The Annals of Statistics
2014, Vol. 42, No. 6, 2202–2242
DOI: 10.1214/14-AOS1243

COVARIATE ASSISTED SCREENING AND ESTIMATION

BY ZHENG TRACY KE[1,2,3], JIASHUN JIN[2] AND JIANQING FAN[1]

University of Chicago, Carnegie Mellon University and Princeton University

Consider a linear model $Y = X\beta + z$, where $X = X_{n,p}$ and $z \sim N(0, I_n)$. The vector β is unknown but is sparse in the sense that most of its coordinates are 0. The main interest is to separate its nonzero coordinates from the zero ones (i.e., variable selection). Motivated by examples in long-memory time series (Fan and Yao [*Nonlinear Time Series: Nonparametric and Parametric Methods* (2003) Springer]) and the change-point problem (Bhattacharya [In *Change-Point Problems (South Hadley, MA, 1992)* (1994) 28–56 IMS]), we are primarily interested in the case where the Gram matrix $G = X'X$ is *nonsparse* but *sparsifiable* by a finite order linear filter. We focus on the regime where signals are both *rare and weak* so that successful variable selection is very challenging but is still possible.

We approach this problem by a new procedure called the *covariate assisted screening and estimation* (CASE). CASE first uses a linear filtering to reduce the original setting to a new regression model where the corresponding Gram (covariance) matrix is sparse. The new covariance matrix induces a sparse graph, which guides us to conduct multivariate screening without visiting all the submodels. By interacting with the signal sparsity, the graph enables us to decompose the original problem into many separated small-size subproblems (if only we know where they are!). Linear filtering also induces a so-called problem of *information leakage*, which can be overcome by the newly introduced *patching* technique. Together, these give rise to CASE, which is a two-stage *screen and clean* [Fan and Song *Ann. Statist.* **38** (2010) 3567–3604; Wasserman and Roeder *Ann. Statist.* **37** (2009) 2178–2201] procedure, where we first identify candidates of these submodels by *patching and screening*, and then re-examine each candidate to remove false positives.

For any procedure $\hat{\beta}$ for variable selection, we measure the performance by the minimax Hamming distance between the sign vectors of $\hat{\beta}$ and β. We show that in a broad class of situations where the Gram matrix is nonsparse but sparsifiable, CASE achieves the optimal rate of convergence. The results are successfully applied to long-memory time series and the change-point model.

Received November 2013; revised April 2014.

[1]Supported in part by NSF Grant DMS-07-04337, the National Institute of General Medical Sciences of the National Institutes of Health Grants R01GM100474 and R01-GM072611.

[2]Supported in part by NSF CAREER award DMS-09-08613.

[3]The major work of this article was completed when Z. Ke was a graduate student at Department of Operations Research and Financial Engineering, Princeton University.

MSC2010 subject classifications. Primary 62J05, 62J07; secondary 62C20, 62F12.

Key words and phrases. Asymptotic minimaxity, graph of least favorables (GOLF), graph of strong dependence (GOSD), Hamming distance, multivariate screening, phase diagram, rare and weak signal model, sparsity, variable selection.

1. Introduction. Consider a linear regression model

$$Y = X\beta + z, \qquad X = X_{n,p}, z \sim N(0, \sigma^2 I_n). \tag{1.1}$$

The vector β is unknown but is sparse, in the sense that only a small fraction of its coordinates is nonzero. The goal is to separate the nonzero coordinates of β from the zero ones (i.e., variable selection). We assume σ, which is the standard deviation of the noise, is known and set $\sigma = 1$ without loss of generality.

In this paper, we assume the Gram matrix

$$G = X'X \tag{1.2}$$

is normalized so that all of the diagonals are 1, instead of n as is often used in the literature. The difference between two normalizations is nonessential, but the signal vector β are different by a factor of $\sqrt{n}$.

We are primarily interested in the cases where:

- the signals (nonzero coordinates of β) are rare (or sparse) and weak;
- the Gram matrix G is *nonsparse* or even ill-posed (but it may be *sparsified* by some simple operations; see details below).

In such cases, the problem of variable selection is new and challenging.

While signal rarity is a well-accepted concept, signal weakness is an important but a largely neglected notion, and many contemporary researches on variable section have been focused on the regime where the signals are *rare but strong*. However, in many scientific experiments, due to the limitation in technology and constraints in resources, the signals are unavoidably weak. As a result, the signals are hard to find, and it is easy to be fooled. Partially, this explains why many published works (at least in some scientific areas) are not reproducible; see, for example, Ioannidis (2005).

We call G *sparse* if each of its rows has relatively few "large" elements, and we call G sparsifiable if G can be reduced to a sparse matrix by some simple operations (e.g., linear filtering or low-rank matrix removal). The Gram matrix plays a critical role in sparse inference, as the sufficient statistics $X'Y \sim N(G\beta, G)$. Examples where G is nonsparse but sparsifiable can be found in the following application areas:

- *Change-point problem*. Recently, driven by researches on DNA copy number variation, this problem has received a resurgence of interest [Niu and Zhang (2012), Olshen et al. (2004), Tibshirani and Wang (2008)]. While existing literature focuses on *detecting* change-points, *locating* change-points is also of major interest in many applications [Andreou and Ghysels (2002), Siegmund (2011), Zhang et al. (2010)]. Consider a change-point model

$$Y_i = \theta_i + z_i, \qquad z_i \overset{\text{i.i.d.}}{\sim} N(0,1), 1 \le i \le p, \tag{1.3}$$

where $\theta = (\theta_1, \ldots, \theta_p)'$ is a piece-wise constant vector with jumps at relatively few locations. Let $X = X_{p,p}$ be the matrix such that $X(i, j) = 1\{j \geq i\}$, $1 \leq i, j \leq p$. We re-parametrize the parameters by

$$\theta = X\beta \qquad \text{where } \beta_k = \theta_k - \theta_{k+1},\ 1 \leq k \leq p-1, \text{ and } \beta_p = \theta_p,$$

so that β_k is nonzero if and only if θ has a jump at location k. The Gram matrix G has elements $G(i, j) = \min\{i, j\}$, which is evidently nonsparse. However, adjacent rows of G display a high level of similarity, and the matrix can be sparsified by a second order adjacent differencing between the rows.

- *Long-memory time series*. We consider using time-dependent data to build a prediction model for variables of interest, $Y_t = \sum_j \beta_j X_{t-j} + \varepsilon_t$, where $\{X_t\}$ is an observed stationary time series and $\{\varepsilon_t\}$ are white noise. In many applications, $\{X_t\}$ is a long-memory process. Examples include volatility process [Fan and Yao (2003), Ray and Tsay (2000)], exchange rates, electricity demands, and river's outflow (e.g., the Nile's). Note that the problem can be reformulated as (1.1), where the Gram matrix $G = X'X$ is asymptotically close to the auto-covariance matrix of $\{X_t\}$ (say, Ω). It is well known that Ω is Toeplitz, the off-diagonal decay of which is very slow and the matrix L^1-norm which diverges as $p \to \infty$. However, the Gram matrix can be sparsified by a first order adjacent differencing between the rows.

Further examples include jump detections in (logarithm) asset prices and time series following a FARIMA model [Fan and Yao (2003)]. Still other examples include the factor models, where G can be decomposed as the sum of a sparse matrix and a low rank (positive semi-definite) matrix. In these examples, G is nonsparse, but it can be sparsified either by adjacent row differencing or low-rank matrix removal.

1.1. *Nonoptimality of L^0-penalization method for rare and weak signals*. When the signals are rare and strong, the problem of variable selection is more or less well understood. In particular, Donoho and Stark (1989) [see also Donoho and Huo (2001)] have investigated the *noiseless case* where they reveal a fundamental phenomenon. In detail, when there is no noise, model (1.1) reduces to $Y = X\beta$. Now, suppose (Y, X) are given, and consider the equation $Y = X\beta$. In the general case where $p > n$, it was shown in Donoho and Stark (1989) that under mild conditions on X, while the equation $Y = X\beta$ has infinitely many solutions, there is *a unique* solution that is *very sparse*. In fact, if X is full rank and this sparsest solution has k nonzero elements, then all other solutions have at least $(n - k + 1)$ nonzero elements; see Figure 1 (left).

In the spirit of Occam's razor, we have reason to believe that this *unique sparse* solution is the ground truth we are looking for. This motivates the well-known method of L^0-penalization, which looks for the sparsest solution where the sparsity is measured by the L^0-norm. In other words, in the noiseless case, the L^0-penalization method is a "fundamentally correct" (but computationally intractable) method.

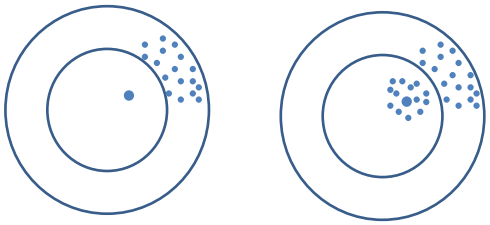

FIG. 1. *Illustration for solutions of $Y = X\beta + z$ in the noiseless case (left; where $z = 0$) and the strong noise case (right). Each dot represents a solution (the large dot is the ground truth), where the distance to the center is the L^0-norm of the solution. In the noiseless case, we only have one very sparse solution, with all other being much denser. In the strong noise case, signals are rare and weak, and we have many very sparse solutions that have comparable sparsity to that of the ground truth.*

In the past two decades, the above observation has motivated a long list of *computable global penalization methods*, including but not limited to the lasso, SCAD, MC+, each of which hopes to produce solutions that approximate that of the L^0-penalization method.

These methods usually use a theoretic framework that contains four intertwined components: "signals are rare but strong," "the true β is the sparsest solution of $Y = X\beta$," "probability of exact recovery is an appropriate loss function" and "L^0-penalization method is a fundamentally correct method."

Unfortunately, the above framework is no longer appropriate when the signals are rare and weak. First, the fundamental phenomenon found in Donoho and Stark (1989) is no longer true. Consider the equation $Y = X\beta + z$, and let β_0 be the ground truth. We can produce many vectors β by perturbing β_0 such that two models $Y = X\beta + z$ and $Y = X\beta_0 + z$ are indistinguishable (i.e., all tests—computable or not—are asymptotically powerless). In other words, the equation $Y = X\beta + z$ may have many *very sparse* solutions, where the ground truth is not necessarily the sparsest one; see Figure 1 (right).

In summary, when signals are rare and weak:

- The situation is more complicated than that considered by Donoho and Stark (1989), and the principle of Occam's razor is less relevant.
- "Exact recovery" is usually impossible, and the Hamming distance between the sign vectors of $\hat{\beta}$ and β is a more appropriate loss function.
- The L^0-penalization method is not "fundamentally correct" if the signals are rare/weak and the Hamming distance is the loss function.

For example, it was shown in Ji and Jin (2012) that in the rare/weak regime, even when X is very simple and when the tuning parameter is ideally set, the L^0-penalization method is not rate optimal in terms of the Hamming distance. See Ji and Jin (2012) for details.

1.2. *Limitation of UPS.* That the L^0-penalization method is rate nonoptimal implies that many other penalization methods (such as the lasso, SCAD, MC+) are also rate nonoptimal in the rare/weak regime.

What could be rate optimal procedures in the rare/weak regime? To address this, Ji and Jin (2012) proposed a method called *univariate penalization screening (UPS)*, and showed that UPS achieves the optimal rate of convergence in Hamming distance under certain conditions.

UPS is a two-stage screen and clean method [Wasserman and Roeder (2009)], at the heart of which is marginal screening. The main challenge that marginal screening faces is the so-called phenomenon of *signal cancellation*, a term coined by Wasserman and Roeder (2009). The success of UPS hinges on relatively strong conditions [e.g., see Genovese et al. (2012)], under which signal cancellation has negligible effects.

1.3. *Advantages and disadvantages of sparsifying.* Motivated by the application examples aforementioned, we are interested in the rare/weak cases where G is nonsparse but can be sparsified by a finite-order linear filtering. That is, if we denote the linear filtering by a $p \times p$ matrix D, then the matrix DG is sparse in the sense that each row has relatively few large entries, and all other entries are relatively small.

In such challenging cases, we should not expect the L^0-penalization method or the UPS to be rate optimal; this motivates us to develop a new approach.

Our strategy is to use sparsifying and so to exploit the sparsity of DG. Multiplying both sides of (1.1) by X' and then by D gives

$$d = DG\beta + N\bigl(0, DGD'\bigr), \qquad d \equiv D\tilde{Y}, \tilde{Y} \equiv X'Y. \tag{1.4}$$

On one hand, sparsifying is helpful for both matrices DG and DGD' are sparse, which can be largely exploited to develop better methods for variable selection. On the other hand, "there is no free lunch," and sparsifying also causes serious issues:

- The post-filtering model (1.4) is not a regular linear regression model.
- If we apply a local method (e.g., UPS, forward/backward regression) to model (1.4), we face so-called challenge of *information leakage*.

In Section 2.4, we carefully explain the issue of information leakage, and discuss how to deal with it.

We remark that while sparsifying can be very helpful, it does not mean that it is trivial to derive optimal procedures from model (1.4). For example, if we apply the L^0-penalization method naively to model (1.4), we then ignore the correlations among the noise, which cannot be optimal. If we apply the L^0-penalization method with the correlation structures incorporated, we are essentially applying it to the original regression model (1.1).

1.4. *Covariate assisted screening and estimation* (*CASE*). To exploit the sparsity in DG and DGD', and to deal with the two aforementioned issues that sparsifying causes, we propose a new variable selection method which we call *covariate assisted screening and estimation* (CASE). The main methodological innovation of CASE is to use linear filtering to create graph sparsity and then to exploit the rich information hidden in the "local" graphical structures among the design variables, which the lasso and many other procedures do not utilize.

At the heart of CASE is *covariate assisted* multivariate screening. Screening is a well-known method of dimension reduction in big data. However, most literature to date has been focused on *univariate screening* or *marginal screening* [Fan and Song (2010), Genovese et al. (2012)]. Extending marginal screening to (brute-force) m-variate screening, $m > 1$, means that we screen all $\binom{p}{m}$ size-m submodels, and has two major concerns:

- *Computational infeasibility*. A brute-force m-variate screening has a computation complexity of $O(p^m)$, which is usually not affordable.
- *Screening inefficiency*. The goal of screening is to remove as many noise entries as we can while retaining most of the signals. When we screen too many submodels than necessary, we have to set the bar higher than necessary to exclude most of the noise entries. As a result, we need signals stronger than necessary in order for them to survive the screening.

To overcome these challenges, CASE uses a new screening strategy called covariance-assisted screening, which excludes most size-m submodels from screening but still manages to retain almost all signals. In detail, we first use the Gram matrix G to construct a sparse graph called *graph of strong dependence* (GOSD). We then include a size-m submodel in our screening list if and only if the m nodes form a connected subgraph of GOSD. This way, we exclude many submodels from the screening by only using information in G, not that in the response vector Y!

The blessing is, when GOSD is sufficiently sparse, it has no more than $L_p p$ connected size-m sub-graphs, where L_p is a generic multi-$\log(p)$ term. Therefore, covariance-assisted screening only visits $L_p p$ submodels, in contrast to $\binom{p}{m}$ submodels the brute-forth screening visits. As a result, covariance-assisted screening is not only computationally feasible, but is also efficient. Now, it would not be a surprise that CASE is a "fundamentally correct" procedure in the rare/weak regime, at least when the GOSD is sufficiently sparse, as in settings considered in this paper; see more discussion below.

1.5. *Objective of the theoretic study.* We now discuss the theoretic component of the paper. The objective of our theoretic study is three-fold:

- to develop a theoretic framework that is appropriate for the regime where signals are rare/weak, and G is nonsparse but is sparsifiable;

- to appreciate the "pros" and "cons" of sparsifying, and to investigate how to fix the "cons";
- to show that CASE is asymptotic minimax and yields an optimal partition of the so-called *phase diagram*.

The phase diagram is a relatively new criterion for assessing the optimality of procedures. Call the two-dimensional space calibrated by the *signal rarity* and *signal strength* the phase space. The phase diagram is the partition of the phase space into different regions where in each of them inference is distinctly different. The notion of phase diagram is especially appropriate when signals are rare and weak.

The theoretic study is challenging for many reasons:

- We focus on a very challenging regime, where signals are rare and weak, and the design matrix is nonsparse or even ill-posed. Such a regime is important from a practical perspective, but has not been carefully explored in the literature.
- The goal of the paper is to develop procedures in the rare/weak regime that are asymptotic minimax in terms of Hamming distance, to achieve which we need to find a lower bound and an upper bound that are both tight. Compared to most works on variable selection where the goal is to find procedures that yield exact recovery for sufficiently strong signals, our goal is comparably more ambitious, and the study it entails is more delicate.
- To derive the phase diagrams in Sections 2.10–2.11, we need explicit forms of the convergence rate of minimax Hamming selection errors. This usually needs very delicate analysis. The study associated with the change-point model is especially challenging and long.

1.6. *Content and notation.* The paper is organized as follows. Section 2 contains the main results of this paper: we formally introduce CASE and establishe its asymptotic optimality. Section 3 contains simulation studies, and Section 4 contains conclusions and discussions.

Throughout this paper, $D = D_{h,\eta}$, $d = D(X'Y)$, $B = DG$, $H = DGD'$ and $\mathcal{G}^*$ denotes the GOSD (in contrast, d_p denotes the degree of GOLF, and H_p denotes the Hamming distance). Also, $\mathbb{R}$ and $\mathbb{C}$ denote the sets of real numbers and complex numbers, respectively, and $\mathbb{R}^p$ denotes the p-dimensional real Euclidean space. Given $0 \le q \le \infty$, for any vector x, $\|x\|_q$ denotes the L^q-norm of x; for any matrix M, $\|M\|_q$ denotes the matrix L^q-norm of M. When $q = 2$, $\|M\|_q$ coincides with the matrix spectral norm; we shall omit the subscript q in this case. When M is symmetric, $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ denote the maximum and minimum eigenvalues of M, respectively. For two matrices M_1 and M_2, $M_1 \succeq M_2$ means that $M_1 - M_2$ is positive semi-definite.

2. Main results. This section is arranged as follows. Sections 2.1–2.6 focus on the model, ideas and the method. In Section 2.1, we introduce the rare and weak signal model. In Section 2.2, we formally introduce the notion of *sparsifiability*. The starting point of CASE is the use of a linear filter. In Section 2.3, we explain how linear filtering helps in variable selection by inducing a sparse graph and an interesting interaction between the graphical sparsity and the signal sparsity. In Section 2.4, we explain that linear filtering also causes a so-called problem of *information leakage*, and discuss how to overcome such a problem by the technique of *patching*. After all these ideas are discussed, we formally introduce the CASE in Section 2.5. In Section 2.6, we discuss the computational complexity and show that CASE is computationally feasible in a broad context.

Sections 2.7–2.9 focus on the asymptotic optimality of CASE. In Section 2.7, we introduce the asymptotic minimax framework where we use Hamming distance as the loss function. In Section 2.8, we study the lower bound for the minimax Hamming risk, and in Section 2.9, we show that CASE achieves the minimax Hamming risk in a broad context.

In Sections 2.10–2.11, we apply our results to long-memory time series and the change-point model. For both of them, we first derive explicit formulas for the convergent rates, and then use the formulas to derive the phase diagrams.

Proofs of results in this section can be found in the supplemental article [Ke, Jin and Fan (2014)], which contains Sections A–C.

2.1. *Rare and weak signal model.* Our primary interest is in the situations where the signals are rare and weak, and where we have *no* information on the underlying structure of the signals. In such situations, it makes sense to use the following *rare and weak* signal model; see Candès and Plan (2009), Donoho and Jin (2008), Jin, Zhang and Zhang (2014). Fix $\varepsilon \in (0,1)$ and $\tau > 0$. Let $b = (b_1, \ldots, b_p)'$ be the $p \times 1$ vector satisfying

$$b_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\varepsilon), \tag{2.1}$$

and let $\Theta_p(\tau)$ be the set of vectors

$$\Theta_p(\tau) = \{\mu \in \mathbb{R}^p : |\mu_i| \geq \tau, 1 \leq i \leq p\}. \tag{2.2}$$

We model β by

$$\beta = b \circ \mu, \tag{2.3}$$

where $\mu \in \Theta_p(\tau)$ and $\circ$ is the Hadamard product (also called the coordinate-wise product). In Section 2.7, we further restrict μ to a subset of $\Theta_p(\tau)$.

In this model, β_i is either 0 or a signal with a strength $\geq \tau$. Since we have no information on where the signals are, we assume that they appear at locations that are randomly generated. We are primarily interested in the challenging case where ε is small and τ is relatively small, so the signals are both rare and weak.

DEFINITION 2.1. We call model (2.1)–(2.3) the rare and weak signal model $\mathrm{RW}(\varepsilon, \tau, \mu)$.

We remark that the theory developed in this paper is not tied to the rare and weak signal model, and applies to more general cases. For example, the main results can be extended to the case where we have some additional information about the underlying structure of the signals (e.g., Ising's model [Ising (1925)]).

2.2. *Sparsifiability, linear filtering and GOSD.* As mentioned before, we are primarily interested in the case where the Gram matrix G can be sparsified by a finite-order linear filtering.

Fix an integer $h \geq 1$ and an $(h+1)$-dimensional vector $\eta = (1, \eta_1, \ldots, \eta_h)'$. Let $D = D_{h,\eta}$ be the $p \times p$ matrix satisfying

$$D_{h,\eta}(i,j) = 1\{i=j\} + \eta_1 1\{i=j-1\} + \cdots + \eta_h 1\{i=j-h\}, \quad 1 \leq i, j \leq p. \tag{2.4}$$

The matrix $D_{h,\eta}$ can be viewed as a linear operator that maps any $p \times 1$ vector y to $D_{h,\eta}y$. For this reason, $D_{h,\eta}$ is also called an order h linear filter [Fan and Yao (2003)].

For $\alpha > 0$ and $A_0 > 0$, we introduce the following class of matrices:

$$\mathcal{M}_p(\alpha, A_0) = \big\{\Omega \in \mathbb{R}^{p\times p} : \Omega(i,i) \leq 1, |\Omega(i,j)| \leq A_0\big(1+|i-j|\big)^{-\alpha}, \quad 1 \leq i, j \leq p\big\}. \tag{2.5}$$

Matrices in $\mathcal{M}_p(\alpha, A_0)$ are not necessarily symmetric.

DEFINITION 2.2. Fix an order h linear filter $D = D_{h,\eta}$. We say that G is sparsifiable by $D_{h,\eta}$ if for sufficiently large p, $DG \in \mathcal{M}_p(\alpha, A_0)$ for some constants $\alpha > 1$ and $A_0 > 0$.

In the long-memory time series model, G can be sparsified by an order 1 linear filter. In the change-point model, G can be sparsified by an order 2 linear filter.

The main benefit of linear filtering is that it induces sparsity in the graph of strong dependence (GOSD) to be introduced below. Recall that the sufficient statistics $\tilde{Y} = X'Y \sim N(G\beta, G)$. Applying a linear filter $D = D_{h,\eta}$ to $\tilde{Y}$ gives

$$d \sim N(B\beta, H), \tag{2.6}$$

where $d = D(X'Y)$, $B = DG$ and $H = DGD'$. Note that no information is lost when we reduce from the model $\tilde{Y} \sim N(G\beta, G)$ to model (2.6), as D is a nonsingular matrix.

At the same time, if G is sparsifiable by $D = D_{h,\eta}$, then both the matrices B and H are sparse, in the sense that each row of either matrix has relatively few large

coordinates. In other words, for a properly small threshold $\delta > 0$ to be determined, let B^* and H^* be the regularized matrices of B and H, respectively,

$$B^*(i,j) = B(i,j)1\{|B(i,j)| \geq \delta\},$$
$$H^*(i,j) = H(i,j)1\{|H(i,j)| \geq \delta\}, \qquad 1 \leq i,j \leq p.$$

It is seen that

$$d \approx N(B^*\beta, H^*), \tag{2.7}$$

where each row of B^* or H^* has relatively few nonzeros. Compared to (2.6), (2.7) is much easier to track analytically, but it contains almost all the information about β.

The above observation naturally motivates the following graph, which we call the *graph of strong dependence* (GOSD).

DEFINITION 2.3. For a given parameter δ, the GOSD is the graph $\mathcal{G}^* = (V, E)$ with nodes $V = \{1, 2, \ldots, p\}$, and there is an edge between i and j when any of the three numbers $H^*(i,j)$, $B^*(i,j)$ and $B^*(j,i)$ is nonzero.

DEFINITION 2.4. A graph $\mathcal{G} = (V, E)$ is called K-sparse if the degree of each node $\leq K$.

The definition of GOSD depends on a tuning parameter δ, the choice of which is not critical, and it is generally sufficient if we choose $\delta = \delta_p = O(1/\log(p))$; see Section B.1 in Ke, Jin and Fan (2014)for details. With such a choice of δ, it can be shown that in a general context, GOSD is K-sparse, where $K = K_\delta$ does not exceed a multi-log(p) term as $p \to \infty$; see Lemma B.1 in Ke, Jin and Fan (2014).

2.3. *Interplay between the graph sparsity and signal sparsity.* With these being said, it remains unclear how the sparsity of $\mathcal{G}^*$ helps in variable selection. In fact, even when $\mathcal{G}^*$ is 2-sparse, it is possible that a node k is connected—through possible long paths—to many other nodes; it is unclear how to remove the effect of these nodes when we try to estimate β_k.

Somewhat surprisingly, the answer lies in an interesting interplay between the signal sparsity and graph sparsity. To see this point, let $S = S(\beta)$ be the support of β, and let $\mathcal{G}_S^*$ be the subgraph of $\mathcal{G}^*$ formed by the nodes in S only. Given the sparsity of $\mathcal{G}^*$, if the signal vector β is also sparse, then it is likely that the sizes of all components of $\mathcal{G}_S^*$ (a component of a graph is a maximal connected subgraph) are uniformly small. This is justified in the following lemma which is proved in Jin, Zhang and Zhang (2014).

LEMMA 2.1. *Suppose $\mathcal{G}^*$ is K-sparse, and the support $S = S(\beta)$ is a realization from $\beta_j \overset{\text{i.i.d.}}{\sim} (1-\varepsilon)\nu_0 + \varepsilon\pi$, where ν_0 is the point mass at 0 and π is any*

distribution with support $\subseteq \mathbb{R} \setminus \{0\}$. *With a probability* (*from randomness of* S) *at least* $1 - p(e\varepsilon K)^{m+1}$, $\mathcal{G}_S^*$ *decomposes into many components with size no larger than* m.

In this paper, we are primarily interested in cases where for large p, $\varepsilon \leq p^{-\vartheta}$ for some parameter $\vartheta \in (0, 1)$ and K is bounded by a multi-log(p) term. In such cases, the decomposability of $\mathcal{G}_S^*$ holds for a finite m, with overwhelming probability.

Lemma 2.1 delineates an interesting picture: The set of signals decomposes into many small-size isolated "signal islands" (if only we know where), each of them is a component of $\mathcal{G}_S^*$ and different ones are disconnected in the GOSD. As a result, the original p-dimensional problem can be viewed as the aggregation of many separated small-size subproblems that can be solved parallelly. This is the key insight of this paper.

Note that the decomposability of $\mathcal{G}_S^*$ attributes to the interplay between the signal sparsity and the graph sparsity, where the latter attributes to the use of linear filtering. The decomposability is not tied to the specific model of β in Lemma 2.1, and holds for much broader situations (e.g., when b is generated by a sparse Ising model [Ising (1925)]).

2.4. *Information leakage and patching*. While it largely facilitates the decomposability of the model, we must note that the linear filtering also induces a so-called problem of *information leakage*. In this section, we discuss how linear filtering causes such a problem and how to overcome it by the so-called technique of *patching*.

The following notation is frequently used in this paper.

DEFINITION 2.5. For $\mathcal{I} \subset \{1, 2, \ldots, p\}$, $\mathcal{J} \subset \{1, \ldots, N\}$ and a $p \times N$ matrix X, $X^{\mathcal{I}}$ denotes the $|\mathcal{I}| \times N$ sub-matrix formed by restricting the rows of X to $\mathcal{I}$, and $X^{\mathcal{I},\mathcal{J}}$ denotes the $|\mathcal{I}| \times |\mathcal{J}|$ sub-matrix formed by restricting the rows of X to $\mathcal{I}$ and columns to $\mathcal{J}$.

Note that when $N = 1$, X is a $p \times 1$ vector, and $X^{\mathcal{I}}$ is an $|\mathcal{I}| \times 1$ vector.

To explain information leakage, we first consider an idealized case where each row of G has $\leq K$ nonzeros. In this case, there is no need for linear filtering, so $B = H = G$ and $d = \tilde{Y}$. Recall that $\mathcal{G}_S^*$ consists of many signal islands, and let $\mathcal{I}$ be one of them. It is seen that

$$d^{\mathcal{I}} \approx N\big(G^{\mathcal{I},\mathcal{I}}\beta^{\mathcal{I}}, G^{\mathcal{I},\mathcal{I}}\big), \tag{2.8}$$

and how well we can estimate $\beta^{\mathcal{I}}$ is captured by the Fisher information matrix $G^{\mathcal{I},\mathcal{I}}$ [Lehmann and Casella (1998)].

Come back to the case where G is nonsparse. Interestingly, despite the strong correlations, $G^{\mathcal{I},\mathcal{I}}$ continues to be the Fisher information for estimating $\beta^{\mathcal{I}}$. However, when G is nonsparse, we must use a linear filtering $D = D_{h,\eta}$ as suggested,

and we have

$$d^{\mathcal{I}} \approx N\big(B^{\mathcal{I},\mathcal{I}}\beta^{\mathcal{I}}, H^{\mathcal{I},\mathcal{I}}\big). \tag{2.9}$$

Moreover, letting $\mathcal{J} = \{1 \le j \le p : D(i, j) \ne 0 \text{ for some } i \in \mathcal{I}\}$, it follows that

$$B^{\mathcal{I},\mathcal{I}}\beta^{\mathcal{I}} = D^{\mathcal{I},\mathcal{J}}G^{\mathcal{J},\mathcal{I}}\beta^{\mathcal{I}}.$$

By the definition of D, $|\mathcal{J}| > |\mathcal{I}|$ and the dimension of the following null space ≥ 1,

$$\text{Null}(\mathcal{I}, \mathcal{J}) = \big\{\xi \in \mathbb{R}^{|\mathcal{J}|} : D^{\mathcal{I},\mathcal{J}}\xi = 0\big\}. \tag{2.10}$$

Compare (2.9) with (2.8), and imagine the oracle situation where we are told the mean vector of $d^{\mathcal{I}}$ in both. The difference is that we can fully recover $\beta^{\mathcal{I}}$ using (2.8), but are not able to do so with only (2.9). In other words, the information containing $\beta^{\mathcal{I}}$ is partially lost in (2.9): if we estimate $\beta^{\mathcal{I}}$ with (2.9) alone, we will never achieve the desired accuracy.

The argument is validated in Lemma 2.2 below, where the Fisher information associated with (2.9) is strictly "smaller" than $G^{\mathcal{I},\mathcal{I}}$; the difference between two matrices can be derived by taking $\mathcal{I}^+ = \mathcal{I}$ and $\mathcal{J}^+ = \mathcal{J}$ in (2.12). We call this phenomenon "information leakage."

To mitigate this, we expand the information content by including data in the neighborhood of $\mathcal{I}$. This process is called "patching." Let $\mathcal{I}^+$ be an extension of $\mathcal{I}$ by adding a few neighboring nodes, and define similarly $\mathcal{J}^+ = \{1 \le j \le p : D(i, j) \ne 0 \text{ for some } i \in \mathcal{I}^+\}$ and $\text{Null}(\mathcal{I}^+, \mathcal{J}^+)$. Assuming that there is no edge between any node in $\mathcal{I}^+$ and any node in $\mathcal{G}_S^* \setminus \mathcal{I}$,

$$d^{\mathcal{I}^+} \approx N\big(B^{\mathcal{I}^+,\mathcal{I}}\beta^{\mathcal{I}}, H^{\mathcal{I}^+,\mathcal{I}^+}\big). \tag{2.11}$$

The Fisher information matrix for $\beta^{\mathcal{I}}$ under model (2.11) is larger than that of (2.9), which is captured in the following lemma.

LEMMA 2.2. *The Fisher information matrix associated with model* (2.11) *is*

$$G^{\mathcal{I},\mathcal{I}} - \big[U\big(U'\big(G^{\mathcal{J}^+,\mathcal{J}^+}\big)^{-1}U\big)^{-1}U'\big]^{\mathcal{I},\mathcal{I}}, \tag{2.12}$$

where U is any $|\mathcal{J}^+| \times (|\mathcal{J}^+| - |\mathcal{I}^+|)$ matrix whose columns form an orthonormal basis of $\text{Null}(\mathcal{I}^+, \mathcal{J}^+)$.

When the size of $\mathcal{I}^+$ becomes appropriately large, the second matrix in (2.12) is small element-wise (and so is negligible) under mild conditions [see details in Lemma A.3 in Ke, Jin and Fan (2014)]. This matrix is usually nonnegligible if we set $\mathcal{I}^+ = \mathcal{I}$ and $\mathcal{J}^+ = \mathcal{J}$ (i.e., without patching).

EXAMPLE 1. We illustrate the above phenomenon with an example where $p = 5000$, G is the matrix satisfying $G(i, j) = [1 + 5|i - j|]^{-0.95}$ for all $1 \le i, j \le p$ and $D = D_{h,\eta}$ with $h = 1$ and $\eta = (1, -1)'$. If $\mathcal{I} = \{2000\}$, then $G^{\mathcal{I},\mathcal{I}} = 1$, but the Fisher information associated with model (2.9) is 0.5. The gap can be substantially narrowed if we patch with $\mathcal{I}^+ = \{1990, 1991, \ldots, 2010\}$, in which case the Fisher information in model (2.11) is 0.904.

Although one of the major effects of information leakage is a reduction in the signal-to-noise ratio, this phenomenon is very different from the well-known "signal cancellation" or "partial faithfulness" in variable selection. "Signal cancellation" is caused by correlations between signal covariates, and CASE overcomes this problem by using multivariate screening. However, "information leakage" is caused by the use of a linear filtering. From Lemma 2.2, we can see that the information leakage appears no matter for what signal vector β. CASE overcomes this problem by the patching technique.

2.5. *Covariate assisted screening and estimation* (*CASE*). In summary, we start from the post-filtering regression model

$$d = D\tilde{Y} \qquad \text{where } \tilde{Y} = X'Y \text{ and } D = D_{h,\eta} \text{ is a linear filter.}$$

We have observed the following:

- *Signal decomposability*. Linear filtering induces sparsity in GOSD, a graph constructed from the Gram matrix G. In this graph, the set of all true signals decomposes into many small-size signal islands, each signal island is a component of GOSD.
- *Information patching*. Linear filtering also causes information leakage, which can be overcome by delicate patching technique.

Naturally, these motivate a two-stage screen and clean approach for variable selection, which we call *covariate assisted screening and estimation* (CASE). CASE contains a *patching and screening* (PS) step and a *patching and estimation* (PE) step.

- *PS-step*. We use sequential χ^2-tests to identify candidates for each signal island. Each χ^2-test is guided by $\mathcal{G}^*$, and aided by a carefully designed patching step. This achieves multivariate screening without visiting all submodels.
- *PE-step*. We re-investigate each candidate with penalized MLE and certain patching technique, in hopes of removing false positives.

For the purpose of patching, the *PS*-step and the *PE*-step use tuning integers ℓ^{ps} and ℓ^{pe}, respectively. The following notation is frequently used in this paper.

DEFINITION 2.6. For any index $1 \le i \le p$, $\{i\}^{ps} = \{1 \le j \le p : |j - i| \le \ell^{ps}\}$. For any subset $\mathcal{I}$ of $\{1, 2, \ldots, p\}$, $\mathcal{I}^{ps} = \bigcup_{i \in \mathcal{I}} \{i\}^{ps}$. Similar notation applies to $\{i\}^{pe}$ and $\mathcal{I}^{pe}$.

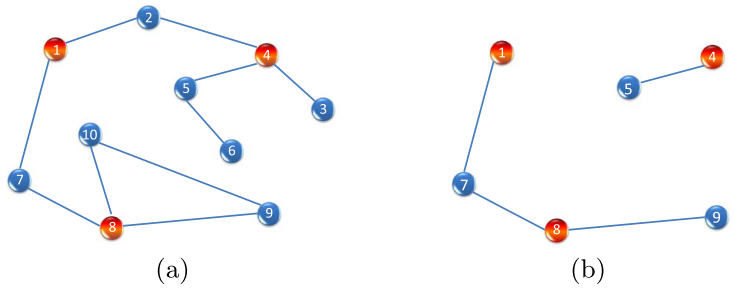

FIG. 2. *Illustration of graph of strong dependence* (*GOSD*). *Red*: *signal nodes*. *Blue*: *noise nodes*. (a) *GOSD with* 10 *nodes*. (b) *Nodes of GOSD that survived the PS-step*.

We now discuss two steps in detail. Consider the *PS*-step first. Fix $m > 1$. Suppose that $\mathcal{G}^*$ has a total of T connected subgraphs with size $\leq m$, which we denote by $\{\mathcal{G}_t\}_{t=1}^T$, arranged in the ascending order of the sizes, with ties breaking lexicographically.

EXAMPLE 2(a). We illustrate this with a toy example, where $p = 10$ and the GOSD is displayed in Figure 2(a). For $m = 3$, GOSD has $T = 30$ connected subgraphs, which we arrange as follows. Note that $\{\mathcal{G}_t\}_{t=1}^{10}$ are singletons, $\{\mathcal{G}_t\}_{t=11}^{20}$ are connected pairs and $\{\mathcal{G}_t\}_{t=21}^{30}$ are connected triplets

$$\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}, \{9\}, \{10\},$$

$$\{1, 2\}, \{1, 7\}, \{2, 4\}, \{3, 4\}, \{4, 5\}, \{5, 6\}, \{7, 8\}, \{8, 9\}, \{8, 10\}, \{9, 10\},$$

$$\{1, 2, 4\}, \{1, 2, 7\}, \{1, 7, 8\}, \{2, 3, 4\}, \{2, 4, 5\}, \{3, 4, 5\}, \{4, 5, 6\}, \{7, 8, 9\},$$

$$\{7, 8, 10\}, \{8, 9, 10\}.$$

Here we examine sequentially only the 30 submodels above to decide whether any variables have additional utilities given the variables recruited before, via χ^2-tests. The first 10 screening problems are just the univariate screening. After that, starting from bivariate screening, we examine the variables given those selected so far. Suppose that we are examining the submodel $\{1, 2\}$. The testing problem depends on how the variables $\{1, 2\}$ are selected in the previous steps. For example, if the variables $\{1, 2, 4, 6\}$ have already been selected in the univariate screening, there is no new recruitment, and we move on to examine the submodel $\{1, 7\}$. If the variables $\{1, 4, 6\}$ have been recruited so far, we need to test if variable $\{2\}$ has additional contributions given variable $\{1\}$. If the variables $\{4, 6\}$ have been recruited in the previous steps, we will examine whether variables $\{1, 2\}$ together have any significant contributions. Therefore, we have never run regression for more than two variables. Similarly, for trivariate screening, we will never run regression for more than 3 variables. Clearly, multivariate screening improves the marginal screening in that it gives signal variables chances to be recruited if they are wrongly excluded by the marginal method.

We now formally describe the procedure. The *PS*-step contains T sub-stages, where we screen $\mathcal{G}_t$ sequentially, $t = 1, 2, \ldots, T$. Let $\mathcal{U}^{(t)}$ be the set of retained indices at the end of stage t, with $\mathcal{U}^{(0)} = \varnothing$ as the convention. For $1 \le t \le T$, the tth sub-stage contains two sub-steps:

- (*Initial step*). Let $\hat{N} = \mathcal{U}^{(t-1)} \cap \mathcal{G}_t$ represent the set of nodes in $\mathcal{G}_t$ that have already been accepted by the end of the $(t-1)$th sub-stage, and let $\hat{F} = \mathcal{G}_t \setminus \hat{N}$ be the set of other nodes in $\mathcal{G}_t$.
- (*Updating step*). Write for short $\mathcal{I} = \mathcal{G}_t$. Fixing a tuning parameter ℓ^{ps} for patching, introduce

$$\begin{aligned} W &= \big(B^{\mathcal{I}^{ps},\mathcal{I}}\big)'\big(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}}\big)^{-1} d^{\mathcal{I}^{ps}}, \\ Q &= \big(B^{\mathcal{I}^{ps},\mathcal{I}}\big)'\big(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}}\big)^{-1}\big(B^{\mathcal{I}^{ps},\mathcal{I}}\big), \end{aligned} \tag{2.13}$$

where W is a random vector and Q can be thought of as the covariance matrix of W. Define $W_{\hat{N}}$, a subvector of W, and $Q_{\hat{N},\hat{N}}$, a submatrix of Q, as follows:

$$\begin{aligned} W_{\hat{N}} &= \big(B^{\mathcal{I}^{ps},\hat{N}}\big)'\big(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}}\big)^{-1} d^{\mathcal{I}^{ps}}, \\ Q_{\hat{N},\hat{N}} &= \big(B^{\mathcal{I}^{ps},\hat{N}}\big)'\big(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}}\big)^{-1}\big(B^{\mathcal{I}^{ps},\hat{N}}\big). \end{aligned} \tag{2.14}$$

Introduce the test statistic

$$T(d, \hat{F}, \hat{N}) = W'Q^{-1}W - W'_{\hat{N}}(Q_{\hat{N},\hat{N}})^{-1}W_{\hat{N}}. \tag{2.15}$$

For a threshold $t = t(\hat{F}, \hat{N})$ to be determined, we update the set of retained nodes by $\mathcal{U}^{(t)} = \mathcal{U}^{(t-1)} \cup \hat{F}$ if $T(d, \hat{F}, \hat{N}) > t$, and let $\mathcal{U}^{(t)} = \mathcal{U}^{(t-1)}$ otherwise. In other words, we accept nodes in $\hat{F}$ only when they have additional utilities.

The *PS*-step terminates at $t = T$. We then write $\mathcal{U}^*_p = \mathcal{U}^{(T)}$ so that

$$\mathcal{U}^*_p = \text{the set of all retained indices at the end of the } PS\text{-step}.$$

In the *PS*-step, as we screen, we accept nodes sequentially. Once a node is accepted in the *PS*-step, it stays there until the end of the *PS*-step; of course, this node could be killed in the *PE*-step. In spirit, this is similar to the well-known forward regression method, but the implementation of two methods are significantly different.

The *PS*-step uses a collection of tuning thresholds

$$\mathcal{Q} = \big\{t(\hat{F}, \hat{N}) : (\hat{F}, \hat{N}) \text{ are defined above}\big\}.$$

A convenient choice for these thresholds is to let $t(\hat{F}, \hat{N}) = 2\tilde{q}\log(p)|\hat{F}|$ for a properly small fixed constant $\tilde{q} > 0$. See Section 2.9 (and also Sections 2.10–2.11) for more discussion on the choices of $t(\hat{F}, \hat{N})$.

In the *PS*-step, we use χ^2-test for screening. This is the best choice when the coordinates of z are Gaussian and have the same variance. When the Gaussian

assumption on z is questionable, we must note that the χ^2-test depends on the Gaussianity of $a'z$ for all p-different a, not on that of z; $a'z$ could be approximately Gaussian by central limit theorem. Therefore, the performance of χ^2-test is relatively robust to non-Gaussianity. If circumstances arise that the χ^2-test is not appropriate (e.g., misspecification of the model, low quantity of the data), we may need an alternative, say, some nonparametric tests. In this case, if the efficiency of the test is nearly optimal, then the screening in the *PS*-step would continue to be successful.

How does the *PS*-step help in variable selection? In Section A in Ke, Jin and Fan (2014), we show that in a broad context, provided that the tuning parameters $t(\hat{F}, \hat{N})$ are properly set, the *PS*-step has two noteworthy properties: the *sure screening* (SS) property and the *separable after screening* (SAS) property. The SS property says that $\mathcal{U}_p^*$ contains all but a negligible fraction of the true signals. The SAS property says that if we view $\mathcal{U}_p^*$ as a subgraph of $\mathcal{G}^*$ (more precisely, as a subgraph of $\mathcal{G}^+$, an expanded graph of $\mathcal{G}^*$ to be introduce below), then this subgraph decomposes into many disconnected components, each having a moderate size.

Together, the SS property and the SAS property enable us to reduce the original large-scale problem to many parallel small-size regression problems, and pave the way for the *PE*-step. See Section A in Ke, Jin and Fan (2014) for details.

EXAMPLE 2(b). We illustrate the above points with the toy example in Example 2(a). Suppose after the *PS*-step, the set of retained indices $\mathcal{U}_p^*$ is $\{1, 4, 5, 7, 8, 9\}$; see Figure 2(b). In this example, we have a total of three signal nodes, $\{1\}$, $\{4\}$ and $\{8\}$, which are all retained in $\mathcal{U}_p^*$ and so the *PS*-step yields sure screening. On the other hand, $\mathcal{U}_p^*$ contains a few nodes of false positives, which will be further cleaned in the *PE*-step. At the same time, viewing it as a subgraph of $\mathcal{G}^*$, $\mathcal{U}_p^*$ decomposes into two disconnected components, $\{1, 7, 8, 9\}$ and $\{4, 5\}$; compare Figure 2(a). The SS property and the SAS property enable us to reduce the original problem of 10 nodes to two parallel regression problems, one with 4 nodes, and the other with 2 nodes.

We now discuss the *PE*-step. Recall that ℓ^{pe} is the tuning parameter for the patching of the *PE*-step, and let $\{i\}^{pe}$ be as in Definition 2.6. The following graph can be viewed as an expanded graph of $\mathcal{G}^*$.

DEFINITION 2.7. Let $\mathcal{G}^+ = (V, E)$ be the graph where $V = \{1, 2, \ldots, p\}$ and there is an edge between nodes i and j when there exist nodes $k \in \{i\}^{pe}$ and $k' \in \{j\}^{pe}$ such that there is an edge between k and k' in $\mathcal{G}^*$.

Recall that $\mathcal{U}_p^*$ is the set of retained indices at the end of the *PS*-step.

DEFINITION 2.8. Fix a graph $\mathcal{G}$ and its subgraph $\mathcal{I}$. We say $\mathcal{I} \trianglelefteq \mathcal{G}$ if $\mathcal{I}$ is a connected subgraph of $\mathcal{G}$, and $\mathcal{I} \triangleleft \mathcal{G}$ if $\mathcal{I}$ is a component (maximal connected subgraph) of $\mathcal{G}$.

Fix $1 \le j \le p$. When $j \notin \mathcal{U}_p^*$, CASE estimates β_j as 0. When $j \in \mathcal{U}_p^*$, viewing $\mathcal{U}_p^*$ as a subgraph of $\mathcal{G}^+$, there is a unique subgraph $\mathcal{I}$ such that $j \in \mathcal{I} \triangleleft \mathcal{U}_p^*$. Fix two tuning parameters u^{pe} and v^{pe}. We estimate $\beta^{\mathcal{I}}$ by minimizing

$$\min_{\theta}\Big\{\frac{1}{2}(d^{\mathcal{I}^{pe}} - B^{\mathcal{I}^{pe},\mathcal{I}}\theta)'(H^{\mathcal{I}^{pe},\mathcal{I}^{pe}})^{-1}(d^{\mathcal{I}^{pe}} - B^{\mathcal{I}^{pe},\mathcal{I}}\theta) + \frac{(u^{pe})^2}{2}\|\theta\|_0\Big\}, \tag{2.16}$$

subject to that θ is an $|\mathcal{I}| \times 1$ vector each of which nonzero coordinate $\ge v^{pe}$, where $\|\theta\|_0$ denotes the L^0-norm of θ. Putting these together gives the final estimator of CASE $\hat{\beta}^{\text{case}} = \hat{\beta}^{\text{case}}(Y; \delta, m, \mathcal{Q}, \ell^{ps}, \ell^{pe}, u^{pe}, v^{pe}, D_{h,\eta}, X, p)$.

CASE uses tuning parameters $(\delta, m, \mathcal{Q}, \ell^{ps}, \ell^{pe}, u^{pe}, v^{pe})$. Earlier in this paper, we have briefly discussed how to choose $(\delta, \mathcal{Q})$. As for m, usually, a choice of $m = 2$ or 3 is sufficient unless the signals are relatively "dense." The choices of $(\ell^{ps}, \ell^{pe}, u^{pe}, v^{pe})$ are addressed in Section 2.9; see also Sections 2.10–2.11.

2.6. *Computational complexity of CASE, comparison with multivariate screening.* The *PS*-step is closely related to the well-known method of marginal screening and has a moderate computational complexity.

Marginal screening selects variables by thresholding the vector d coordinatewise. The method is computationally fast, but it neglects "local" graphical structures, and is thus ineffective. For this reason, in many challenging problems, it is desirable to use *multivariate screening* methods which adapt to "local" graphical structures.

Fix $m > 1$. An m-variate χ^2-screening procedure is one of such desired methods. The method screens all k-tuples of coordinates of d using χ^2-tests, for all $k \le m$, in an exhaustive (brute-force) fashion. Seemingly, the method adapts to "local" graphical structures and could be much more effective than marginal screening. However, such a procedure has a computational cost of $O(p^m)$ [excluding the computational cost for obtaining $X'Y$ from (X, Y); same below] which is usually not affordable when p is large.

The main innovation of the *PS*-step is to use a graph-assisted m-variate χ^2-screening, which is both effective in variable selection and efficient in computation. In fact, the *PS*-step only screens k-tuples of coordinates of d that form a connected subgraph of $\mathcal{G}^*$, for all $k \le m$. Therefore, if $\mathcal{G}^*$ is K-sparse, then there are $\le Cp(eK)^{m+1}$ connected subgraphs of $\mathcal{G}^*$ with size $\le m$; so if $K = K_p$ is no greater than a multi-log(p) term (see Definition 2.10), then the computational complexity of the *PS*-step is only $O(p)$, up to a multi-log(p) term.

EXAMPLE 2(c). We illustrate the difference between the above three methods with the toy example in Example 2(a), where $p = 10$ and the GOSD is displayed in Figure 2(a). Suppose we choose $m = 3$. Marginal screening screens all 10 single nodes of the GOSD. The brute-force m-variate screening screens all k-tuples of indices, $1 \le k \le m$, with a total of $\binom{p}{1} + \cdots + \binom{p}{m} = 175$ such k-tuples. The m-variate screening in the *PS*-step only screens k-tuples that are connected subgraphs of $\mathcal{G}^*$, for $1 \le k \le m$, and in this example, we only have 30 such connected subgraphs.

The computational complexity of the *PE*-step consists two parts. The first part is the complexity of obtaining all components of $\mathcal{U}_p^*$, which is $O(pK)$ and where K is the maximum degree of $\mathcal{G}^+$; note that for settings considered in this paper, $K = K_p^+$ does not exceed a multi-log(p) term [see Lemma B.2 in Ke, Jin and Fan (2014)]. The second part of the complexity comes from solving (2.16), which hinges on the maximal size of $\mathcal{I}$. In Lemma A.2 in Ke, Jin and Fan (2014), we show that in a broad context, the maximal size of $\mathcal{I}$ does not exceed a constant l_0, provided the thresholds $\mathcal{Q}$ are properly set. Numerical studies in Section 3 also support this point. Therefore, the complexity in this part does not exceed $p \cdot 3^{l_0}$. As a result, the computational complexity of the *PE*-step is moderate. Here, the bound $O(pK + p \cdot 3^{l_0})$ is conservative; the actual computational complexity is much smaller than this.

How does CASE perform? In Sections 2.7–2.9, we set up an asymptotic framework and show that CASE is asymptotically minimax in terms of the Hamming distance over a wide class of situations. In Sections 2.10–2.11, we apply CASE to the long-memory time series and the change-point model, and elaborate the optimality of CASE in such models with the so-called *phase diagram*.

2.7. *Asymptotic rare and weak model.* In this section, we add an asymptotic framework to the rare and weak signal model $RW(\varepsilon, \tau, \mu)$ introduced in Section 2.1. We use p as the driving asymptotic parameter and tie (ε, τ) to p through some fixed parameters.

In particular, we fix $\vartheta \in (0, 1)$ and model the sparse parameter ε by

$$\varepsilon = \varepsilon_p = p^{-\vartheta}. \tag{2.17}$$

Note that as p grows, the signal becomes increasingly sparse. It turns out that the most interesting range of signal strength is $\tau = O(\sqrt{\log(p)})$; see, for example, Ji and Jin (2012). For much smaller τ, successful recovery is impossible. For much larger τ, the problem is relatively easy. The critical value of τ depends on ϑ in a complicate way. In light of this, we fix $r > 0$, and let

$$\tau = \tau_p = \sqrt{2r \log(p)}. \tag{2.18}$$

At the same time, recalling that in $RW(\varepsilon, \tau, \mu)$, we require $\mu \in \Theta_p(\tau)$ so that $|\mu_i| \ge \tau$ for all $1 \le i \le p$. Fixing $a > 1$, we now further restrict μ to the following subset of $\Theta_p(\tau)$:

$$\Theta_p^*(\tau_p, a) = \{\mu \in \Theta_p(\tau_p) : \tau_p \le |\mu_i| \le a\tau_p, 1 \le i \le p\}. \tag{2.19}$$

DEFINITION 2.9. We call (2.17)–(2.19) the asymptotic rare and weak model $ARW(\vartheta, r, a, \mu)$.

Requiring the strength of each signal $\le a\tau_p$ is mainly for technical reasons, and hopefully, such a constraint can be removed in the near future. From a practical point of view, since usually we do not have sufficient information on μ, we prefer to have a larger a: we hope that when a is properly large, $\Theta_p^*(\tau_p, a)$ is broad enough, so that neither the optimal procedure nor the minimax risk needs to adapt to a.

Toward this end, we impose some mild regularity conditions on a and the Gram matrix G. Let g be the smallest integer such that

$$g \ge \max\{(\vartheta + r)^2/(2\vartheta r), m\}. \tag{2.20}$$

For any $p \times p$ Gram matrix G and $1 \le k \le p$, let $\lambda_k^*(G)$ be the minimum of the smallest eigenvalues of all $k \times k$ principle sub-matrices of G. Introduce

$$\widetilde{\mathcal{M}}_p(c_0, g) = \{G \text{ is a } p \times p \text{ Gram matrix}, \lambda_k^*(G) \ge c_0, 1 \le k \le g\}. \tag{2.21}$$

For any two subsets V_0 and V_1 of $\{1, 2, \ldots, p\}$, consider the optimization problem

$$\begin{aligned}&\big(\theta_*^{(0)}(V_0, V_1; G), \theta_*^{(1)}(V_0, V_1; G)\big)\\ &\quad = \operatorname{argmin}\big\{\big(\theta^{(1)} - \theta^{(0)}\big)' G\big(\theta^{(1)} - \theta^{(0)}\big)\big\},\end{aligned}$$

up to the constraints that $|\theta_i^{(k)}| \ge \tau_p$ if $i \in V_k$ and $\theta_i^{(k)} = 0$ otherwise, where $k = 0, 1$, and that in the special case of $V_0 = V_1$, the sign vectors of $\theta^{(0)}$ and $\theta^{(1)}$ are unequal. Introduce

$$a_g^*(G) = \max_{\{(V_0, V_1):\, |V_0 \cup V_1| \le g\}} \max\big\{\big\|\theta_*^{(0)}(V_0, V_1; G)\big\|_\infty, \big\|\theta_*^{(1)}(V_0, V_1; G)\big\|_\infty\big\}.$$

The following lemma is elementary, so we omit the proof.

LEMMA 2.3. *For any $G \in \widetilde{\mathcal{M}}_p(c_0, g)$, there is a constant $C = C(c_0, g) > 0$ such that $a_g^*(G) \le C$.*

In this paper, except for Section 2.11 where we discuss the change-point model, we assume

$$G \in \widetilde{\mathcal{M}}(c_0, g), \qquad a > a_g^*(G). \tag{2.22}$$

Under such conditions, $\Theta_p^*(\tau_p, a)$ is broad enough and the minimax risk (to be introduced below) does not depend on a. See Section 2.8 for more discussion.

For any variable selection procedure $\hat{\beta}$, we measure the performance by the Hamming distance

$$h_p(\hat{\beta}; \beta, G) = E\left[\sum_{j=1}^p 1\{\operatorname{sgn}(\hat{\beta}_j) \ne \operatorname{sgn}(\beta_j)\} \Big| X, \beta\right],$$

where the expectation is taken with respect to $\hat{\beta}$. Here, for any $p \times 1$ vector ξ, $\mathrm{sgn}(\xi)$ denotes the sign vector [for any number x, $\mathrm{sgn}(x) = 1, 0, -1$ when $x < 0$, $x = 0$, and $x > 0$ correspondingly].

Under $ARW(\vartheta, r, a, \mu)$, $\beta = b \circ \mu$, so the overall Hamming distance is

$$H_p(\hat{\beta}; \varepsilon_p, \mu, G) = E_{\varepsilon_p}\big[h_p(\hat{\beta}; \beta, G)|X\big],$$

where E_{ε_p} is the expectation with respect to the law of b. Finally, the minimax Hamming distance under $ARW(\vartheta, r, a, \mu)$ is

$$\mathrm{Hamm}_p^*(\vartheta, r, a, G) = \inf_{\hat{\beta}} \sup_{\mu \in \Theta_p^*(\tau_p, a)} H_p(\hat{\beta}; \varepsilon_p, \mu, G).$$

In next section, we will see that the minimax Hamming distance does not depend on a as long as (2.22) holds.

In many recent works, the *probability of exact support recovery* or *oracle property* is used to assess optimality; see, for example, Fan and Li (2001), Fan, Xue and Zou (2014), Zhao and Yu (2006), Zou (2006). However, when signals are rare and weak, exact support recovery is usually impossible, and the Hamming distance is a more appropriate criterion for assessing optimality. In comparison, study on the minimax Hamming distance is not only mathematically more demanding but also scientifically more relevant than that on the oracle property.

2.8. *Lower bound for the minimax Hamming distance.* We view the (global) Hamming distance as the aggregation of "local" Hamming errors. To construct a lower bound for the (global) minimax Hamming distance, the key is to construct lower bounds for "local" Hamming errors. Fix $1 \le j \le p$. The "local" Hamming error at index j is the risk we make among the neighboring indices of j in GOSD, say, $\{k : d(j,k) \le g\}$, where g is as in (2.20) and $d(j,k)$ is the geodesic distance between j and k in the GOSD. The lower bound for such a "local" Hamming error is characterized by an exponent ρ_j^*, which we now introduce.

For any subset $V \subset \{1, 2, \ldots, p\}$, let I_V be the $p \times 1$ vector such that the jth coordinate is 1 if $j \in V$ and 0 otherwise. Fixing two subsets V_0 and V_1 of $\{1, 2, \ldots, p\}$, we introduce

$$\varpi^*(V_0, V_1) = \tau_p^{-2} \min_{\theta^{(0)}, \theta^{(1)}} \big\{\big(\theta^{(1)} - \theta^{(0)}\big)' G\big(\theta^{(1)} - \theta^{(0)}\big)\big\}, \tag{2.23}$$

subject to $\{\theta^{(k)} = I_{V_k} \circ \mu^{(k)} : \mu^{(k)} \in \Theta_p^*(\tau_p, a), k = 0, 1, \mathrm{sgn}(\theta^{(0)}) \ne \mathrm{sgn}(\theta^{(1)})\}$, and let

$$\begin{aligned} \rho(V_0, V_1) &= \max\{|V_0|, |V_1|\}\vartheta \\ &\quad + \frac{1}{4}\left[\left(\sqrt{\varpi^*(V_0, V_1)r} - \frac{|(|V_1| - |V_0|)|\vartheta}{\sqrt{\varpi^*(V_0, V_1)r}}\right)_+\right]^2. \end{aligned} \tag{2.24}$$

The exponent $\rho_j^* = \rho_j^*(\vartheta, r, a, G)$ is defined by

$$\rho_j^*(\vartheta, r, a, G) = \min_{(V_0, V_1):\, j \in V_0 \cup V_1} \rho(V_0, V_1). \tag{2.25}$$

The notation L_p is frequently used in this paper.

DEFINITION 2.10. L_p, as a positive sequence indexed by p, is called a multi-$\log(p)$ term if for any fixed $\delta > 0$, $\lim_{p\to\infty} L_p p^{\delta} = \infty$ and $\lim_{p\to\infty} L_p p^{-\delta} = 0$.

It can be shown that $L_p p^{-\rho_j^*}$ provides a lower bound for the "local" minimax Hamming distance at index j, and that when (2.22) holds, $\rho_j^*(\vartheta, r, a, G)$ does not depend on a; see Lemma 16 in Jin, Zhang and Zhang (2014) for details. In the remaining part of the paper, we will write it as $\rho_j^*(\vartheta, r, G)$ for short.

At the same time, in order for the aggregation of all lower bounds for "local" Hamming errors to give a lower bound for the "global" Hamming distance, we need to introduce *graph of least favorables* (GOLF). Toward this end, recalling g and $\rho(V_0, V_1)$ as in (2.20) and (2.24), respectively, let

$$(V_{0j}^*, V_{1j}^*) = \operatorname*{argmin}_{\{(V_0, V_1):\, j \in V_0 \cup V_1, |V_0 \cup V_1| \le g\}} \rho(V_0, V_1),$$

and when there is a tie, pick the one that appears first lexicographically. We can think (V_{0j}^*, V_{1j}^*) as the "least favorable" configuration at index j.

DEFINITION 2.11. GOLF is the graph $\mathcal{G}^\diamond = (V, E)$ where $V = \{1, 2, \ldots, p\}$ and there is an edge between j and k if and only if $(V_{0j}^* \cup V_{1j}^*) \cap (V_{0k}^* \cup V_{1k}^*) \neq \varnothing$.

The following theorem is similar to Theorem 14 in Jin, Zhang and Zhang (2014), so we omit the proof.

THEOREM 2.1. *Suppose* (2.22) *holds so that* $\rho_j^*(\vartheta, r, a, G)$ *does not depend on the parameter* a *for sufficiently large* p. *As* $p \to \infty$, $\mathrm{Hamm}_p^*(\vartheta, r, a, G) \ge L_p[d_p(\mathcal{G}^\diamond)]^{-1} \sum_{j=1}^p p^{-\rho_j^*(\vartheta, r, G)}$, *where* $d_p(\mathcal{G}^\diamond)$ *is the maximum degree of all nodes in* $\mathcal{G}^\diamond$.

In many examples, including those of primary interest of this paper,

$$d_p(\mathcal{G}^\diamond) \le L_p. \tag{2.26}$$

In such cases, we have the following lower bound:

$$\mathrm{Hamm}_p^*(\vartheta, r, a, G) \ge L_p \sum_{j=1}^p p^{-\rho_j^*(\vartheta, r, G)}. \tag{2.27}$$

2.9. *Upper bound and optimality of CASE.* In this section, we show that in a broad context, provided the tuning parameters are properly set, CASE achieves the lower bound prescribed in Theorem 2.1, up to some L_p terms. Therefore, the lower bound in Theorem 2.1 is tight, and CASE achieves the optimal rate of convergence.

For a given $\gamma > 0$, we focus on linear models with the Gram matrix from

$$\mathcal{M}_p^*(\gamma, g, c_0, A_1) = \widetilde{\mathcal{M}}_p(c_0, g) \cap \mathcal{M}_p(\gamma, A_1),$$

where we recall that the two terms on the right-hand side are defined in (2.5) and (2.21), respectively. The following lemma is proved in Section B in Ke, Jin and Fan (2014).

LEMMA 2.4. *For $G \in \mathcal{M}_p^*(\gamma, g, c_0, A_1)$, the maximum degree of nodes in GOLF satisfies $d_p(\mathcal{G}^\diamond) \le L_p$.*

Combining Lemma 2.4 with Theorem 2.1, the lower bound (2.27) holds.

For any linear filter $D = D_{h,\eta}$, let $\varphi_\eta(z) = 1 + \eta_1 z + \cdots + \eta_h z^h$ be the so-called *characterization polynomial*. We need some regularity conditions:

- *Regularization Condition* A (*RCA*). For any root z_0 of $\varphi_\eta(z)$, $|z_0| \ge 1$.
- *Regularization Condition* B (*RCB*). There are constants $\kappa > 0$ and $c_1 > 0$ such that $\lambda_k^*(DGD') \ge c_1 k^{-\kappa}$ (λ_k^* is as in Section 2.8).

For many well-known linear filters such as adjacent differences, seasonal differences, etc., RCA is satisfied. Also, RCB is only a mild condition since κ can be any positive number. For example, RCB holds in the change-point model and long-memory time series model with certain D matrices. In general, κ is not 0 because when DG is sparse, DGD' is very likely to be approximately singular, and the associated value of λ_k^* can be small when k is large. This is true even for very simple G [e.g., $G = I_p$, $D = D_{1,\eta}$ and $\eta = (1, -1)'$].

At the same time, these conditions can be further relaxed. For example, for the change-point problem, the Gram matrix has barely any off-diagonal decay, and does not belong to $\mathcal{M}_p^*$. Nevertheless, with slight modification in the procedure, the main results continue to hold.

CASE uses tuning parameters $(\delta, m, \mathcal{Q}, \ell^{ps}, \ell^{pe}, u^{pe}, v^{pe})$. The choice of δ is flexible, and we usually set $\delta = 1/\log(p)$. For the main theorem below, we treat m as given. In practice, taking m to be a small integer (say, ≤ 3) is usually sufficient, unless the signals are relatively dense (say, $\vartheta < 1/4$). The choice of ℓ^{ps} and ℓ^{pe} are also relatively flexible, and letting ℓ^{ps} be a sufficiently large constant and ℓ^{pe} be $(\log(p))^\nu$ for some constant $\nu < (1 - 1/\alpha)/(\kappa + 1/2)$ is sufficient, where α is as in Definition 2.2, and κ is as in RCB.

At the same time, in principle, the optimal choices of (u^{pe}, v^{pe}) are

$$u^{pe} = \sqrt{2\vartheta \log p}, \qquad v^{pe} = \sqrt{2r \log p}, \tag{2.28}$$

which depend on the underlying parameters (ϑ, r) that are unknown to us. Despite this, our numeric studies in Section 3 suggest that the choices of (u^{pe}, v^{pe}) are relatively flexible; see Sections 3–4 for more discussions.

Last, we discuss how to choose $\mathcal{Q} = \{t(\hat{F}, \hat{N}) : (\hat{F}, \hat{N})$ are defined as in the *PS*-step$\}$. Let $t(\hat{F}, \hat{N}) = 2q\log(p)$, where $q > 0$ is a constant. It turns out that the main result (Theorem 2.2 below) holds as long as

$$q_0 \le q \le q^*(\hat{F}, \hat{N}), \tag{2.29}$$

where $q_0 > 0$ is an appropriately small constant, and for any subsets (F, N),

$$\begin{aligned} &q^*(F, N) \\ &\quad = \max\bigl\{q : (|F| + |N|)\vartheta + \bigl[\bigl(\sqrt{\tilde{\omega}(F, N)r} - \sqrt{q|F|}\bigr)_+\bigr]^2 \ge \psi(F, N)\bigr\}; \end{aligned} \tag{2.30}$$

here,

$$\begin{aligned} \psi(F, N) &= \frac{(|F| + 2|N|)\vartheta}{2} \\ &\quad + \begin{cases} \dfrac{1}{4}\omega(F, N)r, & |F| \text{ is even}, \\ \dfrac{\vartheta}{2} + \dfrac{1}{4}\bigl[\bigl(\sqrt{\omega(F, N)r} - \vartheta/\sqrt{\omega(F, N)r}\bigr)_+\bigr]^2, & |F| \text{ is odd}, \end{cases} \end{aligned} \tag{2.31}$$

with

$$\omega(F, N) = \min_{\xi \in \mathbb{R}^{|F|} : |\xi_i| \ge 1} \xi'\bigl[G^{F,F} - G^{F,N}\bigl(G^{N,N}\bigr)^{-1}G^{N,F}\bigr]\xi \tag{2.32}$$

and

$$\tilde{\omega}(F, N) = \min_{\xi \in \mathbb{R}^{|F|} : |\xi_i| \ge 1} \xi'\bigl[Q_{F,F} - Q_{F,N}(Q_{N,N})^{-1}Q_{N,F}\bigr]\xi, \tag{2.33}$$

where $Q_{F,N} = (B^{\mathcal{I}^{ps},F})'(H^{\mathcal{I}^{ps},\mathcal{I}^{ps}})^{-1}(B^{\mathcal{I}^{ps},N})$ with $\mathcal{I} = F \cup N$, and $Q_{N,F}$, $Q_{F,F}$ and $Q_{N,N}$ are defined similarly. Compared to (2.13), we see that $Q_{F,N}$, $Q_{F,N}$, $Q_{N,F}$ and $Q_{N,N}$ are all submatrices of Q. Hence, $\tilde{\omega}(F, N)$ can be viewed as a counterpart of $\omega(F, N)$ by replacing the submatrices of $G^{\mathcal{I},\mathcal{I}}$ by the corresponding ones of Q.

From a practical point of view, there is a trade-off in choosing q: a larger q would increase the number of falsely selected variables in the *PS*-step, but would also reduce the computational cost in the *PE*-step. The following is a convenient choice which we recommend in this paper:

$$t(\hat{F}, \hat{N}) = 2\tilde{q}|\hat{F}|\log(p), \tag{2.34}$$

where $0 < \tilde{q} < c_0 r/4$ is a constant, and c_0 is as in $\mathcal{M}_p^*(\gamma, g, c_0, A_1)$.

We are now ready for the main result of this paper.

THEOREM 2.2. *Suppose that for sufficiently large* p, $G \in \mathcal{M}_p^*(\gamma, g, c_0, A_1)$, $D_{h,\eta}G \in \mathcal{M}_p(\alpha, A_0)$ *with* $\alpha > 1$ *and that RCA-RCB hold. Consider* $\hat{\beta}^{\text{case}} = \hat{\beta}^{\text{case}}(Y; \delta, m, \mathcal{Q}, \ell^{ps}, \ell^{pe}, u^{pe}, v^{pe}, D_{h,\eta}, X, p)$ *with the tuning parameters specified above. Then as* $p \to \infty$,

$$\sup_{\mu \in \Theta_p^*(\tau_p, a)} H_p\big(\hat{\beta}^{\text{case}}; \varepsilon_p, \mu, G\big) \le L_p\left[p^{1-(m+1)\vartheta} + \sum_{j=1}^{p} p^{-\rho_j^*(\vartheta, r, G)}\right] + o(1). \tag{2.35}$$

Combine Lemma 2.4 and Theorem 2.2. Given the parameter m is appropriately large, both the upper bound and the lower bound are tight, and CASE achieves the optimal rate of convergence prescribed by

$$\text{Hamm}_p^*(\vartheta, r, a, G) = L_p \sum_{j=1}^{p} p^{-\rho_j^*(\vartheta, r, G)} + o(1). \tag{2.36}$$

Theorem 2.2 is proved in Section A in Ke, Jin and Fan (2014), where we explain the key idea behind the procedure, as well as the selection of the tuning parameters.

2.10. *Application to the long-memory time series model.* The long-memory time series model in Section 1 can be written as a regression model,

$$Y = X\beta + z, \qquad z \sim N(0, I_n),$$

where the Gram matrix G is asymptotically Toeplitz and has slow off-diagonal decays. Without loss of generality, we consider the following idealized case where G is an exact Toeplitz matrix generated by a spectral density f:

$$G(i, j) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos\big(|i - j|\omega\big) f(\omega)\, d\omega, \qquad 1 \le i, j \le p. \tag{2.37}$$

In the literature [Chen, Hurvich and Lu (2006), Moulines and Soulier (1999)], the spectral density for a long-memory process is usually characterized as

$$f(\omega) = \big|1 - e^{\sqrt{-1}\omega}\big|^{-2\phi} f^*(\omega), \tag{2.38}$$

where $\phi \in (0, 1/2)$ is the long-memory parameter, $f^*(\omega)$ is a positive symmetric function that is continuous on $[-\pi, \pi]$ and is twice differentiable except at $\omega = 0$.

In this model, the Gram matrix is nonsparse, but it is sparsifiable. To see the point, let $\eta = (1, -1)'$ and let $D = D_{1,\eta}$ be the first-order adjacent row-differencing. On one hand, since the spectral density f is singular at the origin, it follows from the Fourier analysis that $|G(i, j)| \ge C(1 + |i - j|)^{-(1-2\phi)}$, and hence G is nonsparse. On the other hand, it is seen that

$$B(i, j) = \sqrt{-1} \int_{|j-i|}^{|j-i|+1} \widehat{\omega f(\omega)}(\lambda)\, d\lambda,$$

where we recall that $B = DG$ and note that $\hat{g}$ denotes the Fourier transform of g. Compared to $f(\omega)$, $\omega f(\omega)$ is nonsingular at the origin. Additionally, it is seen that $B \in \mathcal{M}_p(2-2\phi, A)$, where $2-2\phi > 1$, so B is sparse (a similar claim applies to $H = DGD'$). This shows that G is sparsifiable by adjacent row-differencing.

In this example, there is a function $\rho^*_{\mathrm{lts}}(\vartheta, r; f)$ that only depends on (ϑ, r, f) such that

$$\max_{\{j:\log(p)\le j\le p-\log(p)\}} \{|\rho^*_j(\vartheta, r, G) - \rho^*_{\mathrm{lts}}(\vartheta, r; f)|\} \to 0 \qquad \text{as } p \to \infty,$$

where the subscript "lts" stands for long-memory time series. The following theorem can be derived from Theorem 2.2, and is proved in Section B in Ke, Jin and Fan (2014).

THEOREM 2.3. *For a long-memory time series model where* $|(f^*)''(\omega)| \le C|\omega|^{-2}$, *the minimax Hamming distance then satisfies* $\mathrm{Hamm}^*_p(\vartheta, r, G) = L_p p^{1-\rho^*_{\mathrm{lts}}(\vartheta,r;f)}$. *If we apply CASE by letting* $(m+1)\vartheta > \rho^*_{\mathrm{lts}}(\vartheta, r; f)$, $\eta = (1, -1)'$, *and the tuning parameters be set as in Section 2.9, then*

$$\sup_{\mu\in\Theta^*_p(\tau_p,a)} H_p\big(\hat{\beta}^{\mathrm{case}}; \varepsilon_p, \mu, G\big) \le L_p p^{1-\rho^*_{\mathrm{lts}}(\vartheta,r;f)} + o(1).$$

Theorem 2.3 can be interpreted by the so-called *phase diagram*. Phase diagram is a way to visualize settings where the signals are so rare and weak that successful variable selection is simply impossible [Ji and Jin (2012)]. In detail, for a spectral density f and $\vartheta \in (0, 1)$, let $r^*_{\mathrm{lts}}(\vartheta) = r^*_{\mathrm{lts}}(\vartheta; f)$ be the unique solution of $\rho^*_{\mathrm{lts}}(\vartheta, r; f) = 1$. Note that $r = r^*_{\mathrm{lts}}(\vartheta)$ characterizes the minimum signal strength required for exact support recovery with high probability. The following proposition is proved in Section B in Ke, Jin and Fan (2014).

LEMMA 2.5. *Under the conditions of Theorem 2.3, if* $(f^*)''(0)$ *exists, then* $r^*_{\mathrm{lts}}(\vartheta; f)$ *is a decreasing function in* ϑ, *with limits* 1 *and* $\frac{2}{\pi}\int_{-\pi}^{\pi} f^{-1}(\omega)\,d\omega$ *as* $\vartheta \to 1$ *and* $\vartheta \to 0$, *respectively.*

Call the two-dimensional space $\{(\vartheta, r) : 0 < \vartheta < 1, r > 0\}$ the *phase space*. Interestingly, there is a partition of the phase space as follows.

- *Region of no recovery* $\{(\vartheta, r) : 0 < r < \vartheta, 0 < \vartheta < 1\}$. In this region, the minimax Hamming distance $\gtrsim p\varepsilon_p$, where $p\varepsilon_p$ is approximately the number of signals. In this region, the signals are too rare and weak and successful variable selection is impossible.
- *Region of almost full recovery* $\{(\vartheta, r) : \vartheta < r < r^*_{\mathrm{lts}}(\vartheta; f), 0 < \vartheta < 1\}$. In this region, the minimax Hamming distance is much larger than 1 but much smaller than $p\varepsilon_p$. Therefore, the optimal procedure can recover most of the signals but not all of them.

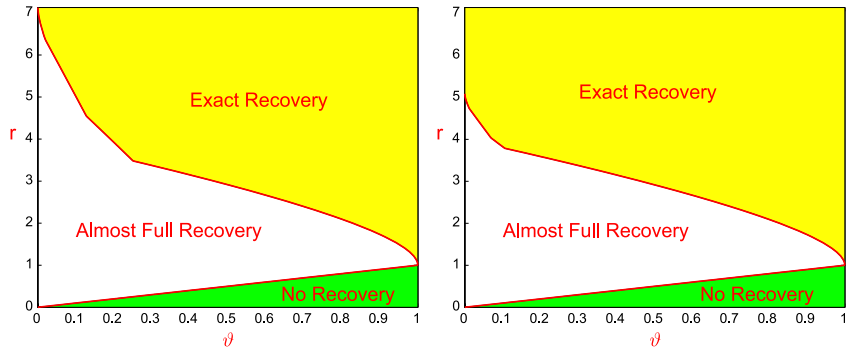

FIG. 3. *Phase diagrams corresponding to the* FARIMA$(0, \phi, 0)$ *process. Left*: $\phi = 0.35$. *Right*: $\phi = 0.25$.

- *Region of exact recovery* $\{(\vartheta, r) : r > r^*_{\mathrm{lts}}(\vartheta; f), 0 < \vartheta < 1\}$. In this region, the minimax Hamming distance is $o(1)$. Therefore, the optimal procedure recovers all signals with probability ≈ 1.

Because of the partition of the phase space, we call this the *phase diagram*.

From time to time, we wish to have a more explicit formula for the rate $\rho^*_{\mathrm{lts}}(\vartheta, r; f)$ and the critical value $r^*_{\mathrm{lts}}(\vartheta; f)$. In general, this is a hard problem, but both quantities can be computed numerically when f is given. In Figure 3, we display the phase diagrams for the *autoregressive fractionally integrated moving average* process (FARIMA) with parameters $(0, \phi, 0)$ [Fan and Yao (2003)], where

$$f^*(\omega) = \frac{\Gamma^2(1-\phi)}{\Gamma(1-2\phi)}. \tag{2.39}$$

Take $\phi = 0.35, 0.25$, for example, $r^*_{\mathrm{lts}}(\vartheta; f) \approx 7.14, 5.08$ for small ϑ.

2.11. *Application to the change-point model.* The change-point model in the Introduction can be viewed as a special case of model (1.1), where β is as in (2.3), and the Gram matrix satisfies

$$G(i, j) = \min\{i, j\}, \qquad 1 \le i, j \le p. \tag{2.40}$$

For technical reasons, it is more convenient *not* to normalize the diagonals of G to 1.

The change-point model can be viewed as an "extreme" case of what is studied in this paper. On one hand, the Gram matrix G is "ill-posed," and each row of G does not satisfy the condition of off-diagonal decay in Theorem 2.2. On the other hand, G has a very special structure which can be largely exploited. In fact, if we sparsify G using the linear filter $D = D_{2,\eta}$, where $\eta = (1, -2, 1)'$, it is seen that $B = DG = I_p$, and $H = DGD'$ is a tri-diagonal matrix with $H(i, j) = 2 \cdot 1\{i = j\} - 1\{|i - j| = 1\} - 1\{i = j = p\}$, which are very simple matrices. For these reasons, we modify the CASE as follows:

- Due to the simple structure of B, we do not need patching in the *PS*-step (i.e., $\ell^{ps} = 0$).
- For the same reason, the choices of thresholds $t(\hat{F}, \hat{N})$ are more flexible than before, and taking $t(\hat{F}, \hat{N}) = 2q\log(p)$ for a proper constant $q > 0$ works.
- Since H is "extreme" (the smallest eigenvalue tends to 0 as $p \to \infty$), we have to modify the *PE*-step carefully.

In detail, the *PE*-step for the change-point model is as follows. Given ℓ^{pe}, let $\mathcal{G}^+$ be as in Definition 2.7. Recall that $\mathcal{U}_p^*$ denotes the set of all retained indices at the end of the *PS*-step. We view $\mathcal{U}_p^*$ as a subgraph of $\mathcal{G}^+$, and let $\mathcal{I} \lhd \mathcal{U}_p^*$ be one of its components. The goal is to split $\mathcal{I}$ into N different subsets

$$\mathcal{I} = \mathcal{I}^{(1)} \cup \cdots \cup \mathcal{I}^{(N)},$$

and for each subset $\mathcal{I}^{(k)}$, $1 \le k \le N$, we construct a patched set $\mathcal{I}^{(k),pe}$. We then estimate $\beta^{\mathcal{I}^{(k)}}$ separately using (2.16). Putting $\beta^{\mathcal{I}^{(k)}}$ together gives our estimate of $\beta^{\mathcal{I}}$.

The subsets $\{(\mathcal{I}^{(k)}, \mathcal{I}^{(k),pe})\}_{k=1}^N$ are recursively constructed as follows. Denote $l = |\mathcal{I}|$, $M = (\ell^{pe}/2)^{1/(l+1)}$, and write

$$\mathcal{I} = \{j_1, j_2, \ldots, j_l\}, \qquad j_1 < j_2 < \cdots < j_l.$$

First, letting k_1 be the largest index such that $j_{k_1} - j_{k_1-1} > \ell^{pe}/M$, define

$$\mathcal{I}^{(1)} = \{j_{k_1}, \ldots, j_l\} \quad \text{and} \quad \mathcal{I}^{(1),pe} = \{j_{k_1} - \ell^{pe}/(2M), \ldots, j_l + \ell^{pe}/2\}.$$

Next, letting $k_2 < k_1$ be the largest index such that $j_{k_2} - j_{k_2-1} > \ell^{pe}/M^2$, define

$$\mathcal{I}^{(2)} = \{j_{k_2}, \ldots, j_{k_1}\}, \qquad \mathcal{I}^{(2),pe} = \{j_{k_2} - \ell^{pe}/(2M^2), \ldots, j_{k_1} + \ell^{pe}/(2M)\}.$$

Continue this process until for some N, $1 \le N \le l$, $k_N = 1$. In this construction, for each $1 \le k \le N$, if we arrange all the nodes of $\mathcal{I}^{(k),pe}$ in the ascending order, then the number of nodes in front of $\mathcal{I}^{(k)}$ is significantly smaller than the number of nodes behind $\mathcal{I}^{(k)}$.

In practice, we introduce a suboptimal but much simpler patching approach as follows. Fix a component $\mathcal{I} = \{j_1, \ldots, j_l\}$ of $\mathcal{U}_p^*$. In this approach, instead of splitting it into smaller sets and patching them separately as in the previous approach, we patch the whole set $\mathcal{I}$ by

$$\mathcal{I}^{pe} = \{i : j_1 - \ell^{pe}/4 < i < j_l + 3\ell^{pe}/4\}, \tag{2.41}$$

and estimate $\beta^{\mathcal{I}}$ using (2.16). Our numeric studies show that two approaches have comparable performances.

Define

$$\rho_{\mathrm{cp}}^*(\vartheta, r) = \begin{cases} \vartheta + r/4, & r/\vartheta \le 6 + 2\sqrt{10}, \\ 3\vartheta + (r/2 - \vartheta)^2/(2r), & r/\vartheta > 6 + 2\sqrt{10}, \end{cases} \tag{2.42}$$

where "cp" stands for change-point. Choose the tuning parameters of CASE such that

$$\ell^{pe} = 2\log(p), \qquad u^{pe} = \sqrt{2\vartheta\log(p)} \quad \text{and} \quad v^{pe} = \sqrt{2r\log(p)}, \tag{2.43}$$

that $(m+1)\vartheta \geq \rho^*_{\mathrm{cp}}(\vartheta, r)$, and that $0 < q < \frac{r}{4}(\sqrt{2}-1)^2$ [recall that we take $t(\hat{F}, \hat{N}) = 2q\log(p)$ for all $(\hat{F}, \hat{N})$ in the change-point setting]. Note that the choice of ℓ^{pe} is different from that in Section 2.5. The main result in this section is the following theorem which is proved in Section B in Ke, Jin and Fan (2014).

THEOREM 2.4. *For the change-point model, the minimax Hamming distance satisfies* $\mathrm{Hamm}^*_p(\vartheta, r, G) = L_p p^{1-\rho^*_{\mathrm{cp}}(\vartheta,r)}$. *Furthermore,* $\hat{\beta}^{\mathrm{case}}$ *with the tuning parameters specified above satisfies*

$$\sup_{\mu\in\Theta^*_p(\tau_p,a)} H_p\big(\hat{\beta}^{\mathrm{case}}; \varepsilon_p, \mu, G\big) \leq L_p p^{1-\rho^*_{\mathrm{cp}}(\vartheta,r)} + o(1).$$

It is noteworthy that the exponent $\rho^*_{\mathrm{cp}}(\vartheta, r)$ has a phase change depending on the ratios of r/ϑ. The insight is, when $r/\vartheta < 6 + 2\sqrt{10}$, the minimax Hamming distance is dominated by the Hamming errors we make in distinguishing between an isolated change-point and a pair of adjacent change-points, and when $r/\vartheta > 6 + 2\sqrt{10}$, the minimax Hamming distance is dominated by the Hamming errors of distinguishing the case of consecutive change-point triplets (say, change-points at $\{j-1, j, j-1\}$) from the case where we do not have a change-point in the middle of the triplets (i.e., the change-points are only at $\{j-1, j+1\}$).

Similarly, the main results on the change-point problem can be visualized with the phase diagram in Figure 4. An interesting point is that it is possible to have

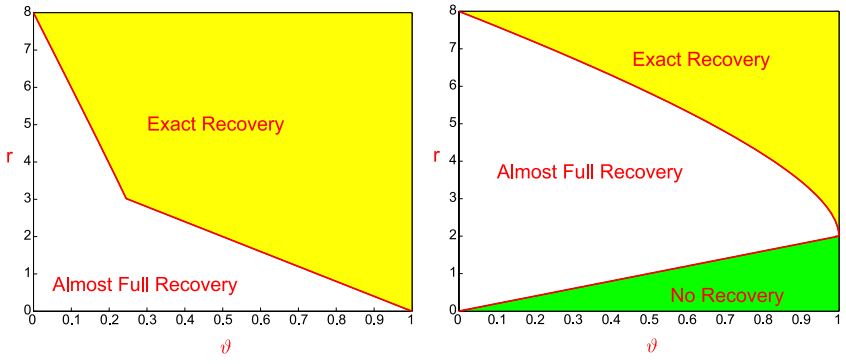

FIG. 4. *Phase diagrams corresponding to the change-point model. Left*: *CASE*; *the boundary is decided by* $(4-10\vartheta)+2\sqrt{(2-5\vartheta)^2-\vartheta^2}$ (*left part*) *and* $4(1-\vartheta)$ (*right part*). *Right*: *hard thresholding*; *the upper boundary is decided by* $2(1+\sqrt{1-\vartheta})^2$ *and the lower boundary is decided by* 2ϑ.

almost full recovery even when the signal strength parameter τ_p is as small as $o(\sqrt{2\log(p)})$. See the proof of Theorem 2.4 for details.

Alternatively, one may apply the following approach to the change-point problem. Treat the liner change-point model as a regression model $Y = X\beta + z$ as in Section 1 (page 2203), and let $W = (X'X)^{-1}X'Y$ be the least-squares estimate. It is seen that $W \sim N(\beta, \Sigma)$, where we note that $\Sigma = (X'X)^{-1}$ is tridiagonal and coincides with H. In this simple setting, a natural approach is to apply a coordinate-wise thresholding $\hat{\beta}_j^{\text{thresh}} = W_j 1\{|W_j| > t\}$ to locate the signals. But this neglects the covariance of W in detecting the locations of the signals and is not optimal even with the ideal choice of thresholding parameter t_0, since the corresponding risk satisfies

$$\sup_{\{\mu \in \Theta_p^*(\tau_p, a)\}} H_p\big(\hat{\beta}^{\text{thresh}}(t_0); \varepsilon_p, \mu, G\big) = L_p p^{1-(r/2+\vartheta)^2/(2r)}.$$

The proof of this is elementary and omitted. The phase diagram of this method is displayed in Figure 4, right panel, which suggests the method is nonoptimal.

Other popular methods in locating multiple change-points include the global methods [Harchaoui and Lévy-Leduc (2010), Olshen et al. (2004), Tibshirani (1996), Yao and Au (1989)] and local methods [e.g., SaRa in Niu and Zhang (2012)]. The global methods are usually computationally expensive and can hardly be optimal due to the strong correlation nature of this problem. Our procedure is related to the local methods but is different in important ways. Our method exploits the graphical structures and uses the GOSD to guide both the screening and cleaning, but SaRa does not utilize the graphical structures and can be shown to be nonoptimal.

To conclude the section, we remark that the change-point model constitutes a special case of the settings we discuss in the paper, where setting some of the tuning parameters is more convenient than in the general case. First, for the change-point model, we can simply set $\delta = 0$ and $\ell^{ps} = 0$. Second, there is an easy-to-compute preliminary estimator available. On the other hand, the performance of CASE is substantially better than the other methods in many situations. We believe that CASE is potentially a very useful method in practice for the change-point problem.

3. Simulations. We conducted a small-scale numeric study where we compare CASE and several popular approaches. The study contains two parts, Section 3.1 and Section 3.2, where we investigate the change-point model and the long-memory time series model, respectively.

In this section, $s_p = p\varepsilon_p$ for convenience. The core tuning parameters for CASE are $(\mathcal{Q}, u^{pe}, v^{pe}, \ell^{ps}, \ell^{pe})$. We streamline these tuning parameters in a way so they only depend on two tuning parameters (s_p, τ_p) (calibrating the sparsity and the minimum signal strength, resp.). Therefore, essentially, CASE only uses two tuning parameters. Our experiments show that the performance of CASE is relatively

insensitive to these two tuning parameters. Furthermore, these two tuning parameters can be set in a data driven fashion, especially in the change-point model. See details below.

We set $m = 2$ so that in the screening stage of CASE, bivariate screening is the highest order screening we use. At least for examples considered here, using a higher-order screening does not have a significant improvement. For long-memory time series, we need a regularization parameter δ (but we do not need it for the change-point model). The guideline for choosing δ is to make sure the maximum degree of GOSD is 15 (say) or smaller. In this section, we choose $\delta = 2.5/\log(p)$. The maximum degree of GOSD is much higher if we choose a much smaller δ; in this case, CASE has similar performance, but is computationally much slower.

3.1. *Change-point model.* In this section, we use model (1.3) to investigate the performance of CASE in identifying multiple change-points. For a given set of parameters (p, ϑ, r, a), we set $\varepsilon_p = p^{-\vartheta}$ and $\tau_p = \sqrt{2r\log(p)}$. First, we generate a $(p-1)\times 1$ vector β by $\beta_j \overset{\text{i.i.d.}}{\sim} (1-\varepsilon_p)\nu_0 + \frac{\varepsilon_p}{2}U(\tau_p, a\tau_p) + \frac{\varepsilon_p}{2}U(-a\tau_p, -\tau_p)$, where $U(s,t)$ is the uniform distribution over $[s,t]$ [when $s = t$, $U(s,t)$ represents the point mass at s]. Next, we construct the mean vector θ in model (1.3) by $\theta_j = \theta_{j-1} + \beta_{j-1}$, $2 \le j \le p$. Last, we generate the data vector Y by $Y \sim N(\theta, I_p)$.

CASE, when applied to the change-point model, requires tuning parameters $(m, \delta, \mathcal{Q}, \ell^{ps}, \ell^{pe}, u^{pe}, v^{pe})$. Denote by $s_p \equiv p\varepsilon_p = p^{1-\vartheta}$ the average number of signals. Given (s_p, τ_p), we determine the tuning parameters as follows: As mentioned earlier, we take $m = 2$, $\delta = 0$ and $\ell^{ps} = 0$. Also, we take $\ell^{pe} = 10\log(p/s_p)$, $u^{pe} = \sqrt{2\log(p/s_p)}$ and $v^{pe} = \tau_p$. $\mathcal{Q}$ contains thresholds $t(F,N)$ for each pair of sets (F,N); we take $t(F,N) = 2q(F,N)\log(p)$ with

$$q(F,N) = 0.8 \times \begin{cases} (r\tilde{\omega} + |F|\vartheta)^2/(4r\tilde{\omega}), & \tilde{\omega} > |F|\vartheta, \\ r\tilde{\omega}, & \tilde{\omega} \le |F|\vartheta, \end{cases} \tag{3.1}$$

where $\vartheta = \log(p/s_p)$, $r = \tau_p^2/(2\log(p))$ and $\tilde{\omega} = \tilde{\omega}(F,N)$ is given in (2.33). With these choices, CASE only depends on two parameters (s_p, τ_p).

Experiment 1a. We compare CASE with the lasso [Tibshirani (1996)], SCAD [Fan and Li (2001)] (penalty shape parameter $a = 3.7$), MC+ [Zhang (2010)] (penalty shape parameter $\gamma = 1.1$) and SaRa [Niu and Zhang (2012)]. For tuning parameters $\lambda > 0$ and $h > 0$ (integer), SaRa takes the following form:

$$\hat{\beta}_i^{\text{SaRa}} = W_i \cdot 1\{|W_i| > \lambda\} \qquad \text{where } W_i = \frac{1}{h}\Bigg(\sum_{j=i+1}^{i+h} Y_j - \sum_{j=i-h+1}^{i} Y_j\Bigg).$$

The tuning parameters for the lasso, SCAD, MC+ and SaRa are ideally set (pretending we know β). For CASE, all tuning parameters depend on (s_p, τ_p), so we implement the procedure using the true values of (s_p, τ_p); this yields slightly inferior results than that of setting (s_p, τ_p) ideally (pretending we know β, as we do in

TABLE 1
Comparison of Hamming errors (*Experiment* 1a; *change-point model*)

ϑ	s_p		τ_p					
			4.0	4.5	5.0	5.5	6.0	6.5
0.3	338.4	CASE	105.8	63.9	37.6	18.5	8.9	4.8
		lasso	371.7	370.0	371.5	370.1	371.5	369.8
		SCAD	370.6	368.3	370.5	368.2	369.3	369.2
		MC+	374.0	372.1	374.3	372.5	373.6	373.1
		SaRa	175.6	144.0	107.8	73.7	49.0	32.3
			3.0	3.5	4.0	4.5	5.0	5.5
0.45	108.3	CASE	50.1	35.5	26.3	20.0	12.8	6.2
		lasso	103.2	104.1	103.8	103.8	104.9	104.3
		SCAD	101.8	102.7	102.1	102.0	102.9	102.5
		MC+	103.7	104.7	104.4	104.3	105.4	104.8
		SaRa	78.9	72.0	66.2	63.4	61.9	60.4
			3.0	3.5	4.0	4.5	5.0	5.5
0.6	30.2	CASE	14.4	11.1	8.9	6.7	5.0	3.9
		lasso	29.3	29.2	29.3	29.7	27.7	29.3
		SCAD	27.7	27.7	27.9	27.4	26.1	27.1
		MC+	29.8	29.8	29.8	30.2	28.4	29.8
		SaRa	20.4	17.0	13.6	10.9	8.6	6.8
			3.0	3.5	4.0	4.5	5.0	5.5
0.75	8.4	CASE	3.5	2.9	2.4	1.8	1.6	1.3
		lasso	8.2	8.3	8.5	8.8	8.0	8.5
		SCAD	6.8	7.0	7.0	6.9	6.6	6.6
		MC+	8.7	8.8	9.1	9.2	8.7	9.1
		SaRa	5.2	4.5	3.8	3.0	2.4	2.0

the lasso, SCAD, MC+ and SaRa), so our comparison in this setting is fair. Note that even when (s_p, τ_p) are given, it is unclear how to set the tuning parameters of the lasso, SCAD, MC+ and SaRa.

Fix $p = 5000$ and $a = 1$. We let ϑ range in $\{0.3, 0.45, 0.6, 0.75\}$ and τ_p range in $\{3, 3.5, \ldots, 6.5\}$. The parameters fall into the regime where exact-recovery is impossible. Table 1 reports the average Hamming errors of 100 independent repetitions. We see that CASE consistently outperforms other methods, especially when ϑ is small, that is, signals are less sparse.

We also observe that three *global penalization methods*, the lasso, SCAD and MC+, perform unsatisfactorily, with Hamming errors comparable to the expected number of signals s_p. It suggests that the *global penalization methods* are not appropriate for the change-point model when the signals are rare and weak. Similar conclusions can be drawn in most experiments in this section. To save space, we only report results of the lasso, SCAD and MC+ in this experiment.

TABLE 2
Comparison of Hamming errors (Experiment 1b). *"adCASE" stands for adaptive CASE, where* (s_p, τ_p) *are estimated from SaRa* [*where* (λ, h) *are set by BIC*]

ϑ	s_p		τ_p					
			4.0	4.5	5.0	5.5	6.0	6.5
0.3	338.4	CASE	105.8	63.9	37.6	18.5	8.9	4.8
		adCASE	100.3	63.6	37.8	18.6	8.9	4.8
		SaRa	190.7	162.0	131.3	98.0	68.2	47.1
			3.0	3.5	4.0	4.5	5.0	5.5
0.45	108.3	CASE	50.1	35.5	26.3	20.0	12.8	6.2
		adCASE	48.6	33.9	26.0	20.8	16.6	9.7
		SaRa	86.1	76.7	71.4	66.7	65.0	62.8
			3.0	3.5	4.0	4.5	5.0	5.5
0.6	30.2	CASE	14.4	11.1	8.9	6.7	5.0	3.9
		adCASE	14.0	11.0	8.8	6.5	4.8	3.4
		SaRa	35.7	28.5	24.1	19.9	15.8	11.9
			3.0	3.5	4.0	4.5	5.0	5.5
0.75	8.4	CASE	3.5	2.9	2.4	1.8	1.6	1.3
		adCASE	3.7	3.0	2.2	1.8	1.5	1.3
		SaRa	13.3	11.5	8.0	5.2	4.0	2.9

Experiment 1b. In this experiment, we investigate the performance of CASE with (s_p, τ_p) estimated by SaRa; we call this the *adaptive CASE*. In detail, we estimate (s_p, τ_p) by $\hat{s}_p = \sum_{j=1}^{p} 1\{\hat{\beta}_j^{\mathrm{SaRa}} \neq 0\}$ and $\hat{\tau}_p = \mathrm{median}(\{|\hat{\beta}_j^{\mathrm{SaRa}}| : \hat{\beta}_j^{\mathrm{SaRa}} \neq 0\})$, where the tuning parameters (λ, h) of SaRa are determined by minimizing $\mathrm{BIC}(\hat{\beta}) = \frac{1}{2}\|Y - X\hat{\beta}\|^2 + \log(p) \cdot \|\hat{\beta}\|_0$; this is a slight modification of Bayesian information criteria (BIC).

For experiment, we use the same setting as in Experiment 1a. Table 2 reports the average Hamming errors of CASE, SaRa and the adaptive CASE based on 100 independent repetitions. First, the adaptive CASE [CASE but (s_p, τ_p) are estimated by SaRa] has a very similar performance to CASE. Second, although the adaptive CASE uses SaRa as the preliminary estimator, its performance is substantially better than that of SaRa (and other methods in the same setting; see Experiment 1a).

Experiment 2. In this experiment, we consider the post-filtering model, model (1.4), associated with the change-point model, and illustrate that the seeming simplicity of this model (where $DG = I_p$, and DGD' is tri-diagonal) does not mean it is a trivial setting for variable selection. In particular, if we naively apply the L^0/L^1-penalization to the post-filtering model, we end up with naive soft/hard thresholding; we illustrate our point by showing that CASE significantly outperforms naive thresholding (since we use Hamming distance as the loss function, there is no difference between soft and hard thresholding). For both CASE and naive thresholding, we set tuning parameters assuming (s_p, τ_p) as known.

TABLE 3
Comparison of Hamming errors (Experiment 2; change-point model), $p = 10^6$. "nHT" stands for naive hard thresholding

			τ_p								
ϑ	s_p		**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
0.35	7943	CASE	956.7	332.6	117.5	49.1	24.1	13.9	10.6	7.7	7.3
		nHT	4430.5	2381.3	1085.8	418.1	139.7	41.9	11.0	2.5	0.5
0.50	1000	CASE	195.3	68.8	20.8	5.0	1.3	0.7	0.4	0.1	0.2
		nHT	767.9	489.0	250.8	105.3	38.4	12.4	3.5	0.7	0.2
0.75	32	CASE	9.3	3.1	2.3	0.4	0.1	0.1	0.1	0.0	0.0
		nHT	31.1	25.6	15.7	8.3	3.2	1.8	0.5	0.0	0.0

The threshold of naive thresholding is set as $(r + 2\vartheta)^2/(2r) \cdot \log(p)$, where $\vartheta = \log(p/s_p)$ and $r = \tau_p^2/(2\log(p))$; this threshold choice is known as theoretically optimal.

Fix $p = 10^6$ and $a = 1$ (so that the signals have equal strengths). Let ϑ range in $\{0.35, 0.5, 0.75\}$, and τ_p range in $\{5, \ldots, 13\}$. Table 3 reports the average Hamming errors of 50 independent repetitions, which show that CASE outperforms naive thresholding in most cases, especially when ϑ is small or τ_p is small. It suggests that the post-filtering model remains largely nontrivial, and to deal with it, we need sophisticated methods.

Experiment 3. In this experiment, we fix $(p, \vartheta, \tau_p) = (5000, 0.50, 4.5)$, and let a range in $\{1, 1.5, \ldots, 3\}$ (so signals may have different strengths). We investigate a case where the signals have the "half-positive-half-negative" sign pattern, that is, $\beta_j \overset{\text{i.i.d.}}{\sim} (1-\varepsilon_p)\nu_0 + \frac{\varepsilon_p}{2}U(\tau_p, a\tau_p) + \frac{\varepsilon_p}{2}U(-a\tau_p, -\tau_p)$, and a case where the signals have the "all-positive" sign pattern, that is, $\beta_j \overset{\text{i.i.d.}}{\sim} (1-\varepsilon_p)\nu_0 + \varepsilon_p U(\tau_p, a\tau_p)$. We compare CASE with SaRa for different values of a and sign patterns (we do not include the lasso, SCAD, MC+ in this particular experiment, for at least for the experiments reported above, they are inferior to SaRa). The tuning parameters for both CASE and SaRa are set ideally as in Experiment 1a. The results of 50 independent repetitions are reported in Table 4, which suggest that CASE uniformly outperforms SaRa for various values of a and the two sign patterns.

3.2. *Long-memory time series model.* In this section, we investigate long-memory time series, focusing on the FARIMA$(0, \phi, 0)$ process [Fan and Yao (2003)], where ϕ is the long-memory parameter. We let $X = G^{1/2}$ where G is constructed according to (2.37)–(2.39). For β generated in ways to be specified, we let $Y \sim N(X\beta, I_p)$.

CASE uses tuning parameters $(m, \delta, \mathcal{Q}, \ell^{ps}, \ell^{pe}, u^{pe}, v^{pe})$, which are set in the same way as in the change-point model, except for two differences. First, different

TABLE 4
Comparison of Hamming errors (Experiment 3; change-point model) for different choices of a (the ratio between the maximum and minimum signal strengths) and for two sign patterns "half–half" and "all positive." $p = 5000$, $\vartheta = 0.5$, $s_p = 70.7$ and $\tau_p = 4.5$

		a				
		1	**1.5**	**2**	**2.5**	**3**
Half–half	CASE	14.26	6.32	5.50	4.78	4.56
	SaRa	24.98	18.96	16.56	14.00	12.50
All-positive	CASE	13.44	6.18	4.90	5.38	4.14
	SaRa	24.26	18.58	16.80	13.66	12.12

from that in the change-point model, we need a regularization parameter δ which we set as $2.5/\log(p)$. Second, we take $\ell^{ps} = \ell^{pe}/2$.

Experiment 4a. In this experiment, we compare CASE with the lasso, SCAD (shape parameter $a = 3.7$) and MC+ (shape parameter $\gamma = 2$). Fixing $p = 5000$ and $\phi = 0.35$, we let ϑ range in $\{0.35, 0.45, 0.55\}$, and let τ_p range in $\{4, \ldots, 8\}$. For each pair of (ϑ, τ_p), we generate the vector β by $\beta_j \overset{\text{i.i.d.}}{\sim} (1-\varepsilon_p)\nu_0 + \frac{\varepsilon_p}{2}\nu_{\tau_p} + \frac{\varepsilon_p}{2}\nu_{-\tau_p}$. Similar to that in Experiment 1a, the tuning parameters of CASE are set assuming (s_p, τ_p) as known, and the tuning parameters of the lasso, SCAD and MC+ are set ideally to minimize the Hamming error (assuming β is known). By similar argument as in Experiment 1a, the comparison is fair. Table 5 reports the average Hamming errors based on 100 independent repetitions. The results suggest that CASE outperforms the lasso and SCAD, and has a comparable performance to that of MC+.

Experiment 4b. We investigate the setting where "signal cancellation" is more severe than that in Experiment 4b. Toward this end, we use the same setting as in Experiment 4a, except for that β is generated in a way that signals appear in adjacent pairs with opposite signs, $(\beta_{2j-1}, \beta_{2j}) \overset{\text{i.i.d.}}{\sim} (1-\varepsilon_p)\nu_{(0,0)} + \varepsilon_p \nu_{(\tau_p, -\tau_p)}$, $1 \le j \le p/2$, where $\nu_{(a,b)}$ is the point mass at $(a, b) \in \mathbb{R}^2$. Hamming errors based on 100 repetitions are reported in Table 6, suggesting that CASE significantly outperforms all the other methods.

It is noteworthy that MC+ behaves much more unsatisfactorily here than in Experiment 4a, and the main reason is that MC+ does not adequately address "signal cancellation." In contrast, one of the major advantages of CASE is that it addresses adequately the "signal cancellation"; this is why it has satisfactory performance in both Experiments 4a and 4b.

Experiment 5. In some of the experiments above, we set the tuning parameters of CASE assuming (s_p, τ_p) as known. It is therefore interesting to investigate how the misspecification of (s_p, τ_p) affects the performance of CASE. Fix $p = 5000$ and $\phi = 0.35$. We consider two combinations of (ϑ, τ_p): $(\vartheta, \tau_p) = (0.35, 6), (0.55, 5)$.

TABLE 5
Comparison of Hamming errors (*Experiment* 4a). *The Gram matrix is the population covariance matrix of the* FARIMA$(0, \phi, 0)$ *process with* $\phi = 0.35$. $p = 5000$

			τ_p				
ϑ	s_p		**4**	**5**	**6**	**7**	**8**
0.35	253.7	CASE	118.0	60.7	26.3	9.5	4.3
		lasso	145.2	91.6	60.2	37.4	26.0
		SCAD	140.6	87.0	42.8	19.5	8.0
		MC+	108.6	50.2	20.4	7.4	2.6
0.45	108.3	CASE	60.3	27.7	11.8	4.0	1.9
		lasso	65.6	40.0	23.2	13.5	7.7
		SCAD	64.0	37.7	19.6	9.2	3.9
		MC+	52.0	23.6	8.6	3.0	1.0
0.55	46.2	CASE	27.9	13.4	4.3	1.4	0.5
		lasso	27.8	16.0	8.0	3.9	2.1
		SCAD	27.0	15.2	7.0	3.1	1.2
		MC+	23.4	10.6	3.1	0.7	0.2

The vector β is generated in the same way as in Experiment 4b, with the signals appearing in adjacent pairs. We fix one parameter of (s_p, τ_p) and mis-specify the other [since s_p is not on the same scale as τ_p, the results are reported based on the misspecification of (ϑ, τ_p), instead of (s_p, τ_p); recall here $s_p = p^{1-\vartheta}$]. We then apply CASE with tuning parameters set based on the misspecified values of

TABLE 6
Comparison of Hamming errors (*Experiment* 4b)

			τ_p				
ϑ	s_p		**4**	**5**	**6**	**7**	**8**
0.35	253.7	CASE	138.6	60.8	23.3	7.2	1.8
		lasso	223.0	158.9	97.9	54.8	27.1
		SCAD	257.5	156.8	95.1	52.1	25.1
		MC+	206.7	129.2	68.6	33.4	13.6
0.45	108.3	CASE	75.7	36.4	13.3	3.7	0.9
		lasso	100.0	84.7	58.4	32.2	15.9
		SCAD	99.2	83.2	56.6	30.6	14.9
		MC+	98.1	76.0	44.8	21.5	8.9
0.55	46.2	CASE	38.6	20.0	8.9	3.6	1.0
		lasso	45.4	40.1	31.0	20.6	10.9
		SCAD	45.0	39.4	30.1	19.6	9.9
		MC+	44.9	38.4	26.3	14.8	6.8

TABLE 7
Hamming errors for CASE, applied when (s_p, τ_p) *are misspecified as* $p^{1-\tilde{\vartheta}}$ *and* $\tilde{\tau}_p$, *respectively* (*Experiment* 5). *The Gram matrix is the population covariance matrix of the* FARIMA$(0, \phi, 0)$ *process with* $\phi = 0.35$. $p = 5000$

$\vartheta = 0.35, \tau_p = 6$	$\tilde{\vartheta}$	0.2	0.25	0.3	0.35	0.4	0.45	0.5
$s_p = 253.7$		27.8	24.8	23.2	23.2	24.5	26.3	48.9
	$\tilde{\tau}_p$	4	5	5.5	6	6.5	7	8
		47.3	30.2	25.3	23.2	23.9	26.9	42.7
$\vartheta = 0.55, \tau_p = 5$	$\tilde{\vartheta}$	0.4	0.45	0.5	0.55	0.6	0.65	0.7
$s_p = 46.2$		21.8	19.0	19.3	19.8	21.7	25.5	25.4
	$\tilde{\tau}_p$	3	4	4.5	5	5.5	6	7
		23.8	22.2	20.8	19.8	21.0	23.9	29.0

(s_p, τ_p). Table 7 reports the average Hamming errors of 50 independent repetitions, which is a rather flat function of misspecified values of ϑ (with τ_p fixed) or of misspecified values of τ_p (with ϑ fixed). In comparison, the Hamming errors of the lasso are 97.9 and 40.1 in the two settings, respectively, with the tuning parameter ideally set as in Experiment 1a. This suggests that CASE is relatively insensitive to the misspecification of (s_p, τ_p), and outperforms the lasso as long as the misspecification of (s_p, τ_p) is not severe.

Experiment 6. In this experiment, we continue to investigate the effect of "signal cancellation," where we compare CASE with the lasso using several choices of signal patterns (and so the levels of "signal cancellation" vary). Fix $p = 4998$, $\phi = 0.35$, $\vartheta = 0.75$, and let τ_p range in $\{5, \ldots, 10\}$. The experiment contains two parts. In the first part, we partition $\{1, 2, \ldots, p\}$ into $p/2$ blocks, where each block contains two adjacent indices. In $(1 - \varepsilon_p)$ fraction of the blocks, $\beta_j = 0$. In the ε_p fraction of the blocks, we have either that $\beta_j = \tau_p$ for both j in the block (we denote the sign pattern by "++"), or that $\beta_j = \tau_p$ if j is the first index in the block and $\beta_j = -\tau_p$ otherwise (sign pattern "+−"). In the second part, we partition $\{1, 2, \ldots, p\}$ into $p/3$ blocks (block size is 3), and generate β in a similar fashion, but with four different sign patterns in each block: "+ + +," "+ + −," "+ − +" and "+ − −." Hamming errors based on 50 independent repetitions are reported in Figure 5. The results suggest that (a) when the sign patterns are "++" or "+ + +," signal cancellation has negligible effects, and CASE is inferior to the lasso, (b) when the sign patterns are "+−," "+ + −," "+ − +" and "+ − −," the effect of signal cancellation kicks in, and CASE outperforms the lasso in most cases. This is consistent with our theoretic insight that CASE adequately addresses signal cancellation, but the lasso does not.

4. Discussion. Variable selection when the Gram matrix G is nonsparse is a challenging problem. We approach this problem by first sparsifying G with a finite

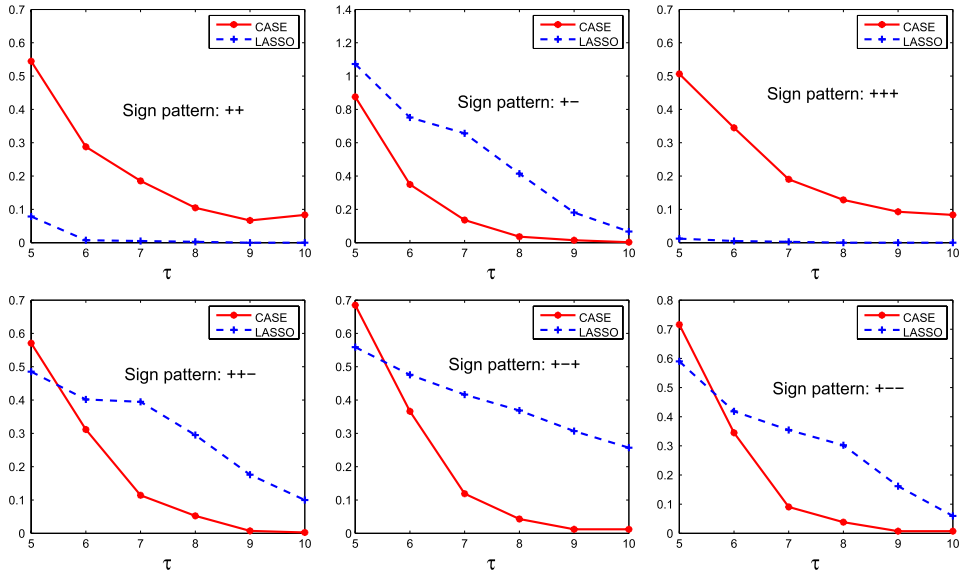

FIG. 5. *Comparison of Hamming errors* (*Experiment* 6). *The Gram matrix is the population covariance matrix of the* FARIMA$(0, \phi, 0)$ *process with* $\phi = 0.35$. $p = 5000$, $\vartheta = 0.75$ *and* $s_p = 32$. *Signals appear in adjacent pairs* (*Panels* 1–2) *or adjacent triplets* (*Panels* 3–6), *with different sign patterns specified in the panels*, *associated with different level of "signal cancellation." CASE outperforms lasso when signal cancellation is severe*.

order linear filter, and then constructing a sparse graph GOSD. The key insight is that, in the post-filtering data, the true signals live in many small-size components that are disconnected in GOSD, but we do not know where. We propose CASE as a new approach to variable selection. This is a two-stage screen and clean method, where we first use a covariance-assisted multivariate screening to identify candidates for such small-size components, and then re-examine each candidate with penalized least squares. In both stages, to overcome the problem of information leakage, we employ a delicate patching technique.

We develop an asymptotic framework focusing on the regime where the signals are rare and weak so that successful variable selection is challenging but is still possible. We show that CASE achieves the optimal rate of convergence in Hamming distance across a wide class of situations where G is nonsparse but sparsifiable. Such optimality cannot be achieved by many popular methods, including but not limited to the lasso, SCAD, MC+ and Dantzig selector. When G is nonsparse, these methods are not expected to behave well even when the signals are strong. We have successfully applied CASE to two different applications: the change-point problem and the long-memory times series.

Compared to the well-known method of marginal screening [Fan and Song (2010), Wasserman and Roeder (2009)], CASE employs a covariance-assisted multivariate screening procedure, so that it is theoretically more effective than

marginal screening, with only a moderate increase in the computational complexity. CASE is closely related to the graphical lasso [Friedman, Hastie and Tibshirani (2008), Meinshausen and Bühlmann (2006)], which also attempts to exploit the graph structure. However, the setting considered here is very different from that in Friedman, Hastie and Tibshirani (2008) and Meinshausen and Bühlmann (2006), and our emphasis on optimality is also very different.

The paper is closely related to the recent work Jin, Zhang and Zhang (2014) [see also Ji and Jin (2012)], but is different in important ways. The work in Jin, Zhang and Zhang (2014) is motivated by recent literature of compressive sensing and genetic regulatory network, and is largely focused on the case where the Gram matrix G is sparse in an unstructured fashion. The current work is motivated by the recent interest on DNA-copy number variation and long-memory time series, and is focused on the case where there are strong dependence between different design variables, so G is usually nonsparse and sometimes ill-posed. To deal with the strong dependence, we have to use a finite-order linear filter and delicate patching techniques. Additionally, the current paper also studies applications to the long-memory time series and the change-point problem which have not been considered in Jin, Zhang and Zhang (2014). Especially, the studies on the change-point problem encompasses very different and very delicate analysis on both the derivation of the lower bound and upper bound which we have not seen before in the literature. For these reasons, the two papers have very different scopes and techniques, and the results in one paper cannot be deduced from those in the other.

In this paper, we are primarily interested in the linear model, model (1.1), but CASE is applicable in much broader settings. For example, in model (1.1), we assume that the coordinates of z have the same variance σ^2, and σ is known (and so without loss of generality, we assume $\sigma = 1$). When σ is unknown, the main results in this paper continue to hold, provided that we can estimate σ consistently [say, except for a probability of $o(p^{-2})$, there is an estimate $\hat{\sigma}$ such that $|\hat{\sigma}/\sigma - 1| = o(1)$]. Such an estimator can be obtained by adapting the scaled-lasso approach by Sun and Zhang (2012) or the refitted cross validation by Fan, Guo and Hao (2012) to the post-filtering model (1.4). Correspondingly, we need to modify the tuning parameters of CASE slightly. For example, in the *PS*-step, $\mathcal{Q}$ is replaced by $\hat{\sigma}^2\mathcal{Q} \equiv \{\hat{\sigma}^2 t(F, N)\}$, and in the *PE*-step, u^{pe} and v^{pe} are replaced by $\hat{\sigma}u^{pe}$ and $\hat{\sigma}v^{pe}$, respectively.

Also, in model (1.1), we have assumed that the coordinates of z are Gaussian distributed. Such an assumption can also be relaxed. In fact, in the core of CASE is the analysis of low-dimensional sub-vectors of $\tilde{Y} = X'Y$, where we note that each coordinate of $\tilde{Y}$ has the form of $b_0 + a'z$ for some constant b_0 and $n \times 1$ nonstochastic vector a. Note that a only depends on the design matrix and the index of the coordinate of $\tilde{Y}$ (so there are p different vectors a at most). Essentially, the Gaussian assumption is only required for $a'z$ for all p different choices of a. Note that even when z is non-Gaussian, $a'z$ could be approximately Gaussian by central limit theorem; this holds, for example, for the long-memory time series

considered in the paper. As a result, the Gaussian assumption on z can be largely relaxed.

The main results in this paper can be extended in many other directions. For example, we have used a rare and weak signal model where the signals are randomly generated from a two-component mixture. The main results continue to hold if we choose to use a much more general model, as long as the signals live in small-size isolated "islands" in the post-filtering data, where each island is a connected subgraph of GOSD.

Also, we have been focused on the change-point model and the long-memory time series model, where the post-filtering matrices have polynomial off-diagonal decay and are sparse in a structured fashion. CASE can be extended to more general settings, where the sparsity of the post-filtering matrices are unstructured, provided that we modify the patching technique accordingly: the patching set can be constructed by including nodes which are connected to the original set through a short-length path in the GOSD.

Still another extension is that the Gram matrix can be sparsified by an operator D, but D is not necessary linear filtering. To apply CASE to this setting, we need to design specific patching technique. For example, when D^{-1} is sparse, for a given $\mathcal{I}$, we can construct $\mathcal{I}^{pe} = \{j : |D^{-1}(i, j)| > \delta_1, \text{for some } i \in \mathcal{I}\}$, where δ_1 is a chosen threshold.

The paper is closely related to recent literature on DNA copy number variation and financial data analysis, and it is of interest to further investigate such connections. To save space, we leave explorations along this line to the future.

A key component of our approach is the notion of "sparsifiability," meaning that we can make the Gram matrix G sparse by some simple operations. Usually, to find such a simple operation, we need a good understanding about the structure of G. In some applications, it is not hard to find such an operation:

- In compressive sensing or genome-wide association study (GWAS), where the rows of the design matrix X are i.i.d. samples from a p-dimensional distribution of zero means and a sparse covariance matrix Σ. In Compressive Sensing, Σ is usually proportional to the identity matrix, and in GWAS, Σ is a banded matrix. In such examples, G is already sparse, and so sparsifiable.
- The current paper is largely motivated by the change-point problem and long-memory time series models, where G can be sparsifiable by a liner filter D.
- Another example of sparsifiability is that G is the sum of a symmetric low-rank matrix L and a sparse matrix S; when spectral norm of S is much smaller than the smallest nonzero eigenvalue of L, G is sparsifiable by principle component analysis (PCA).

In more complicated case where we have little understanding about the structure G, how to sparsify it with a simple operation is a nontrivial problem, though we can always try linear filtering, PCA or both. We leave the study along this line for future.

The notion of "sparsifiability" may apply to both nonrandom design models and random design models. For a random-design model where rows of X are i.i.d. samples from a p-dimensional distribution with zero means and covariance matrix Σ, that G is sparsifiable usually means that Σ is sparsifiable. In more general case, how to sparsify the Gram matrix G is a nontrivial problem, and we leave such discussions to the future work.

Acknowledgments. The authors would like to thank Ning Hao, Philippe Rambour and David Siegmund for helpful pointers and comments.

SUPPLEMENTARY MATERIAL

Supplement to "Covariate assisted screening and estimation" (DOI: 10.1214/14-AOS1243SUPP; .pdf). Owing to space constraints, the technical proofs are relegated a supplementary document. It contains Sections A–C.

REFERENCES

ANDREOU, E. and GHYSELS, E. (2002). Detecting multiple breaks in financial market volatility dynamics. *J. Appl. Econometrics* **17** 579–600.

BHATTACHARYA, P. K. (1994). Some aspects of change-point analysis. In *Change-Point Problems* (*South Hadley, MA*, 1992). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **23** 28–56. IMS, Hayward, CA. MR1477912

CANDÈS, E. J. and PLAN, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* **37** 2145–2177. MR2543688

CHEN, W. W., HURVICH, C. M. and LU, Y. (2006). On the correlation matrix of the discrete Fourier transform and the fast solution of large Toeplitz systems for long-memory time series. *J. Amer. Statist. Assoc.* **101** 812–822. MR2281249

DONOHO, D. L. and HUO, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47** 2845–2862. MR1872845

DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **105** 14790–14795.

DONOHO, D. L. and STARK, P. B. (1989). Uncertainty principles and signal recovery. *SIAM J. Appl. Math.* **49** 906–931. MR0997928

FAN, J., GUO, S. and HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74** 37–65. MR2885839

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** 819–849. MR3210988

FAN, J. and YAO, Q. (2003). *Nonlinear Time Series*: *Nonparametric and Parametric Methods*. Springer, New York. MR1964455

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GENOVESE, C. R., JIN, J., WASSERMAN, L. and YAO, Z. (2012). A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.* **13** 2107–2143. MR2956354

HARCHAOUI, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493. MR2796565

IOANNIDIS, J. P. (2005). Why most published research findings are false. *PLoS Medicine* **2** e124.
ISING, E. (1925). A contribution to the theory of ferromagnetism. *Z. Phys* **31** 253–258.
JI, P. and JIN, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.* **40** 73–103. MR3013180
JIN, J., ZHANG, C.-H. and ZHANG, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *J. Mach. Learn. Res.* **15** 2723–2772.
KE, Z. T., JIN, J. and FAN, J. (2014). Supplement to "Covariate assisted screening and estimation." DOI:10.1214/14-AOS1243SUPP.
LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York. MR1639875
MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363
MOULINES, E. and SOULIER, P. (1999). Broadband log-periodogram regression of time series with long-range dependence. *Ann. Statist.* **27** 1415–1439. MR1740105
NIU, Y. S. and ZHANG, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat.* **6** 1306–1326. MR3012531
OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572.
RAY, B. K. and TSAY, R. S. (2000). Long-range dependence in daily stock volatilities. *J. Bus. Econom. Statist.* **18** 254–262.
SIEGMUND, D. O. (2011). Personal communication.
SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267–288. MR1379242
TIBSHIRANI, R. and WANG, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* **9** 18–29.
WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689
YAO, Y.-C. and AU, S. T. (1989). Least-squares estimation of a step function. *Sankhyā Ser. A* **51** 370–381. MR1175613
ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701
ZHANG, N. R., SIEGMUND, D. O., JI, H. and LI, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97** 631–645. MR2672488
ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449
ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

Z. T. KE
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: zke@galton.uchicago.edu

J. JIN
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: jiashun@stat.cmu.edu

J. FAN
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: jqfan@princeton.edu