Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Model-Free Variable Selection with Matrix-Valued Predictors

Zeda Li & Yuexiao Dong

To cite this article: Zeda Li & Yuexiao Dong (2020): Model-Free Variable Selection with Matrix-Valued Predictors, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2020.1806854

To link to this article: https://doi.org/10.1080/10618600.2020.1806854

Model-Free Variable Selection with Matrix-Valued Predictors

Zeda Li

Baruch College, The City University of New York

and

Yuexiao Dong

Department of Statistical Science, Temple University

Corresponding author Zeda Li zeda.li@baruch.cuny.edu

Abstract

We introduce a novel framework for model-free variable selection with matrix-valued predictors. To test the importance of rows, columns, and submatrices of the predictor matrix in terms of predicting the response, three types of hypotheses are formulated under a unified framework. The asymptotic properties of the test statistics under the null hypothesis are established and a permutation testing algorithm is also introduced to approximate the distribution of the test statistics. A maximum ratio criterion (MRC) is proposed to facilitate the model-free variable selection. Unlike the traditional stepwise regression procedures that require calculating *p*-values at each step, the MRC is a non-iterative procedure that does not require *p*-value calculation and is guaranteed to achieve variable selection consistency under mild conditions. Performance of the proposed method is evaluated in extensive simulations and demonstrated through the analysis of an electroencephalography (EEG) data. Supplementary materials for this article are available online.

Keywords: Matrix normal distribution; Matrix-valued predictors; Maximum ratio criterion; Model-free variable selection; Permutation test

1 Introduction

Regression analysis with scalar-valued responses and vector-valued predictors has been the focus in traditional multivariate analysis. Datasets with matrix-valued predictor X and scalar response Y, however, are becoming more common in contemporary data analysis. See, for example, the electroencephalography (EEG) data (Li et al., 2010), the neuroimaging data (Zhou et al., 2013), and the longitudinal biomarker data (Pfeiffer et al., 2012). Our motivating application comes from a study of multivariate EEG, which is used to measure electrophysiological activity simultaneously across multiple

regions, or channels, of the brain. The study considers two groups of subjects: an alcoholic group and a control group. Each subject is exposed to a stimulus while the 64-channel EEG is recorded at 256 time points within a second, resulting in a 64 × 256 matrix-valued predictor. The goal of our analysis is to identify rows and columns of the predictor matrix that are important to discriminate the subjects from the two distinct groups.

To analyze datasets with matrix-valued predictors, an increasing number of methods have been proposed in the sufficient dimension reduction literature. Li et al. (2010) introduced the central dimension folding subspace that aims to reduce predictor's row and column dimensions and to retain full regression information of $F_{Y|X}$, where $F_{Y|X}$ denotes the conditional distribution of Y given X. Xue and Yin (2014) focused on the central mean dimension folding subspace and proposed to reduce the dimensions of the matrix-valued predictors while preserving the regression mean information of Y given X. Dimension folding for functionals of the conditional distribution $F_{Y|X}$, which includes the conditional mean as a special case, was studied in Xue and Yin (2015). Motivated by the ensemble sufficient dimension approach for vector-valued predictors proposed by Yin and Li (2011), Xue et al. (2016) introduced the ensemble sufficient dimension folding methods for matrix-valued predictors. Ding and Cook (2014) proposed extensions of the principal component analysis and principal fitted components for matrix-valued predictors. Important connections between the classical slice inverse regression (Li, 1991) and the central dimension folding subspace are investigated in Ding and Cook (2015a,b). All of the aforementioned methods were proposed to reduce the predictor's row (or column) dimensionality by finding linear combinations of the rows (or columns) in a model-free manner without selecting particular rows (or columns), and they are model-free in the sense that no parametric link function between Y and X is assumed.

Although sufficient dimension reduction methods for matrix-valued predictors have been studied extensively, existing variable selection methods for matrix-valued predictors are relatively few. Zhao and Leng (2014) proposed a

penalized regression method for matrix-valued predictors, which can be used to select important rows and columns of the predictor matrix. However, this method is under the linear model assumption and thus is not model-free. More recently, Wang (2016) proposed using the regularized sufficient dimension folding to perform variable selection. In addition to these methods, existing variable selection methods for vector-valued predictors, such as trace pursuit (Yu et al., 2016a) and distance correlation screening (Li et al., 2012), can also be applied to a dataset with matrix-valued predictors by vectorizing the predictor. However, as noted by Li et al. (2010) and shown in our simulation studies, vectorizing a matrix-valued predictor may lose the original data structure, reduces the estimation accuracy, and hinders the ability to recover the important rows and columns.

In this article, we present a general framework for model-free variable selection with scalar response Y and matrix-valued predictor X. A unified hypothesis testing approach is developed to test the importance of rows, columns, or any submatrix of X in terms of predicting Y. Asymptotic tests and permutation tests are proposed to approximate the distribution of the test statistics under the null hypothesis. Furthermore, a novel maximum ratio criterion (MRC) is introduced to facilitate the model-free variable selection. Sequential procedures, such as stepwise regression or backward elimination, can be combined with the newly proposed permutation test to implement the model-free variable selection. However, these sequential tests could be very time-consuming due to its iterative nature and the requirement of p-value calculation at each iteration. The MRC, on the other hand, is a non-iterative procedure that does not require p-value calculation.

The rest of the article is organized as follows. Section 2 presents a general model for the proposed hypothesis tests and MRC. Section 3 describes the procedures for the active row recovery. Section 4 introduces the procedures for testing the contribution of submatrices. Section 5 presents the results of extensive simulation studies. The proposed framework is applied to real data in Section 6. Section 7 concludes the paper with some discussions and future directions. The procedures for the active column recovery are discussed in

the Appendix. Details of matrix-valued normal distribution and proofs are given in the supplementary materials.

2 The General Framework

Suppose the response variable $Y \in \mathbb{R}$ and predictor $X \in \mathbb{R}^{p\times q}$ have the following general relationship:

$$Y = g(X) + \varepsilon, \quad (1)$$

where $g : \mathbb{R}^{p\times q} \mapsto \mathbb{R}$ is an unknown function, ε is independent of X, and $E(\varepsilon) = 0$. We assume that X follows the matrix normal distribution, which is denoted as $X \sim N_{p,q}(\mu, U, V)$ with $\mu \in \mathbb{R}^{p\times q}, U \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{q\times q}$. Let $\mathrm{tr}(\cdot)$ be the trace of a matrix and $\mu = E(X)$. Then, the row covariance matrix is $U = E\{(X-\mu)(X-\mu)^{\mathrm{T}}\}/\mathrm{tr}(V)$, and the column covariance matrix is $V = E\{(X-\mu)^{\mathrm{T}}(X-\mu)\}/\mathrm{tr}(U)$. The formal definition and properties of the matrix normal distribution can be found in De Waal (1985) and Gupta and Nagar (2000). Without loss of generality, we assume μ is a $p \times q$ zero matrix, such that $X \sim N_{p,q}(0_{p\times q}, U, V)$.

Let $\mathcal{I}_{\mathrm{row}} = \{1, \cdots, p\}$ be the full index set of rows and $X_{j,\cdot}$ be the jth row of X for $j = 1, \cdots, p$. Define the active row set $\mathcal{A}$ as

$$\mathcal{A} = \left\{ j \in \mathcal{I}_{\mathrm{row}} : Y \text{ depends on } X_{j,\cdot} \text{ in model (1)} \right\}.$$

Similarly, let $\mathcal{I}_{\mathrm{col}} = \{1, \cdots, q\}$ be the full index set of columns and $X_{\cdot,k}$ be the kth column of X for $k = 1, \cdots, q$. Define the active column set $\mathcal{B}$ as

$$\mathcal{B} = \left\{ k \in \mathcal{I}_{\mathrm{col}} : Y \text{ depends on } X_{\cdot,k} \text{ in model (1)} \right\}.$$

Based on the active row and column predictors, model (1) can be expressed as

$$Y = g^{*}(X_{\mathcal{A},\mathcal{B}}) + \varepsilon,$$

where $g^*: \mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|} \mapsto \mathbb{R}$ with $|\cdot|$ denoting the cardinality of a set, and $X_{\mathcal{A},\mathcal{B}}$ denotes the submatrix of X that contains the active rows indexed by $\mathcal{A}$ and the active columns indexed by $\mathcal{B}$. Under model (1), Y depends on X only through $X_{\mathcal{A},\mathcal{B}}$. Our goal of model-free variable selection is to recover $X_{\mathcal{A},\mathcal{B}}$ without assuming the forms of $g(\cdot)$ and $g^*(\cdot)$.

3 Procedures to Recover the Active Rows

In this section, we introduce procedures to recover the active row set $\mathcal{A}$. The population-level development of the test statistics for an individual row is provided in Section 3.1, while the corresponding sample level test statistics and its asymptotic distribution under the null hypothesis are discussed in Section 3.2. The permutation test procedure is introduced in Section 3.3. The MRC to recover $\mathcal{A}$ is presented in Section 3.4. We formulate the column hypotheses in Section 3.5. Since the asymptotic test for the column hypotheses, the permutation test for the column hypotheses, and using MRC for active column recovery are parallel to the development of the row hypotheses, the details of the column hypotheses are provided in the Appendix.

3.1 Population Level Development

Let $X_{j,\cdot}, j = 1, \cdots, p$, be the jth row of X and $X_{-j,\cdot} \in \mathbb{R}^{(p-1)\times q}$ be the matrix that includes all but the jth row of X. To test the importance of $X_{j,\cdot}$ in terms of predicting Y, we consider the following row hypotheses

$$H_{0,\{j\}}^{\text{row}}: Y \perp\!\!\!\perp X \mid X_{-j,\cdot} \text{ v.s. } H_{a,\{j\}}^{\text{row}}: \text{Y is not independent of } X \text{ given } X_{-j,\cdot}, \qquad (2)$$

where "$\perp\!\!\!\perp$" denotes independence. Under the null hypothesis, $H_{0,\{j\}}^{\text{row}}$, the response Y depends on X only through $X_{-j,\cdot}$, which indicates that $X_{j,\cdot}$ is not important to predict the response Y. In the special case of $q = 1$, X becomes a p-dimensional vector, and (2) is equivalent to testing the importance of one component of X given the other $p - 1$ predictors. This special case is known as the marginal coordinate test (Cook, 2004).

Recall that U is the row covariance matrix of X. Let $U_{-j,-j} \in \mathbb{R}^{(p-1)\times(p-1)}$ be the submatrix of U that excludes the jth row and the jth column of U. Define the following quantity

$$\delta_j^{\text{row}} = \text{tr}(M) - \text{tr}(M_{-j,\cdot}), \qquad (3)$$

where $M = U^{-1}E(XY)V^{-1}E^{\text{T}}(XY)$ and $M_{-j,\cdot} = U_{-j,-j}^{-1}E(X_{-j,\cdot}Y)V^{-1}E^{\text{T}}(X_{-j,\cdot}Y)$. As we will see in Proposition 1, the trace difference δ_j^{row} is the key quantity to test the importance of the jth row of X. Intuitively, $\text{tr}(M)$ captures the strength of the relationship between Y and X, $\text{tr}(M_{-j,\cdot})$ captures the strength of the relationship between Y and $X_{-j,\cdot}$, and the trace difference measures the effect of $X_{j,\cdot}$ on Y in the presence of all the other rows.

Some notations are needed before we state the main results. Denote $U_{j,j} \in \mathbb{R}$ as the element in the jth row and jth column of U. Similarly, we define $U_{-j,j} \in \mathbb{R}^{p-1}$ and $U_{j,-j} = U_{-j,j}^{\text{T}}$. Furthermore, let $U_{j,j|-j} = U_{j,j} - U_{j,-j}U_{-j,-j}^{-1}U_{-j,j}$ and $R_{j|-j} = X_{j,\cdot} - U_{j,-j}U_{-j,-j}^{-1}X_{-j,\cdot}$.

Proposition 1. *Suppose* $X \sim N_{p,q}(0_{p\times q}, U, V)$. *Then*

1. $\delta_j^{\text{row}} = U_{j,j|-j}^{-1}E(R_{j|-j}Y)V^{-1}E^{\text{T}}(R_{j|-j}Y)$.
2. $\delta_j^{\text{row}} = 0$ *under* $H_{0,\{j\}}^{\text{row}}$.

Part 1 of Proposition 1 provides the explicit formula to calculate δ_j^{row} in (3). Part 2 of Proposition 1 indicates that δ_j^{row} is zero when $X_{j,\cdot}$ is not important. The principle of using trace difference to test predictors' contributions was first proposed by Yu et al. (2016a). However, our proposal has two major differences from the trace pursuit method in Yu et al. (2016a). First, Yu et al. (2016a) introduced the trace pursuit methods based on the sliced inverse regression (Li, 1991), the sliced average variance estimation (Cook and Weisberg, 1991), and the directional regression (Li and Wang, 2007). Our proposed test, on the other hand, has root in ordinary least squares (Li and

Duan, 1989). Second, Yu et al. (2016a) addressed variable-selection for vector-valued predictors, while we focus on the matrix-valued predictors.

It should be noted that our proposal for model-free variable selection is closely related to sufficient dimension reduction with matrix-valued predictors. More specifically, consider $\Phi_1 \in \mathbb{R}^{p\times d_p}$ and $\Phi_2 \in \mathbb{R}^{q\times d_q}$, such that Φ_1 and Φ_2 have the smallest possible column dimensions d_p and d_q to satisfy $Y @@ \perp\!\!\!\perp X \mid \Phi_1^{\mathrm{T}} X \Phi_2$. Then $\mathrm{Span}(\Phi_2)\otimes\mathrm{Span}(\Phi_1)$ is called the central dimension folding space (Li et al., 2010), where Span denotes the column space and $\otimes$ denotes the Kronecker product. In particular, $\Phi_1^{\mathrm{T}} X$ achieves dimension reduction of the rows by finding linear combinations of the rows of X. Under the normality assumption, it can be shown that $\mathrm{Span}\{U^{-1}E(XY)\}\subseteq\mathrm{Span}(\Phi_1)$. Recall that $M = U^{-1}E(XY)V^{-1}E^{\mathrm{T}}(XY)$, and it follows that $\mathrm{Span}(M)\subseteq\mathrm{Span}(\Phi_1)$. The effectiveness of using M to recover active rows is directly related to the fact that M can be used to estimate $\mathrm{Span}(\Phi_1)$ of the central dimension folding space.

3.2 Asymptotic Test for an Individual Row

Let $\{(X^{(i)},Y^{(i)}),\ i=1,\ldots,n\}$ be an i.i.d. sample of (X,Y). The sample level test statistic for $H_{0,\{j\}}^{\mathrm{row}}$ is denoted as $\hat{\delta}_j^{\mathrm{row}} = \mathrm{tr}(\hat{M}) - \mathrm{tr}(\hat{M}_{-j,\cdot})$, where $\hat{M} = \hat{U}^{-1}E_n(XY)\hat{V}^{-1}E_n^{\mathrm{T}}(XY)$ and $\hat{M}_{-j,\cdot} = \hat{U}_{-j,-j}^{-1}E_n(X_{-j,\cdot}Y)\hat{V}^{-1}E_n^{\mathrm{T}}(X_{-j,\cdot}Y)$ are the sample estimates of M and $M_{-j,\cdot}$, respectively. We obtain $\hat{U}$ and $\hat{V}$ through the maximum likelihood estimators, and the details are provided in the supplementary materials. The sample estimators $E_n(XY)$ and $E_n(X_{-j,\cdot}Y)$ are straightforward to compute. For example, $E_n(XY) = n^{-1}\sum_{i=1}^{n}(X^{(i)}-\bar{X})Y^{(i)}$ with $\bar{X} = n^{-1}\sum_{i=1}^{n}X^{(i)}$. The asymptotic distribution of $\hat{\delta}_j^{\mathrm{row}}$ under the null hypothesis is given in Theorem 1.

Theorem 1. *Suppose* $X \sim N_{p,q}(0_{p\times q}, U, V)$ *and all moments involved are finite. Then*

$$n\hat{\delta}_j^{\text{row}} \xrightarrow{D} U_{j,j|-j}^{-1} \sum_{\ell=1}^{q} d_\ell \chi_\ell^2(1)$$

under $H_{0,\{j\}}^{\text{row}}$*, where "*$\xrightarrow{D}$*" means convergence in distribution,* $\chi_\ell^2(1)$*, for* $\ell = 1, \ldots, q$*, are independent chi-square distribution with one degree of freedom,* $d_1 \geq \ldots \geq d_q$ *are the eigenvalues of* $V^{-1/2}\Lambda V^{-1/2}$*, and the exact form of* Λ *is provided in the supplementary materials.*

The asymptotic distribution in Theorem 1 needs to be estimated in practice. Specifically, let $\hat{U}_{j,j|-j}, \hat{V}$ and $\hat{\Lambda}$ be the sample estimates of $U_{j,j|-j}, V$ and Λ, respectively, and let $\hat{d}_1 \geq \ldots \geq \hat{d}_q$ be eigenvalues of $\hat{V}^{-1/2}\hat{\Lambda}\hat{V}^{-1/2}$. The asymptotic distribution of $n\hat{\delta}_j^{\text{row}}$ under $H_{0,\{j\}}^{\text{row}}$ can be approximated by $\hat{U}_{j,j|-j}^{-1} \sum_{\ell=1}^{q} \hat{d}_\ell \chi_\ell^2(1)$. Denote $\Theta = (\hat{d}_1, \ldots, \hat{d}_q)^{\mathrm{T}} \in \mathbb{R}^q$ and let $W \in \mathbb{R}^{N\times q}$ be i.i.d. samples of $\chi^2(1)$. Then, $\hat{U}_{j,j|-j}^{-1} W\Theta$, which is a *N*-dimensional vector, contains *N* realizations of the approximated asymptotic distribution of $n\hat{\delta}_j^{\text{row}}$ under $H_{0,\{j\}}^{\text{row}}$. The proportion of the elements in $\hat{U}_{j,j|-j}^{-1} W\Theta$ that is greater than $n\hat{\delta}_j^{\text{row}}$ can be treated as the approximated *p*-value. For a given significance level α, we reject $H_{0,\{j\}}^{\text{row}}$ if this *p*-value is smaller than α.

3.3 Permutation Test for an Individual Row

In addition to the asymptotic test, a permutation test can be used as an alternative to test the row contributions. In order to obtain the *p*-value for the row hypotheses (2), we need to approximate the distribution of $\hat{\delta}_j^{\text{row}}$ under $H_{0,\{j\}}^{\text{row}} : Y \perp\!\!\!\perp X \mid X_{-j,\cdot}$. This is possible due to the following key observation.

Proposition 2. *Suppose* $X \sim N_{p,q}(0_{p\times q}, U, V)$*. Then* $Y \perp\!\!\!\perp R_{j|-j}$ *under* $H_{0,\{j\}}^{\text{row}}$*.*

Proposition 2 states that the conditional independence $Y \perp\!\!\!\perp X \mid X_{-j,\cdot}$ and the normality of X implies the unconditional independence between Y and $R_{j|-j}$. Given the fact that $\delta_j^{\text{row}} = U_{j,j|-j}^{-1} E(R_{j|-j}Y)V^{-1}E^{\mathrm{T}}(R_{j|-j}Y)$ from part 1 of Proposition 1 and the independence between Y and $R_{j|-j}$ under the null, we consider the following row permutation (RP) testing algorithm.

RP.1 Based on the original sample $\{(X^{(i)}, Y^{(i)}) : i = 1, \ldots, n\}$, calculate $\hat{\delta}_j^{\text{row}} = \text{tr}(\hat{M}) - \text{tr}(\hat{M}_{-j,\cdot})$.

RP.2 Fix $\{X^{(i)} : i = 1, \ldots, n\}$. For $b = 1, \ldots, B$, denote $\{Y_{[b]}^{(i)} : i = 1, \ldots, n\}$ as the bth random permutation of $\{Y^{(i)} : i = 1, \ldots, n\}$. Then calculate $\hat{\delta}_{j,[b]}^{\text{row}}$ based on the permuted sample $\{(X^{(i)}, Y_{[b]}^{(i)}) : i = 1, \ldots, n\}$.

RP.3 Calculate the p-value $p_j^{\text{row}} = B^{-1} \sum_{b=1}^{B} I\left(\hat{\delta}_{j,[b]}^{\text{row}} > \hat{\delta}_j^{\text{row}}\right)$, where $I(\cdot)$ is indicator function. For a given significance level α, reject $H_{0,\{j\}}^{\text{row}} : Y \perp\!\!\!\perp X \mid X_{-j,\cdot}$ if $p_j^{\text{row}} < \alpha$.

The permutation testing algorithm is very easy to implement, as it only involves calculating the trace differences between $\hat{M}$ and $\hat{M}_{-j,\cdot}$ based on the original sample and the permuted samples. Using the permutation test for model-free variable selection was first studied in Dong et al. (2016). In the case of vector-valued predictors, Dong et al. (2016) proved the equivalence between the permutation test and the asymptotic test. One can follow similar steps to prove the equivalence in our setting of matrix-valued predictors, and the details are omitted. The effectiveness of the permutation test and its comparison to the asymptotic test are examined through simulation studies in Section 5.

3.4 Maximum Ratio Criterion to Recover the Active Row Set

To facilitate recovering of the active row set $\mathcal{A}$, we propose a novel maximum ratio criterion (MRC). Recall that $\hat{\delta}_j^{\text{row}}$, $j = 1, \cdots, p$, is the sample level test

statistics. Let $\hat{\delta}^{\text{row}}_{(1)} > \hat{\delta}^{\text{row}}_{(2)} > \cdots > \hat{\delta}^{\text{row}}_{(p)}$ be the ordered test statistics for the p rows. We consider the ratio of consecutively ranked statistics $\hat{\delta}^{\text{row}}_{(j)} / \hat{\delta}^{\text{row}}_{(j+1)}$. Due to the different orders of magnitude for $\hat{\delta}^{\text{row}}_{j}$ between $j \in \mathcal{A}$ and $j \notin \mathcal{A}$, we expect that, for large sample size n, the top-ranked test statistics belong to the active row predictors, while the bottom-ranked statistics correspond to the inactive row predictors. As n goes to infinity, the ratio $\hat{\delta}^{\text{row}}_{(j)} / \hat{\delta}^{\text{row}}_{(j+1)}$ is expected to be maximized when $\hat{\delta}^{\text{row}}_{(j)}$ corresponds to the active predictor with the smallest test statistics, and $\hat{\delta}^{\text{row}}_{(j+1)}$ corresponds to the inactive predictor with the largest test statistics.

Formally, we assume the cardinality of the active set is $|\mathcal{A}| = c_a$ with $c_a < p$. Let u_j be the subscript of the original test statistic such that it matches the jth order statistic, or $\hat{\delta}^{\text{row}}_{u_j} = \hat{\delta}^{\text{row}}_{(j)}$. Then, we estimate the active set $\mathcal{A}$ by

$$\mathcal{A} = \{u_1, u_2, \cdots, u_{\hat{c}_a}\}, \text{ where } \hat{c}_a = \operatorname{argmax}_{j=1,\cdots,p-1} \left\{ \hat{\delta}^{\text{row}}_{(j)} / \hat{\delta}^{\text{row}}_{(j+1)} \right\}.$$

Unlike the traditional stepwise regression procedures that require calculating p-values at each step, MRC is a non-iterative procedure that does not require p-value calculation. MRC also avoids selecting a predefined significant level α, which is required for the sequential test procedures. The consistency of $\mathcal{A}$ is shown in Theorem 2.

Theorem 2. *Suppose Y and X follow model (1) and* $X \sim N_{p,q}(0_{p\times q}, U, V)$. *Assume that the jth row of* $\Psi = U^{-1}E(XY)V^{-1}$ *is not equal to 0 for any* $j \in \mathcal{A}$. *Then, for fixed p and q,* $\Pr(\mathcal{A} = \mathcal{A}) \to 1$ *as* $n \to \infty$.

We remark that MRC is not applicable with $c_a = p$, or when all the rows are active. Before we discuss the assumption about Ψ, we state Stein's Lemma for matrix normal distribution in the next Proposition.

Proposition 3. *Let* $X \sim N_{p,q}(0_{p\times q}, U, V)$ *and all moments involved are finite. Then, for any function* $f : \mathbb{R}^{p\times q} \mapsto \mathbb{R}$*, we have* $E\{\partial f(X)/\partial X\} = U^{-1}E\{f(X)X\}V^{-1}$.

Apply Proposition 3 to $f(X) = E(Y \mid X)$, and we get

$$E\{\partial E(Y \mid X)/\partial X\} = U^{-1}E(XY)V^{-1} = \Psi. \qquad (4)$$

From (4), we know the *j*th row of Ψ is $E\{\partial E(Y \mid X)/\partial X_{j,\cdot}\}$. Consider the case when there is a symmetric link function between Y and element in the *j*th row of *X*. For example, let $E(Y \mid X) = X_{j,1}^2$. It can be shown that $\delta_j^{\text{row}} = 0$ even though $H_{a,\{j\}}^{\text{row}}$ is true. MRC will no longer work in this case as it is likely to claim $j \notin \mathcal{A}$ while the opposite is true. On the other hand, the *j*th row of Ψ becomes 0 from (4). By making the assumption that the *j*th row of Ψ is nonzero for any $j \in \mathcal{A}$, we are excluding the pathological case with symmetric link functions to guarantee the selection consistency of MRC.

3.5 Column Hypotheses

Parallel to the row hypotheses in (2), we now define the column hypotheses. For $k = 1, \cdots, q$, denote $X_{\cdot,-k} \in \mathbb{R}^{p\times(q-1)}$ as the submatrix without the *k*th column of *X*. We consider the following column hypotheses

$$H_{0,\{k\}}^{\text{col}} : Y \perp\!\!\!\perp X \mid X_{\cdot,-k} \text{ v.s. } H_{a,\{k\}}^{col} : \text{Y is not independent of } X \text{ given } X_{\cdot,-k}.$$

Under $H_{0,\{k\}}^{\text{col}}$, we see that *Y* depends on *X* only through $X_{\cdot,-k}$, which means the *k*th column $X_{\cdot,k}$ is not important. The asymptotic test, the permutation test, and the MRC for the column hypotheses are provided in the Appendix.

4 Hypotheses Test for the Contribution of a Submatrix

Recall from Section 2 that $\mathcal{I}_{\text{row}} = \{1, \cdots, p\}$ denotes the full row index set and $\mathcal{I}_{\text{col}} = \{1, \cdots, q\}$ denotes the full column index set. Let $X_{\mathcal{D},\mathcal{F}}$ be the submatrix of

X indexed by row subset $\mathcal{D} \subseteq \mathcal{I}_{\text{row}}$ and column subset $\mathcal{F} \subseteq \mathcal{I}_{\text{col}}$. Consider the following block hypotheses

$$H_{0,\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}} : Y \perp\!\!\!\perp X \mid X_{\mathcal{D},\mathcal{F}} \text{ v.s. } H_{\text{a},\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}} : Y \text{ is not independent of } X \text{ given } X_{\mathcal{D},\mathcal{F}}.$$

Under the null $H_{0,\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}}$, Y depends on X only through $X_{\mathcal{D},\mathcal{F}}$. Under the null row hypotheses $H_{0,\{j\}}^{\text{row}}$ and the null column hypotheses $H_{0,\{k\}}^{\text{col}}$ defined in Section 3, the response Y depends on X only through $X_{-j,\cdot}$ and $X_{\cdot,-k}$, respectively. Note that $\mathcal{D}$ correspond to $-j$ and $\mathcal{F}$ correspond to $-k$. To be consistent with the notation $H_{0,\{j\}}^{\text{row}}$ and $H_{0,\{k\}}^{\text{col}}$, we use $H_{0,\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}}$ in (6).

Let $U_{\mathcal{D},\mathcal{D}}$ be the submatrix of U indexed by $\mathcal{D}$. Define $V_{\mathcal{F},\mathcal{F}}$ as the submatrix of V and $M_{\mathcal{D},\mathcal{F}} = U_{\mathcal{D},\mathcal{D}}^{-1} E(X_{\mathcal{D},\mathcal{F}} Y) V_{\mathcal{F},\mathcal{F}}^{-1} E^{\mathrm{T}}(X_{\mathcal{D},\mathcal{F}} Y)$. Then, we define $\delta_{\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}} = \operatorname{tr}(M) - \operatorname{tr}(M_{\mathcal{D},\mathcal{F}})$. Similar to Proposition 1, we have the following results

Proposition 4. *Suppose* $X \sim N_{p,q}(0_{p\times q}, U, V)$. *Then* $\delta_{\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}} = 0$ *under* $H_{0,\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}}$.

It should be noted that the explicit formula to calculate $\delta_{\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}}$ can be derived following similar steps in the proof of Proposition 1. The asymptotic distribution of $\hat{\delta}_{\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}}$ can also be derived following Theorem 1. However, the derivation is tedious and the result is difficult to use in practice. Instead of the asymptotic test, we advocate using the permutation test for the block hypotheses (6). As argued by Dong et al. (2016), the permutation test is an attractive alternative to the asymptotic test when the exact asymptotic distribution is not readily available or in a complicated format. We conclude this section with the sample-level block permutation (BP) testing algorithm.

BP.1 Based on the original sample $\{(X^{(i)}, Y^{(i)}) : i = 1, \ldots, n\}$, calculate $\hat{\delta}_{\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}} = \operatorname{tr}(\hat{M}) - \operatorname{tr}(\hat{M}_{\mathcal{D},\mathcal{F}})$.

BP.2 Fix $\{X^{(i)}: i=1,\ldots,n\}$. For $b=1,\ldots,B$, denote $\{Y_{[b]}^{(i)}: i=1,\ldots,n\}$ as the bth random permutation of $\{Y^{(i)}: i=1,\ldots,n\}$. Then calculate $\hat{\delta}^{\text{block}}_{\{\mathcal{D},\mathcal{F}\}^-,[b]}$ based on the permuted sample $\{(X^{(i)}, Y_{[b]}^{(i)}): i=1,\ldots,n\}$.

BP.3 Calculate the p-value $p^{\text{block}}_{\{\mathcal{D},\mathcal{F}\}^-} = B^{-1}\sum_{b=1}^{B} I\left(\hat{\delta}^{\text{block}}_{\{\mathcal{D},\mathcal{F}\}^-,[b]} > \hat{\delta}^{\text{block}}_{\{\mathcal{D},\mathcal{F}\}^-}\right)$, where $I(\cdot)$ is the indicator function. For a given significance level α, reject $H^{\text{block}}_{0,\{\mathcal{D},\mathcal{F}\}^-}: Y \perp\!\!\!\perp X \mid X_{\mathcal{D},\mathcal{F}}$ if $p^{\text{block}}_{\{\mathcal{D},\mathcal{F}\}^-} < \alpha$.

5 Simulation Studies

In this section, the finite sample performances of the proposed model-free variable selection procedures are evaluated by synthetic datasets. In Section 5.1, we compare the proposed matrix-based tests with the existing test after vectorization of the predictor matrix. In Section 5.2, we assess the performance of the proposed row testing procedures when the matrix normality assumption is violated. In Section 5.3, we evaluate the performance of the submatrix permutation testing procedure. In Section 5.4, the effectiveness of the MRC for active set recovery is investigated. In Section 5.5, an additional model with binary responses and larger predictors dimensions is considered. All simulations are repeated 1000 times.

5.1 Tests for Row Hypotheses

We consider the following two models

$$\text{I}: Y = 3\sin(X_{1,1}) + 3\sin(X_{2,2}) + X_{1,2} + X_{2,1} + \epsilon,$$
$$\text{II}: Y = \text{sgn}(X_{1,1} + X_{p,q})\exp(0.2X_{1,q} + 0.2X_{p,1}) + \epsilon,$$

where $X \in \mathbb{R}^{p\times q} \sim N_{p,q}(0,U,V)$, and ϵ, independent of X, is generated from the standard normal distribution. The elements in the jth row and kth column of U and V are $0.5^{|j-k|}$ for $j,k=1,\ldots,10$. We set predictor dimension to be $p=q=10$ and fix sample size as n = 200.

For each simulated dataset, three methods are applied to test the row hypotheses: the proposed asymptotic test, the proposed permutation test, and the trace pursuit sliced inverse regression-based test (TP-SIR) of Yu et al. (2016a). TP-SIR is originally designed for testing important elements in vector-valued predictors. To test $H_{0,\{j\}}^{\text{row}}: Y \perp\!\!\!\perp X \mid X_{-j,\cdot}$ through TP-SIR, we first vectorize X and then perform TP-SIR to test the importance of $X_{j,1},\ldots,X_{j,q}$ one by one. If at least one of these q tests is rejected, then we reject $H_{0,\{j\}}^{\text{row}}$. For fair comparison, Bonferroni correction is applied for TP-SIR. The frequencies of rejecting the null hypothesis for $j \in \mathcal{A}^c$ are the estimated Type-I error rates, while the frequencies of rejecting the null for $j \in \mathcal{A}$ are the estimated powers, which are boldfaced for easy reference. In particular, for Model —, $\mathcal{A} = \{1,2\}$ and $\mathcal{A}^c = \{3,4,5,6,7,8,9,10\}$; for Model II, $\mathcal{A} = \{1,10\}$ and $\mathcal{A}^c = \{2,3,4,5,6,7,8,9\}$.

Results of testing the row hypotheses are presented in Table 1. We report the frequency that $H_{0,\{j\}}^{\text{row}}$ is rejected based on 1000 repetitions. Three nominal levels $\alpha = 0.01$, 0.05, and 0.1 are considered. Overall, the asymptotic tests and the permutation tests have similar results, and they outperform TP-SIR in both models. For our proposed asymptotic test and permutation test, the estimated Type-I error rates are close to the nominal levels in both models, and the estimated powers become 1 or close to 1 when α increases. For TP-SIR, the estimated powers are much smaller, and the estimated Type-I error rates are always larger than the true nominal level.

5.2 Marginal Transformations for Non-normal Predictors

Theoretical development of the proposed testing procedures depends on the normality assumption of the predictors. This assumption may be violated in real applications. A common practice in the literature to deal with non-normal predictors is marginal transformation. See, for example, Wang et al. (2014), Mai and Zou (2015), and Dong et al. (2016). We note that the aforementioned papers all assume that the non-normal predictors become normal after the marginal predictor transformation. In this section, we consider the setting

when the original predictor is not matrix-normal, and it becomes matrix-normal after marginal predictor transformations. For $p = q = 10$, we denote the marginally transformed predictors as $W = \{\omega(X_{j,k})\}_{j,k=1}^{10}$ such that $W \sim N_{10,10}(0, U, V)$. Here $\omega : \mathbb{R} \mapsto \mathbb{R}$ is a marginal transformation of the original predictor, and the covariances U and V are the same as those in Section 5.1. We consider three cases for the marginal distribution of $X_{j,k}$, $j,k = 1,\ldots,10$. In Case 1, $X_{j,k} \sim U(-3,3)$ has the uniform distribution, and the transformation is $\omega(\cdot) = \Phi^{-1}\{F_u(\cdot)\}$, where F_u is the $U(-3,3)$ distribution function and Φ is the standard normal distribution function. In Case 2, $X_{j,k} \sim t(2)$ follows the t distribution with 2 degrees of freedom, and the corresponding transformation is $\omega(\cdot) = \Phi^{-1}\{F_t(\cdot)\}$, where F_t is the $t(2)$ distribution function. In Case 3, $X_{j,k} \sim \mathcal{C}(0,1)$ follows the standard Cauchy distribution, and the corresponding transformation is $\omega(\cdot) = \Phi^{-1}\{F_c(\cdot)\}$, where F_c is the standard Cauchy distribution function. The response is then generated from

$$\text{III}: Y = 3\sin\{\omega(X_{1,1})\} + 3\sin\{\omega(X_{2,2})\} + \omega(X_{1,2}) + \omega(X_{2,1}) + \epsilon,$$

where ϵ is standard normal and is independent of X.

Suppose we observe an i.i.d sample $\{X^{(i)}, Y^{(i)}\}, i = 1,\cdots,n$ from Model III and we want to test the row hypotheses $H_{0,\{j\}}^{\text{row}} : Y \perp\!\!\!\perp X \mid X_{-j,\cdot}$. On one hand, we carry out the proposed tests without predictor transformation. On the other hand, we apply the marginal predictor transformations on X, and then carry out the proposed tests based on the marginally transformed predictors. More specifically, let $F_{j,k}^{(n)}(x) = \sum_{\ell=1}^{n} \frac{1}{n+1} I(X_{j,k}^{(\ell)} < x)$ be the empirical distribution function of $X_{j,k}$. The transformed predictor becomes $W^{(i)} = \{W_{j,k}^{(i)}\}_{j,k=1}^{10}$, where $W_{j,k}^{(i)} = \Phi^{-1}\{F_{j,k}^{(n)}(X_{j,k}^{(i)})\}$ for $i = 1,\ldots,n$.

Based on sample size n = 200 and nominal level α = 0.05, we report the frequencies of rejecting $H_{0,\{j\}}^{\text{row}}$ in Table 2. The entries that correspond to true

$H_{0,\{j\}}^{\text{row}}$ are boldfaced. For cases 1 and 2, both the asymptotic test and the permutation test perform well with or without the predictor transformation in terms of the power. In terms of the estimated Type-I error rate, the permutation test is close to the nominal level with or without the transformation, while the asymptotic test gets closer to $\alpha = 0.05$ after the transformation. For case 3 when the elements of X have the Cauchy distribution, the permutation test again performs reasonably well with or without the transformation. The asymptotic test does not work without the transformation and improves significantly after the transformation.

5.3 Permutation Test for Submatrix Hypotheses

In this section, we evaluate the permutation test for the block hypothesis $H_{0,\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}}$ based on models I and II. The submatrix indices $\mathcal{D}$ and $\mathcal{F}$ are user-specified in practice. Let $\mathcal{F}^c$ be the complement of $\mathcal{F}$ in the full column index set $\mathcal{I}_{\text{col}}$. In our simulation, we fix $p = q = 10, \mathcal{D} = \{1, 2, 10\}$, and vary $\mathcal{F}$ such that $\mathcal{F}^c = \{k\}, k = 1, \cdots, 10$. The corresponding submatrix is denoted as $X_{\mathcal{D},-k}, k = 1, \cdots, 10$. For example, $X_{\mathcal{D},-1} \in \mathbb{R}^{3\times 9}$ denotes the the submatrix that contains the 1st, 2nd and the 10th rows and every column except for the 1st column of X. For Model I, $\mathcal{A} = \{1, 2\}$ and $\mathcal{B} = \{1, 2\}$. Thus, the frequencies of rejecting $Y \perp\!\!\!\perp X \mid X_{\mathcal{D},-k}$ for k = 1 and 2 correspond to the estimated powers, and the frequencies of rejecting $Y \perp\!\!\!\perp X \mid X_{\mathcal{D},-k}$ for k from 3 to 10 are the estimated Type-I error rates. For Model II, $\mathcal{A} = \{1, 10\}$ and $\mathcal{B} = \{1, 10\}$. Therefore, the frequencies of rejecting $Y \perp\!\!\!\perp X \mid X_{\mathcal{D},-k}$ for k = 1 and 10 correspond to the estimated powers, and the frequencies of rejecting $Y \perp\!\!\!\perp X \mid X_{\mathcal{D},-k}$ for k from 2 to 9 are the estimated Type-I error rates.

Consider three nominal levels $\alpha = 0.01$, 0.05, 0.1 and two sample sizes n = 200 and 500. We summary the frequencies of rejecting $H_{0,\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}}$ in Table 3. The entries that correspond to true $H_{0,\{\mathcal{D},\mathcal{F}\}^-}^{\text{block}}$ are boldfaced. The estimated Type-I error rates are close to the true nominal levels for both models. The estimated power for Model I is reasonable with n = 200, and increase to 1 with

n = 500. For Model II, the estimated power is low when sample size is 200, especially for $\alpha = 0.01$. The estimated powers become much closer to 1 when sample size increases to 500.

5.4 MRC for Active Row Set Recovery

We now examine the performance of MRC for recovering the active row set in models I and II. A stepwise algorithm proposed by Yu et al. (2016a) to recover the active predictor set with vector-valued predictors is included for comparison. To adapt the TP-SIR test for active row set recovery, we first vectorize the matrix-valued predictor, and then apply the stepwise TP-SIR (S-TP-SIR) algorithm. If at least one element from a particular row is selected by S-TP-SIR, then the estimated active row set will include this row. Let $\mathcal{A}_{(\ell)}$ be the estimated active row set in the ℓ th repetition, $\ell = 1, \cdots, 1000$. We report the frequencies of the jth row being included in $\mathcal{A}_{(\ell)}$ for j = 1, 2, 10. Note that $X_{1,\cdot}$ is active for both models, $X_{2,\cdot}$ is only active for Model I, and $X_{10,\cdot}$ is only active for Model II. Additionally, we also compute the average model size $MS = \sum_{\ell=1}^{1000} |\mathcal{A}_{(\ell)}| / 1000$, the under-fitted frequency $UF = \sum_{\ell=1}^{1000} I(\mathcal{A} \not\subseteq \mathcal{A}_{(\ell)}) / 1000$, the correctly-fitted frequency $CF = \sum_{\ell=1}^{1000} I(\mathcal{A} = \mathcal{A}_{(\ell)}) / 1000$, and the overfitted frequency $OF = \sum_{\ell=1}^{1000} I(\mathcal{A} \subset \mathcal{A}_{(\ell)}) / 1000$.

For n = 200, 500, and 1000, we summarize the results in Table 4. The entries that correspond to the active rows are boldfaced for easy reference. For both models, the performances of MRC and S-TP-SIR are somewhat comparable with n = 200. For n = 500 and 1000, while both MRC and S-TP-SIR select the active rows with frequency 1 or close to 1, the correctly-fitted frequency of MRC is much larger than that of S-TP-SIR. More importantly, the correctly-fitted frequency of MRC increases as n increases and becomes 1 (for Model I) or very close to 1 (for Model II) when n = 1000. S-TP-SIR, on the other hand, does not consistently recover the true active row set, as the correctly-fitted frequency does not approach 1 when n increases.

5.5 MRC for the Binary Response

In this section, we consider using MRC for active row recovery when the response is binary. This model has been used in Li et al. (2010). Let Y be a Bernoulli random variable with $\Pr(Y=1)=0.5$. Let the conditional distribution of X given Y be matrix-valued normal with the conditional means

$$E(X\,|\,Y=0)=0_{p\times q},\ E(X\,|\,Y=1)=\begin{pmatrix} I_2 & 0_{2\times(q-2)} \\ 0_{(p-2)\times 2} & 0_{(p-2)\times(q-2)} \end{pmatrix},$$

and the conditional variances

$$\mathrm{Var}(X_{j,k}\,|\,Y=0)=1,\ \mathrm{Var}(X_{j,k}\,|\,Y=1)=\begin{cases} 2 & \text{if } (j,k)\in\mathcal{G} \\ 1 & \text{if } (j,k)\notin\mathcal{G} \end{cases}.$$

Here $j=1,\ldots,p, k=1,\ldots,q, \mathcal{G}$ denotes the combined index sets $(\{1,2\},\{1,2\}), (j,k)\in\mathcal{G}$ means $j\in\{1,2\}$ and $k\in\{1,2\}$, and $(j,k)\notin\mathcal{G}$ means $j\notin\{1,2\}$ or $k\notin\{1,2\}$. From Example 1 of Li et al. (2010), we know the active row set is $\mathcal{A}=\{1,2\}$. We fix sample size as n = 200, and consider predictor dimensions to be $p=q=50$ or p = 200 and q = 50. Recall from Proposition 1 that the key quantity δ_j^{row} involves precision matrices U^{-1} and V^{-1}. To estimate the precision matrices with large p and q, we apply the ℓ_1 penalized method in Cai et al. (2011) and Yin and Li (2012). Although the conditional distribution of X given Y is normal, and marginal distribution of X is not matrix-normal in this example. As we have seen in Section 5.2, predictor transformation can improve the performance of the row hypotheses test in the presence of non-normal predictors. To evaluate the performance of active row set recovery, we compare MRC with or without the marginal predictor transformation introduced in Section 5.2.

Table 5 reports the simulation results for active row set recovery. We observe that MRC without predictor transformation works reasonably well, and MRC with predictor transformation can further improve the performance. In the case of p = 50, the correctly-fitted frequency improves from 0.927 to 0.982 after marginal predictor transformation. In the very challenging case of p = 200 and

q = 50, the improvement of the correctly-fitted frequency is more significant (from 0.853 to 0.970).

6 Analysis of the EEG Data

We analyze the EEG dataset that was introduced in Section 1. The dataset is obtained from the UCI machine learning repository *https://archive.ics.uci.edu/ml/datasets/EEG+Database* and is associated with a study that investigates the genetic predisposition and tendency for alcoholism. The data contains 122 subjects and they are categorized into two groups: an alcoholic group of 77 subjects and a control group of 45 subjects. Each subject was asked to complete a total of 120 trials. During each trial, subjects were exposed to a stimulus while the 64 channels of EEG signals for an epoch of one second (256 time points with 256 HZ sampling rate) were recorded. For each subject, we take the average of EEG signals across trials. The EEG signal produced after averaging across trials is called an event-related potential (ERP). The final dataset contains matrix-valued predictors $X \in \mathbb{R}^{64\times 256}$ and binary responses. This dataset has small sample size n = 122 relative to large p = 64 and q = 256, while our theory requires diverging n with fixed p and q. The asymptotic tests no longer apply in this setting, and the computation for the permutation test is prohibitive. The MRC is still applicable as it does not require the calculation of any p-values. Computing the test statistics will suffice. We use the ℓ_1 penalized method that was discussed in Section 5.5 to estimate the precision matrices U^{-1} and V^{-1}. Before applying the MRC, we apply the marginal variable transformation that was introduced in Section 5.2.

First, we apply the MRC to recover the active columns. Recall that the columns correspond to the time points, and the rows correspond to different channels. Taking the row average across the 64 channels, we get the individual row average ERP as a curve through 256 time points. For 45 individuals who are in the control group, the mean of the row average curves is referred to as the average control group ERP curve. The average alcoholic group ERP curve is obtained similarly. We plot the two group average ERP

curves in Figure 1. Along the time axis of Figure 1, the active columns selected by MRC are indicated by the short vertical lines. It is clear that the most significant differences between the two curves happen in the interval between 200ms to 500ms, which coincides with the time points selected by MRC. Furthermore, previous medical studies have shown that the amplitude of ERP for alcoholic-impacted subjects could be lower than the control between 300ms to 700ms (Desmedt, 1980).

Next, we apply MRC to recover the active rows. For all individuals within the same group, we consider their ERP values at a fixed channel at 400ms, which is within the 200ms to 500ms time interval discovered in Figure 1. After taking the average of these individual ERP values, we get the group average ERP value at a fixed channel at 400ms. Since each channel corresponds to a different location of the human brain, we can draw the group average ERP map consisting of 64 different channels. The control group's average ERP map at 400ms and the alcoholic group's average ERP map at 400ms are provided in panel (b) and panel (c) of Figure 2, respectively. Different colors correspond to different ERP values. We display the channels (filled dots) selected by MRC in panel (a) of Figure 2. The channels with the most significant color contrast between panels (b) and (c) largely coincide with the channels selected by MRC in panel (a). We further notice that the significant locations are in the parietal (back) and the central area of the brain, which is consistent with the findings in the biomedical literature (Porjesz and Begleiter, 2003).

7 Concluding Remarks

We present novel procedures for model-free variable selection with matrix-valued predictors. The row hypotheses, the column hypotheses, and the submatrix hypotheses are introduced in a unified framework. Asymptotic tests and permutation tests are proposed to approximate the null distribution of the sample test statistics. MRC is used to facilitate active row (column) set recovery. Marginal predictor transformation is considered in the presence of non-normal predictors. The following extensions are worth consideration in

the future. First, following Yu et al. (2016a), where different sufficient dimension reduction methods with vector-valued predictors are adapted for various model-free variable selection procedures, more variable selection methods with matrix-valued predictors can be motivated from existing matrix-valued sufficient dimension reduction methods. Second, the selection consistency of MRC requires fixed predictor dimension p and q. In the case of vector-valued predictors, penalized regression methods such as Li and Yin (2008) as well as ranking-based methods such as Yu et al. (2016b) and Baranowski et al. (2020) have been shown to be effective in high-dimensional settings. These methods may be applied for consistent active row (column) set recovery for high-dimensional matrix-valued predictors. Last but not least, although we have shown the effectiveness of marginal predictor transformation with non-normal predictors in simulations, the corresponding theoretical development is worth further investigation.

Appendix: Procedures to Recover the Active Columns

Parallel to the development in Section 3, we discuss the procedures to recover active columns. For $k = 1, \cdots, q$, denote $X_{\cdot,-k} \in \mathbb{R}^{p\times(q-1)}$ as the submatrix without the kth column of X. Consider column hypotheses

$$H_{0,\{k\}}^{\text{col}} : Y \perp\!\!\!\perp X \mid X_{\cdot,-k} \text{ v.s. } H_{a,\{k\}}^{\text{col}} : Y \text{ is not independent of } X \text{ given } X_{\cdot,-k}.$$

Under $H_{0,\{k\}}^{\text{col}}$, we see that Y depends on X only through $X_{\cdot,-k}$, which means the kth column $X_{\cdot,k}$ is not important. For the column covariance matrix V, let $V_{k,k} \in \mathbb{R}$ be the element in the kth row and kth column of V. Similarly, we can define $V_{-k,-k} \in \mathbb{R}^{(q-1)\times(q-1)}$, $V_{k,-k} \in \mathbb{R}^{1\times(q-1)}$, and $V_{-k,k} \in \mathbb{R}^{(q-1)\times 1}$. Further, let $V_{k,k|-k} = V_{k,k} - V_{k,-k}V_{-k,-k}^{-1}V_{-k,k}$ and $C_{k|-k} = X_{\cdot,k} - X_{\cdot,-k}V_{-k,-k}^{-1}V_{-k,k}$. Recall that $M = U^{-1}E(XY)V^{-1}E^{\mathrm{T}}(XY)$ from Section 3. Define $M_{\cdot,-k} = U^{-1}E(X_{\cdot,-k}Y)V_{-k,-k}^{-1}E^{\mathrm{T}}(X_{\cdot,-k}Y)$ and $\delta_k^{\text{col}} = \text{tr}(M) - \text{tr}(M_{\cdot,-k})$. Similar to Proposition 1, we have the following results.

Proposition 5. *Suppose* $X \sim N_{p,q}(0_{p\times q}, U, V)$. *Then*

1. $\delta_k^{\text{col}} = V_{k,k|-k}^{-1} E^{\mathrm{T}}(C_{k|-k}Y) U^{-1} E(C_{k|-k}Y)$.
2. $\delta_k^{\text{col}} = 0$ *under* $H_{0,\{k\}}^{\text{col}}$.

At the sample level, the test statistic for the column hypotheses (5) is $\hat{\delta}_k^{\text{col}} = \text{tr}(\hat{M}) - \text{tr}(\hat{M}_{\cdot,-k})$, where $\hat{M} = \hat{U}^{-1} E_n(XY) \hat{V}^{-1} E_n^{\mathrm{T}}(XY)$ and $\hat{M}_{\cdot,-k} = \hat{U}^{-1} E_n(X_{\cdot,-k}Y) \hat{V}_{-k,-k}^{-1} E_n^{\mathrm{T}}(X_{\cdot,-k}Y)$. The asymptotic distribution of $\hat{\delta}_k^{\text{col}}$ under the null is provided in Theorem 3.

Theorem 3. *Suppose* $X \sim N_{p,q}(0_{p\times q}, U, V)$ *and all the moments involved are finite. Then*

$$n\hat{\delta}_k^{\text{col}} \xrightarrow{D} V_{k,k|-k}^{-1} \sum_{\ell=1}^{p} \tau_\ell \chi_\ell^2(1)$$

under $H_{0,\{k\}}^{\text{col}}$*, where "*$\xrightarrow{D}$*" means convergence in distribution,* $\chi_\ell^2(1)$ *are independent chi-square with one degree of freedom for* $\ell = 1, \ldots, p$*,* $\tau_1 \geq \ldots \geq \tau_p$ *are the eigenvalues of* $U^{-1/2}\Gamma U^{-1/2}$*, and the exact form of* Γ *is in the supplementary.*

The asymptotic null distribution in Theorem 3 has to be approximated in practice. The details are similar to the discussions in Section 3.2 and thus omitted. The permutation test for an individual column is based on the following observation.

Proposition 6. *Suppose* $X \sim N_{p,q}(0_{p\times q}, U, V)$*. Then* $Y \perp\!\!\!\perp C_{k|-k}$ *under* $H_{0,\{k\}}^{\text{col}} : Y \perp\!\!\!\perp X \mid X_{\cdot,-k}$.

Lastly, from Proposition 6, the sample level column permutation (CP) test algorithm is outlined below:

CP.1 Based on the original sample $\{(X^{(i)}, Y^{(i)}) : i = 1, \ldots, n\}$, calculate $\hat{\delta}_k^{\text{col}} = \text{tr}(\hat{M}) - \text{tr}(\hat{M}_{\cdot,-k})$.

CP.2 Fix $\{X^{(i)}: i=1,\ldots,n\}$. For $b=1,\ldots,B$, denote $\{Y_{[b]}^{(i)}: i=1,\ldots,n\}$ as the *b*th random permutation of $\{Y^{(i)}: i=1,\ldots,n\}$. Then calculate $\hat{\delta}_{k,[b]}^{\text{col}}$ based on the permuted sample $\{(X^{(i)}, Y_{[b]}^{(i)}): i=1,\ldots,n\}$.

CP.3 Calculate the *p*-value $p_k^{\text{col}} = B^{-1}\sum_{b=1}^{B} I\left(\hat{\delta}_{k,[b]}^{\text{col}} > \hat{\delta}_k^{\text{col}}\right)$, where $I(\cdot)$ is indicator function. For a given significance level α, reject $H_{0,\{k\}}^{\text{col}}: Y \perp\!\!\!\perp X \mid X_{\cdot,-k}$ if $p_k^{\text{col}} < \alpha$.

Denote $\hat{\delta}_{(1)}^{\text{col}} > \hat{\delta}_{(2)}^{\text{col}} > \cdots > \hat{\delta}_{(q)}^{\text{col}}$ as the ordered test statistics for the q columns. Let the cardinality of the active column set be $|\mathcal{B}| = c_b$ with $c_b < q$. For $k=1,\ldots,q$, let v_k be the subscript of the original test statistic such that it matches the kth order statistic, or $\hat{\delta}_{v_k}^{\text{col}} = \hat{\delta}_{(k)}^{\text{col}}$. Then we estimate the active set $\mathcal{B}$ by MRC

$$\mathcal{B} = \{v_1, v_2, \cdots, v_{\hat{c}_b}\}, \text{ where } \hat{c}_b = \operatorname{argmax} k = 1,\cdots,q-1 \left\{\hat{\delta}_{(k)}^{\text{col}} / \hat{\delta}_{(k+1)}^{\text{col}}\right\}.$$

The consistency of $\mathcal{B}$ is provided in the next Theorem.

Theorem 4. *Suppose Y and X follow model (1), and* $X \sim N_{p,q}(0_{p\times q}, U, V)$. *Furthermore, assume for any* $k \in \mathcal{B}$, *the kth column of* $\Psi = U^{-1}E(XY)V^{-1}$ *is not equal to 0. Then, for fixed p and q,* $\Pr(\mathcal{B} = \mathcal{B}) \to 1$ *as* $n \to \infty$.

Supplementary Materials

Supplementary materials are available online, including a pdf file that includes additional simulation studies, details of the matrix normal distribution, and proofs. R code for implementing the proposed procedures are provided.

References

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2020), "Ranking-based variable selection for high-dimensional data," *Statistica Sinica*, In Press.

Cai, T., Liu, W., and Luo, X. (2011), "A constrained ℓ_1 minimization approach to sparse precision matrix estimation," *Journal of the American Statistical Association*, 106, 594–607.

Cook, D. (2004), "Testing predictor contributions in sufficient dimension reduction," *Annals of Statistics*, 32, 1062–1092.

Cook, R. D. and Weisberg, S. (1991), "Sliced inverse regression for dimension reduction: comment," *Journal of the American Statistical Association*, 86, 328–332.

De Waal, D. J. (1985), *Matrix-valued distributions*, John Wiley & Sons, Inc., pp. 485–501.

Desmedt, J. E. (1980), "P300 in serial tasks: an essential post-decision closure mechanism," *Progress in Brain Research*, 54, 682–686.

Ding, S. and Cook, R. (2014), "Dimension folding PCA and PFC for matrix–valued predictors," *Statistica Sinica*, 24, 463–523.

— (2015a), "Higher-order sliced inverse regressions," *Wiley Interdisciplinary Reviews: Computational Statistics*, 7, 249–257.

— (2015b), "Tensor sliced inverse regression," *Journal of Multivariate Analysis*, 133, 216–231.

Dong, Y., Yang, C., and Yu, Z. (2016), "On permutation tests for predictor contribution in sufficient dimension reduction," *Journal of Multivariate Analysis*, 149, 81–91.

Gupta, A. K. and Nagar, D. K. (2000), *Matrix variate distributions*, Boca Raton, FL: Chapman & Hall.

Li, B., Kim, M. K., and Altman, N. (2010), "On dimension folding of matrix- or array-valued statistical objects," *The Annals of Statistics*, 38, 1094–1121.

Li, B. and Wang, S. (2007), “On directional regression for dimension reduction,” *Journal of the American Statistical Association*, 102, 997–1008.

Li, K. C. (1991), “Sliced inverse regression for dimension reduction,” *Journal of the American Statistical Association*, 86, 316–327.

Li, K. C. and Duan, N. (1989), “Regression analysis under link violation,” *The Annals of Statistics*, 17, 1009–1052.

Li, L. and Yin, X. (2008), “Sliced Inverse Regression with Regularizations,” *Biometrics*, 64, 124–131.

Li, R., Zhong, W., and Zhu, L. (2012), “Feature screening via distance correlation learning,” *Journal of the American Statistical Association*, 107, 1129–1139.

Mai, Q. and Zou, H. (2015), “Nonparametric variable transformation in sufficient dimension reduction,” *Technometrics*, 57, 1–10.

Pfeiffer, R. M., Forzani, L., and Bura, E. (2012), “Sufficient dimension reduction for longitudinally measured predictors,” *Statistics in Medicine*, 31, 2414–2427.

Porjesz, B. and Begleiter, H. (2003), “Alcoholism and human electrophysiology,” *Alcohol Research & Health: the Journal of the National Institute on Alcohol Abuse and Alcoholism*, 27, 153–160.

Wang, T., Guo, X., Zhu, L. X., and Xu, P. (2014), “Transformed sufficient dimension reduction,” *Biometrika*, 101, 815–829.

Wang, Y. (2016), “Sufficient dimension folding, variable selection and its inference,” Ph.D. thesis, The University of Georgia.

Xue, Y. and Yin, X. (2014), “Sufficient dimension folding for regression mean function,” *Journal of Computational and Graphical Statistics*, 23, 1028–1043.

— (2015), “Sufficient dimension folding for a functional of conditional distribution of matrix- or array-valued objects,” *Journal of Nonparametric Statistics*, 27, 253–269.

Xue, Y., Yin, X., and Jiang, X. (2016), “Ensemble sufficient dimension folding methods for analyzing matrix-valued data,” *Computational Statistics and Data Analysis*, 103, 193–205.

Yin, J. and Li, H. (2012), “Model selection and estimation in the matrix normal graphical model,” *Journal of Multivariate Analysis*, 107, 119 – 140.

Yin, X. and Li, B. (2011), “Sufficient dimension reduction based on an ensemble of minimum average variance estimators,” *Annals of Statistics*, 39, 3392–3416.

Yu, Z., Dong, Y., and Li Xing, Z. (2016a), “Trace Pursuit: A general framework for model-free variable selection,” *Journal of the American Statistical Association*, 111, 813–821.

Yu, Z., Dong, Y., and Shao, J. (2016b), “On marginal sliced inverse regression for ultrahigh dimensional model-free feature selection,” *Annals of Statistics*, 44, 2594–2623.

Zhao, J. and Leng, C. (2014), “Structured LASSO for regression with matrix covariates,” *Statistica Sinica*, 24, 799–814.

Zhou, H., Li, L., and Zhu, H. (2013), “Tensor regression with applications in neuroimaging data analysis,” *Journal of American Statistics Association*, 108, 540–552.

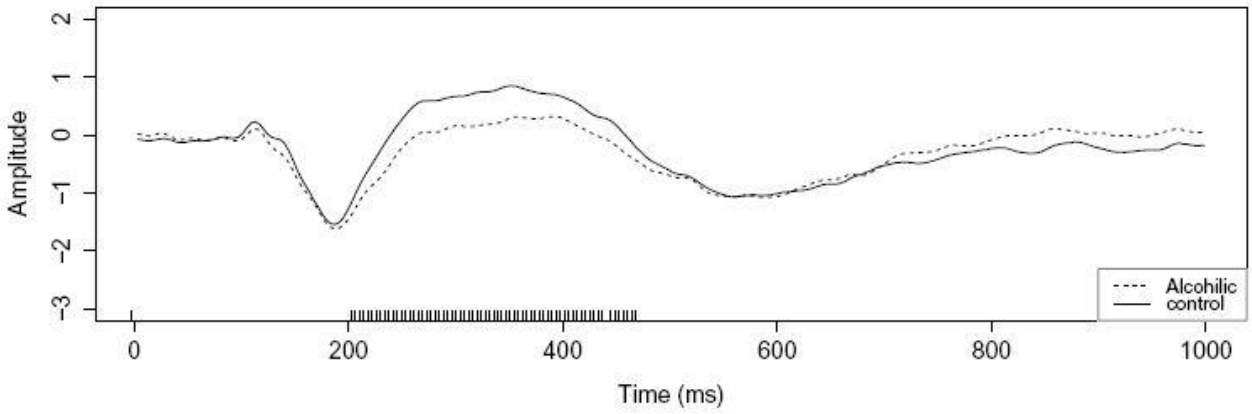

Fig. 1 Group average ERP curves, with selected active time points denoted by the vertical lines at the bottom.

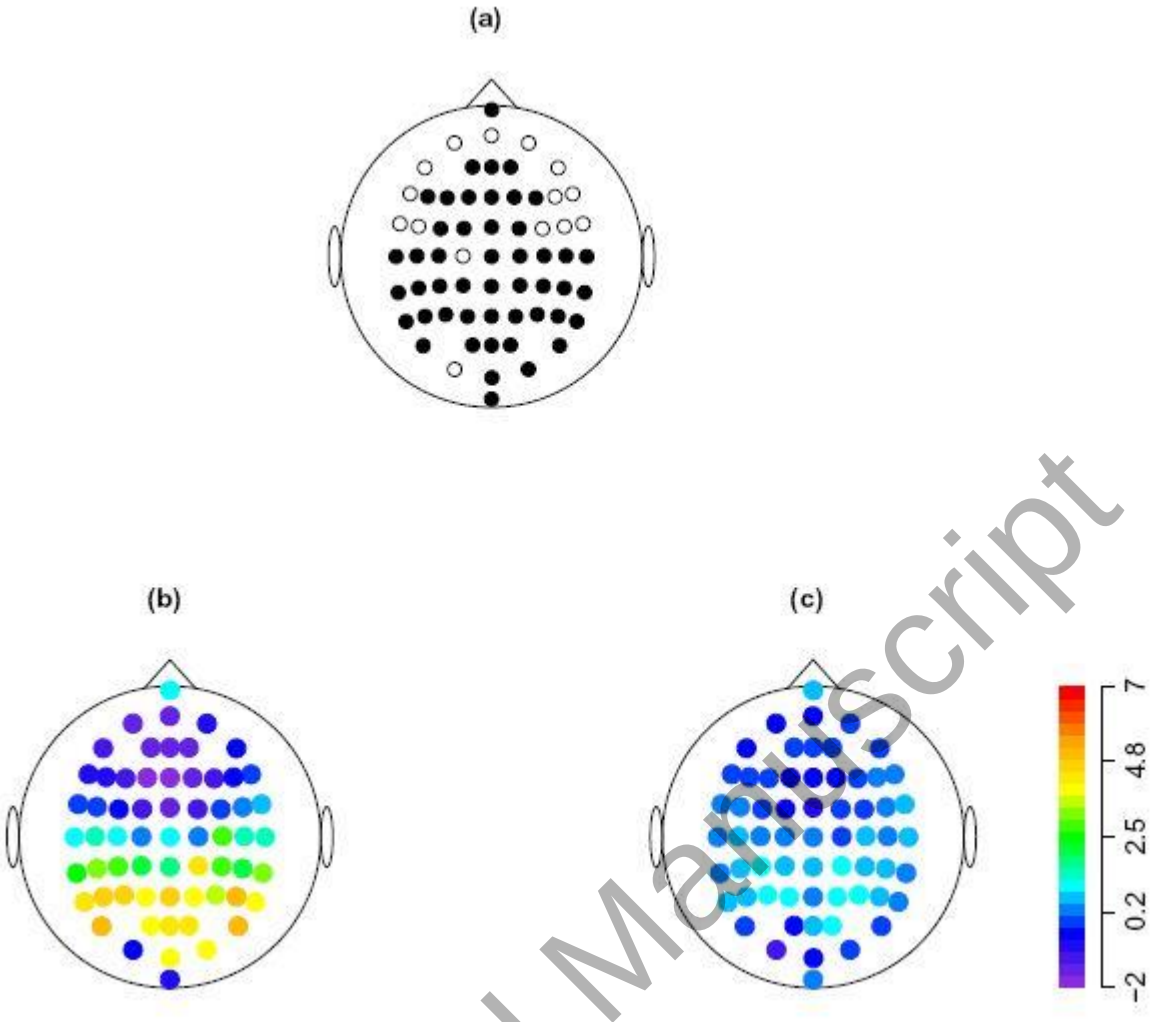

Fig. 2 (a) Channels selected by the MRC (filled dots); (b) control group average ERP map at 400ms; (c) alcoholic group average ERP map at 400ms. Panel (b) and panel (c) are color coded for different ERP values.

Table 1 Simulation results of testing row hypotheses for Model I and Model II. Based on 1000 repetitions, frequencies of rejecting $H_{0,\{j\}}^{\text{row}}$ are reported.

Model	α	Test	$X_{1,\cdot}$	$X_{2,\cdot}$	$X_{3,\cdot}$	$X_{4,\cdot}$	$X_{5,\cdot}$	$X_{6,\cdot}$	$X_{7,\cdot}$	$X_{8,\cdot}$	$X_{9,\cdot}$	$X_{10,\cdot}$
—	.01	Asymptotic	**1**	**.984**	.014	.009	.011	.007	.010	.010	.008	.006
		Permutation	**1**	**.992**	.013	.012	.006	.008	.014	.007	.011	.012
		TP-SIR	**.560**	**.441**	.052	.052	.045	.039	.049	.044	.038	.045
	.05	Asymptotic	**1**	**.998**	.051	.047	.039	.051	.051	.044	.065	.043
		Permutation	**1**	**.997**	.050	.052	.057	.048	.055	.045	.048	.042
		TP-SIR	**.877**	**.753**	.189	.173	.191	.169	.169	.182	.171	.175
	.10	Asymptotic	**1**	**.999**	.106	.103	.113	.108	.111	.097	.117	.097
		Permutation	**1**	**1**	.106	.097	.084	.086	.093	.095	.106	.099
		TP-SIR	**.952**	**.876**	.323	.323	.339	.293	.321	.322	.295	.311
II	.01	Asymptotic	**.829**	.008	.007	.005	.008	.005	.013	.006	.017	**.819**
		Permutation	**.872**	.012	.007	.015	.010	.010	.012	.012	.014	**.871**
		TP-SIR	**.491**	.050	.063	.066	.054	.061	.058	.053	.054	**.491**
	.05	Asymptotic	**.950**	.042	.043	.057	.044	.038	.054	.035	.044	**.939**
		Permutation	**.962**	.058	.045	.058	.047	.055	.058	.056	.058	**.966**
		TP-SIR	**.676**	.139	.153	.166	.142	.164	.145	.139	.153	**.713**
	.10	Asymptotic	**.980**	.104	.096	.111	.086	.095	.102	.086	.094	**.974**
		Permutation	**.979**	.107	.098	.115	.085	.100	.096	.106	.104	**.986**
		TP-SIR	**.753**	.214	.237	.236	.209	.226	.220	.209	.221	**.797**

Table 2 Simulation results of testing row hypotheses for Model III. Based on 1000 repetitions and $\alpha = 0.05$, frequencies of rejecting $H_{0,\{j\}}^{\text{row}}$ are reported.

Case	Test	Transformation	$X_{1,\cdot}$	$X_{2,\cdot}$	$X_{3,\cdot}$	$X_{4,\cdot}$	$X_{5,\cdot}$	$X_{6,\cdot}$	$X_{7,\cdot}$	$X_{8,\cdot}$	$X_{9,\cdot}$	$X_{10,\cdot}$
1	Asymptotic	No	**1**	**1**	.038	.037	.035	.023	.033	.028	.034	.036
		Yes	**1**	**1**	.043	.057	.041	.048	.054	.036	.049	.010
	Permutation	No	**1**	**1**	.057	.053	.042	.046	.047	.059	.070	.067
		Yes	**1**	**1**	.050	.049	.054	.057	.043	.037	.049	.049
2	Asymptotic	No	**.987**	**.875**	.005	.012	.009	.017	.015	.007	.016	.015
		Yes	**1**	**.999**	.046	.054	.037	.037	.049	.031	.048	.046
	Permutation	No	**1**	**.996**	.068	.062	.053	.043	.047	.049	.044	.061
		Yes	**1**	**.997**	.052	.051	.057	.052	.051	.051	.053	.047
3	Asymptotic	No	**0**	**0**	0	0	0	0	0	0	0	0
		Yes	**.950**	**.881**	.007	.015	.010	.007	.008	.011	.008	.012
	Permutation	No	**1**	**.999**	.110	.051	.052	.057	.052	.048	.051	.056
		Yes	**1**	**1**	.089	.051	.055	.065	.047	.036	.056	.054

Table 3 Simulation results of testing submatrix hypotheses for Model I and Model II. Based on 1000 repetitions, frequencies of rejecting $H_{0,\{\mathcal{D},\mathcal{F}\}^{\top}}^{\text{block}}$ are reported.

Model	n	α	$X_{\mathcal{D},-1}$	$X_{\mathcal{D},-2}$	$X_{\mathcal{D},-3}$	$X_{\mathcal{D},-4}$	$X_{\mathcal{D},-5}$	$X_{\mathcal{D},-6}$	$X_{\mathcal{D},-7}$	$X_{\mathcal{D},-8}$	$X_{\mathcal{D},-9}$	$X_{\mathcal{D},-10}$
—	200	.01	**.804**	**.629**	.011	.010	.015	.011	.012	.010	.011	.012
		.05	**.938**	**.820**	.052	.051	.050	.057	.056	.052	.051	.050
		.10	**.969**	**.896**	.095	.088	.088	.090	.108	.101	.096	.089
	500	.01	**1**	**1**	.014	.010	.012	.013	.012	.017	.015	.014
		.05	**1**	**1**	.055	.043	.056	.061	.063	.047	.058	.066
		.10	**1**	**1**	.120	.111	.111	.134	.120	.120	.098	.102
II	200	.01	**.318**	.008	.010	.010	.008	.006	.011	.004	.006	**.338**
		.05	**.561**	.048	.051	.046	.054	.049	.048	.046	.049	**.582**
		.10	**.683**	.091	.100	.095	.099	.108	.101	.095	.102	**.716**
	500	.01	**.956**	.012	.010	.013	.009	.013	.011	.011	.011	**.953**
		.05	**.995**	.055	.048	.050	.052	.050	.042	.046	.044	**.988**
		.10	**.999**	.092	.102	.102	.102	.099	.104	.099	.092	**.998**

Table 4 Simulation results of active row set recovery for Model I and Model II. Based on 1000 repetitions, frequencies of the 1st row, the 2nd row and the 10th row being selected, the average model size (MS), the underfitted frequency (UF), the correctly-fitted frequency (CF), and the overfitted frequency (OF) are reported.

Model	*n*	Method	$X_{1,\cdot}$	$X_{2,\cdot}$	$X_{10,\cdot}$	*MS*	*UF*	*CF*	*OF*
—	200	S-TP-SIR	**1**	**1**	0.072	2.464	0	0.608	0.392
		MRC	**0.984**	**0.918**	0.059	2.371	0.098	0.809	0.093
	500	S-TP-SIR	**1**	**1**	0.067	2.457	0	0.619	0.381
		MRC	**1**	**0.996**	0.002	2.013	0.004	0.992	0.004
	1000	S-TP-SIR	**1**	**1**	0.060	2.393	0	0.650	0.350
		MRC	**1**	**1**	0	2	0	1	0
II	200	S-TP-SIR	**0.964**	0.144	**0.939**	3.045	0.066	0.269	0.665
		MRC	**0.890**	0.198	**0.802**	3.455	0.208	0.448	0.344
	500	S-TP-SIR	**1**	0.111	**1**	2.739	0	0.434	0.566
		MRC	**0.993**	0.025	**0.989**	2.196	0.018	0.938	0.044
	1000	S-TP-SIR	**1**	0.112	**1**	2.813	0	0.418	0.582
		MRC	**1**	0.002	**1**	2.014	0	0.998	0.002

Table 5 Simulation results of MRC for active row set recovery with binary response. Based on 1000 repetitions, frequencies of the 1st row and the 2nd row being selected, the average model size (MS), the under-fitted frequency (UF), the correctly-fitted frequency (CF), and the overfitted frequency (OF) are reported.

Dimensions	Transformation	$X_{1,\cdot}$	$X_{2,\cdot}$	*MS*	*UF*	*CF*	*OF*
$p = 50$, $q = 50$	No	0.962	0.984	2.331	0.054	0.927	0.019
	Yes	0.992	0.992	2.030	0.016	0.982	0.002
$p = 200$, $q = 50$	No	0.955	0.943	4.626	0.102	0.853	0.045
	Yes	0.984	0.982	1.976	0.024	0.970	0.006