Biometrika (2017), **104**, 4, *pp.* 785–800
Printed in Great Britain
doi: 10.1093/biomet/asx052
Advance Access publication 9 October 2017

Partition-based ultrahigh-dimensional variable screening

By JIAN KANG

Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109, U.S.A.
jiankang@umich.edu

HYOKYOUNG G. HONG

Department of Statistics and Probability, Michigan State University, 619 Red Cedar Rd, East Lansing, Michigan 48823, U.S.A.
hhong@msu.edu

and YI LI

Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, Michigan 48109, U.S.A.
yili@umich.edu

Summary

Traditional variable selection methods are compromised by overlooking useful information on covariates with similar functionality or spatial proximity, and by treating each covariate independently. Leveraging prior grouping information on covariates, we propose partition-based screening methods for ultrahigh-dimensional variables in the framework of generalized linear models. We show that partition-based screening exhibits the sure screening property with a vanishing false selection rate, and we propose a data-driven partition screening framework with unavailable or unreliable prior knowledge on covariate grouping and investigate its theoretical properties. We consider two special cases: correlation-guided partitioning and spatial location-guided partitioning. In the absence of a single partition, we propose a theoretically justified strategy for combining statistics from various partitioning methods. The utility of the proposed methods is demonstrated via simulation and analysis of functional neuroimaging data.

Some key words: Correlation-based variable screening; Partition; Spatial variable screening; Ultrahigh-dimensional variable screening.

1. Introduction

Biotechnological advances have resulted in an explosion of ultrahigh-dimensional data, where the dimension of the data can be of exponential order in the sample size. Because of high computational cost and poor numerical stability, ultrahigh-dimensional data have long defied existing regularization approaches designed for high-dimensional data analysis (Tibshirani, 1996; Fan & Li, 2001; Zou & Hastie, 2005; Meinshausen & Bühlmann, 2006; Yuan & Lin, 2006; Zhao & Yu, 2006; Zou, 2006; Candès & Tao, 2007; Zhang & Lu, 2007; Huang et al., 2008; Zou & Zhang, 2009; Meinshausen & Bühlmann, 2010; Wang & Leng, 2012). An overarching goal of ultrahigh-dimensional data analytics is to effectively reduce the dimension of covariates.

Sure independence screening (Fan & Lv, 2008) has been extended to generalized linear models (Fan & Fan, 2008; Fan et al., 2009; Fan & Song, 2010), generalized additive models (Fan et al., 2012) and proportional hazards models (Zhao & Li, 2012; Gorst-Rasmussen & Scheike, 2013; Hong et al., 2016a; Li et al., 2016). By extending screening criteria that are solely based on marginal correlations between the outcome and predictors, a variety of statistics that account for dependence between predictors have been proposed to improve screening accuracy and robustness (Hall & Miller, 2009; Zhu et al., 2011; Cho & Fryzlewicz, 2012; Li et al., 2012; Cui et al., 2015). In particular, high-dimensional ordinary least squares projection (Wang & Leng, 2016), which uses the generalized inverse of the design matrix in lieu of marginal correlations, has good theoretical properties and high computational efficiency.

In many cases, scientists have knowledge about important predictors from previous research. For example, neuroimaging studies have identified voxel-level imaging predictors clustered in certain brain regions that are linked to brain functions or diseases. Genome-wide association studies have detected single nucleotide polymorphisms that are strongly associated with clinical outcomes. However, most variable screening approaches are not designed to make use of such information.

As an alternative to marginal screening approaches, conditional sure independence screening methods have been developed for generalized linear models (Barut et al., 2016) and proportional hazards models (Hong et al., 2016a). By including important predictors, conditional screening ranks the marginal utility of each variable after adjusting for variables in the conditioning set.

Partitioning biomarkers into smaller groups according to biological knowledge or other useful information may facilitate variable selection. In biological studies, leveraging information about groups of weak predictors is often useful because such predictors may have a non-trivial impact on outcomes as a group, and without considering the group structure these features might be missed. We exemplify the merit of using the grouping structure with a simple example.

Suppose that we want to identify the important associations between the outcome Y and $X_1, \ldots, X_{1000}$, where $Y = 0{\cdot}5X_1 - X_2 + \epsilon$ with $\epsilon \sim N(0, 1{\cdot}6)$ and $(X_1, \ldots, X_{1000})$ follows a multivariate normal distribution with mean zero, unit marginal variance and correlation $\mathrm{corr}(X_j, X_k) = 0{\cdot}5$ for any $j \neq k \in \{1, \ldots, 1000\}$. To screen for important variables, marginal screening would fit 1000 variate regression models, $Y = X_j\beta_j + \tilde{\epsilon}$ for $j = 1, \ldots, 1000$, and use $\hat{\beta}_j$, the estimate of β_j, as the screening statistic. Suppose that we partition these 1000 predictors into 200 groups such that the group membership index sets are $\mathcal{S}_1 = (1, \ldots, 5), \ldots, \mathcal{S}_{200} = (996, \ldots, 1000)$. An alternative screening approach would fit 200 multivariate regression models along the group partition, $Y = \sum_{j \in \mathcal{S}_g} X_j\beta_j + \tilde{\epsilon}$ for $g = 1, \ldots, 200$, and use the corresponding $\hat{\beta}_j$ as the screening statistic. We examine the performances of the two approaches based on 300 samples and 400 replicates. Figure 1 shows plots of the densities of $\hat{\beta}_1$ and $\hat{\beta}_2$ and the mixture density of $\hat{\beta}_3, \ldots, \hat{\beta}_{1000}$ for both approaches. Due to signal cancellation, the univariate regression introduces large biases in estimating β_1 and β_2, causing considerable overlap between the distribution of $\hat{\beta}_1$ or $\hat{\beta}_2$ and those of the estimates for the noise variables, whereas groupwise multivariate regression separates the distribution of $\hat{\beta}_1$ or $\hat{\beta}_2$ from the distributions of the others.

This example motivates partition-based screening, which is based on a partition of covariates using prior knowledge. Our work generalizes the univariate framework of sure independence screening (Fan & Lv, 2008) and its group version (Niu et al., 2011). Under mild conditions, partition-based screening exhibits good theoretical properties. A new functional operator, generalized linear conditional expectation, is introduced to help establish sure screening properties. When prior grouping information is available, we show that the screening accuracy of partition-based screening is superior to that of competing methods. In the absence of prior grouping information,

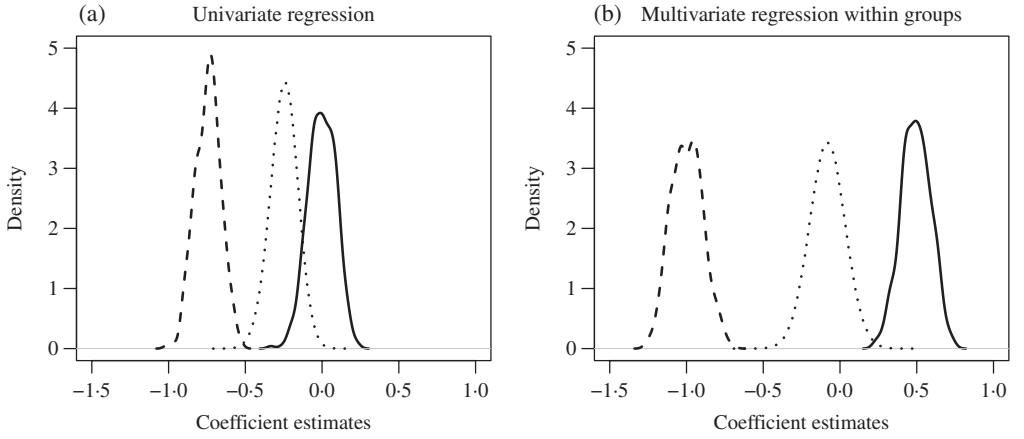

Fig. 1. Simple example simulations: (a) summary distributions of 400 replicates of coefficient estimates in 1000 univariate regression model fits; (b) summary of 200 group-specific multivariate regression model fits. Plotted are the estimated densities of $\hat{\beta}_1$ (solid) and $\hat{\beta}_2$ (dashed) and the estimated mixture densities of $\hat{\beta}_3, \ldots, \hat{\beta}_{1000}$ (dotted); true coefficients are $\beta_1 = 0{\cdot}5$, $\beta_2 = -1$ and $\beta_3 = \cdots = \beta_{1000} = 0$.

we propose correlation-based screening and spatial partition-based screening, which make the proposed methods applicable to a wide range of problems.

2. Partition-based variable screening

Suppose that we have n independent samples $D = \{(X_i, Y_i),\ i = 1, \ldots, n\}$, where Y_i is an outcome and $X_i = (X_{i,1}, \ldots, X_{i,p})^{\mathrm{T}}$ is a collection of p predictors for the ith sample. Assume without loss of generality that all the covariates have been standardized so that $E(X_{i,j}) = 0$ and $E(X_{i,j}^2) = 1$. We consider a class of generalized linear models by assuming that the conditional density of Y_i given X_i belongs to a linear exponential family,

$$\pi(Y_i \mid X_i) = \exp\{Y_i(\beta_0 + X_i^{\mathrm{T}}\beta) - b(\beta_0 + X_i^{\mathrm{T}}\beta) + A(Y_i, X_i)\}, \tag{1}$$

where $A(\cdot\,,\cdot)$ and $b(\cdot)$ are known functions, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ represent the coefficients of the predictors, and β_0 is an intercept, regarded as a nuisance parameter. Let $\mathcal{M} = \{j : \beta_j \neq 0\}$. We assume that $b(\cdot)$ is twice continuously differentiable, with a nonnegative second derivative $b''(\cdot)$. For a nonrandom function $f(\cdot)$ and a sequence of independent random variables ξ_i $(i = 1, \ldots, n)$, let $E_n\{f(\xi)\} = n^{-1}\sum_{i=1}^n f(\xi_i)$ be the empirical mean of $\{f(\xi_i)\}_{i=1}^n$, which are independent replicates of $f(\xi)$. The loglikelihood function is

$$\ell(\beta_0, \beta; D) = \frac{1}{n}\sum_{i=1}^{n} l(\beta_0 + X_i^{\mathrm{T}}\beta, Y_i) = E_n\{l(\beta_0 + X^{\mathrm{T}}\beta, Y)\}, \tag{2}$$

where $l(\theta, y) = y\theta - b(\theta)$. We assume that $\{X_{ij}, X_i, Y_i\}$ are independently and identically distributed copies of $\{X_j, X, Y\}$. When $p < n$, the maximum likelihood estimator of β, denoted by $\hat{\beta}_{\mathrm{MLE}}$, can be obtained by maximizing $\ell(\beta_0, \beta; D)$. When $p \geqslant n$, regularization estimation is often performed under an assumption of sparsity among predictors. When p is of exponential order in n, a popular approach for reducing the dimensionality is screening.

First, we consider a simple case where the covariates can be partitioned into G disjoint groups in accordance with known information. Denote by g_j the group membership of variable X_j. Let

$X_g^* = \{X_j : g_j = g\}$ be the collection of predictors in group g, where $g \in \{1, \ldots, G\}$. Additionally, let $\beta_g^* = \{\beta_j : g_j = g\}$ represent the corresponding coefficients and let $\beta_{g,0}$ be the group-specific intercept in the model. Denote their estimates by $\hat{\beta}_g^* = \{\hat{\beta}_j : g_j = g\}$ and $\hat{\beta}_{g,0}^*$, respectively. For predictor j with $g_j = g$, the partition-based screening statistic is defined as

$$(\hat{\beta}_{g,0}^*, \hat{\beta}_g^*) = \underset{(\beta_{g,0}, \beta_g^*)}{\arg\max}\ E_n\{l(\beta_{g,0} + X_g^{*\mathrm{T}}\beta_g^*, Y)\}.$$

We call $\hat{\beta}_g^*$ the partition-based screening statistic. Then, for a chosen thresholding parameter γ, the set of indices selected by our proposed partition-based screening is $\hat{\mathcal{M}}_\gamma = \{j : |\hat{\beta}_j| \geqslant \gamma\}$.

When $g_j = j$ $(j = 1, \ldots, p)$, partition-based screening encompasses sure independence screening as a special case.

3. SURE SCREENING PROPERTIES

Let $(\Omega, \mathcal{F}, \mathrm{pr})$ be the probability space for all random variables considered in this paper. Let $\mathbb{R}^d$ be a d-dimensional Euclidean vector space for some positive integer d. Denote by $E(\cdot)$, $\mathrm{var}(\cdot)$ and $\mathrm{cov}(\cdot, \cdot)$ the expectation, variance and covariance operators associated with $(\Omega, \mathcal{F}, \mathrm{pr})$. For any vector $a = (a_1, \ldots, a_p) \in \mathbb{R}^p$, let $a_{\mathcal{C}} = (a_j, j \in \mathcal{C})^{\mathrm{T}}$ be the subvector with elements indexed by $\mathcal{C}$. Let $\|a\|_d = (\sum_{j=1}^p |a|_j^d)^{1/d}$ be the L_d-norm for any vector $a \in \mathbb{R}^p$, and denote the Euclidean norm by $\|a\|$ when no confusion is likely to arise. Let $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ be the smallest and largest eigenvalues of the matrix M, respectively.

We start with population-level parameters for the discussion of sure screening properties. Let

$$(\bar{\beta}_{g,0}^*, \bar{\beta}_g^*) = \underset{(\beta_{g,0}, \beta_g^*)}{\arg\max}\ E\{l(\beta_{g,0} + X_g^{*\mathrm{T}}\beta_g^*, Y)\}, \tag{3}$$

where $\bar{\beta}_g^* = \{\bar{\beta}_j : g_j = g\}$ is the population version of $\hat{\beta}_g^*$. We first establish conditions to ensure that if $|\beta_j|$ exceeds a threshold, then $|\bar{\beta}_j|$ will exceed a certain constant. Write $\beta_{-j}^* = \{\beta_l : g_l = g_j,\ l \neq j\}^{\mathrm{T}}$ and $X_{-j}^* = \{X_l : g_l = g_j,\ l \neq j\}^{\mathrm{T}}$. With (2), $\bar{\beta}_j$ satisfies the score equations

$$E\{b'(\bar{\beta}_{g,0}^* + X_{-j}^{*\mathrm{T}}\bar{\beta}_{-j}^* + X_j\bar{\beta}_j)(1, X_g^{*\mathrm{T}})^{\mathrm{T}}\} = E\{Y(1, X_g^{*\mathrm{T}})^{\mathrm{T}}\}. \tag{4}$$

To derive the theoretical properties of the proposed methods, we introduce a functional operator on random variables.

DEFINITION 1. *For two random variables $\zeta : \Omega \to \mathbb{R}$ and $\xi : \Omega \to \mathbb{R}^p$, let $h : \mathbb{R} \to \mathbb{R}$ be a continuous link function. The generalized linear conditional expectation of ζ given ξ is*

$$E_h(\zeta \mid \xi) = h(\alpha_0 + \alpha^{\mathrm{T}}\xi), \tag{5}$$

where $(\alpha_0, \alpha^{\mathrm{T}})^{\mathrm{T}}$ is the solution to the equation $E[\{\zeta - h(\alpha_0 + \alpha^{\mathrm{T}}\xi)\}(1, \xi^{\mathrm{T}})^{\mathrm{T}}] = 0$.

The generalized linear conditional expectation measures how ξ can explain ζ through a generalized linear model, where ζ is regarded as the outcome variable and ξ as the predictors. It can also be interpreted as the best prediction of ζ using ξ based on a generalized linear model, leading to an alternative measure of the dependence between ζ and ξ. The generalized linear conditional

expectation may depend on the choice of link functions, and it is equivalent to the conditional expectation if the true conditional distribution of ζ given ξ is specified by the corresponding generalized linear model. The introduction of (5) facilitates the development of partition-based screening and its theoretical properties, and extends the linear conditional expectation proposed by Barut et al. (2016) and Hong et al. (2016a). Some basic properties are summarized below.

LEMMA 1. *Let ζ and ξ be random variables in $(\Omega, \mathcal{F}, \mathrm{pr})$.*

(i) *When $h(x) = 1(x) = x$, $E_h(\zeta \mid \xi)$ is unique and has a closed-form expression. Moreover, $E_1(\zeta \mid \xi) = E(\zeta) + \mathrm{cov}(\zeta, \xi)\mathrm{var}(\xi)^{-1}\{\xi - E(\xi)\}$ and $E\{E_1(\zeta \mid \xi)\xi\} = E(\zeta\xi)$.*
(ii) *When the conditional distribution of ζ given ξ belongs to a linear exponential family, i.e., $f(\zeta \mid \xi) = \exp\{\zeta(\gamma_0 + \gamma^{\mathrm{T}}\xi) - b(\gamma_0 + \gamma^{\mathrm{T}}\xi) + A(\zeta, \xi)\}$, then $h(x) = b'(x)$ and $E_{b'}(\zeta \mid \xi) = E(\zeta \mid \xi) = b'(\gamma_0 + \gamma^{\mathrm{T}}\xi)$.*
(iii) *For any h, we have $E\{E_h(\zeta \mid \xi)\} = E(\zeta)$.*

These properties immediately imply the following result.

THEOREM 1. *Suppose that the solution to (4) is unique. For $j = 1, \ldots, p$, the partition-based regression parameter $\bar{\beta}_j$ equals 0 if and only if $E_{b'}(Y \mid X^*_{-j}) = E_{b'}(Y \mid X^*_{-j}, X_j)$.*

The sufficient part of Theorem 1 implies that if the generalized linear conditional expectation of the response given all the predictors within group g_j does not involve X_j, then the regression coefficient β_j will be vanishing, implying that unimportant variables would have smaller fitted coefficients.

To ensure the sure screening property at the population level, the important variables $\{X_j, j \in \mathcal{M}\}$ should be conditionally associated with Y given other variables within the same group $\{X^*_{-j}, j \in \mathcal{M}\}$. The following conditions are required.

Condition 1. For $j \in \mathcal{M}$, there exist $c_0 > 0$ and $\kappa < 1/2$ such that

$$\left|E\left[X_j\{E_{b'}(Y \mid X^*_{-j}, X_j) - E_{b'}(Y \mid X^*_{-j})\}\right]\right| > c_0 n^{-\kappa}.$$

Condition 2. The derivative $b'(\theta)$ satisfies a Lipschitz condition, i.e., there exists an $L > 0$ such that $|b'(\theta_1) - b'(\theta_2)| < L|\theta_1 - \theta_2|$ for all $\theta_1, \theta_2 \in \mathbb{R}$.

Condition 3. There exists a constant $M > 0$ such that $E(X_j^2) \leqslant M$ for all j.

Condition 1 provides a lower bound on the generalized linear dependence between each active covariate X_j and Y conditional on other covariates within the same group, justifying the use of group partitions to retain true signals. Linear regression, logistic regression and probit regression all satisfy Condition 2.

THEOREM 2. *If Conditions 1–3 hold, then there exists $c_2 > 0$ such that $\min_{j \in \mathcal{M}} |\bar{\beta}_j| > c_2 n^{-\kappa}$.*

To establish sure screening properties, we need regularity conditions (Fan & Song, 2010; Barut et al., 2016); see Conditions A·1–A·6 in the Supplementary Material.

THEOREM 3. *Let $S_g = \sum_{j=1}^{p} I(g_j = g)$ be the size of group g. Assume that Conditions* A·1–A·6 *in the Supplementary Material hold and that $Q_{g,n} = n^{1-2\kappa}(R_n r_{g,n})^{-2} \to \infty$ as $n \to \infty$ for all $g = 1, \ldots, G$.*

(i) *With c_2 as in Theorem 2, there exists a positive constant c_3 such that*

$$\mathrm{pr}\Big(\max_{1\leqslant j\leqslant p}|\hat{\beta}_j-\bar{\beta}_j|\geqslant c_2n^{-\kappa}/2\Big)\leqslant\sum_{g=1}^{G}S_g\exp\big(-c_3Q_{g,n}\big)+nr_2\exp(-r_0R_n^{\alpha}),$$

where $r_2=\sum_{g=1}^{G}S_g(S_gr_1+s_1)$.

(ii) *If Conditions 1–3 hold, then with $\gamma=c_4n^{-\kappa}$ and $c_4\leqslant c_2/2$, we have*

$$\mathrm{pr}\Big(\mathcal{M}\subset\hat{\mathcal{M}}_{\gamma}\Big)\geqslant 1-\sum_{j\in\mathcal{M}}\exp(-c_3Q_{g_j,n})-nr_3\exp(-r_0R_n^{\alpha}),$$

where $r_3=\sum_{j\in\mathcal{M}}(S_{g_j}r_1+s_1)$.

For logistic regression, the Lipschitz constant $r_{g,n}$ is bounded. Therefore, the optimal rate for R_n is of order $n^{(1-2\kappa)/(2+\alpha)}$, ensuring that $Q_{g,n}$ is of the same order as R_n^{α}. This also implies that the partition-based screening method can handle group sizes of order $\log S_g=o(n^{(1-2\kappa)\alpha/(\alpha+2)})$ $(g=1,\ldots,G)$. The same optimal rate and a similar order of dimensionality can be achieved for logistic regression by sure independence screening (Fan & Song, 2010) and conditional sure independence screening (Barut et al., 2016).

To provide an upper bound on the number of selected variables, we need the following additional conditions:

Condition 4. $\sum_{j=1}^{p}\mathrm{var}(X_j\beta_j)$ and $b''(\theta)$ are both bounded for all θ and β.

Condition 5. Let $g=g_j$ and $\Omega_j=E[\delta_j(1,X_g^{*\mathrm{T}})^{\mathrm{T}}(1,X_g^{*\mathrm{T}})]$ with

$$\delta_j=\frac{b'(\bar{\beta}_{g,0}^{*}+X_{-j}^{*\mathrm{T}}\bar{\beta}_{-j}^{*}+X_j\bar{\beta}_j^{*})-b'(\tilde{\beta}_{g,0}^{*}+X_{-j}^{*\mathrm{T}}\tilde{\beta}_{-j}^{*})}{\bar{\beta}_{g,0}^{*}-\tilde{\beta}_{g,0}^{*}+X_{-j}^{*\mathrm{T}}(\bar{\beta}_{-j}^{*}-\tilde{\beta}_{-j}^{*})+X_j\bar{\beta}_j^{*}},$$

where $(\tilde{\beta}_{g,0}^{*},\tilde{\beta}_{-j}^{*})=\arg\max_{(\beta_0,\beta_{-j}^{*})}E\{l(\beta_0+X_{-j}^{*\mathrm{T}}\beta_{-j}^{*},Y)\}$. Then there exists a $K_1>0$ such that $\lambda_{\min}(\Omega_j)>K_1$ for all $j=1,\ldots,p$.

Condition 6. Assume that $\|U\|^2=o(V)$ where $U=(U_1,\ldots,U_p)^{\mathrm{T}}$ with

$$U_j=E\big\{E_1(X_j\mid X_{-j}^{*},X_g^{*},g\neq g_j)(\beta_0-\tilde{\beta}_{g_j,0}^{*}+X^{\mathrm{T}}\beta-X_{-j}^{*\mathrm{T}}\tilde{\beta}_{-j}^{*})\big\}$$

and $V=\sup_{1\leqslant j\leqslant p}E[\{X_j-E_1(X_j\mid X_{-j}^{*},X_g^{*},g\neq g_j)\}^2]$. Here $(\beta_0,\beta^{\mathrm{T}})^{\mathrm{T}}$ are the parameters that generate the data.

For a linear model, Condition 5 becomes $\lambda_{\min}[E\{(1,X_g^{*\mathrm{T}})^{\mathrm{T}}(1,X_g^{*\mathrm{T}})\}]>K_1$ for all g, which is a mild condition. In Condition 6, $U_j=0$ for the linear model when $\mathrm{cov}(X)$ has a block-diagonal structure over a group partition, i.e., $\mathrm{cov}(X_g^{*},X_{g'}^{*})=0$ for $g\neq g'$, because $E_1(X_j\mid X_{-j}^{*},X_g^{*},g\neq g_j)=E_1(X_j\mid X_{-j}^{*})$ and $E\{X_{-j}^{*}(\beta_0-\tilde{\beta}_{g_j,0}^{*}+X^{\mathrm{T}}\beta-X_{-j}^{*\mathrm{T}}\tilde{\beta}_{-j}^{*})\}=E[X_{-j}^{*}\{b'(\beta_0+X^{\mathrm{T}}\beta)-b'(\tilde{\beta}_{g_j,0}^{*}+X_{-j}^{*\mathrm{T}}\tilde{\beta}_{-j}^{*})\}]=0$.

Condition 6, which requires $\|U\|$ to be bounded, may be restrictive. This condition holds if the number of nonzero coefficients is finite or if the correlations among different partitions shrink

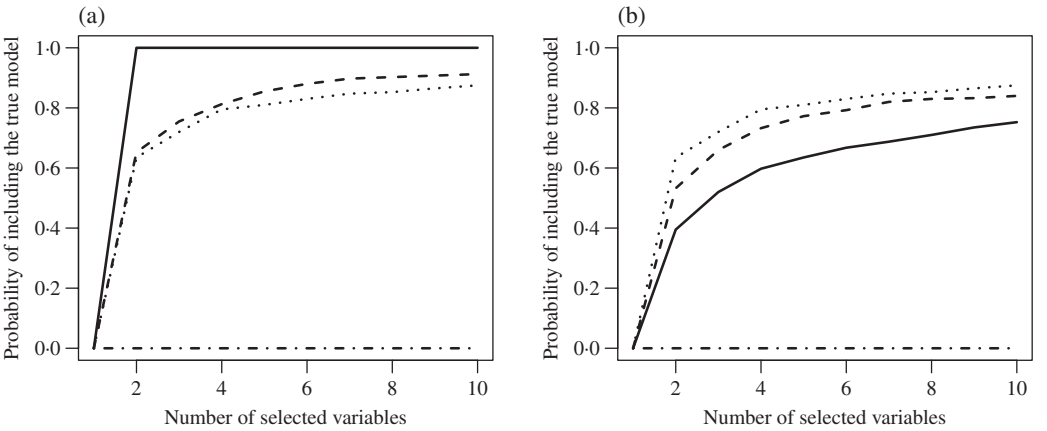

Fig. 2. Variable screening accuracy in the example in § 1 by sure independence screening (dot-dash), high-dimensional ordinary least squares projection (dotted), partition-based screening without goodness-of-fit adjustment (dashed) and partition-based screening with goodness-of-fit adjustment (solid). Panel (a) shows results with the group partitions defined in § 1 where group 1 includes the two true predictors; panel (b) displays results with a random group partition where the two true predictors are not in the same group.

as $p \to \infty$. Condition 6 can be viewed as rigid, even though the group structure is natural in many biomedical applications. To overcome this difficulty, one could first perform a principal component analysis on X, and then apply the proposed procedures to the residuals of X after projecting them to a set of variables with the largest loadings on the leading eigenvectors (Hong et al., 2016a). See the Supplementary Material for more details.

THEOREM 4. *With γ, c_3 and r_2 as in Theorem 3, if Conditions 4–6 and* A·1–A·6 *hold, then as $n \to \infty$,*

$$\mathrm{pr}\{|\hat{\mathcal{M}}_\gamma| \leqslant O(n^{2\kappa} V)\} \geqslant 1 - \sum_{g=1}^{G} S_g \exp\left(-c_3 Q_{g,n}\right) - nr_2 \exp(-r_0 R_n^{\alpha}).$$

4. EXTENSIONS OF PARTITION-BASED SCREENING

4·1. *Goodness-of-fit adjustment*

One difficulty in the proposed partition-based screening is that the coefficient estimates from group-specific models may not be comparable because different models may have various degrees of goodness-of-fit. Under the generalized linear model framework, we propose to adjust for the goodness-of-fit by weighting the screening statistics using the deviance ratio $\Psi_g \in (0, 1)$ (Friedman et al., 2010), which is the fraction of null deviance explained by the covariates in group g and is equal to R^2 in the linear model. In other words, we weight the partition-based screening statistic as $\hat{\beta}_j^a = \Psi_{g_j} \hat{\beta}_j$ for predictor j and redefine the selected index set as $\hat{\mathcal{M}}_\gamma^a = \{j : |\hat{\beta}_j^a| > \gamma\}$. However, the performance of such a procedure may be sensitive to grouping. For instance, in the example in § 1, goodness-of-fit adjustment can improve the model selection accuracy when all the true predictors are in the same group, as in Fig. 2(a). In contrast, when the true predictors are in separate groups, there is no improvement; see Fig. 2(b) and the Supplementary Material.

4·2. *Data-driven partition*

When prior partitioning information is unavailable, we may use information from the data, including correlations between predictors and the spatial locations. Let us denote the data-driven partition by $\hat{\mathcal{G}}$. We derive two procedures for determining $\hat{\mathcal{G}}$ based on such information.

The data-driven partition can be determined by the covariance or correlation structure of the design matrix. We propose a simple correlation-guided partition procedure. We use the correlation between covariates to define a $p \times p$ distance matrix, denoted by $\Delta = (d_{j,k})$, where $d_{j,k} = 1 - |\mathrm{corr}(X_k, X_j)|$ $(1 \leqslant k, j \leqslant p)$. We apply the nearest-neighbour chain algorithm (Murtagh, 1983), a standard hierarchical clustering algorithm, to Δ to obtain an estimate of $\hat{\mathcal{G}}$, where the number of groups, $\hat{G}$, can be determined by controlling the corresponding maximum group size $\max_g \hat{S}_g$ for the generalized linear model fitting. Based on our experience, choosing $\max_g \hat{S}_g \propto n^{1/2}$ can lead to good performance. When the correlation matrix has a block-diagonal structure and the correlations within each block are high, this procedure can correctly identify the block structure. Covariate-assisted variable screening (Ke et al., 2014) and graphlet screening (Jin et al., 2014) are also based on the covariance structure of covariates, but these procedures focus on linear regression.

Spatial regression models have often been used in environmental health and neuroimaging studies, where a spatial location is attached to each covariate. For example, in the scalar-on-image regression problem for brain imaging, where the spatial location of each voxel is in a standard three-dimensional brain template, the imaging intensities at different voxels are usually considered as potential predictors for clinical outcomes. It is generally believed that spatially close predictors tend to have stronger correlations and may have more similar effects on the outcome (Wang et al., 2017). Therefore spatial location information can be useful in determining the partition for variable screening. Specifically, model-based clustering (Fraley & Raftery, 2002) and k-means clustering (Jain, 2010) can be used to assign each predictor to a fixed number of clusters or spatial locations, typically determined by controlling the corresponding maximum group size in a similar fashion to correlation-guided partitioning.

Using a partition $\hat{\mathcal{G}}$ determined by data, we can also establish the following theoretical results for data-driven partition-based screening, which can be proved by conditioning on the event $\{\hat{\mathcal{G}} = \mathcal{G}\}$ and using Theorems 1–4.

THEOREM 5. *Suppose that $\hat{\mathcal{G}}$ is a consistent estimator of $\mathcal{G}$ satisfying Conditions 1–6 and* A·1–A·6*; that is,* $\lim_{n\to\infty} \mathrm{pr}(\hat{\mathcal{G}} = \mathcal{G}) = 1$. *With the same γ, c_3, r_2 and r_3 as in Theorem 3, as $n \to \infty$ we have*

$$\mathrm{pr}\Big(\mathcal{M} \subset \hat{\mathcal{M}}_\gamma\Big) \geqslant \mathrm{pr}(\hat{\mathcal{G}} = \mathcal{G}) - \sum_{j\in\mathcal{M}} \exp(-c_3 Q_{g_j,n}) - nr_3(-r_0 R_n^\alpha),$$

$$\mathrm{pr}\Big\{|\hat{\mathcal{M}}_\gamma| \leqslant O(n^{2\kappa} V)\Big\} \geqslant \mathrm{pr}(\hat{\mathcal{G}} = \mathcal{G}) - \sum_{g=1}^{G} S_g \exp(-c_3 Q_{g,n}) - nr_2(-r_0 R_n^\alpha).$$

4·3. *Combined partition-based screening*

Several partitioning rules for covariates may exist, but none is clearly superior. For example, in neuroimaging, brain atlases, such as Talairach–Tournoux, Harvard–Oxford, Eickoff–Zilles and automatic anatomical labelling, have variable partitioning of brain regions. In genome-wide association studies, different sources of information, including the locations of genes that harbour

single nucleotide polymorphisms and linkage disequilibrium between single nucleotide polymorphisms, can be integrated to determine the groups. Using multiple sources of partitioning, we propose a strategy for combining screening statistics from different partitions and establish its theoretical properties.

DEFINITION 2. *Suppose that we have $K < \infty$ partitions and that partition k has $G^{(k)}$ groups and group indices $\mathcal{G}^{(k)} = (g_1^{(k)}, \ldots, g_p^{(k)})^{\mathrm{T}}$. Let $S_g^{(k)} = \sum_{j=1}^{p} I(g_j^{(k)} = g)$ and let $\hat{\beta}^{(k)} = (\hat{\beta}_1^{(k)}, \ldots, \hat{\beta}_p^{(k)})^{\mathrm{T}}$ be the screening statistics for partition k. The combined partition-based screening statistic is $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^{\mathrm{T}}$ with $\tilde{\beta}_j = \max_{1 \leqslant k \leqslant K} |\hat{\beta}_j^{(k)}|$. Given a thresholding parameter γ, the selected index set is $\tilde{\mathcal{M}}_\gamma = \{j : \tilde{\beta}_j \geqslant \gamma\}$, which is referred to as combined partition-based screening selection.*

THEOREM 6. *Suppose that Conditions 4–6 and A·1–A·6 hold for all partitions k ($k = 1, \ldots, K$), and take $\gamma = c_5 n^{-\kappa}$ for some constant c_5.*

(i) *If there exists an $l \in \{1, \ldots, K\}$ such that $\mathcal{G}^{(l)}$ satisfies Conditions 1–3, then $\lim_{n\to\infty} \mathrm{pr}(\mathcal{M} \subset \tilde{\mathcal{M}}_\gamma) = 1$.*

(ii) *Let $V^{(k)}$ be the V term in Condition 6 for $\mathcal{G}^{(k)}$. Then combined partition-based screening controls the false positive rates; that is,*

$$\lim_{n\to\infty} \mathrm{pr}\left\{ |\tilde{\mathcal{M}}_\gamma| \leqslant O\left(n^{2\kappa} \sum_{k=1}^{K} V^{(k)} \right) \right\} = 1.$$

Theorem 6 suggests that combined partition-based screening has sure screening properties, even when some partition-based screening procedures do not satisfy Conditions 1–3, which are in general difficult to verify. Moreover, Conditions 4–6 and A·1–A·6 are true for many generalized linear models. Thus, combined partition-based screening can extract useful prior knowledge about partitions and maintain good theoretical properties.

4·4. *Choice of thresholding parameters*

The thresholding parameter γ is critical to the performance of the variable screening procedure. Overestimating γ will inflate false positive rates and underestimating γ will hinder sure screening. We define the expected false positive rate $\mathrm{EFPR}_\gamma = E(|\hat{\mathcal{M}}_\gamma \cap \mathcal{M}^{\mathrm{c}}|/|\mathcal{M}^{\mathrm{c}}|)$, where $\mathcal{M}^{\mathrm{c}} = \{j : j \notin \mathcal{M}\}$. To control EFPR_γ, we resort to higher-criticism t statistics (Zhao & Li, 2012; Barut et al., 2016; Hong et al., 2016a). We introduce $\hat{\mathcal{M}}_\tau^{\#} = \{j : I_j(\hat{\beta}_j)^{1/2} |\hat{\beta}_j| \geqslant \tau\}$, where $I_j(\hat{\beta}_j)$ is the element that corresponds to β_j in the information matrix $I_{g_j}(\beta_{g_j}^*)$. The key idea is to select γ such that $\hat{\mathcal{M}}_\gamma$ is of the same size as $\hat{\mathcal{M}}_\tau^{\#}$, where τ is chosen to control the expected false positive rate. Under Conditions 4 and 5 and Conditions A·1–A·6 and B in the Supplementary Material, we have the following theorem.

THEOREM 7. *For a given false positive number q, take $\tau = \Phi^{-1}\{1 - q/(2p)\}$, where Φ denotes the standard normal distribution function. Set $\gamma = |\hat{\beta}_{(s)}|$, where $s = |\mathcal{M}_\tau^{\#}|$ and $\{(1), \ldots, (p)\}$ is a permutation of $\{1, \ldots, p\}$ such that $|\hat{\beta}_{(1)}| > \cdots > |\hat{\beta}_{(p)}|$. Then there exist $N_7 > 0$ and $c_7 > 0$ such that for any $n > N_7$, $\mathrm{EFPR}_\gamma \leqslant q/p + c_7 n^{-1/2}$.*

For $\tilde{\mathcal{M}}_\gamma$, the set of indices selected by combined partition-based screening in § 4·3, we choose $\gamma = |\hat{\beta}_{(s)}|$ with $s = \min_{1 \leqslant k \leqslant K} s^{(k)}$. Here $s^{(k)}$ is the size of the higher-criticism t-tests for partition k.

5. SIMULATION STUDIES

We conducted simulation studies to compare the model selection accuracy of partition-based screening with that of existing variable screening methods. We generated the covariates X from multivariate normal distributions and specified the true coefficient β with four different settings.

Setting 1: $(n,p) = (200, 5000)$ and $\beta = (3,3,3,3,3,-7{\cdot}5,0_{p-6}^{\mathrm{T}})^{\mathrm{T}}$. Thus $\mathcal{M} = \{1,\ldots,6\}$. The covariance structure of X is compound symmetric with unit variance and correlation 0·5, i.e., $\mathrm{cov}(X) = 0{\cdot}5I_p + 0{\cdot}51_p1_p^{\mathrm{T}}$.

Setting 2: $(n,p) = (200, 5000)$ and $\beta_j = 3(-1)^j I(j \leqslant 10)$ for $j \in \{1,\ldots,p\}$. Thus $\mathcal{M} = \{1,\ldots,10\}$. The covariance structure of X is that of a block first-order autoregression model with unit variance and correlation 0·9, i.e., $\mathrm{cov}(X_j, X_{j'}) = 0{\cdot}9^{|j-j'|}$ for any $j \neq j' \in \mathcal{B}_k$, where $\mathcal{B}_k = \{j \in \mathbb{Z} : 100k - 99 \leqslant j \leqslant 100k\}$ for $k = 1,\ldots,50$, and $\mathrm{cov}(X_j, X_{j'}) = 0$ otherwise.

Setting 3: $(n,p) = (200, 5000)$ and $\beta_j = (-1)^j I(j \leqslant 10)$ for $j \in \{1,\ldots,p\}$. Thus $\mathcal{M} = \{1,\ldots,10\}$. The covariance structure of X is block compound symmetric with unit variance and correlation 0·9, i.e., $\mathrm{cov}(X_j, X_{j'}) = 0{\cdot}9$ for any $j \neq j' \in \mathcal{B}_k$, where $\mathcal{B}_k = \{j \in \mathbb{Z} : 100(k-1)+1 \leqslant j \leqslant 100k\}$ for $k = 1,\ldots,50$, and $\mathrm{cov}(X_j, X_{j'}) = 0$ otherwise.

Setting 4: $(n,p) = (500, 10\,000)$. We first define $\mathcal{S}$ as a collection of 100×100 equally spaced grid points on $[0,1]^2$. Specifically, set $\mathcal{S} = \{s_j\}_{j=1}^p$ with $s_j = 0{\cdot}01(l,k), j = (100 - 0{\cdot}5l)(l - 1) + k - l$ for $1 \leqslant l,k \leqslant 100$, and $\mathcal{S} = \bigcup_{g=1}^{100} \mathcal{S}_g$, where $\mathcal{S}_g \cap \mathcal{S}_{g'} = \emptyset$ for any $g \neq g' \in \{1,\ldots,G\}$. All the s_j were clustered into 100 exclusive spatially contiguous regions using a k-means clustering algorithm. Set $\beta_j = 3(-1)^j$ if $s_j \in \mathcal{S}_1$ and $\beta_j = 0$ otherwise. The covariance structure of X is exponentially decaying over space, $\mathrm{cov}(X_j, X_{j'}) = \exp(-10\|s_j - s_{j'}\|_2)$ for any $j \neq j' \in \mathcal{S}$. For example, $\mathrm{cov}(X_j, X_{j'}) = 0{\cdot}9$ when $\|s_j - s_{j'}\|_2 = 0{\cdot}01$ and $\mathrm{cov}(X_j, X_{j'}) < 0{\cdot}05$ when $\|s_j - s_{j'}\|_2 > 0{\cdot}3$. This configuration was designed to mimic the spatial data with an active set $\mathcal{M} = \{j : s_j \in \mathcal{S}_1\}$.

Given X and β generated from each of the above settings, we generated Y from a linear regression model and a logistic regression model. For the linear regression model, we set the variance of random errors so that the theoretical R^2 is equal to 0·9. We replicated our simulation 200 times and evaluated the performance using the following criteria: probability of including the true model, minimum model size, and true positive rate.

DEFINITION 3. *Every partition $\mathcal{G}$ belongs to at least one of the following types:*

(i) *a size-reduced partition if there exists a group $g \in \mathcal{G}$ that contains all active covariates, which means $\mathcal{M} \subset \{j : g_j = g\}$;*
(ii) *an optimal partition if there exists a group $g \in \mathcal{G}$ that is a collection of all active covariates, which means $\mathcal{M} = \{j : g_j = g\}$;*
(iii) *a misspecified partition if there does not exist a group $g \in \mathcal{G}$ such that $\mathcal{M} \subset \{j : g_j = g\}$.*

Each partition is either a size-reduced partition or a misspecified partition. An optimal partition must be a size-reduced partition. Neither the misspecified partition, the size-reduced partition, nor the optimal partition, is unique in general. For each non-optimal reduced partition, there exists at least a group containing $\mathcal{M}$ that can be further reduced in size while containing $\mathcal{M}$.

We assessed the performance of the proposed methods under various partition types. In Setting 1, the size-reduced partition $\mathcal{G}^{\mathrm{red}} = (g_1^{\mathrm{red}},\ldots,g_{357}^{\mathrm{red}})$ was specified with 357

groups and each group had 14 members except for group 357 which had 16 members; the group size was approximately $n^{1/2}$. For covariate $j \in \{1, \ldots, 4998\}$, its group label was assigned to be $g_j^{\mathrm{red}} = \sum_{g=1}^{357} gI(14g - 13 \leqslant j \leqslant 14g)$, and for $j = 4999$ or 5000, $g_j^{\mathrm{red}} = 357$, where $\mathcal{M} \subset \{j : g_j^{\mathrm{red}} = 1\}$ for each setting. The misspecified partitions $\mathcal{G}^{\mathrm{mis}_1} = (g_1^{\mathrm{mis}_1}, \ldots, g_p^{\mathrm{mis}_1})^{\mathrm{T}}$ and $\mathcal{G}^{\mathrm{mis}_2} = (g_1^{\mathrm{mis}_2}, \ldots, g_p^{\mathrm{mis}_2})^{\mathrm{T}}$ were respectively sampled from $\mathrm{pr}(g_j^{\mathrm{mis}_1} = g) = 1/178$ $(g = 1, \ldots, 178)$ and $\mathrm{pr}(g_j^{\mathrm{mis}_2} = g) = 1/357$ $(g = 1, \ldots, 357)$. In this setting, variable X_6 is marginally unimportant but conditionally important. Some studies (Barut et al., 2016; Hong et al., 2016a,b) have shown that conditional sure independence screening performs much better than sure independence screening in terms of retaining X_6. In Settings 2 and 3, where the correlation matrix for covariates is block diagonal, we focused primarily on the performance of partition-based screening under reduced partitions and correlation-guided partition-based screening with the same specifications as in Setting 1. The partition determined by the estimated correlation structure is denoted by $\mathcal{G}^{\mathrm{cor}}$. In Setting 4, the optimal partition $\mathcal{G}^{\mathrm{opt}}$ was designed as follows. For each covariate j, $g_j^{\mathrm{opt}} = \sum_{g=1}^{100} gI(s_j \in \mathcal{S}_g)$, where $\mathcal{M} = \{j : g_j^{\mathrm{opt}} = 1\} = \mathcal{S}_1$. To generate $\mathcal{G}^{\mathrm{red}}$, we combined groups 1 and 2 while keeping other groups intact, i.e., for each covariate j, $g_j^{\mathrm{red}} = I(s_j \in \mathcal{S}_1 \cup \mathcal{S}_2) + \sum_{g=2}^{99} gI(s_j \in \mathcal{S}_{g+1})$, where $\mathcal{M} \subset \{j : g_j^{\mathrm{opt}} = 1\} = \mathcal{S}_1 \cup \mathcal{S}_2$. To form the misspecified partitions in Setting 4, we split $\mathcal{S}_1$ into two adjacent but mutually exclusive subregions $\mathcal{S}_{1,1}$ and $\mathcal{S}_{1,2}$ such that $\mathcal{S}_1 = \mathcal{S}_{1,1} \cup \mathcal{S}_{1,2}$. We considered two different misspecified partitions, denoted by $\mathcal{G}^{\mathrm{mis}_1}$ and $\mathcal{G}^{\mathrm{mis}_2}$, where $g_j^{\mathrm{mis}_1} = \sum_{g=1}^{2} gI(s_j \in \mathcal{S}_{1,g}) + \sum_{g=3}^{101} gI(s_j \in \mathcal{S}_{g-1})$ and $g_j^{\mathrm{mis}_2} = I(s_j \in \mathcal{S}_{1,g}) + 2I(s_j \in \mathcal{S}_{1,2} \cup \mathcal{S}_2) + \sum_{g=3}^{100} g(s_j \in \mathcal{S}_g)$. Figure 3 is a graphical representation of Setting 4.

For further comparison, we investigated the performance of sure independence screening (Fan & Lv, 2008) and high-dimensional ordinary least squares projection (Wang & Leng, 2016) in Settings 1–4. In addition, conditional sure independence screening (Barut et al., 2016) with a conditioning variable X_1 was considered for Settings 1–3. To evaluate the performance of high-dimensional ordinary least squares projection for logistic regression, we modified it to accommodate generalized linear models with a ridge penalty by specifying the tuning parameter to be 1. Sure and conditional sure independence screening results were obtained using the R (R Development Core Team, 2017) package SIS, while high-dimensional ordinary least squares projection was implemented using the R package screening. To make different methods comparable, we chose γ so that the number of selected indices was equal to the sample size and computed the true positive rate and the probability that the selected indices include the true model, following Fan & Lv (2008).

Table 1 summarizes the simulation results. In Settings 1–3, partition-based screening performs best for linear and logistic regression. In Setting 4, the performance of spatial-oriented reduced partition screening is almost the same as spatial-oriented optimal partition screening in linear regression, and is close to spatial-oriented optimal partition screening in logistic regression. This indicates that even when an optimal partition is not available, a size-reduced partition is a good alternative. In Settings 2 and 3, correlation-guided partition screening produces better selection accuracy than all three existing methods, and is comparable to partition-based screening. Thus, when there is insufficient prior knowledge to determine a size-reduced partition but the covariate variables have a block-diagonal correlation structure up to permutations, data-driven partition-based screening can yield improved selection accuracy. In Setting 1, where the covariance structure of covariates is compound symmetric, correlation-guided partition screening does not yield more accurate selection than high-dimensional ordinary least squares projection and conditional sure independence screening. In this case we examined the performance of partition-based screening with randomly generated misspecified partitions, as well as combined partition-based

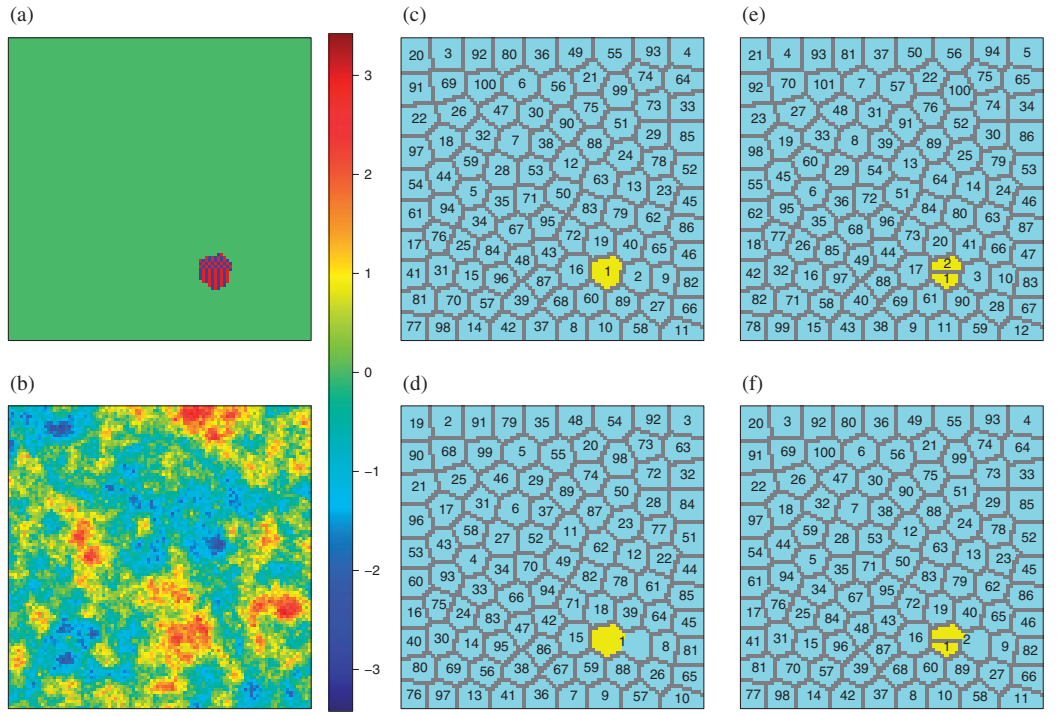

Fig. 3. Model designs and configurations of different partitions on $\mathcal{S}$, the 100×100 equally spaced grid points in two-dimensional space $[0, 1]^2$, in Setting 4 of the simulation study for spatial variable screening: (a) the true coefficient β_j, which takes only three possible values, -3 in blue, 3 in red and 0 in green; (b) one set of simulated predictors X_j over space from a Gaussian random field on $\mathcal{S}$ with covariance kernel $\exp(-10\|s - s'\|_2)$ for $s, s' \in \mathcal{S}$; (c) the partition $\{\mathcal{S}_g\}_{g=1}^{100}$ in Setting 4 to define the true nonzero coefficients, with $\beta_j \neq 0$ if and only if $s_j \in \mathcal{S}_1$, in yellow. Panels (c), (d), (e) and (f) respectively represent the optimal partition $\mathcal{G}^{\text{opt}}$, the size-reduced partition $\mathcal{G}^{\text{red}}$, and two misspecified partitions, $\mathcal{G}^{\text{mis}_1}$ and $\mathcal{G}^{\text{mis}_2}$, for the corresponding partition-based screening for selecting variables in Setting 4; in panels (c)–(f), $\mathcal{S}_1$ is coloured yellow.

screening that combines five or ten different random partition-based screening statistics. The results indicate that combined partition-based screening with ten random $\mathcal{G}^{\text{mis}_2}$ is slightly better than combined partition-based screening with five random $\mathcal{G}^{\text{mis}_2}$ and outperforms partition-based screening with $\mathcal{G}^{\text{mis}_1}$ only or with $\mathcal{G}^{\text{mis}_2}$ only. Therefore, in the absence of an optimal partition, combining multiple partitions for variable screening fares better than relying on a single partition. Moreover, it produces better results than high-dimensional ordinary least squares projection and sure independence screening in Setting 1. The advantages of combined partition-based screening are obvious in Setting 4, where the accuracy of spatial-oriented partition screening with $\mathcal{G}^{\text{red}}$ can be improved by combining it with two misspecified partition-based screening statistics. Thus, combining various screening statistics from multiple sources of partitions, even though they may have been misspecified, appears to be a useful strategy.

6. APPLICATION

We applied the proposed methods to analyse resting-state functional magnetic resonance imaging data from the Autism Brain Imaging Data Exchange study (Di Martino et al., 2014). The primary goal of this study was to understand how brain activity is associated with autism spectrum disorder, a disease with substantial heterogeneities among children. Functional magnetic resonance imaging measures blood oxygen levels linked to neural activity, and resting-state

Table 1. *Model selection accuracy of variable screening methods for linear and logistic regression in Settings* 1–4

Setting 1	Linear regression			Logistic regression		
	PIT (%)	MMS	TPR (%)	PIT (%)	MMS	TPR (%)
SIS	10	2273	84	12	2590	80
HOLP	91	24	98	53	173	90
CSIS	97	11	100	73	62	95
PartS ($\mathcal{G}^{\text{red}}$)	100	7	100	100	6	100
PartS ($\mathcal{G}^{\text{mis}_1}$)	84	36	97	46	226	88
PartS ($\mathcal{G}^{\text{mis}_2}$)	87	34	98	60	145	91
CombPartS (5 $\mathcal{G}^{\text{mis}_2}$)	93	25	99	64	115	93
CombPartS (10 $\mathcal{G}^{\text{mis}_2}$)	96	22	99	74	105	96
CorrPartS ($\mathcal{G}^{\text{cor}}$)	78	64	96	47	219	88
Setting 2	Linear regression			Logistic regression		
	PIT (%)	MMS	TPR (%)	PIT (%)	MMS	TPR (%)
SIS	0	3982	56	0	4243	41
HOLP	0	3721	59	0	4231	43
CSIS	69	52	96	0	311	90
PartS ($\mathcal{G}^{\text{red}}$)	100	10	100	100	10	100
CorrPartS ($\mathcal{G}^{\text{cor}}$)	91	10	98	88	10	98
Setting 3	Linear regression			Logistic regression		
	PIT (%)	MMS	TPR (%)	PIT (%)	MMS	TPR (%)
SIS	0	3671	20	100	1367	60
HOLP	65	110	100	16	560	90
CSIS	6	3033	70	6	2265	70
PartS ($\mathcal{G}^{\text{red}}$)	100	11	100	100	10	100
CorrPartS ($\mathcal{G}^{\text{cor}}$)	62	101	100	18	426	80
Setting 4	Linear regression			Logistic regression		
	PIT (%)	MMS	TPR (%)	PIT (%)	MMS	TPR (%)
SIS	0	9890	17	0	9860	26
HOLP	0	9886	18	0	9876	27
SpatPartS ($\mathcal{G}^{\text{opt}}$)	100	100	100	100	100	100
SpatPartS ($\mathcal{G}^{\text{red}}$)	100	100	100	73	174	100
SpatPartS ($\mathcal{G}^{\text{mis}_1}$)	0	8232	65	0	9033	65
SpatPartS ($\mathcal{G}^{\text{mis}_2}$)	0	9190	55	0	9269	74
CombPartS ($\mathcal{G}^{\text{red}}$, $\mathcal{G}^{\text{mis}_1}$, $\mathcal{G}^{\text{mis}_2}$)	100	100	100	80	174	100

SIS, sure independence screening; HOLP, high-dimensional ordinary least squares projection; CSIS, conditional sure independence screening; PartS, partition-based screening; CombPartS, combined partition-based screening; CorrPartS, correlation-guided partition screening; SpatPartS, spatial-oriented partition screening; MMS, median minimum size of the selected models that are required to have a sure screening; TPR, average true positive rate; PIT, estimated probability of including all true predictors in the top n selected predictors.

functional magnetic resonance imaging measures brain activity only when the brain is not performing any tasks. This study aggregated 20 resting-state functional magnetic resonance imaging datasets from 17 experiment sites. For each subject, the resting-state functional magnetic resonance imaging signal was recorded for each voxel in the brain over multiple time-points. Standard imaging pre-processing steps (Di Martino et al., 2014) included motion correction, slice-timing correction, and spatial smoothing. The entire brain was registered into the 3 mm standard Montreal Neurological Institute space, which consists of 38 547 voxels in 90 brain regions defined by the automated anatomical labelling system (Hervé et al., 2012). After removal of missing values, the complete dataset included 819 subjects, consisting of 378 patients and 441 age-matched

Table 2. *Eight automated anatomical labelling regions with more than* 60 *voxels that are selected by combined partition-based screening*

Selected region	Voxel counts	Median rank	Selected region	Voxel counts	Median rank
Frontal_Mid_R	95	5	Frontal_Sup_L	71	14
Temporal_Mid_R	78	27	Frontal_Mid_L	71	47
Temporal_Mid_L	75	154	Frontal_Sup_R	66	35
Precuneus_L	73	56	Precuneus_R	65	41

controls. To select imaging biomarkers for autism spectrum disorder risk prediction, we considered the fractional amplitude of low-frequency fluctuations (Zou et al., 2008), defined as the ratio of the power spectrum for frequencies 0·01–0·08 Hz to the entire frequency range. This measure has been widely used as a voxel-wise measure of the intrinsic functional brain architecture derived from resting-state functional magnetic resonance imaging data (Zuo et al., 2010).

We constructed a spatial logistic regression model that has clinical diagnosis of autism spectrum disorder as the outcome and the voxel-wise fractional amplitudes of frequency fluctuations as imaging predictors, adjusting for age at scan, sex and intelligence quotient. Because imaging predictors on the risk of autism spectrum disorder are spatially clustered and sparse (Liu & Calhoun, 2014), the primary aim of this study is to identify imaging biomarkers among 38 547 voxel-level fractional amplitudes of low-frequency fluctuations that predict the autism spectrum disorder risk. We applied partition-based screening by using anatomical information, correlation among imaging predictors and spatial information. Specifically, we considered the following methods: brain region partition-based screening on 90 brain regions; correlation-guided partition screening, which partitions the 38 547 voxels into G groups using the clustering algorithm introduced in § 4·2, where G is taken to be 256, 128, 64, 32, 16 or 8; spatial-oriented partition screening, which partitions the 38 547 voxels into 1024 equal-sized regions where the voxels are spatially contiguous within each region; and combined partition-based screening that combines all of the above. We also applied high-dimensional ordinary least squares projection for logistic regression with the ridge penalty, as implemented in the R package screening.

To assess the performance of the different methods, we used ten-fold crossvalidation, randomly splitting the data into ten equal-sized subsets. We applied the variable screening methods to the training dataset and obtained a set of selected voxels, based on which we made a prediction about the disease status in the testing dataset using logistic regression with the elastic net penalty, as implemented in the R package glmnet. We repeated this ten times and computed the crossvalidation accuracy. Among all the methods, combined partition-based screening achieved the smallest crossvalidation prediction error, 37%, and high-dimensional ordinary least squares projection had the largest crossvalidation prediction error, 48%. All the other partition-based screening methods achieved a prediction error of approximately 40%. More details and the receiver operating characteristic curves are given in the Supplementary Material.

Next, we applied combined partition-based screening to the entire dataset, using the method in § 4·4 to determine the threshold by taking an upper bound on the expected false positive rate to be 0·20. A total of 6142 important voxels were selected. Eight regions with more than 60 selected voxels are reported in Table 2, along with the median rank of voxel-specific screening statistics within each region. These regions are known to be involved in specific brain functions related to autism (Friederici et al., 2003; Japee et al., 2015).

7. DISCUSSION

The method proposed in this paper can be improved. First, our framework requires that the size of each partition group be less than the sample size to make (3) sensible. If this condition is

not met, penalized likelihood methods such as the lasso can be applied, though these approaches may involve the selection of tuning parameters and the correction of biases due to penalization. A simple but efficient remedy would be to further refine the groups randomly. This refining procedure can be performed multiple times, and the resulting screening statistics can be combined using the rule in § 4·3. Second, although this paper has focused on non-overlapping partitions for ease of theoretical development, our screening framework can accommodate overlapping partitions. According to the combination rule in § 4·3, for each predictor that is covered by more than one partition, we can simply choose the screening statistic with larger value. Third, the time complexity of correlation-guided partition screening is $O(p^2)$, mainly due to the need to compute the correlation matrix and clustering predictors. To compute the correlations among ultrahigh-dimensional predictors more efficiently, we suggest adopting parallel computing techniques. To speed up the clustering of predictors, we propose to threshold the correlation matrix and generate a binary matrix, regarded as the adjacency matrix of an undirected graph. The connected components corresponding to group partitions can be obtained by using the breadth-first or depth-first search algorithms with time complexity between $O(p)$ and $O(p^2)$.

Acknowledgement

The authors are grateful to the editor, the associate editor and the referees for their constructive comments and suggestions, which have substantially improved the manuscript. Kang was supported in part by the U.S. National Institutes of Health. Hong was supported in part by the U.S. National Security Agency. Li was supported in part by the U.S. National Institutes of Health and the National Natural Science Foundation of China.

Supplementary material

Supplementary material available at *Biometrika* online includes theoretical results and further results and figures for the simulation study.

References

Barut, E., Fan, J. & Verhasselt, A. (2016). Conditional sure independence screening. *J. Am. Statist. Assoc.* **111**, 1266–77.

Candès, E. & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–51.

Cho, H. & Fryzlewicz, P. (2012). High dimensional variable selection via tilting. *J. R. Statist. Soc.* B **74**, 593–622.

Cui, H., Li, R. & Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *J. Am. Statist. Assoc.* **110**, 630–41.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M. et al. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molec. Psychiat.* **19**, 659–67.

Fan, J. & Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Ann. Statist.* **36**, 2605–37.

Fan, J., Feng, Y. & Song, R. (2012). Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Am. Statist. Assoc.* **106**, 544–57.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc.* B **70**, 849–911.

Fan, J., Samworth, R. & Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **10**, 2013–38.

Fan, J. & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38**, 3567–604.

FRALEY, C. & RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Am. Statist. Assoc.* **97**, 611–31.

FRIEDERICI, A. D., RÜSCHEMEYER, S.-A., HAHNE, A. & FIEBACH, C. J. (2003). The role of left inferior frontal and superior temporal cortex in sentence comprehension: Localizing syntactic and semantic processes. *Cereb. Cortex* **13**, 170–7.

FRIEDMAN, J. H., HASTIE, T. J. & TIBSHIRANI, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* **33**, 1–22.

GORST-RASMUSSEN, A. & SCHEIKE, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *J. R. Statist. Soc.* B **75**, 217–45.

HALL, P. & MILLER, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comp. Graph. Statist.* **18**, 533–50.

HERVÉ, P.-Y., RAZAFIMANDIMBY, A., VIGNEAU, M., MAZOYER, B. & TZOURIO-MAZOYER, N. (2012). Disentangling the brain networks supporting affective speech comprehension. *NeuroImage* **61**, 1255–67.

HONG, H. G., KANG, J. & LI, Y. (2016a). Conditional screening for ultra-high dimensional covariates with survival outcomes. *Lifetime Data Anal.* doi: 10.1007/s10985-016-9387-7.

HONG, H. G., WANG, L. & HE, X. (2016b). A data-driven approach to conditional screening of high-dimensional variables. *Stat* **5**, 200–12.

HUANG, J., HOROWITZ, J. L. & MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587–613.

JAIN, A. K. (2010). Data clustering: 50 years beyond k-means. *Pat. Recog. Lett.* **31**, 651–66.

JAPEE, S., HOLIDAY, K., SATYSHUR, M. D., MUKAI, I. & UNGERLEIDER, L. G. (2015). A role of right middle frontal gyrus in reorienting of attention: A case study. *Front. Syst. Neurosci.* **9**. doi: 10.3389/fnsys.2015.00023.

JIN, J., ZHANG, C.-H. & ZHANG, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *J. Mach. Learn. Res.* **15**, 2723–72.

KE, T., JIN, J. & FAN, J. (2014). Covariance assisted screening and estimation. *Ann. Statist.* **42**, 2202–42.

LI, G., PENG, H., ZHANG, J. & ZHU, L. (2012). Robust rank correlation based screening. *Ann. Statist.* **40**, 1846–77.

LI, J., ZHENG, Q., PENG, L. & HUANG, Z. (2016). Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics* **72**, 1145–54.

LIU, J. & CALHOUN, V. D. (2014). A review of multivariate analyses in imaging genetics. *Front. Neuroinfo.* **8**, 1–11.

MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–62.

MEINSHAUSEN, N. & BÜHLMANN, P. (2010). Stability selection (with Discussion). *J. R. Statist. Soc.* B **72**, 417–73.

MURTAGH, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *Comp. J.* **26**, 354–9.

NIU, Y., HAO, N. & AN, L. (2011). Detection of rare functional variants using group ISIS. *BMC Proc.* **5**, S108.

R DEVELOPMENT CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc.* B **58**, 267–88.

WANG, H. & LENG, C. (2012). Unified LASSO estimation by least squares approximation. *J. Am. Statist. Assoc.* **102**, 1039–48.

WANG, X. & LENG, C. (2016). High dimensional ordinary least squares projection for screening variables. *J. R. Statist. Soc.* B **78**, 589–611.

WANG, X., ZHU, H. & FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2017). Generalized scalar-on-image regression models via total variation. *J. Am. Statist. Assoc.* doi: 10.1080/01621459.2016.1194846.

YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49–67.

ZHANG, H. H. & LU, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika* **94**, 691–703.

ZHAO, P. & YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541–63.

ZHAO, S. D. & LI, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Mult. Anal.* **105**, 397–411.

ZHU, L.-P., LI, L., LI, R. & ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Am. Statist. Assoc.* **696**, 1464–75.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

ZOU, H. & HASTIE, T. J. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B **67**, 301–20.

ZOU, H. & ZHANG, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37**, 1733–51.

ZOU, Q.-H., ZHU, C.-Z., YANG, Y., ZUO, X.-N., LONG, X.-Y., CAO, Q.-J., WANG, Y.-F. & ZANG, Y.-F. (2008). An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF. *J. Neurosci. Meth.* **172**, 137–41.

ZUO, X.-N., DI MARTINO, A., KELLY, C., SHEHZAD, Z. E., GEE, D. G., KLEIN, D. F., CASTELLANOS, F. X., BISWAL, B. B. & MILHAM, M. P. (2010). The oscillating brain: Complex and reliable. *NeuroImage* **49**, 1432–45.

[*Received on* 18 *August* 2016. *Editorial decision on* 3 *August* 2017]