Statistica Sinica

Response Variable Selection in Multivariate Linear Regression

Kshitij Khare and Zhihua Su

University of Florida

Abstract: In this article, we discuss response variable selection and subsequent estimation of the regression coefficients in multivariate linear regression. Because of the asymmetric roles of the predictors and responses in regression, response variable selection is markedly different from the usual predictor variable selection. When a response is inferred to have coefficients zero, it should not be simply removed from subsequent estimation. Instead we analyze its relationship with the responses that have nonzero coefficients, which we call the dynamic responses. If it is correlated with the dynamic responses given all other responses, it should be retained to improve the estimation efficiency of the nonzero coefficients, as an ancillary statistic. Otherwise, it can be removed from further inference (leading to significant resource savings in high-dimensional settings), and we call it a static response. Therefore, we can classify the responses into three categories: the dynamic responses, the ancillary responses, and the static responses. We derive an algorithm to identify these response variables, and provide an estimator of the regression coefficients based on the selection result. Applications on synthetic and real data illustrate the efficacy of the proposed response variable selection

procedure in both low and high dimensional settings. Consistency of the variable selection procedures and asymptotic properties of the estimators are established both for the large sample setting and the high-dimensional small sample setting.

Key words and phrases: Response variable selection, High-dimensional data, Group sparsity, Oracle property.

1. Introduction

Consider the standard multivariate linear regression

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{Y} \in \mathbb{R}^r$ is the multivariate response vector, $\mathbf{X} \in \mathbb{R}^p$ contains the predictors with mean $\boldsymbol{\mu}_{\mathbf{X}}$ and positive definite covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}}$, the error vector $\boldsymbol{\varepsilon}$ has mean $\mathbf{0}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. The errors and the predictors are independent of each other. We use n to denote the sample size. In this article, we assume that $n > p$ since the main focus of this article is response variable selection. If $n < p$, we can apply any method in predictor variable selection to reduce the dimensionality of the predictors and make $p < n$. However, we do allow the number of responses r to be greater than the sample size n.

Motivation. Response variable selection is motivated by many applications where multiple outputs/responses are measured along with predic-

tors, and a classification of response variables based on their relationship with the predictors is of interest. For example, in the development of a new medicine, many clinical or hematological characteristics of a patient are measured. It is of scientific interest to identify which characteristics change after the intake of the medicine. In economics, it might be of strategic importance to find out which industrial sectors are affected by a government policy, such as imposing a tariff on an imported good like bauxite. In particular, we categorize response variables into *dynamic*, *ancillary* and *static* variables. The rigorous definitions are provided in Section 2, but we briefly discuss the intuitive underpinnings and motivation here. For dynamic response variables, the corresponding regression coefficient vector (row of $\boldsymbol{\beta}$) has at least one non-zero component. Identification of these variables is of scientific interest in various applications. Let $\mathcal{D}$ denote the set of indices of all dynamic responses, and let $\boldsymbol{\beta}_{\mathcal{D}}$ denote the regression coefficients of dynamic responses. Once the dynamic responses have been identified, one might be tempted to exclude/discard the non-dynamic response variables form the estimation process, but these variables might still carry information about $\boldsymbol{\beta}_{\mathcal{D}}$ through their correlations with the dynamic variables. Non-dynamic response variables which are correlated with the dynamic response variables (given all other response variables) are defined

as ancillary responses. Identification of ancillary responses is important as it reduces the asymptotic variance of the MLE for $\boldsymbol{\beta}_{\mathcal{D}}$ (see Proposition 1). All other non-dynamic responses are defined as static responses. Static responses carry no information about $\boldsymbol{\beta}_{\mathcal{D}}$ and can be eliminated from further analysis. The categorization of non-dynamic responses into ancillary and static responses helps researchers avoid collection of the static responses in future experiments, thereby resulting in time/resource savings.

One might argue that just including all the non-dynamic responses in the estimation of $\boldsymbol{\beta}_{\mathcal{D}}$ avoids the extra selection effort for ancillary responses while getting us the same estimation efficiency. This is fine when the number of responses r is smaller than the sample size n. However, *in high-dimensional settings*, where the number of response variables (and likely the number of non-dynamic response variables) is comparable to or larger than the sample size, inclusion of all non-dynamic responses creates several methodological and computational complications and is not advisable.

Connections with existing literature. Compared to predictor variable selection, the literature on response variable selection is surprisingly limited. The standard method is to test if the regression coefficients for each response equal to zero, adjusting for multiple testing, e.g. Benjamini and Yekutieli (2001). The response variables with zero regression coeffi-

cients are usually discarded after selection. An and Zhang (2017) uses a double group-lasso penalty to perform simultaneous selection of predictors and responses, but the responses are treated as if they were uncorrelated, i.e, the covariance structure among elements in $\mathbf{Y}$ is not used.

There is a rich body of literature which leverages generalized estimating equations (GEE) for improved estimation of regression coefficients by accounting for correlated responses in longitudinal data and repeated measurement data settings, see Lipsitz et al. (1994); Ballinger (2004); Leung et al. (2009) and the references therein. A high-dimensional adaptation of these methods in Wang et al. (2012) imposes generic sparsity in the regression coefficients through penalization. There is also a growing body of literature for joint sparse estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}^{-1}$, see Peng et al. (2009); Rothman et al. (2010); Yin and Li (2011); Deshpande et al. (2019); Ha et al. (2020); Li et al. (2021) and the references therein. To the best of our knowledge, these methods either aim for parameter reduction through imposition of general sparsity patterns in $\boldsymbol{\beta}$ and/or $\boldsymbol{\Omega}$, or for selection of "master" predictor variables using column sparsity in $\boldsymbol{\beta}$.

However, these methods do not provide tools for our key goal of identifying dynamic, ancillary and static responses using specific and structured sparsity in $\boldsymbol{\beta}$ and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ (see equation (2.6) below). *While improved*

efficiency of the regression coefficient estimates is a shared goal with this literature, a key contribution/novelty of the proposed approach is the potential scientific insights through identification of dynamic responses and the future computational and resource savings resulting from the identification of ancillary/static responses (as discussed above).

Outline of the paper. In this article, we propose a two-step procedure for response variable selection taking the covariance among the responses into account - the first step identifies the dynamic variables, and the second-step identifies the ancillary variables. We then perform the estimation of the regression coefficients based on the selection results. The paper is organized as follows. In Section 2 we formally define the three categories of response variables, and derive various technical results which support the motivations for response variable selection discussed above. In Sections 3.1-3.3, we provide details of the proposed selection procedure for the low-dimensional setting ($n \geq r$) and derive its asymptotic properties. In Sections 3.4-3.5 we consider methodology for the challenging high-dimensional setting ($n < r$) and derive the corresponding asymptotic properties. Detailed experimental validation is provided in Section 4.1 (simulated data) and Section 4.2 (real data). Proofs of the technical results, implementation details, additional simulations and future research directions are provided in the supplement.

2. Categories of response variables

In this section, we introduce three categories of responses and discuss about estimation of the coefficients $\boldsymbol{\beta}$ after selection. The three categories of the responses are defined based on the different roles they played in estimation.

A natural purpose of response variable selection is to identify the responses with nonzero coefficients, and those with zero coefficients.

Definition 1. Under the multivariate linear regression model (1.1), if a response has a regression coefficient vector with at least one non-zero component, we call it dynamic response.

Let $\mathcal{D}$ be a subset of $\{1, \ldots, r\}$ which contains the indices of all dynamic responses and let $r_{\mathcal{D}}$ be its cardinality. We use $\mathbf{Y}_{\mathcal{D}} \in \mathbb{R}^{r_{\mathcal{D}}}$ to denote the vector of dynamic response, and $\mathbf{Y}_{-\mathcal{D}} \in \mathbb{R}^{r-r_{\mathcal{D}}}$ to denote the responses whose coefficient vectors have identically zero components. Without loss of generality, $\mathbf{Y}$ can be written as $\mathbf{Y} = (\mathbf{Y}_{\mathcal{D}}^T, \mathbf{Y}_{-\mathcal{D}}^T)^T$, and the regression coefficients have corresponding partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{D}}^T, \mathbf{0})^T$. Each row in $\boldsymbol{\beta}_{\mathcal{D}}$ is nonzero. Then the linear regression model (1.1) has the structure

$$\begin{pmatrix} \mathbf{Y}_{\mathcal{D}} \\ \mathbf{Y}_{-\mathcal{D}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{D}} \\ \boldsymbol{\alpha}_{-\mathcal{D}} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{D}} \\ \mathbf{0} \end{pmatrix} \mathbf{X} + \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathcal{D}} \\ \boldsymbol{\varepsilon}_{-\mathcal{D}} \end{pmatrix}, \quad \text{var} \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathcal{D}} \\ \boldsymbol{\varepsilon}_{-\mathcal{D}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{D}} & \boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}} \\ \boldsymbol{\Sigma}_{-\mathcal{D},\mathcal{D}} & \boldsymbol{\Sigma}_{-\mathcal{D}} \end{pmatrix}. \tag{2.2}$$

Suppose that the data consists of n independent and identically distributed (IID) observations $(\mathbf{Y}_i, \mathbf{X}_i)$, where $\mathbf{Y}_i$ is sampled from the conditional distribution of $\mathbf{Y} \mid \mathbf{X}_i$, $i = 1, \cdots, n$. The following proposition (Proposition 2 in Su et al. (2016)) indicates that after selection, although $\mathbf{Y}_{-\mathcal{D}}$ has zero coefficients, it can improve the efficiency in the estimation of $\boldsymbol{\beta}_{\mathcal{D}}$ via its correlation with $\mathbf{Y}_{\mathcal{D}}$. Let $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$ and $\widetilde{\boldsymbol{\beta}}_{-\mathcal{D}}$ be the ordinary least squares (OLS) estimators of the coefficients from the regression of $\mathbf{Y}_{\mathcal{D}}$ on $\mathbf{X}$ and $\mathbf{Y}_{-\mathcal{D}}$ on $\mathbf{X}$. Note that the OLS estimators do not account for the error correlations. It is worth noting that the multivariate regression model in (1.1) can be thought of as a special case of the seemingly unrelated regression (SUR) model (Zellner, 1962) with common predictors across all responses. In such a setting, the generalized least squares (GLS) estimate of regression coefficients is exactly the same as the OLS estimate (Amemiya, 1985, Page 197). Let $\mathbf{R}_{\mathcal{D}}$ be the residuals from the regression of $\mathbf{Y}_{\mathcal{D}}$ on $\mathbf{X}$, and $\mathbf{R}_{-\mathcal{D}}$ the residuals from the regression of $\mathbf{Y}_{-\mathcal{D}}$ on $\mathbf{X}$. The operator vec$(\cdot)$ stacks a matrix into a vector columnwise, and $\otimes$ stands for the Kronecker product.

Proposition 1. *Assume that the errors are normally distributed in model (2.2) and $\mathcal{D}$ is given. The maximum likelihood estimator of $\boldsymbol{\beta}_{\mathcal{D}}$ under model (2.2) is $\widehat{\boldsymbol{\beta}}_{\mathcal{D}} = \widetilde{\boldsymbol{\beta}}_{\mathcal{D}} - \widetilde{\boldsymbol{\beta}}_{\mathcal{D}|-\mathcal{D}}\widetilde{\boldsymbol{\beta}}_{-\mathcal{D}}$, where $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}|-\mathcal{D}}$ is the OLS estimator of the coefficients from the regression of $\mathbf{R}_{\mathcal{D}}$ on $\mathbf{R}_{-\mathcal{D}}$. The asymptotic distribution*

of $\widehat{\boldsymbol{\beta}}_{\mathcal{D}}$ *is given by*

$$\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{D}})-\mathrm{vec}(\boldsymbol{\beta}_{\mathcal{D}})\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_1), \quad \mathbf{V}_1 = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\otimes(\boldsymbol{\Sigma}_{\mathcal{D}}-\boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}}\boldsymbol{\Sigma}_{-\mathcal{D}}^{-1}\boldsymbol{\Sigma}_{-\mathcal{D},\mathcal{D}}).$$

Recall that $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$ *is the maximum likelihood estimator of* $\boldsymbol{\beta}_{\mathcal{D}}$ *under the model*

$$\mathbf{Y}_{\mathcal{D}} = \boldsymbol{\alpha}_{\mathcal{D}} + \boldsymbol{\beta}_{\mathcal{D}}\mathbf{X} + \boldsymbol{\varepsilon}_{\mathcal{D}}, \quad \mathrm{var}(\boldsymbol{\varepsilon}_{\mathcal{D}}) = \boldsymbol{\Sigma}_{\mathcal{D}}.$$

The asymptotic distribution of $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$ *is given by*

$$\sqrt{n}\{\mathrm{vec}(\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}) - \mathrm{vec}(\boldsymbol{\beta}_{\mathcal{D}})\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_2), \quad \mathbf{V}_2 = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathcal{D}}.$$

Moreover,

$$\mathbf{V}_2 - \mathbf{V}_1 = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathcal{D}}^{1/2}\boldsymbol{\rho}\boldsymbol{\Sigma}_{\mathcal{D}}^{1/2},$$

where $\boldsymbol{\rho} = \boldsymbol{\Sigma}_{\mathcal{D}}^{-1/2}\boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}}\boldsymbol{\Sigma}_{-\mathcal{D}}^{-1}\boldsymbol{\Sigma}_{-\mathcal{D},\mathcal{D}}\boldsymbol{\Sigma}_{\mathcal{D}}^{-1/2}$*, and the eigenvalues of* $\boldsymbol{\rho}$ *are squared canonical correlations between* $\mathbf{Y}_{\mathcal{D}}$ *and* $\mathbf{Y}_{-\mathcal{D}}$ *given* $\mathbf{X}$*.*

The normality assumption in Proposition 1 is just for getting explicit forms of the asymptotic variance, which facilitates the comparison. Similar results can be derived under non-normal errors, but the expression of $\mathbf{V}_1$ and $\mathbf{V}_2$ can be much more complicated. Proposition 1 suggests that $\widehat{\boldsymbol{\beta}}_{\mathcal{D}}$ is a more efficient estimator for $\boldsymbol{\beta}_{\mathcal{D}}$ than $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$, which only uses $\mathbf{Y}_{\mathcal{D}}$. The efficiency gain increases with the canonical correlation between $\mathbf{Y}_{\mathcal{D}}$ and $\mathbf{Y}_{-\mathcal{D}}$. This is an important difference between response variable selection and predictor variable selection. In predictor variable selection, if a predictor has

regression coefficients zero, we exclude it from the model, because it is more efficient than retaining it in the model. But in response variable selection, because that $\mathbf{Y}_{-\mathcal{D}}$ carries information on $\boldsymbol{\beta}_{\mathcal{D}}$ through its correlation with $\mathbf{Y}_{\mathcal{D}}$, $\mathbf{Y}_{-\mathcal{D}}$ should be used in the construction of the estimator of $\boldsymbol{\beta}_{\mathcal{D}}$ to improve efficiency. A generalization of Proposition 1 to a setting where $r_{\bar{\mathcal{D}}}$ remains fixed, but the total number of responses r is allowed to grow with n is provided in Supplemental Section S2.

In applications where $\mathbf{Y}$ is high dimensional, it is possible that $\mathbf{Y}_{-\mathcal{D}}$ is also high dimensional, and only part of $\mathbf{Y}_{-\mathcal{D}}$ carries information on $\boldsymbol{\beta}_{\mathcal{D}}$. The other part of $\mathbf{Y}_{-\mathcal{D}}$ has regression coefficients zero and does not provide information on $\boldsymbol{\beta}_{\mathcal{D}}$, we can safely eliminate them from model (2.2), and researchers do not need to take the time and efforts to measure $\mathbf{Y}_{-\mathcal{D}}$ in future experiments. To distinguish these two types of responses, we define ancillary and static responses.

Definition 2. If a response variable has zero regression coefficients, and is independent of the dynamic responses $\mathbf{Y}_{\mathcal{D}}$ given all the other response variables, we call it static response. If a response variable has zero regression coefficients, but is not independent of $\mathbf{Y}_{\mathcal{D}}$ given all the other response variables, we call it ancillary response.

Let $\mathcal{A}$ and $\mathcal{S}$ be subsets of $\{1, \ldots, r\}$ that contain the indices of all ancil-

lary and static responses respectively. Let $r_{\mathcal{A}}$ and $r_{\mathcal{S}}$ denote the cardinalities of $\mathcal{A}$ and $\mathcal{S}$. Then we have $r_{\mathcal{D}} + r_{\mathcal{A}} + r_{\mathcal{S}} = r$. Based on Definition 2, we have $\mathbf{Y}_{\mathcal{D}} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{S}} \mid (\mathbf{Y}_{\mathcal{A}}, \mathbf{X})$. Proposition 2 indicates that the static responses do not improve the estimation efficiency of $\boldsymbol{\beta}_{\mathcal{D}}$.

Proposition 2. *Assume that $\mathcal{D}$, $\mathcal{A}$ and $\mathcal{S}$ are known, and $\mathbf{Y}_{\mathcal{D}} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{S}} \mid (\mathbf{Y}_{\mathcal{A}}, \mathbf{X})$. Suppose that the errors are normally distributed in the following two models (2.3) and (2.4), where*

$$\begin{pmatrix} \mathbf{Y}_{\mathcal{D}} \\ \mathbf{Y}_{\mathcal{A}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{D}} \\ \boldsymbol{\alpha}_{\mathcal{A}} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{D}} \\ \mathbf{0} \end{pmatrix} \mathbf{X} + \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathcal{D}} \\ \boldsymbol{\varepsilon}_{\mathcal{A}} \end{pmatrix}, \quad \text{var} \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathcal{D}} \\ \boldsymbol{\varepsilon}_{\mathcal{A}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{D}} & \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{A}} \\ \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{D}} & \boldsymbol{\Sigma}_{\mathcal{A}} \end{pmatrix}, \tag{2.3}$$

and

$$\begin{pmatrix} \mathbf{Y}_{\mathcal{D}} \\ \mathbf{Y}_{\mathcal{A}} \\ \mathbf{Y}_{\mathcal{S}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{D}} \\ \boldsymbol{\alpha}_{\mathcal{A}} \\ \boldsymbol{\alpha}_{\mathcal{S}} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{D}} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \mathbf{X} + \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathcal{D}} \\ \boldsymbol{\varepsilon}_{\mathcal{A}} \\ \boldsymbol{\varepsilon}_{\mathcal{S}} \end{pmatrix}, \quad \text{var} \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathcal{D}} \\ \boldsymbol{\varepsilon}_{\mathcal{A}} \\ \boldsymbol{\varepsilon}_{\mathcal{S}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{D}} & \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{A}} & \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{S}} \\ \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{D}} & \boldsymbol{\Sigma}_{\mathcal{A}} & \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{S}} \\ \boldsymbol{\Sigma}_{\mathcal{S},\mathcal{D}} & \boldsymbol{\Sigma}_{\mathcal{S},\mathcal{A}} & \boldsymbol{\Sigma}_{\mathcal{S}} \end{pmatrix}. \tag{2.4}$$

Let $\widehat{\boldsymbol{\beta}}_{\mathcal{D},1}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{D},2}$ be the maximum likelihood estimator of $\boldsymbol{\beta}_{\mathcal{D}}$ under models (2.3) and (2.4) respectively. Then $\widehat{\boldsymbol{\beta}}_{\mathcal{D},1} = \widetilde{\boldsymbol{\beta}}_{\mathcal{D}} - \widetilde{\boldsymbol{\beta}}_{\mathcal{D}|\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{D},2} = \widetilde{\boldsymbol{\beta}}_{\mathcal{D}} - \widetilde{\boldsymbol{\beta}}_{\mathcal{D}|(\mathcal{A},\mathcal{S})}\widetilde{\boldsymbol{\beta}}_{(\mathcal{A},\mathcal{S})}$. The asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathcal{D},i}$, $i = 1, 2$, is given by

$$\sqrt{n}\{\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{D},i}) - \text{vec}(\boldsymbol{\beta}_{\mathcal{D}})\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}), \quad \mathbf{V} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes (\boldsymbol{\Sigma}_{\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}\boldsymbol{\Sigma}_{\mathcal{A},\mathcal{D}}). \tag{2.5}$$

The forms of $\widehat{\boldsymbol{\beta}}_{\mathcal{D},1}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{D},2}$ can be obtained from Proposition 1 by replacing $-\mathcal{D}$ by $\mathcal{A}$ and $(\mathcal{A}, \mathcal{S})$. Proposition 2 suggests after response variable selection, we only need to use $\mathbf{Y}_{\mathcal{D}}$ and $\mathbf{Y}_{\mathcal{A}}$ for estimation; the static responses $\mathbf{Y}_{\mathcal{S}}$ can be eliminated. Proposition 3 gives an equivalent form of $\widehat{\boldsymbol{\beta}}_{\mathcal{D},1}$. Let $\mathbf{R}_{\mathcal{D}|\mathcal{A}}$ be the residuals from the regression of $\mathbf{Y}_{\mathcal{D}}$ on $\mathbf{Y}_{\mathcal{A}}$, and $\mathbf{R}_{\mathbf{X}|\mathcal{A}}$ the residuals of $\mathbf{X}$ on $\mathbf{Y}_{\mathcal{A}}$.

Proposition 3. *Assume that the error vector $\boldsymbol{\varepsilon}$ has finite second moments in model (2.3), and $\mathcal{D}$ and $\mathcal{A}$ are known. Let $\widehat{\boldsymbol{\beta}}_{\mathcal{D},3}$ be the regression coefficients from the regression of $\mathbf{R}_{\mathcal{D}|\mathcal{A}}$ on $\mathbf{R}_{\mathbf{X}|\mathcal{A}}$, then we have $\widehat{\boldsymbol{\beta}}_{\mathcal{D},3} = \widehat{\boldsymbol{\beta}}_{\mathcal{D},1}$.*

Proposition 3 indicates that after selection, the estimator of $\boldsymbol{\beta}_{\mathcal{D}}$ can be obtained by conditioning both $\mathbf{Y}_{\mathcal{D}}$ and $\mathbf{X}$ on $\mathbf{Y}_{\mathcal{A}}$, and then estimate the regression coefficients. The responses in $\mathbf{Y}_{\mathcal{A}}$ serve as the ancillary statistic, based on which we give its name.

Proposition 4 also provides an alternative way to obtain the estimator of $\boldsymbol{\beta}_{\mathcal{D}}$ by regressing $\mathbf{Y}_{\mathcal{D}}$ on $\mathbf{X}$ and $\mathbf{Y}_{\mathcal{A}}$. This is in the same spirit of the added variable plot in Cook and Weisberg (1982).

Proposition 4. *Under model* (2.6)*, let $(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$ be the OLS estimator for $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ in the following model*

$$\mathbf{Y}_{\mathcal{D}} = \boldsymbol{\mu} + \boldsymbol{\beta}_1 \mathbf{X} + \boldsymbol{\beta}_2 \mathbf{Y}_{\mathcal{A}} + \boldsymbol{\varepsilon}^*,$$

where the error vector $\boldsymbol{\varepsilon}^*$ *has mean* $\mathbf{0}$ *and finite second moments. Then* $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_{\mathcal{D},3} = \widehat{\boldsymbol{\beta}}_{\mathcal{D},1}$.

Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ be the precision matrix of $\boldsymbol{\varepsilon}$. Based on three categories of responses, $\boldsymbol{\Omega}$ can be partitioned according to $\mathcal{D}$, $\mathcal{A}$ and $\mathcal{S}$. Since $\mathbf{Y}_{\mathcal{D}} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{S}} \mid (\mathbf{Y}_{\mathcal{A}}, \mathbf{X})$ implies $\boldsymbol{\Omega}_{\mathcal{D},\mathcal{S}} = \mathbf{0}$, model (1.1) can then be written as

$$\begin{pmatrix} \mathbf{Y}_{\mathcal{D}} \\ \mathbf{Y}_{\mathcal{A}} \\ \mathbf{Y}_{\mathcal{S}} \end{pmatrix} = \boldsymbol{\alpha} + \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{D}} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \mathbf{X} + \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathcal{D}} \\ \boldsymbol{\varepsilon}_{\mathcal{A}} \\ \boldsymbol{\varepsilon}_{\mathcal{S}} \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{\mathcal{D}} & \boldsymbol{\Omega}_{\mathcal{D},\mathcal{A}} & \mathbf{0} \\ \boldsymbol{\Omega}_{\mathcal{A},\mathcal{D}} & \boldsymbol{\Omega}_{\mathcal{A}} & \boldsymbol{\Omega}_{\mathcal{A},\mathcal{S}} \\ \mathbf{0} & \boldsymbol{\Omega}_{\mathcal{S},\mathcal{A}} & \boldsymbol{\Omega}_{\mathcal{S}} \end{pmatrix}. \tag{2.6}$$

Note that no columns in $\boldsymbol{\Omega}_{\mathcal{D},\mathcal{A}}$ is zero. From (2.6), it is easy to see the roles of the three categories of responses, i.e., the dynamic responses $\mathbf{Y}_{\mathcal{D}}$ have nonzero coefficients $\boldsymbol{\beta}_{\mathcal{D}}$, the ancillary responses $\mathbf{Y}_{\mathcal{A}}$ have zero coefficients but improve the efficiency in the estimation of $\boldsymbol{\beta}_{\mathcal{D}}$, and the static responses $\mathbf{Y}_{\mathcal{S}}$ have zero coefficients and do not provide information for the estimation of $\boldsymbol{\beta}_{\mathcal{D}}$. The selection of $\mathcal{D}$, $\mathcal{A}$ and $\mathcal{S}$ is based on the structure of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ in (2.6) and will be discussed in Section 3. Before we proceed, we first introduce a property of model (2.6), which will be used to select $\mathcal{A}$ and $\mathcal{S}$.

Proposition 5. *Assume that the error vector* $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_{\mathcal{D}}^T, \boldsymbol{\varepsilon}_{\mathcal{A}}^T, \boldsymbol{\varepsilon}_{\mathcal{S}}^T)^T$ *has finite second moments and has covariance structure as in (2.6). Then the regression coefficients* $\mathbf{B}_{\mathcal{D}|(\mathcal{A},\mathcal{S})} = (\mathbf{B}_{\mathcal{D}|\mathcal{A}}, \mathbf{B}_{\mathcal{D}|\mathcal{S}})$ *of the following regression model*

$$\boldsymbol{\varepsilon}_{\mathcal{D}} = \mathbf{B}_{\mathcal{D}|(\mathcal{A},\mathcal{S})} \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathcal{A}} \\ \\ \boldsymbol{\varepsilon}_{\mathcal{S}} \end{pmatrix} + \mathbf{e} \tag{2.7}$$

satisfy that $\mathbf{B}_{\mathcal{D}|\mathcal{S}} = \mathbf{0}$ *and each column in* $\mathbf{B}_{\mathcal{D}|\mathcal{A}}$ *is nonzero.*

Proposition 5 implies that identification of the zero block in $\boldsymbol{\Omega}$ can be converted to another response variable selection problem in which we only need to identify the dynamic and non-dynamic responses.

3. Response Variable Selection

3.1. Construction of objective functions

We first discuss variable selection with fixed r and a large sample. Recall that the data consists of n independent and identically distributed (IID) observations $(\mathbf{Y}_i, \mathbf{X}_i)$, where $\mathbf{Y}_i$ is sampled from the conditional distribution of $\mathbf{Y} \mid \mathbf{X}_i$, $i = 1, \cdots, n$. Let $\mathbb{Y}$ denote the $n \times r$ matrix whose ith row is $\mathbf{Y}_i^T$, $\mathbb{X}$ denote the $n \times p$ matrix whose ith row is $\mathbf{X}_i^T$, $\mathbf{1}_n$ be an n dimension column vector of 1's and tr denote the trace of a matrix. The log likelihood of $\mathbf{Y}_i \mid \mathbf{X}_i$, $i = 1, \ldots, n$, is given by

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Omega}) = -\frac{nr}{2}\log(2\pi) + \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\operatorname{tr}\big\{(\mathbb{Y} - \mathbf{1}_n\boldsymbol{\alpha}^T - \mathbb{X}\boldsymbol{\beta}^T)\boldsymbol{\Omega}(\mathbb{Y} - \mathbf{1}_n\boldsymbol{\alpha}^T - \mathbb{X}\boldsymbol{\beta}^T)^T\big\}.$$

After some straightforward calculations, $\boldsymbol{\alpha}$ is estimated as $\widehat{\boldsymbol{\alpha}} = \bar{\mathbf{Y}} - \boldsymbol{\beta}\bar{\mathbf{X}}$, where $\bar{\mathbf{Y}} = \sum_{i=1}^n \mathbf{Y}_i/n$ and $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$ are the sample means of $\mathbf{Y}$ and

$\mathbf{X}$. Substituting $\widehat{\boldsymbol{\alpha}}$ to the log likelihood $l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Omega})$, we obtain the objective function for $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$

$$f(\boldsymbol{\beta}, \boldsymbol{\Omega}) = -\log|\boldsymbol{\Omega}| + \frac{1}{n}\operatorname{tr}\left\{(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\beta}^T)\boldsymbol{\Omega}(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\beta}^T)^T\right\}, \tag{3.8}$$

where $\mathbb{Y}_c \in \mathbb{R}^{n\times r}$ and $\mathbb{X}_c \in \mathbb{R}^{n\times p}$ are centered data matrices, i.e., the ith row of $\mathbb{Y}_c$ is $(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$ and ith row of $\mathbb{X}_c$ is $(\mathbf{X}_i - \bar{\mathbf{X}})^T$. Based on the objective function (3.8), the sets $\mathcal{D}$, $\mathcal{A}$ and $\mathcal{S}$ can be estimated in two steps.

Step 1. The goal of this step is to estimate $\mathcal{D}$. For this purpose, we need to induce row-wise sparsity in the matrix $\boldsymbol{\beta}$, and the group lasso penalty (Yuan and Lin, 2006) is a natural choice. According to Wang and Leng (2008) and Nardi and Rinaldo (2008), if we have an identical penalty parameter λ for each group, the estimator may lack selection consistency and estimation efficiency. Therefore we add a weight w_i to make the penalty in each group proportional to $1/\|\widehat{\boldsymbol{\beta}}_{i\cdot}\|^\gamma$ for $\gamma > 0$, where $\widehat{\boldsymbol{\beta}}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ and $\|\cdot\|$ is the Euclidean norm. This adaptive approach is also used in adaptive lasso (Zou, 2006), sparse reduced-rank regression (Chen and Huang, 2012) and sparse sufficient dimension reduction (Chen et al., 2010). To be more specific, we solve the following optimization problem

$$\begin{aligned} f_1(\boldsymbol{\beta}) = &\log|\mathbf{S}_{\mathbf{Y}|\mathbf{X}}| + \frac{1}{n}\operatorname{tr}\left\{(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\beta}^T)\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\beta}^T)^T\right\} \\ &+ \lambda_1\sum_{i=1}^{r} w_i\|\boldsymbol{\beta}_{i\cdot}\|, \end{aligned} \tag{3.9}$$

where $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ is the sample covariance matrix of the residuals from the OLS fit of $\mathbf{Y}$ on $\mathbf{X}$, $\boldsymbol{\beta}_{i\cdot}$ denotes the ith row of $\boldsymbol{\beta}$, $w_i = 1/\|\widetilde{\boldsymbol{\beta}}_{i\cdot}\|^{\gamma_1}$, where $\widetilde{\boldsymbol{\beta}}$ is the OLS estimator of $\boldsymbol{\beta}$, γ_1 and λ_1 are tuning parameters. Note that the group lasso penalty $\lambda_1 \sum_{i=1}^{r} w_i \|\boldsymbol{\beta}_{i\cdot}\|$ induces row-wise sparsity in $\boldsymbol{\beta}$ as desired. Suppose we obtain $\widehat{\boldsymbol{\beta}}_{\text{step1}}$ as a minimizer of $f_1(\boldsymbol{\beta})$. Then, we set $\widehat{\mathcal{D}} = \{j : (\widehat{\boldsymbol{\beta}}_{\text{step1}})_{j\cdot} \neq \mathbf{0}\}$. The responses that have at least one nonzero regression coefficient are in $\mathbf{Y}_{\widehat{\mathcal{D}}}$, and $r_{\widehat{\mathcal{D}}}$ is the cardinality of $\widehat{\mathcal{D}}$. The response variables that have all zero regression coefficients are either $\mathbf{Y}_{\widehat{\mathcal{A}}}$ or $\mathbf{Y}_{\widehat{\mathcal{S}}}$, which will be decided by Step 2.

Step 2. The goal of this step is to estimate $\mathcal{A}$ and $\mathcal{S}$. Proposition 5 indicates that a difference between the ancillary and static response is whether the corresponding column in $\mathbf{B}_{\mathcal{D}|(\mathcal{A},\mathcal{S})}$ is zero. Let $\mathbf{R} = \mathbb{Y}_c - \mathbb{X}_c \widehat{\boldsymbol{\beta}}_{\text{step1}}^T$ denote the residuals from Step 1. According to the estimated $\widehat{\mathcal{D}}$ from Step 1, $\mathbf{R}$ is partitioned as $\mathbf{R} = (\mathbf{R}_{\widehat{\mathcal{D}}}, \mathbf{R}_{-\widehat{\mathcal{D}}})$. We regress $\mathbf{R}_{\widehat{\mathcal{D}}}$ on $\mathbf{R}_{-\widehat{\mathcal{D}}}$ and use the group lasso penalty to induce column-wise sparsity in $\mathbf{B}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}$, leading to the following objective function

$$\begin{aligned} f_2(\mathbf{B}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}) &= \log|\mathbf{S}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}| + \frac{1}{n}\operatorname{tr}\left\{\left(\mathbf{R}_{\widehat{\mathcal{D}}} - \mathbf{R}_{-\widehat{\mathcal{D}}}\mathbf{B}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}^T\right)\mathbf{S}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}^{-1}\left(\mathbf{R}_{\widehat{\mathcal{D}}} - \mathbf{R}_{-\widehat{\mathcal{D}}}\mathbf{B}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}^T\right)^T\right\} \\ &+ \lambda_2 \sum_{i=1}^{r-r_{\widehat{\mathcal{D}}}} \tilde{w}_i \|\mathbf{B}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S}),\cdot i}\| \end{aligned} \tag{3.10}$$

where $\mathbf{S}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}$ is the sample covariance matrix of the residuals from the re-

gression of $\mathbf{R}_{\widehat{\mathcal{D}}}$ on $\mathbf{R}_{-\widehat{\mathcal{D}}}$, the weights are $\tilde{w}_i = 1/\|\widetilde{\mathbf{B}}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S}),\cdot i}\|^{\gamma_2}$, $\widetilde{\mathbf{B}}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}$ is the OLS estimator from the regression of $\mathbf{R}_{\widehat{\mathcal{D}}}$ on $\mathbf{R}_{-\widehat{\mathcal{D}}}$, $\widetilde{\mathbf{B}}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S}),\cdot i}$ denotes the ith column of $\widetilde{\mathbf{B}}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}$, and γ_2 and λ_2 are tuning parameters. Suppose $\widehat{\mathbf{B}}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S}),\text{step2}}$ is obtained as a minimizer of $f_2(\mathbf{B}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})})$, then $\mathbf{Y}_{\widehat{\mathcal{A}}}$ contains the responses whose corresponding columns in $\widehat{\mathbf{B}}_{\mathcal{D}|(\mathcal{A},\mathcal{S}),\text{step2}}$ are nonzero, and $r_{\widehat{\mathcal{A}}}$ is the cardinality of $\widehat{\mathcal{A}}$. The static responses in $\mathbf{Y}_{\widehat{\mathcal{S}}}$ are estimated as the responses whose corresponding columns in $\widehat{\mathbf{B}}_{\mathcal{D}|(\mathcal{A},\mathcal{S}),step2}$ are zero, and $r_{\widehat{\mathcal{S}}}$ is the cardinality of $\widehat{\mathcal{S}}$.

After Step 1 and Step 2, $\boldsymbol{\beta}$ is estimated as $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}^T, \mathbf{0})^T$, where $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}} = \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}} - \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}|\widehat{\mathcal{A}}}\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}}$ as discussed in Proposition 2, where $\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$, $\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}|\widehat{\mathcal{A}}}$ and $\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}}$ are OLS estimators. In other words, $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$ is the OLS estimator that uses information from both the dynamic responses and the ancillary responses.

3.2. Computational algorithm

Algorithm for **Step 1**: We estimate $\boldsymbol{\beta}$ one row at a time. For a fixed j, $j = 1, \ldots, r$, it can be shown that minimizing f_1 with respect to $\boldsymbol{\beta}_{j\cdot}$ is equivalent to minimizing the function

$$\frac{1}{n}\left\{(\mathbf{S}^{-1}_{\mathbf{Y}|\mathbf{X}})_{jj}\left(\mathbb{Y}_{c,\cdot j} - \mathbb{X}_c\boldsymbol{\beta}_{j\cdot}^T\right)^T\left(\mathbb{Y}_{c,\cdot j} - \mathbb{X}_c\boldsymbol{\beta}_{j\cdot}^T\right) + \sum_{k\neq j} 2(\mathbf{S}^{-1}_{\mathbf{Y}|\mathbf{X}})_{jk}(\mathbb{Y}_{c,\cdot k} - \mathbb{X}_c\boldsymbol{\beta}_{k\cdot}^T)^T(\mathbb{Y}_{c,\cdot j} - \mathbb{X}_c\boldsymbol{\beta}_{j\cdot}^T)\right\}$$

$$+ \lambda_1 w_j \|\boldsymbol{\beta}_{j\cdot}\| \tag{3.11}$$

with respect to $\boldsymbol{\beta}_{j\cdot}$, where $\mathbb{Y}_{c,\cdot k}$ denotes the kth column of $\mathbb{Y}_c$. Note that the function in (3.11) is a non-differentiable convex function of $\boldsymbol{\beta}_{j\cdot}$. Minimization of such functions (quadratic form in vector plus its ℓ_2-norm) has been considered in Foygel and Drton (2010), Puig et al. (2009) and Simon et al. (2013) in the context of group lasso. In particular, Simon et al. (2013) provides a reasonably fast majorize-minimize algorithm to solve this minimization problem. This approach has been implemented in the R package *SGL*, and we use it to solve for $\boldsymbol{\beta}_{j\cdot}$ in (3.11).

Algorithm for **Step 2**: The optimization problem in Step 2 is the same as that in Simon et al. (2013), with $\mathbf{S}^{-1/2}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}\mathbf{R}_{\widehat{\mathcal{D}}}$ being their $\mathbf{Y}$ and $\mathbf{S}^{-1/2}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}\mathbf{B}_{\mathcal{D}|(\widehat{\mathcal{A}},\widehat{\mathcal{S}})}$ being the coefficients. Note that a column in $\mathbf{S}^{-1/2}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}\mathbf{B}_{\mathcal{D}|(\widehat{\mathcal{A}},\widehat{\mathcal{S}})}$ is zero if and only if the corresponding column in $\mathbf{B}_{\mathcal{D}|(\widehat{\mathcal{A}},\widehat{\mathcal{S}})}$ is zero.

Remark 1. Simon et al. (2013) studies the "multi-response group-lasso" problem, and provides an iterative algorithm for minimizing the objective function

$$\frac{1}{n}\operatorname{tr}\left\{\left(\mathbb{Y}_c-\mathbb{X}_c\boldsymbol{\beta}^T\right)^T\left(\mathbb{Y}_c-\mathbb{X}_c\boldsymbol{\beta}^T\right)\right\}+\lambda_1\sum_{k=1}^{r}\left\|(\boldsymbol{\beta}^T)_{k\cdot}\right\|, \tag{3.12}$$

where $(\boldsymbol{\beta}^T)_{k\cdot}$ is the kth row of $\boldsymbol{\beta}^T$. See also Argyriou et al. (2007) and Obozinski et al. (2007). However, this iterative algorithm presented in Simon et al. (2013) is not applicable in the context of (3.9). There are

two notable differences between the minimization problems in (3.9) and (3.12). Firstly, in (3.9), we use the group-lasso penalty on the rows of $\boldsymbol{\beta}$ with the purpose of response variable selection, whereas in (3.12), a group-lasso penalty is used for the columns of $\boldsymbol{\beta}$ for the purpose of predictor variable selection. Secondly, unlike (3.12), the trace term in (3.9) contains the term $\boldsymbol{\Omega}$, since we consider a multi-response regression model with a general covariance structure.

3.3. Theoretical Properties

In this section, we establish variable selection consistency and oracle property of the estimator $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$ in the fixed r setting. Let $\bar{\mathcal{D}}$, $\bar{\mathcal{A}}$ and $\bar{\mathcal{S}}$ denote the true sets of dynamic, ancillary and static responses, respectively, $\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}}$ the true regression coefficients of dynamic responses, and $\bar{\boldsymbol{\Sigma}}$ the true error covariance matrix. Let $\bar{P}$ denote the probability measure corresponding to the true data generating model ((2.6) with the true parameters introduced above). For consistency in the fixed r setting, normality of the true error distribution is not needed, and we will only assume that the errors are IID and have finite fourth moments under $\bar{P}$.

Theorem 1. *Suppose* $n^{1/2}\lambda_i \to 0$ *and* $n^{(1+\gamma_i)/2}\lambda_i \to \infty$ *for* $i = 1, 2$. *Then*

1. *(Dynamic response selection consistency)* $\bar{P}(\widehat{\mathcal{D}} = \bar{\mathcal{D}}) \to 1$ *as* $n \to \infty$.

2. *(Ancillary response selection consistency)* $\bar{P}(\widehat{\mathcal{A}} = \bar{\mathcal{A}}) \to 1$ *as* $n \to \infty$.

3. *(Estimation consistency)* $\|\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}) - \mathrm{vec}(\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}})\| = O_{\bar{P}}(n^{-1/2})$.

Theorem 1 indicates that the estimator $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$ is $\sqrt{n}$-consistent, and our variable selection procedure discussed in §3.1 is consistent.

To discuss the optimal estimation rate, we need to first introduce the oracle model for response variable selection. If we know about the oracle information on which responses are dynamic, ancillary and static, the oracle model is model (2.3). Note that the oracle model includes the dynamic and ancillary responses, but not the static responses. The oracle estimator of $\boldsymbol{\beta}_{\bar{\mathcal{D}}}$ is $\widehat{\boldsymbol{\beta}}_{\bar{\mathcal{D}},\mathrm{oracle}} = \widetilde{\boldsymbol{\beta}}_{\bar{\mathcal{D}}} - \widetilde{\boldsymbol{\beta}}_{\bar{\mathcal{D}}|\bar{\mathcal{A}}}\widetilde{\boldsymbol{\beta}}_{\bar{\mathcal{A}}}$. The asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\bar{\mathcal{D}},\mathrm{oracle}}$ is the same as that of $\widehat{\boldsymbol{\beta}}_{\bar{\mathcal{D}},1}$ in Proposition 2; see (2.5). Note that while $\bar{P}(\widehat{\mathcal{D}} = \bar{\mathcal{D}}) \to 1$, $\widehat{\mathcal{D}}$ and $\bar{\mathcal{D}}$ may differ at some sample points. Hence, we define $\|\mathbf{u} - \mathbf{v}\| := \sqrt{\sum_{i=1}^{a}(u_i - v_i)^2 + \sum_{i=a+1}^{b} v_i^2}$ if $\mathbf{u} \in \mathbb{R}^a, \mathbf{v} \in \mathbb{R}^b$ with $a < b$ for the following result.

Theorem 2. *Assume that the conditions in Theorem 1 hold, then* $\|\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}) - \mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\bar{\mathcal{D}},\mathrm{oracle}})\| = o_{\bar{P}}(n^{-1/2})$.

Theorem 2 suggests that the estimator $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$ has the same convergence rate and asymptotic variance as the oracle estimator. Thus it has the oracle property.

3.4. Response variable selection in high dimensional setting

In the high dimensional setting, we allow r to grow with n, and denote r as r_n. In this section, we discuss adjustments to the selection algorithm under this setting.

Note that $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}$ in **Step 1** is singular when $n < r_n$. Hence, an estimator of $\mathbf{\Sigma}^{-1}_{\mathbf{Y}|\mathbf{X}}$ is needed for the objective function in (3.9). There are a number of precision matrix estimators that adapt to the high-dimensional setting, including constrained l_1-minimization estimator (Cai et al., 2011, CLIME), lasso penalized D-trace estimator (Zhang and Zou, 2014), scaled lasso estimator (Sun and Zhang, 2013), and convex correlation selection estimator (Khare et al., 2015, CONCORD). We adopt the CONCORD estimator since it computes fast and recovers the sparsity pattern with high accuracy. Let ω_{ij} denote the (i,j)th element of $\mathbf{\Omega}$, and let $\mathbf{R}_{i\cdot}$ denote the ith column of the residual matrix $\mathbf{R} \in \mathbb{R}^{n\times r}$ from the OLS regression of $\mathbf{Y}$ on $\mathbf{X}$. Then the CONCORD estimator of $\mathbf{\Omega}$, denoted by $\widehat{\mathbf{\Omega}}$, is minimizer of the objective function

$$Q_{\text{con}}(\mathbf{\Omega}) = -\sum_{i=1}^{r} n \log \omega_{ii} + \frac{1}{2}\sum_{i=1}^{r} \|\omega_{ii}\mathbf{R}_{i\cdot} + \sum_{j\neq i} \omega_{ij}\mathbf{R}_{j\cdot}\|^2 + \lambda \sum_{1\leq i\neq j\leq r} |\Omega_{ij}| \tag{3.13}$$

over the space of positive definite matrices, for an appropriately chosen penalty parameter λ. The CONCORD estimator is implemented in the

R package *gconcord*. Then we place $\mathbf{S}^{-1}_{\mathbf{Y}|\mathbf{X}}$ by $\widehat{\mathbf{\Omega}}$ in (3.9), and obtain the objective function

$$\tilde{f}_1(\boldsymbol{\beta}) = -\log|\widehat{\mathbf{\Omega}}| + \frac{1}{n}\operatorname{tr}\{(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\beta}^T)\widehat{\mathbf{\Omega}}(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\beta}^T)^T\} + \lambda_1\sum_{i=1}^{r_n} w_i\|\boldsymbol{\beta}_{i\cdot}\|, \tag{3.14}$$

then we follow the same algorithm for **Step 1** in Section 3.2 to estimate $\mathcal{D}$.

In **Step 2**, the matrix $\mathbf{S}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}$ in (3.10) is singular if $n < r_n - r_{\widehat{\mathcal{D}}}$. Moreover, because $\mathbf{S}^{-1}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}$ does not exist, the OLS estimator $\widetilde{\mathbf{B}}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}$ in the weights $\tilde{w}_i$ does not exist either. To resolve the issues, we also turn to the CONCORD estimator $\widehat{\mathbf{\Omega}}$. Since $\mathbf{S}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}$ estimates $\mathbf{\Omega}^{-1}_{\mathcal{D}}$, we use the corresponding block in $\widehat{\mathbf{\Omega}}$, i.e. $\widehat{\mathbf{\Omega}}_{\mathcal{D}}$, to replace $\mathbf{S}^{-1}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}$ in (3.10). Note that $\mathbf{B}_{\mathcal{D}|(\mathcal{A},\mathcal{S})} = -\mathbf{\Omega}^{-1}_{\mathcal{D}}(\mathbf{\Omega}_{\mathcal{D},\mathcal{A}}, \mathbf{\Omega}_{\mathcal{D},\mathcal{S}})$ (see the proof of Proposition 5), and we initialize $\mathbf{B}_{\mathcal{D}|(\mathcal{A},\mathcal{S})}$ by $-\widehat{\mathbf{\Omega}}^{-1}_{\mathcal{D}}\widehat{\mathbf{\Omega}}_{\mathcal{D},-\mathcal{D}}$. The objective function is obtained as

$$\begin{aligned}\tilde{f}_2(\mathbf{B}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}) &= -\log|\widehat{\mathbf{\Omega}}_{\widehat{\mathcal{D}}}| + \frac{1}{n}\operatorname{tr}\left\{\left(\mathbf{R}_{\widehat{\mathcal{D}}} - \mathbf{R}_{-\widehat{\mathcal{D}}}\mathbf{B}^T_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}\right)\widehat{\mathbf{\Omega}}_{\widehat{\mathcal{D}}}\left(\mathbf{R}_{\widehat{\mathcal{D}}} - \mathbf{R}_{-\widehat{\mathcal{D}}}\mathbf{B}^T_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}\right)^T\right\} \\ &+ \lambda_2\sum_{i=1}^{r_n - r_{\widehat{\mathcal{D}}}}\tilde{w}_i\|\mathbf{B}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S}),\cdot i}\|\end{aligned} \tag{3.15}$$

Then $\mathcal{A}$ and $\mathcal{S}$ are estimated following **Step 2** in Section 3.2.

3.5. Response selection consistency in high-dimensional setting

In this section, we establish the consistency of the response variable selection procedure in Section 3.4 and the asymptotic properties of the estimator

of $\boldsymbol{\beta}_{\mathcal{D}}$ when r_n tends to infinity with n. Let $\bar{\mathcal{D}}$, $\bar{\mathcal{A}}$ and $\bar{\mathcal{S}}$ denote the true sets of dynamic, ancillary and static responses, respectively, $\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}}$ the true regression coefficients of dynamic responses, and $\bar{\boldsymbol{\Sigma}}$ denote the true covariance matrix of the errors, and $\bar{\boldsymbol{\Omega}} = \bar{\boldsymbol{\Sigma}}^{-1}$. Note that the dimensions of $\bar{\boldsymbol{\Sigma}}$ and $\bar{\boldsymbol{\Omega}}$ increase with n, but the dependence is suppressed for the simplicity of notation. Mild regularity assumptions needed to establish the following result are provided and discussed in S8 of the Supplement due to space constraints. They include sub-Gaussianity of errors, uniform boundedness of eigenvalues of $\bar{\boldsymbol{\Sigma}}$ (Assumption 1), incoherence and minimum signal size conditions for consistency of $\hat{\boldsymbol{\Omega}}$ (Assumptions 2 and 3), rate of growth of true number of dynamic and ancillary variables (Assumption 4), minimum signal size assumptions corresponding to Step 1 and Step 2 of the procedure (Assumptions 5-6), and assumptions controlling the group-specific penalty parameters in Step 1 and Step 2 (Assumptions 7-8). In particular, these assumptions allow r to increase at a faster rate (almost sub-exponentially) compared to n.

Theorem 3. *Under Assumptions 1-8 (provided in the Supplement), the following holds for every $\eta > 0$.*

1. *(Dynamic response selection consistency) Let $\widehat{\boldsymbol{\beta}}_{step1}$ denote the solution to (3.14), and $\widehat{\mathcal{D}} = \{j : \widehat{\boldsymbol{\beta}}_{step1,j\cdot} \neq \mathbf{0}\}$. Then $\widehat{\mathcal{D}} = \bar{\mathcal{D}}$ with*

probability at least $1 - 6r_n^{-\eta}$ *for large enough* n *(depending on* η*).*

2. *(Ancillary and static response selection consistency) Let* $\widehat{\mathbf{B}}$ *denote the solution to (3.15), and* $\widehat{\mathcal{A}} = \{j : \widehat{\mathbf{B}}_{\cdot j} \neq \mathbf{0}\}$*. Then, for large enough* n*,* $\widehat{\mathcal{A}} = \bar{\mathcal{A}}$ *and* $\widehat{\mathcal{S}} = \bar{\mathcal{S}}$ *with probability at least* $1 - 22r_n^{-\eta}$ *for large enough* n *(depending on* η*).*

Theorem 3 establishes the selection consistency of three categories of the response variables. As a direct consequence of the selection consistency, asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathcal{D}}$ is given in Theorem 4 (proof in Supplement).

Theorem 4. *Assume that the conditions in Theorem 3 hold, the errors are normally distributed, and* $r_{\bar{\mathcal{D}}}$ *is fixed as* n *grows. Then*

$$\sqrt{n}\{\mathrm{vec}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}) - \mathrm{vec}(\bar{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}})\} \xrightarrow{d} N(0, \mathbf{V}), \quad \mathbf{V} = \bar{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1} \otimes (\bar{\boldsymbol{\Sigma}}_{\mathcal{D}} - \bar{\boldsymbol{\Sigma}}_{\mathcal{D},\mathcal{A}} \bar{\boldsymbol{\Sigma}}_{\mathcal{A}}^{-1} \bar{\boldsymbol{\Sigma}}_{\mathcal{A},\mathcal{D}}).$$

Theorem 4 implies that $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$ also has the same asymptotic distribution as the oracle estimator $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}},\mathrm{oracle}}$ when r_n grows with n.

4. Data analysis

4.1. Simulation

This simulation focuses on the high-dimensional setting where $n < r$. We fixed $n = 50$, $p = 8$, $r_{\mathcal{D}} = 6$ and $r_{\mathcal{A}} = 2$. The response dimension r was

ranged from 200 to 1000. Elements in $\boldsymbol{\beta}_{\mathcal{D}}$ were independent $N(0, 0.5^2)$ variates, the intercept was $\boldsymbol{\alpha} = \mathbf{0}$. The covariance matrix $\boldsymbol{\Sigma}$ was generated such that the squared largest canonical correlation between $\mathbf{Y}_{\mathcal{D}}$ and $\mathbf{Y}_{\mathcal{A}}$ is about 0.9 for all r. The details on generating $\boldsymbol{\Sigma}$ is included in S13 of the Supplement. We generated $\mathbf{X}$ from $N_p(0, 0.5^2\mathbf{I}_p)$ and $N_p(0, 0.25^2\mathbf{I}_p)$ to represent different signal strengths. We also generated $\mathbf{X}$ from $N_p(0, (\mathbf{I}_p + 1_p 1_p^T)/8)$ to represent correlated predictors, where 1_p denotes a p-dimensional vector of 1. Tuning for parameters is detailed in Section S11 of the Supplement. For each setting, we simulated 200 replications, and evaluated the selection performance by true positive rates $\text{TPR}_{\mathcal{D}}$, $\text{TPR}_{\mathcal{A}}$ and $\text{TPR}_{\mathcal{S}}$ for all three categories of the responses: $\text{TPR}_{\mathcal{D}} = |\bar{\mathcal{D}} \cap \widehat{\mathcal{D}}|_c/|\bar{\mathcal{D}}|_c$, $\text{TPR}_{\mathcal{A}} = |\bar{\mathcal{A}} \cap \widehat{\mathcal{A}}|_c/|\bar{\mathcal{A}}|_c$ and $\text{TPR}_{\mathcal{S}} = |\bar{\mathcal{S}} \cap \widehat{\mathcal{S}}|_c/|\bar{\mathcal{S}}|_c$, where for a set S, $|S|_c$ denotes its cardinality. We added precision measures $\text{PPV}_{\mathcal{D}}$, $\text{PPV}_{\mathcal{A}}$ and $\text{PPV}_{\mathcal{S}}$ for sensitivity analysis, where $\text{PPV}_{\mathcal{D}} = |\bar{\mathcal{D}} \cap \widehat{\mathcal{D}}|_c/|\widehat{\mathcal{D}}|_c$, i.e. the ratio of true positive over the sum of true positive and false positive. The measures $\text{PPV}_{\mathcal{A}}$ and $\text{PPV}_{\mathcal{S}}$ are defined accordingly. We measured the efficiency gain of a randomly picked element, say β_{ij}, by the efficiency ratio R_{ij} defined as

$$R_{ij} = \frac{\text{var}(\tilde{\beta}_{ij})}{\text{var}(\hat{\beta}_{ij})}, \tag{4.16}$$

where $\text{var}(\tilde{\beta}_{ij})$ and $\text{var}(\hat{\beta}_{ij})$ are the variances of the OLS estimator $\tilde{\beta}_{ij}$ and our estimator $\hat{\beta}_{ij}$ calculated based on 200 replications. Then R_{median} is

the median of all the R_{ij} for the nonzero elements in $\boldsymbol{\beta}$. The results are in Table 1. Both the TPR and PPV measures show that the variable selection procedure can identify the dynamic and ancillary responses quite well when r is much larger than n. A weaker signal slightly reduces the efficiency gains, but does not have a large impact on the results. The correlated predictors do not seem to have any obvious negative effect on variable selection or efficiency gains. We also investigated the ratio of the MSE's. The measure R^{all}_{MSE} computes the median of $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F^2 / \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F^2$ (over the 200 replications), where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Since $\widetilde{\boldsymbol{\beta}}$ is the OLS estimator using all the responses, it is not sparse, and all the errors on the sparse and non-sparse parts of $\boldsymbol{\beta}$ accumulate. On the other hand, because of the consistency of the response selection procedure stated in Theorem 3, when $\widehat{\boldsymbol{\beta}}$ correctly identifies the zero elements in $\boldsymbol{\beta}$, the sparse part of $\boldsymbol{\beta}$ does not contribute in the MSE, except for a few false positive cases. When r is large, the sparse part of $\boldsymbol{\beta}$ is also large, which has a big contribution for $\widetilde{\boldsymbol{\beta}}$. Thus ratios R^{all}_{MSE} are very large. We also investigated $R^{\mathcal{D}}_{\text{MSE}}$, which is similar to R^{all}_{MSE}, but only focuses on the nonzero part of $\boldsymbol{\beta}$ and is defined as the median of $\|\widetilde{\boldsymbol{\beta}}_{\mathcal{D}} - \boldsymbol{\beta}_{\mathcal{D}}\|_F^2 / \|\widehat{\boldsymbol{\beta}}_{\mathcal{D}} - \boldsymbol{\beta}_{\mathcal{D}}\|_F^2$ (over the 200 replications). The ratios are still significantly greater than 1, indicating the response variable selection procedure indeed improves the

estimation performance.

Table 1: Summary of selection and estimation performance when $r >> n$

r	200	300	500	1000	200	300	500	1000	200	300	500	1000
	$\mathbf{X} \sim N_p(0, 0.5^2\mathbf{I}_p)$				$\mathbf{X} \sim N_p(0, 0.25^2\mathbf{I}_p)$				$\mathbf{X} \sim N_p(0, \frac{1}{8}(1_p 1_p^T + \mathbf{I}_p))$			
$\text{TPR}_{\mathcal{D}}$	0.998	1.000	1.000	1.000	0.949	0.994	0.998	1.000	0.992	0.996	1.000	1.000
$\text{TPR}_{\mathcal{A}}$	0.985	0.978	0.953	0.898	0.970	0.975	0.950	0.898	0.980	0.978	0.953	0.898
$\text{TPR}_{\mathcal{S}}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\text{PPV}_{\mathcal{D}}$	0.997	0.997	0.999	0.999	0.997	0.999	0.999	0.999	0.997	1.000	0.997	0.996
$\text{PPV}_{\mathcal{A}}$	0.998	1.000	1.000	0.991	0.907	0.990	0.996	0.991	0.979	0.993	1.000	0.991
$\text{PPV}_{\mathcal{S}}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
R_{median}	4.054	3.945	3.350	2.687	2.587	3.347	3.171	2.687	3.308	3.115	3.380	2.651
R^{all}_{MSE}	99.12	140.20	206.09	370.40	87.63	138.18	205.06	370.40	98.80	140.73	201.52	363.26
$R^{\mathcal{D}}_{\text{MSE}}$	4.440	3.934	3.029	2.391	3.188	3.891	2.993	2.391	4.356	3.834	3.030	2.410

An additional simulation for low dimension setting is in Section S12 of the Supplement. The simulation shows the consistency of variable selection on all three categories of the responses as well as the estimator of $\boldsymbol{\beta}$. It also demonstrates the estimator from the two-stage selection procedure is more efficient than the estimator of $\boldsymbol{\beta}_{\mathcal{D}}$ that uses the oracle information of the true dynamic responses. In other words, the efficiency gains from the ancillary responses can be more substantial to offset the cost of the selection of all three categories.

4.2. Applications

Glioblastoma multiforme (GBM) is the most aggressive type of brain cancer with a median survival time of 15 months (Shea et al., 2016). A dataset from

the Cancer Genome Atlas (TCGA) Research network contains expression values for various microRNA and genes on 192 patients with GBM. Following Wang (2015); Molstad (2019), we chose a subset of 20 microRNA with the largest median absolute deviation, and a subset of 500 genes similarly. MicroRNAs are known to contribute to the development of GBM by binding to target messenger RNAs and regulating gene expressions (Xiong et al., 2019). While there is an abundance of Gene Expression Profiling (GEP) data in the post-genomic era, microRNA expression data is not as prevalent. Hence, methods for imputation of microRNA values given gene expression values are useful for understanding the role of microRNAs in disease pathogenesis when only gene expression data are available (see Kuo et al. (2012)). Consequently, several papers in the statistical literature (see Lee and Liu (2012); Wang (2015); Molstad (2019)) have considered a multivariate regression model with the microRNA expressions as response variables and the gene expressions as predictors. Also, identification of dynamic, ancillary and static responses might help identify functionally relevant miRNAs for GBM, and shed light on the internal dependence structure of the miRNA expressions. Since the number of predictors is larger than the sample size, before applying the response variable selection procedure, we reduced the dimension of predictors by two types of procedures: multi-response lasso

(Simon et al., 2013) and principal component analysis (PCA).

The R package *glmnet* was used to perform predictor variable selection with multi-response lasso, and 31 genes were selected. Hence we have $r = 20, p = 31$ and $n = 192$. Then we performed the response variable selection using the algorithm in Section 3.2. Two microRNAs were identified as dynamic: miR-124a and miR-219. The role of miR-124a in inhibiting the proliferation of GBM has been discussed in Silber et al. (2008), and the close association of miR-219 with GBM is discussed in Xiong et al. (2019). Six microRNAs are identified as ancillary miR-136, miR-338, miR-34a, miR-377, miR-7 and miR801, others are identified as static. We also reduced the dimension of the predictors using PCA, and kept 34 principal components, which explains 80% of the total variation in 500 genes. After performing response variable selection, the same two microRNAs (miR-124a and miR-219) are identified as dynamic. Eight microRNAs are identified as ancillary – the six ancillary microRNAs mentioned before and two additional microRNAs: miR-204 and miR-370.

We also computed the OLS estimator $\widetilde{\boldsymbol{\beta}}$ of the regression coefficients.

For additional validation, we explored miRNA and target gene pairs identified using data for other diseases such as neural tube defects (Stingo et al., 2010), but did not find an overlap with the current GBM based setting.

Note that the OLS estimator is computed using the entire response vector $\mathbf{Y}$. To compare the estimation efficiency, we bootstrapped the residuals for 200 times to compute the bootstrap standard deviations for each element in $\boldsymbol{\beta}_{\widehat{\mathcal{D}}}$ for both $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$, the OLS estimator, and the proposed estimator $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$ (with the predictors selected by multi-response lasso), then we computed the ratio R_{ij} in (4.16). The ratios ranged from 1.43 to 1.90, which implies that to achieve the same efficiency, the OLS estimator needs at least $1.43^2 \approx 2$ times the original sample size. To test the prediction performance, we randomly split the data into two parts of equal size. Half of the data is used as the training set and the other half is used as the testing set. The prediction error is computed as

$$\text{Prediction error} = \sqrt{\frac{1}{n}\sum_{j=1}^{2}\sum_{i\in\text{test set j}}(\mathbf{Y}_i - \widehat{\mathbf{Y}}_{i,\text{predict}})^T(\mathbf{Y}_i - \widehat{\mathbf{Y}}_{i,\text{predict}})}.$$

Then the prediction error is averaged over 100 random splits. The estimator of $\boldsymbol{\beta}$ after the response variable selection is $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}^T, \mathbf{0})^T$. Compared to the OLS estimator $\widetilde{\boldsymbol{\beta}}$, the estimator $\widehat{\boldsymbol{\beta}}$ reduces the prediction error by 8.38%. When the reduced set of predictors is chosen using PCA, the efficiency ratio R_{ij} ratios ranged from 1.47 to 2.86, and the estimator $\boldsymbol{\beta}$ computed after the response variable selection reduced the prediction error by 12.72% compared to the OLS estimator $\widetilde{\boldsymbol{\beta}}$.

We now demonstrate response variable selection in a high-dimensional

setting on a breast cancer data set (Chin et al., 2006), which is included in the R package *PMA*. The data set contains gene expression profiles and comparative genomic hybridization (CGH) measurements for all 23 chromosomes from 89 patients. Previous studies reveal that DNA copy number alteration is associated to the development or progression of human breast tumor (Pollack et al., 2002). CGH is a molecular cytogenetic method for detecting DNA copy number alteration in tumor cells, and measures the DNA copy number in several spots along a chromosome (Witten et al., 2009). There is a close association between the gene expression profiling data and the CGH measurements. Models which predict copy number alteration (CNA) values based on gene expression profiling data can be useful for imputing CNA for relevant analyses with datasets where only gene expression profiling data is available (Geng et al., 2011). In particular, following Chen et al. (2013); Lian et al. (2015); Molstad and Rothman (2016), we use multivariate linear regression with CGH measurements being the response variables and gene expression profiles as the predictor variables. Both the predictor and response variables are standardized. Chen et al. (2013) focused on chromosome 21 and Lian et al. (2015) focused on chromosome 18. We include the results for all 23 chromosomes. Each chromosome has 66 to 1942 gene expression profiles. So p is larger than $n = 89$ for most

chromosomes. Using multi-response lasso to select a common small set of predictors for all the response variables is not appropriate in this setting, since gene expressions around a region are generally expected to be more informative of the corresponding CNA values than expressions at more distant sites. This insight is also supported by earlier analyses in Chen et al. (2013); Molstad and Rothman (2016). Hence, instead we applied PCA to the predictors and due to the small sample size, we retained the smallest number of components that explain 70% of the variation. We then applied the response variable selection procedure in Section 3.4 with the chosen PCA components as the predictors.

To summarize, 23 response variable selection procedures were performed, corresponding to data for each of the 23 chromosomes. The response variable selection results are summarized in Table 2. For some chromosomes all responses are chosen as dynamic, while for some others all responses are chosen as static (entire $\boldsymbol{\beta}$ estimated as zero). For others, a non-trivial mix of the three categories is obtained. For example, for Chromosome 9, the CGH measurements at 36 chromosomal spots, including 2644, 12628, 35800, etc, were chosen as dynamic, the CGH measurements at 7 chromosomal spots, including 13369, 33163, 36175, etc, were chosen as ancillary and the other 64 responses were chosen as static. As discussed in the introduction, the

removal of a large number of static responses can stabilize the subsequent $\boldsymbol{\beta}_{\mathcal{D}}$ estimation in high-dimensional settings, and also lead to cost savings in future data collection.

Table 2: Selection of three categories of responses for breast cancer data

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12
Dynamic	136	0	126	0	0	76	19	137	36	0	96	42
Ancillary	0	0	2	0	0	3	45	1	7	0	83	12
Static	0	72	0	167	98	0	97	0	64	124	0	37
Chromosome	13	14	15	16	17	18	19	20	21	22	23	
Dynamic	58	76	0	0	84	51	0	63	0	18	0	
Ancillary	0	0	0	0	0	0	0	12	0	0	0	
Static	0	0	67	61	3	0	41	36	44	0	55	

The prediction error of the OLS estimator $\widetilde{\boldsymbol{\beta}}$ and the proposed estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}^T, \mathbf{0})^T$ were also compared. The prediction error is computed by cross validation averaged over 500 random splits of the data. Results are included in Table 3. Take chromosome 9 as an example, it has 107 DNA copy-number variations and gene expression profiles for 706 genes. Seventeen gene expression PCA components accounted for 70% of the variation, thus we have $r = 107, p = 17$. The OLS estimator $\widetilde{\boldsymbol{\beta}}$ has prediction error 1.90. In this example 36 responses are selected as dynamic, 7 responses are selected as ancillary and 64 responses are selected as static. We set the

coefficients of the dynamic responses as $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}} = \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}} - \widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}|\widehat{\mathcal{A}}}\widetilde{\boldsymbol{\beta}}_{\widehat{\mathcal{A}}}$ and others as 0, and the prediction error is 1.74 (an 8.42% reduction). For chromosome 11, 96 responses are selected as dynamic, 83 responses are selected as ancillary (no static responses). Since we are fitting a regression with $\mathbf{X}$ and $\mathbf{Y}_{\widehat{\mathcal{A}}}$ as predictors (see Proposition 4), the sample size 44 in the training dataset is too small for the regression, and we set the dynamic response coefficients as their OLS estimators, and the rest of the coefficients as zero. This still achieves a 8.14% gain in prediction error compared to the OLS estimator. Table 3 demonstrates that the proposed response variable selection proce-

Table 3: Improvement of prediction error for breast cancer data

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12
Prediction error	0.00%	24.46%	0.33%	20.22%	15.43%	0.68%	18.60%	0.04%	8.42%	17.72%	8.14%	6.73%
Chromosome	13	14	15	16	17	18	19	20	21	22	23	
Prediction error	0.00%	0.00%	14.72%	3.91%	0.35%	0.00%	23.44%	4.80%	2.23%	0.00%	30.58%	

dure can significantly improve the prediction error compared to the OLS estimator in a practical setting with $r_n > n$.

5. Acknowledgement

The authors would like to thank the careful review and helpful suggestions from the AE and two reviewers. The authors greatly appreciate Wonyul Lee and Yufeng Liu for sharing the glioblastoma multiforme cancer data.

This research of Zhihua Su is supported in part by Simons Foundation grant 632688.

References

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge MA.

An, B. and B. Zhang (2017). Simultaneous selection of predictors and responses for high dimensional multivariate linear regression. *Statistics & Probability Letters 127*, 173–177.

Argyriou, A., T. Evgeniou, and M. Pontil (2007). Multi-task feature learning. In *Advances in Neural Information Processing Systems*, Volume 19, pp. 41–48.

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods 7*(2), 127–150.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics 29*(4), 1165–1188.

Cai, T., W. Liu, and X. Luo (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*(494), 594–607.

Chen, K., H. Dong, and K.-S. Chan (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika 100*(4), 901–920.

Chen, L. and J. Z. Huang (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Amer. Statist. Assoc. 107*(500), 1533–1545.

Chen, X., C. Zou, and R. D. Cook (2010). Coordinate-independent sparse sufficient dimension

reduction and variable selection. *The Annals of Statistics 38*(6), 3696–3723.

Chin, K., S. DeVries, J. Fridlyand, P. T. Spellman, R. Roydasgupta, W.-L. Kuo, A. Lapuk, R. M. Neve, Z. Qian, and T. Ryder (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell 10*(6), 529–541.

Cook, R. and S. Weisberg (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

Deshpande, S. K., V. Ročková, and E. I. George (2019). Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics 28*(4), 921–931.

Foygel, R. and M. Drton (2010). Exact block-wise optimization in group lasso and sparse group lasso for linear regression. *arXiv preprint arXiv:1010.3320*.

Geng, H., J. Iqbal, W. C. Chan, and H. H. Ali (2011). Virtual cgh: an integrative approach to predict genetic abnormalities from gene expression microarray data applied in lymphoma. *BMC medical genomics 4*(1), 32.

Ha, M. J., F. C. Stingo, and V. Baladandayuthapani (2020). Bayesian structure learning in multi-layered genomic networks. *J. Amer. Statist. Assoc. 116*, 1–33.

Khare, K., S.-Y. Oh, and B. Rajaratnam (2015). A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77*(4), 803–825.

Kuo, T.-Y., E. Hsi, I.-P. Yang, P.-C. Tsai, J.-Y. Wang, and S.-H. H. Juo (2012). Computational analysis of mrna expression profiles identifies microrna-29a/c as predictor of colorectal cancer early recurrence. *PLoS One 7*, e31587.

Lee, W. and Y. Liu (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *Journal of multivariate analysis 111*, 241–255.

Leung, D. H. Y., Y.-G. Wang, and M. Zhu (2009, 04). Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method. *Biostatistics 10*(3), 436–445.

Li, Y., J. Datta, B. A. Craig, and A. Bhadra (2021). Joint mean–covariance estimation via the horseshoe. *Journal of Multivariate Analysis 183*, 104716.

Lian, H., S. Feng, and K. Zhao (2015). Parametric and semiparametric reduced-rank regression with flexible sparsity. *Journal of Multivariate Analysis 136*, 163–174.

Lipsitz, S. R., G. M. Fitzmaurice, E. J. Orav, and N. M. Laird (1994). Performance of generalized estimating equations in practical situations. *Biometrics 50*(1), 270–278.

Molstad, A. J. (2019). Insights and algorithms for the multivariate square-root lasso. *arXiv preprint arXiv:1909.05041*.

Molstad, A. J. and A. J. Rothman (2016). Indirect multivariate response linear regression. *Biometrika 103*(3), 595–607.

Nardi, Y. and A. Rinaldo (2008). On the asymptotic properties of the group lasso estimator for

linear models. *Electronic Journal of Statistics 2*, 605–633.

Obozinski, G., B. Taskar, and M. Jordan (2007). Joint covariate selection for grouped classification. *Department of Statistics, U. of California, Berkeley, TR 743*.

Peng, J., J. Zhu, A. Bergamaschi, W. Han, D. Noh, J. Pollack, and P. Wang (2009). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics 41*, 53–77.

Pollack, J. R., T. Sørlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A.-L. Børresen-Dale, and P. O. Brown (2002). Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences 99*(20), 12963–12968.

Puig, A. T., A. Wiesel, and A. O. Hero (2009). A multidimensional shrinkage-thresholding operator. In *Statistical Signal Processing, IEEE/SP 15th Workshop*, pp. 113–116. IEEE.

Rothman, A., E. Levina, and J. Zhu (2010). Sparse multivariate regression with covariate estimation. *Journal of Computational and Graphical Statistics 19*, 947–962.

Shea, A., V. Harish, Z. Afzal, J. Chijioke, H. Kedir, S. Dusmatova, A. Roy, M. Ramalinga, B. Harris, and J. Blancato (2016). Micrornas in glioblastoma multiforme pathogenesis and therapeutics. *Cancer medicine 5*(8), 1917–1946.

Silber, J., D. A. Lim, C. Petritsch, A. I. Persson, A. K. Maunakea, M. Yu, S. R. Vandenberg, D. G. Ginzinger, C. D. James, and J. F. Costello (2008). mir-124 and mir-137 inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor

stem cells. *BMC medicine 6*(1), 14.

Simon, N., J. Friedman, and T. Hastie (2013). A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv preprint arXiv:1311.6529*.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics 22*(2), 231–245.

Stingo, F., Y. Chen, M. Vannucci, M. Barrier, and P. Mirkes (2010). A bayesian graphical modeling approach to microrna regulatory network inference. *Ann. Appl. Stat. 4*, 2024–2028.

Su, Z., G. Zhu, X. Chen, and Y. Yang (2016). Sparse envelope model: efficient estimation and response variable selection in multivariate linear regression. *Biometrika 103*(3), 579–593.

Sun, T. and C.-H. Zhang (2013). Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research 14*(1), 3385–3418.

Wang, H. and C. Leng (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis 52*(12), 5277–5286.

Wang, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statistica Sinica*, 831–851.

Wang, L., J. Zhou, and A. Qu (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics 68*(2), 353–360.

Witten, D. M., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with

applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*(3), 515–534.

Xiong, D.-D., W.-Q. Xu, R.-Q. He, Y.-W. Dang, G. Chen, and D.-Z. Luo (2019). In silico analysis identified mirna-based therapeutic agents against glioblastoma multiforme. *Oncology reports 41*(4), 2194–2208.

Yin, J. and H. Li (2011). A sparse conditional gaussian graphical model for analysis of genetic genomics data. *Annals of Applied Statistics 5*, 2630–2650.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B 68*(1), 49–67.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions. *Journal of the American Statistical Association 57*, 348–368.

Zhang, T. and H. Zou (2014). Sparse precision matrix estimation via lasso penalized d-trace loss. *Biometrika 101*(1), 103–120.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.

Authors affiliation: Department of Statistics, University of Florida; E-mail: {kdkhare,zhihuasu}@ufl.edu