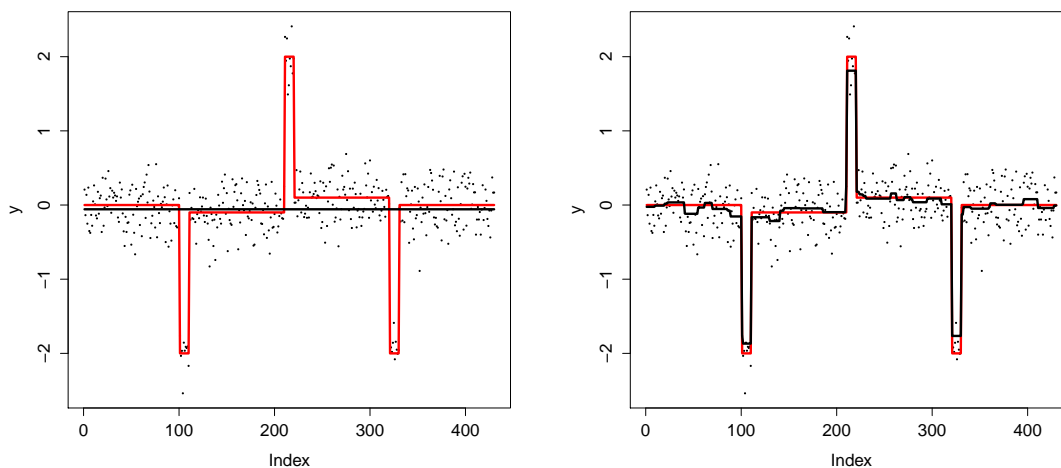
FIG 3. *FLSA solutions. Left panel: the FLSA solution(in black lines) with tuning parameter selected by minimizing the pattern loss of $\hat{\mu}^*$; Right panel: the FLSA solution(in black lines) with tuning parameter selected by minimizing the ℓ_2 error between $\hat{\mu}^*$ and μ^*. The red lines are the expected signal sequence.*

μ^*. By examining the pattern loss of the solutions on the path, we find that there is one solution (in fact any solution on that linear part between the one chosen and the one at the previous breakpoint) has 0 loss, which means it recovers the signal pattern exactly. We report this solution in Figure 4.

Note that the reported estimate 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.

The above example just gives one data set to compare the performances of the FLSA and preconditioned fused Lasso. We now randomly draw 1000 datasets and compare the approximate probability (denoted as P) that there exists a λ such that the pattern of the signals can be completely recovered. The results are as follows.

- FLSA: $P \approx 0$.
- Preconditioned Fused Lasso $P \approx 0.926$.

This example again illustrates the strength of our algorithm in pattern recovery of blocky signals. Nevertheless, as intuitively, it loses power like other recovering algorithms when the noise level becomes stronger and makes it difficult to tell the boundaries between the blocks. Our theorem also reflects this relationship between recovery probability and noise level. In the next example, we change σ from 0.1 to 0.4 and compute the probability of
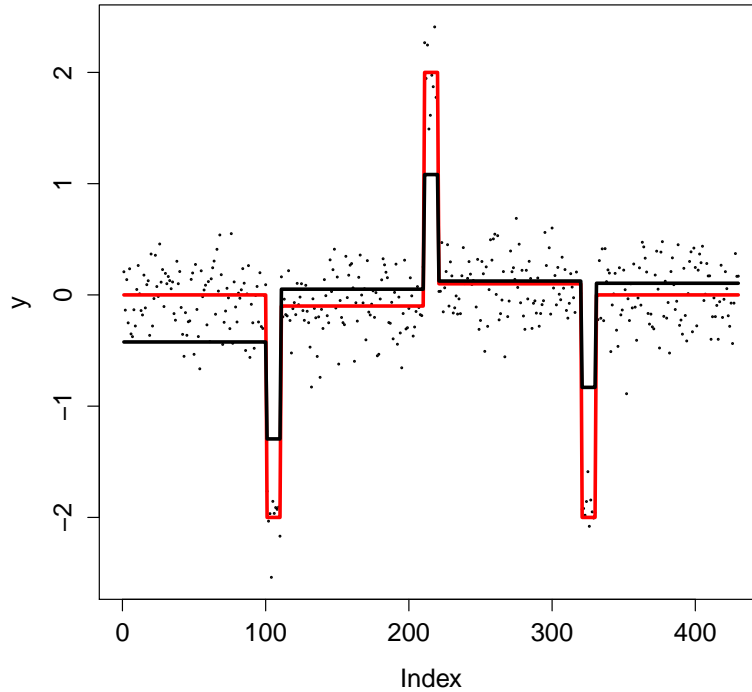
FIG 4. *The preconditioned fused Lasso estimator (in black lines). Tuning parameter is selected by finding a solution which exactly recovers the signal pattern. The red line is the expected signal sequence.*

pattern recovery. For each σ, we randomly draw 1000 datasets following model described in Equation (24) and obtain the estimated probability via the proportion that there exists a λ such that the pattern of the signals can be completely recovered. The estimated probabilities are reported in Figure 5, from which we see that the probability of pattern recovery under small noise is extremely high but this cannot hold when the signals are corrupted by stronger noise, which makes the boundaries between groups vague and hard to distinguish.

6. Conclusions and Discussions. In this paper we provided more understanding of the FLSA and shed some light on the insight of 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 re-
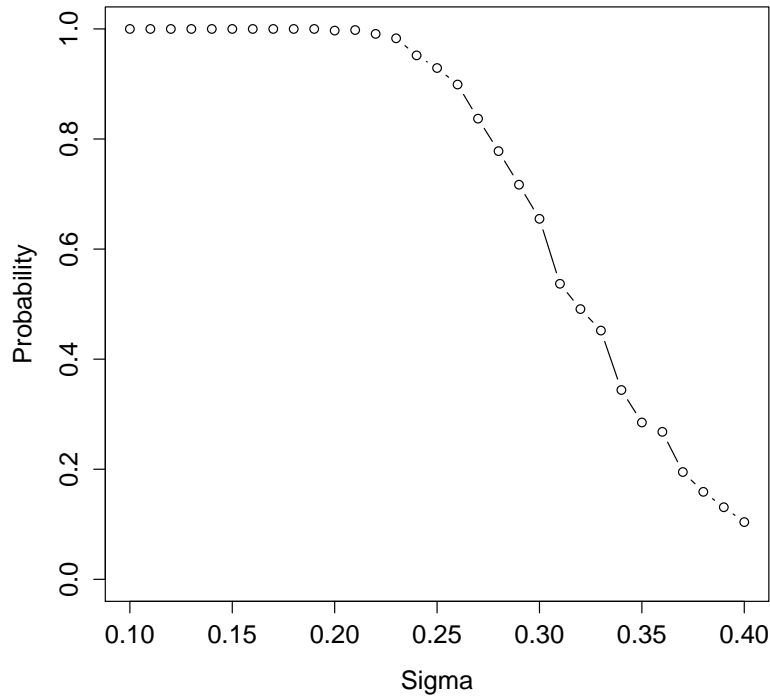
FIG 5. *The estimated probability of pattern recovery under different noise level for the preconditioned fused Lasso. Each point is estimated with* 1000 *randomly generated datasets.*

sults on the preconditioned fused Lasso. The result implies that when the noise is weak, the preconditioned fused Lasso can recover the signal pattern with high probability. We also found that the preconditioned fused Lasso is sensitive to the scale of the noise level. Simulation studies confirmed our findings.

One may argue that we only considered the pattern recovery using the fusion regularization parameter λ_2 and that the sparsity regularization parameter λ_1 can be used to adjust to the right pattern. However, remember that the main purpose of introducing λ_1 is for the sparsity of the model. It is not statistically reasonable to use this regularization parameter only to recover the blocks.

If considered in the context of both sparsity and block recovery, this is impossible in most time. Using the example above, we claim that the pattern and sparsity cannot be recovered at the same time by FLSA. We know from Friedman et al. [2007] that as long as two parameters are fused together for some $\lambda_2^{(0)}$, it will be fused for all $\lambda_2 > \lambda_2^{(0)}$. This

implies inversely that if two $\hat{\mu}_i's$ are partitioned for some big λ_2, they will not be fused for all $\tilde{\lambda}_2 < \lambda_2$. Let us focus on the first partition as λ_2 decreases from some large value when all the estimated parameters are fused together. We found it happens at Point 210, which was not a jump point in the original signal sequence μ^*. Then Point 209 ($\hat{\mu}_{209} \approx -0.063$) and Point 210($\hat{\mu}_{210} \approx -0.052$) can never be fused together when λ_2 goes down. The only way to make them together is to do drag both of them to zero in the soft-thresholding step. But they are nonzero signals($\mu^*_{209} = \mu^*_{210} = -0.1$) and the sparsity recovery will be clearly violated if doing so.

We claim that a good pattern recovery will facilitate 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.

Appendix

We prove some of our theorems in the appendix.

APPENDIX A: PROOF OF THEOREM 2

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 3.   For the linear model $Y = \mathbf{X}\beta^* + \epsilon$, 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 Equation (4) which satisfies $\hat{\beta}(\lambda) =_s \beta^*$, if and only if the following two conditions hold

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

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

where the vector inequality and equality are taken elementwise. Moreover, if the inequality (25) holds strictly, then

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

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

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

Remarks. As in Wainwright [2009], we state an equivalent condition for (25). 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 (25) holds if and only if

(28)
$$\mathcal{M}(V) = \left\{ \max_{j \in S^c} |V_j| \leq \lambda \right\}$$

holds.

With Lemma 3 and the above comments, now we prove Theorem 2. 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 (25) (a necessary condition) is violated with probability greater than $1/2$.

In the proof of Theorem 3, we need an algebra result as follows.

LEMMA 4. *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

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

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 very 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 $\ell \geq j + 1$,

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

$\square$

APPENDIX B: PROOF OF LEMMA 2

To prove Lemma 2, 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.*

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

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 $X^T X$. Hence it suffices to show that all the eigenvalues of $X^T X$ are greater or equal to 0.25.

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

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

By Lemma 4, 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 $u \in \mathbb{R}^{n \times 1}$,

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

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

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

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.

$\square$

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 be

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

where X_c is the centered version of X with the j-th column of X_c be $X_j - \bar{X}_j$. Let the second moments of the data set 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 data set, and T its corresponding matrix of non-central second moments. Let $\lambda_j(\cdot)$ be the j-th largest eigenvalue of a matrix. Then*

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

Lemma 6 can be found on page 5 in Cadima and Jolliffe [2009].
With the results from Lemma 5 and Lemma 6, we now prove Lemma 2.

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.

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

Let $\sigma_j(\cdot)$ denote the j-th largest singular value of a matrix. By Lemma 5, the smallest singular value is not less than 0.5, that is $\sigma_j(X) \geq 0.5, \forall j = 1, \ldots, n$. 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 Equation (10). Let $\sigma_j(\cdot)$ denote the j-th largest singular value of a matrix. By Lemma 6, we have

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

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

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

$\square$

REFERENCES

J. Cadima and I. Jolliffe. On relationships between uncentred and column-centred principal component analysis. Pak J Statist, 25(4):473–503, 2009.

D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. Information Theory, IEEE Transactions on, 52(1):6–18, 2006.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456):1348–1360, 2001.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. The Annals of Applied Statistics, 1(2):302–332, 2007.

E. Greenshtein and Y.A. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. Bernoulli, 10(6):971–988, 2004.

H. Hoefling. A path algorithm for the fused lasso signal approximator. Journal of Computational and Graphical Statistics, 19(4):984–1006, 2010.

J. Jia and K. Rohe. Preconditioning to comply with the irrepresentable condition. arXiv preprint arXiv:1208.5584, 2012.

J. Jia, K. Rohe, and B. Yu. The lasso under heteroscedasticity. arXiv preprint arXiv:1011.1026, 2010.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. Annals of Statistics, pages 1356–1378, 2000.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics, 34(3):1436–1462, 2006.

A. Rinaldo. Properties and refinements of the fused lasso. The Annals of Statistics, 37(5B):2922–2952, 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for cgh data using the fused lasso. Biostatistics, 9(1):18–29, 2008.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1):91–108, 2004.

J.A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. Information Theory, IEEE Transactions on, 52(3):1030–1051, 2006.

M.J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. Information Theory, IEEE Transactions on, 55(12):5728–5741, 2009.

C.H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. The Annals of Statistics, 36(4):1567–1594, 2008.

P. Zhao and B. Yu. On model selection consistency of lasso. Journal of Machine Learning Research, 7(2): 2541, 2006.

Department of Mathematics
University of Pennsylvania
Philadelphia, PA 19104.
E-mail: junyangq@sas.upenn.edu

School of Mathematical Sciences
and Center for Statistical Science, LMAM, LMEQF
Peking University
Beijing 100871, China
E-mail: jzjia@math.pku.edu.cn