Statistica Sinica **34** (2024), 1325-1345
doi:https://doi.org/10.5705/ss.202022.0127

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 the subsequent estimation of the regression coefficients in multivariate linear regression. Because of the asymmetric roles of the predictors and responses in a regression, response variable selection differs markedly from the usual predictor variable selection. When a response is inferred to have a coefficient of zero, it should not simply be removed from subsequent estimation. Instead, we should analyze its relationship with the responses that have nonzero coefficients, which we call 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 responses into three categories: dynamic responses, ancillary responses, and 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 using synthetic and real data illustrate the efficacy of the proposed response variable selection procedure in both low- and high-dimensional settings. Lastly, we establish the consistency of the variable selection procedures and the asymptotic properties of the estimators for both the large-sample setting and the high-dimensional small-sample setting.

Key words and phrases: Group sparsity, high-dimensional data, oracle property, response variable selection.

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 $\boldsymbol{\Sigma}_{\mathbf{X}}$, and 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. Furthermore, we assume that $n > p$, because our primary focus is response variable selection. If $n < p$, we can use any predictor variable selection method

*Corresponding author.

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 applications in which we measure multiple outputs/responses and predictors, and wish to classify the response variables based on their relationship with the predictors. For example, when developing a new medicine, many clinical or hematological characteristics of a patient are measured. Here, it is of scientific interest to identify which characteristics change after taking the medicine. In economics, it might be of strategic importance to determine which industrial sectors are affected by a government policy, such as imposing a tariff on an imported good, such as bauxite. In particular, we categorize response variables as *dynamic*, *ancillary*, or *static* variables. 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 nonzero component. Identifying 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 the dynamic responses. Once the dynamic responses have been identified, one might be tempted to exclude/discard the nondynamic response variables from the estimation process. However, these variables might still carry information about $\boldsymbol{\beta}_{\mathcal{D}}$ through their correlations with the dynamic variables. Nondynamic response variables that are correlated with the dynamic response variables (given all other response variables) are defined as ancillary responses. Identifying ancillary responses is important, because this reduces the asymptotic variance of the MLE for $\boldsymbol{\beta}_{\mathcal{D}}$ (see Proposition 1). All other nondynamic responses are defined as static responses. Static responses carry no information about $\boldsymbol{\beta}_{\mathcal{D}}$, and can be eliminated from further analysis. Categorizing the nondynamic responses as ancillary or static responses helps researchers avoid collecting the static responses in future experiments, thus saving time and other resources.

One might argue that simply including all nondynamic responses in the estimation of $\boldsymbol{\beta}_{\mathcal{D}}$ avoids the extra selection effort for ancillary responses, while yielding 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 nondynamic response variables) is comparable to or larger than the sample size, including all nondynamic responses creates several methodological and computational complications, and thus is not advisable.

Connections with existing literature. Compared with that on predictor variable selection, the literature on response variable selection is surprisingly limited. The standard method is to test whether the regression coefficients for each response are equal to zero, adjusting for multiple testing; see, for example, Benjamini and Yekutieli (2001). The response variables with zero regression coefficients are usually discarded after selection. An and Zhang (2017) use a double group-lasso penalty to simultaneously select the predictors and the responses. However, they treat the responses as uncorrelated, and so do not use the covariance structure among the elements in $\mathbf{Y}$.

Numerous works use generalized estimating equations (GEEs) to improve the estimation of the regression coefficients by accounting for correlated responses in longitudinal data and repeated measurement data settings; see Lipsitz et al. (1994), Ballinger (2004), Leung, Wang and Zhu (2009), and the references therein. A high-dimensional adaptation of these methods in Wang, Zhou and Qu (2012) imposes generic sparsity in the regression coefficients through penalization. There is also a growing body of literature on the joint sparse estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}^{-1}$; see Peng et al. (2009), Rothman, Levina and Zhu (2010), Yin and Li (2011), Deshpande, Ročková and George (2019), Ha, Stingo and Baladandayuthapani (2020), Li et al. (2021), and the references therein. To the best of our knowledge, these methods either reduce the number of parameters by imposing general sparsity patterns in $\boldsymbol{\beta}$ and/or $\boldsymbol{\Omega}$, or select "master" predictor variables using the column sparsity in $\boldsymbol{\beta}$.

However, these methods do not provide a way to identify dynamic, ancillary, and static responses using the specific and structured sparsity in $\boldsymbol{\beta}$ and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ (see equation (2.5) below). *Although we share the goal of improving the efficiency of regression coefficient estimates, the proposed approach may offer scientific insights as a result of identifying dynamic responses, and save computational and other resources by identifying ancillary/static responses (as discussed above).*

Outline of the paper. We propose a two-step procedure for selecting the response variables, while considering the covariance among the responses. The first step identifies the dynamic variables, and the second step identifies the ancillary variables. We then estimate the regression coefficients based on the selection results. The remainder of the paper is organized as follows. In Section 2, we formally define the three categories of response variables, and derive various technical results that support the motivations for the response variable selection discussed above. In Sections 3.1–3.3, we discuss 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 a methodology for the challenging high-dimensional setting ($n < r$), and derive the corresponding asymptotic properties. A detailed experimental validation is provided in Section 4.1 (simulated data) and Section 4.2 (real data). The proofs of the technical results, implementation details, additional simulations, and future research directions are provided in the Supplementary Material.

2. Categories of response variables

In this section, we introduce three categories of responses, and discuss estimating the coefficients $\boldsymbol{\beta}$ after the selection. The three categories are defined based on the roles they play in the 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 nonzero component, we call it a dynamic response.

Let $\mathcal{D}$ be a subset of $\{1, \ldots, r\}$ that 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 responses, and $\mathbf{Y}_{-\mathcal{D}} \in \mathbb{R}^{r-r_{\mathcal{D}}}$ to denote those responses with coefficient vectors that 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

$$\begin{pmatrix} \mathbf{Y}_{\mathcal{D}} \\ \mathbf{Y}_{-\mathcal{D}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_{\mathcal{D}} \\ \boldsymbol{\alpha}_{-\mathcal{D}} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{D}} \\ \mathbf{0} \end{pmatrix} \mathbf{X} + \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathcal{D}} \\ \boldsymbol{\varepsilon}_{-\mathcal{D}} \end{pmatrix}, \quad \mathrm{var}\begin{pmatrix} \boldsymbol{\varepsilon}_{\mathcal{D}} \\ \boldsymbol{\varepsilon}_{-\mathcal{D}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{D}} & \boldsymbol{\Sigma}_{\mathcal{D},-\mathcal{D}} \\ \boldsymbol{\Sigma}_{-\mathcal{D},\mathcal{D}} & \boldsymbol{\Sigma}_{-\mathcal{D}} \end{pmatrix}. \tag{2.1}$$

Suppose that the data consist of n independent and identically distributed (i.i.d.) observations $(\mathbf{Y}_i, \mathbf{X}_i)$, where $\mathbf{Y}_i$ is sampled from the conditional distribution of $\mathbf{Y} \mid \mathbf{X}_i$, for $i = 1, \ldots, n$. The following proposition (Proposition 2 in Su et al. (2016)) indicates that after selection, although $\mathbf{Y}_{-\mathcal{D}}$ has zero coefficients, it improves the efficiency of 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 regressions of $\mathbf{Y}_{\mathcal{D}}$ on $\mathbf{X}$ and $\mathbf{Y}_{-\mathcal{D}}$ on $\mathbf{X}$, respectively. Note that the OLS estimators do not account for the error correlations. Furthermore, 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 the regression coefficients is the same as the OLS estimate Amemiya (1985, p.197). Let $\mathbf{R}_{\mathcal{D}}$ be the residuals from the regression of $\mathbf{Y}_{\mathcal{D}}$ on $\mathbf{X}$, and $\mathbf{R}_{-\mathcal{D}}$ be the residuals from the regression of $\mathbf{Y}_{-\mathcal{D}}$ on $\mathbf{X}$. The operator $\mathrm{vec}(\cdot)$ 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* (2.1) *and $\mathcal{D}$ is given. The maximum likelihood estimator of $\boldsymbol{\beta}_{\mathcal{D}}$ under model* (2.1) *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 ensures explicit forms of the asymptotic variance, which facilitates the comparison. Similar results can be derived under nonnormal errors, but the expressions for $\mathbf{V}_1$ 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 is $\widetilde{\boldsymbol{\beta}}_{\mathcal{D}}$, which uses only $\mathbf{Y}_{\mathcal{D}}$. 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 that are zero, it is excluded from the model, because this is more efficient than retaining it in the model. However, in response variable selection, because $\mathbf{Y}_{-\mathcal{D}}$ carries information on $\boldsymbol{\beta}_{\mathcal{D}}$ through its correlation with $\mathbf{Y}_{\mathcal{D}}$, we use $\mathbf{Y}_{-\mathcal{D}}$ to construct 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 the Supplementary Material, Section S2.

When $\mathbf{Y}$ is high-dimensional, $\mathbf{Y}_{-\mathcal{D}}$ may also be high-dimensional, and only part of $\mathbf{Y}_{-\mathcal{D}}$ may carry information on $\boldsymbol{\beta}_{\mathcal{D}}$. The other part of $\mathbf{Y}_{-\mathcal{D}}$ has regression coefficients that are zero and does not provide information on $\boldsymbol{\beta}_{\mathcal{D}}$, and thus can safely be eliminated from model (2.1), removing the need to measure $\mathbf{Y}_{-\mathcal{D}}$ in future experiments. To distinguish between these two types of responses, we define ancillary and static responses.

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

Let $\mathcal{A}$ and $\mathcal{S}$ be subsets of $\{1, \ldots, r\}$ that contain the indices of all ancillary and static responses, respectively. Let $r_{\mathcal{A}}$ and $r_{\mathcal{S}}$ denote the cardinalities of $\mathcal{A}$ and $\mathcal{S}$, respectively. 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 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* (2.2) *and* (2.3)*, where*

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

and

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

Let $\widehat{\boldsymbol{\beta}}_{\mathcal{D},1}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{D},2}$ be the maximum likelihood estimators of $\boldsymbol{\beta}_{\mathcal{D}}$ under models (2.2) *and* (2.3)*, 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 $\widehat{\boldsymbol{\beta}}_{\mathcal{D},i}$, for $i = 1, 2$, is given by*

$$\sqrt{n}\{\text{vec}(\widehat{\boldsymbol{\beta}}_{\mathcal{D},i}) - \text{vec}(\boldsymbol{\beta}_{\mathcal{D}})\} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}), \quad \mathbf{V} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes (\boldsymbol{\Sigma}_{\mathcal{D}} - \boldsymbol{\Sigma}_{\mathcal{D},\mathcal{A}} \boldsymbol{\Sigma}_{\mathcal{A}}^{-1} \boldsymbol{\Sigma}_{\mathcal{A},\mathcal{D}}). \quad (2.4)$$

The forms of $\widehat{\boldsymbol{\beta}}_{\mathcal{D},1}$ and $\widehat{\boldsymbol{\beta}}_{\mathcal{D},2}$ can be obtained from Proposition 1 by replacing $-\mathcal{D}$ with $\mathcal{A}$ and $(\mathcal{A}, \mathcal{S})$, respectively. Proposition 2 suggests that after the response variable selection, we need only use $\mathbf{Y}_{\mathcal{D}}$ and $\mathbf{Y}_{\mathcal{A}}$ for the 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 $\mathbf{R}_{\mathbf{X}|\mathcal{A}}$ be the residuals from the regression of $\mathbf{X}$ on $\mathbf{Y}_{\mathcal{A}}$.

Proposition 3. *Assume that the error vector $\boldsymbol{\varepsilon}$ has finite second moments in model* (2.2)*, and that $\mathcal{D}$ 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 estimating the regression coefficients. The responses in $\mathbf{Y}_{\mathcal{A}}$ serve as the ancillary statistic, hence its name.

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

Proposition 4. *Under model* (2.5)*, 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 $\mathcal{S}$. Because $\mathbf{Y}_{\mathcal{D}} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{S}} \mid (\mathbf{Y}_{\mathcal{A}}, \mathbf{X})$ implies $\boldsymbol{\Omega}_{\mathcal{D},\mathcal{S}} = \mathbf{0}$, model (1.1) can then be written as

$$\begin{pmatrix} \mathbf{Y}_{\mathcal{D}} \\ \mathbf{Y}_{\mathcal{A}} \\ \mathbf{Y}_{\mathcal{S}} \end{pmatrix} = \boldsymbol{\alpha} + \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{D}} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \mathbf{X} + \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathcal{D}} \\ \boldsymbol{\varepsilon}_{\mathcal{A}} \\ \boldsymbol{\varepsilon}_{\mathcal{S}} \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{\mathcal{D}} & \boldsymbol{\Omega}_{\mathcal{D},\mathcal{A}} & \mathbf{0} \\ \boldsymbol{\Omega}_{\mathcal{A},\mathcal{D}} & \boldsymbol{\Omega}_{\mathcal{A}} & \boldsymbol{\Omega}_{\mathcal{A},\mathcal{S}} \\ \mathbf{0} & \boldsymbol{\Omega}_{\mathcal{S},\mathcal{A}} & \boldsymbol{\Omega}_{\mathcal{S}} \end{pmatrix}. \tag{2.5}$$

Note that no columns in $\boldsymbol{\Omega}_{\mathcal{D},\mathcal{A}}$ are zero. From (2.5), 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 (2.5), and is discussed further in Section 3. Before we proceed, we first introduce a property of model (2.5) that we use 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 a covariance structure as in* (2.5). *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 regression model*

$$\boldsymbol{\varepsilon}_{\mathcal{D}} = \mathbf{B}_{\mathcal{D}|(\mathcal{A},\mathcal{S})} \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathcal{A}} \\ \boldsymbol{\varepsilon}_{\mathcal{S}} \end{pmatrix} + \mathbf{e} \tag{2.6}$$

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 identifying the zero block in $\boldsymbol{\Omega}$ can be converted to a response variable selection problem in which we need only identify the dynamic and nondynamic 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 consist of n i.i.d. observations $(\mathbf{Y}_i, \mathbf{X}_i)$, where $\mathbf{Y}_i$ is sampled from the conditional distribution of $\mathbf{Y} \mid \mathbf{X}_i$, for $i = 1, \ldots, n$. Let $\mathbb{Y}$ denote an $n \times r$ matrix in which the ith row is $\mathbf{Y}_i^T$, $\mathbb{X}$ denote an $n \times p$ matrix in which the ith row is $\mathbf{X}_i^T$, $\mathbf{1}_n$ be an n-dimension column vector of ones, and tr denote the trace of a matrix. The log likelihood of $\mathbf{Y}_i \mid \mathbf{X}_i$, for $i = 1, \ldots, n$, is given by

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

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

$$f(\boldsymbol{\beta}, \boldsymbol{\Omega}) = -\log|\boldsymbol{\Omega}| + \frac{1}{n}\operatorname{tr}\left\{(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\beta}^T)\boldsymbol{\Omega}(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\beta}^T)^T\right\}, \tag{3.1}$$

where $\mathbb{Y}_c \in \mathbb{R}^{n\times r}$ and $\mathbb{X}_c \in \mathbb{R}^{n\times p}$ are centered data matrices; that is, 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 (3.1), 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, Zou and Cook (2010)). Specifically, we solve the following optimization problem:

$$\begin{aligned} f_1(\boldsymbol{\beta}) = &\log|\mathbf{S}_{\mathbf{Y}|\mathbf{X}}| + \frac{1}{n}\operatorname{tr}\left\{(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\beta}^T)\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\beta}^T)^T\right\} \\ &+ \lambda_1\sum_{i=1}^{r} w_i\|\boldsymbol{\beta}_{i\cdot}\|, \end{aligned} \tag{3.2}$$

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 $\boldsymbol{\beta}$, and 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 is determined in Step 2.

Step 2. The goal of this step is to estimate $\mathcal{A}$ and $\mathcal{S}$. Proposition 5 indicates that a difference between ancillary and static responses 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:

$$\begin{aligned} f_2(\mathbf{B}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}) = & \log|\mathbf{S}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}| \\ & + \frac{1}{n}\operatorname{tr}\Big\{\Big(\mathbf{R}_{\widehat{\mathcal{D}}} - \mathbf{R}_{-\widehat{\mathcal{D}}}\mathbf{B}^T_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}\Big)\mathbf{S}^{-1}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}\Big(\mathbf{R}_{\widehat{\mathcal{D}}} - \mathbf{R}_{-\widehat{\mathcal{D}}}\mathbf{B}^T_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}\Big)^T\Big\} \\ & + \lambda_2 \sum_{i=1}^{r-r_{\widehat{\mathcal{D}}}} \tilde{w}_i \|\mathbf{B}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S}),\cdot I}\|, \end{aligned} \tag{3.3}$$

where $\mathbf{S}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}$ is the sample covariance matrix of the residuals from the regression 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 $\widehat{\mathbf{B}}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S}),\mathrm{step2}}$ 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 in which the corresponding columns in $\widehat{\mathbf{B}}_{\mathcal{D}|(\mathcal{A},\mathcal{S}),\mathrm{step2}}$ 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 in which the 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 $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}^T_{\widehat{\mathcal{D}}}, \mathbf{0})^T$, 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 an 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

$$\begin{aligned} \frac{1}{n}\Bigg\{ & (\mathbf{S}^{-1}_{\mathbf{Y}|\mathbf{X}})_{jj}\left(\mathbb{Y}_{c,\cdot j} - \mathbb{X}_c\boldsymbol{\beta}^T_{j\cdot}\right)^T\left(\mathbb{Y}_{c,\cdot j} - \mathbb{X}_c\boldsymbol{\beta}^T_{j\cdot}\right) \\ & + \sum_{k\neq j} 2(\mathbf{S}^{-1}_{\mathbf{Y}|\mathbf{X}})_{jk}(\mathbb{Y}_{c,\cdot k} - \mathbb{X}_c\boldsymbol{\beta}^T_{k\cdot})^T(\mathbb{Y}_{c,\cdot j} - \mathbb{X}_c\boldsymbol{\beta}^T_{j\cdot})\Bigg\} + \lambda_1 w_j\|\boldsymbol{\beta}_{j\cdot}\| \end{aligned} \tag{3.4}$$

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 (3.4) is a nondifferentiable convex function of $\boldsymbol{\beta}_{j\cdot}$. Minimizing such functions (quadratic form in vector plus its ℓ_2-norm) is considered in Foygel and Drton (2010), Puig, Wiesel and Hero (2009), and Simon et al. (2013) in the context of a group lasso. In particular, Simon et al. (2013) provide a reasonably fast majorize-minimize algorithm to solve this minimization problem. This approach is implemented in the R package *SGL*, and we use it to solve for $\boldsymbol{\beta}_{j\cdot}$ in (3.4).

Algorithm for Step 2: The optimization problem in Step 2 is the same as that in Simon, Friedman and Hastie (2013), with $\mathbf{S}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}^{-1/2}\mathbf{R}_{\widehat{\mathcal{D}}}$ being their $\mathbf{Y}$ and $\mathbf{S}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}^{-1/2}\mathbf{B}_{\mathcal{D}|(\widehat{\mathcal{A}},\widehat{\mathcal{S}})}$ being the coefficients. Note that a column in $\mathbf{S}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}^{-1/2}\mathbf{B}_{\mathcal{D}|(\widehat{\mathcal{A}},\widehat{\mathcal{S}})}$ 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, Friedman and Hastie (2013) studies the "multi-response group-lasso" problem, and provides an iterative algorithm for minimizing the objective function

$$\frac{1}{n}\operatorname{tr}\left\{\left(\mathbb{Y}_c-\mathbb{X}_c\boldsymbol{\beta}^T\right)^T\left(\mathbb{Y}_c-\mathbb{X}_c\boldsymbol{\beta}^T\right)\right\}+\lambda_1\sum_{k=1}^{r}\left\|(\boldsymbol{\beta}^T)_{k\cdot}\right\|, \tag{3.5}$$

where $(\boldsymbol{\beta}^T)_{k\cdot}$ is the kth row of $\boldsymbol{\beta}^T$; see also Argyriou, Evgeniou and Pontil (2007) and Obozinski, Taskar and Jordan (2007). However, the iterative algorithm presented in Simon, Friedman and Hastie (2013) is not applicable in the context of (3.2). There are two notable differences between the minimization problems in (3.2) and (3.5). First, in (3.2), we use the group-lasso penalty on the rows of $\boldsymbol{\beta}$ for the response variable selection, whereas in (3.5), we use a group-lasso penalty for the columns of $\boldsymbol{\beta}$ for the predictor variable selection. Second, unlike (3.5), the trace term in (3.2) contains the term $\boldsymbol{\Omega}$, because we consider a multi-response regression model with a general covariance structure.

3.3. Theoretical properties

In this section, we establish the 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, let $\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}}$ be the true regression coefficients of dynamic responses, and let $\bar{\boldsymbol{\Sigma}}$ be the true error covariance matrix. Let $\bar{P}$ denote the probability measure corresponding to the true data-generating model ((2.5), with the true parameters introduced above). For consistency in the fixed r setting, we do not require the normality of the true error distribution, and thus assume only that the errors are i.i.d. 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: $\|\operatorname{vec}(\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}})-\operatorname{vec}(\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}})\|=O_{\bar{P}}(n^{-1/2})$.*

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

To discuss the optimal estimation rate, we first introduce the oracle model for response variable selection. If we know the oracle information on which the responses are dynamic, ancillary, or static, the oracle model is model (2.2). 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 $\widehat{\boldsymbol{\beta}}_{\bar{\mathcal{D}},\text{oracle}} = \widetilde{\boldsymbol{\beta}}_{\bar{\mathcal{D}}} - \widetilde{\boldsymbol{\beta}}_{\bar{\mathcal{D}}|\bar{\mathcal{A}}}\widetilde{\boldsymbol{\beta}}_{\bar{\mathcal{A}}}$. The asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\bar{\mathcal{D}},\text{oracle}}$ is the same as that of $\widehat{\boldsymbol{\beta}}_{\bar{\mathcal{D}},1}$ in Proposition 2; see (2.4). 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 $\mathbf{u} \in \mathbb{R}^a$ and $\mathbf{v} \in \mathbb{R}^b$, with $a < b$, for the following result.

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

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 a high-dimensional setting

In a 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, we need an estimator of $\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{X}}^{-1}$ for the objective function in (3.2). Several precision matrix estimators can be adapted to a high-dimensional setting, including the constrained l_1-minimization estimator (Cai, Liu and Luo (2011, CLIME)), lasso penalized D-trace estimator (Zhang and Zou (2014)), scaled lasso estimator (Sun and Zhang (2013)), and convex correlation selection estimator (Khare, Oh and Rajaratnam (2015, CONCORD)). Here, we use the CONCORD estimator, because it computes quickly and recovers the sparsity pattern with high accuracy. Let ω_{ij} denote the (i, j)th element of $\boldsymbol{\Omega}$, 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 $\boldsymbol{\Omega}$, denoted by $\widehat{\boldsymbol{\Omega}}$, is the minimizer of the objective function

$$Q_{\text{con}}(\boldsymbol{\Omega}) = -\sum_{i=1}^{r} n \log \omega_{ii} + \frac{1}{2}\sum_{i=1}^{r} \|\omega_{ii}\mathbf{R}_{i\cdot} + \sum_{j\neq i} \omega_{ij}\mathbf{R}_{j\cdot}\|^2 + \lambda \sum_{1\leq i\neq j\leq r} |\Omega_{ij}| \quad (3.6)$$

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 replace $\mathbf{S}_{\mathbf{Y}|\mathbf{X}}^{-1}$ with $\widehat{\boldsymbol{\Omega}}$ in (3.2), and obtain the objective function

$$\tilde{f}_1(\boldsymbol{\beta}) = -\log|\widehat{\boldsymbol{\Omega}}| + \frac{1}{n}\text{tr}\{(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\beta}^T)\widehat{\boldsymbol{\Omega}}(\mathbb{Y}_c - \mathbb{X}_c\boldsymbol{\beta}^T)^T\} + \lambda_1 \sum_{i=1}^{r_n} w_i\|\boldsymbol{\beta}_{i\cdot}\|. \quad (3.7)$$

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 (3.3) 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 these problems, we again use the CONCORD estimator $\widehat{\boldsymbol{\Omega}}$. Because $\mathbf{S}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}$ estimates $\boldsymbol{\Omega}_{\mathcal{D}}^{-1}$, we use the corresponding block in $\widehat{\boldsymbol{\Omega}}$, that is, $\widehat{\boldsymbol{\Omega}}_{\mathcal{D}}$, to replace $\mathbf{S}^{-1}_{\widehat{\mathcal{D}}|-\widehat{\mathcal{D}}}$ in (3.3). Note that $\mathbf{B}_{\mathcal{D}|(\mathcal{A},\mathcal{S})} = -\boldsymbol{\Omega}_{\mathcal{D}}^{-1}(\boldsymbol{\Omega}_{\mathcal{D},\mathcal{A}}, \boldsymbol{\Omega}_{\mathcal{D},\mathcal{S}})$ (see the proof of Proposition 5), and we initialize $\mathbf{B}_{\mathcal{D}|(\mathcal{A},\mathcal{S})}$ by $-\widehat{\boldsymbol{\Omega}}_{\mathcal{D}}^{-1}\widehat{\boldsymbol{\Omega}}_{\mathcal{D},-\mathcal{D}}$. The objective function is obtained as

$$\begin{aligned}\tilde{f}_2(\mathbf{B}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}) = &- \log|\widehat{\boldsymbol{\Omega}}_{\widehat{\mathcal{D}}}| \\ &+ \frac{1}{n}\operatorname{tr}\left\{\left(\mathbf{R}_{\widehat{\mathcal{D}}} - \mathbf{R}_{-\widehat{\mathcal{D}}}\mathbf{B}^T_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}\right)\widehat{\boldsymbol{\Omega}}_{\widehat{\mathcal{D}}}\left(\mathbf{R}_{\widehat{\mathcal{D}}} - \mathbf{R}_{-\widehat{\mathcal{D}}}\mathbf{B}^T_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S})}\right)^T\right\} \\ &+ \lambda_2 \sum_{i=1}^{r_n - r_{\widehat{\mathcal{D}}}} \tilde{w}_i \|\mathbf{B}_{\widehat{\mathcal{D}}|(\mathcal{A},\mathcal{S}),\cdot I}\|. \end{aligned} \tag{3.8}$$

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

3.5. Response selection consistency in a 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, $\bar{\boldsymbol{\beta}}_{\bar{\mathcal{D}}}$ be the true regression coefficients of the dynamic responses, $\bar{\boldsymbol{\Sigma}}$ 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 we suppress the dependence to simplify the notation. Mild regularity assumptions needed to establish the following result are provided and discussed in S8 of the Supplementary Material, owing to space constraints. They include the sub-Gaussianity of the errors, uniform boundedness of the eigenvalues of $\bar{\boldsymbol{\Sigma}}$ (Assumption 1), incoherence and minimum signal size conditions for the consistency of $\hat{\boldsymbol{\Omega}}$ (Assumptions 2 and 3), rates of growth of the true numbers 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) than that of n.

Theorem 3. *Under Assumptions* 1–8 *(provided in the Supplementary Material), the following hold for every $\eta > 0$:*

1. *(Dynamic response selection consistency) Let $\widehat{\boldsymbol{\beta}}_{step1}$ denote the solution to* (3.7)*, 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* (3.8), *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 the three categories of the response variables. As a direct consequence of the selection consistency, the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathcal{D}}$ is given in Theorem 4 (proof in the Supplementary Material).

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 $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}$ 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 fix $n = 50$, $p = 8$, $r_{\mathcal{D}} = 6$, and $r_{\mathcal{A}} = 2$. The response dimension r ranges from 200 to 1,000. Elements in $\boldsymbol{\beta}_{\mathcal{D}}$ are independent $N(0, 0.5^2)$ variates, and the intercept is $\boldsymbol{\alpha} = \mathbf{0}$. The covariance matrix $\boldsymbol{\Sigma}$ is 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. How to generate $\boldsymbol{\Sigma}$ is discussed in S13 of the Supplementary Material. We generate $\mathbf{X}$ 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 generate $\mathbf{X}$ from $N_p(0, (\mathbf{I}_p + 1_p1_p^T)/8)$ to represent correlated predictors, where 1_p denotes a p-dimensional vector of ones. We discuss tuning the parameters in Section S11 of the Supplementary Material. For each setting, we simulate 200 replications, and evaluate the selection performance by using the true positive rates $\mathrm{TPR}_{\mathcal{D}}$, $\mathrm{TPR}_{\mathcal{A}}$, and $\mathrm{TPR}_{\mathcal{S}}$ for the three categories of responses: $\mathrm{TPR}_{\mathcal{D}} = |\bar{\mathcal{D}} \cap \widehat{\mathcal{D}}|_c/|\bar{\mathcal{D}}|_c$, $\mathrm{TPR}_{\mathcal{A}} = |\bar{\mathcal{A}} \cap \widehat{\mathcal{A}}|_c/|\bar{\mathcal{A}}|_c$, and $\mathrm{TPR}_{\mathcal{S}} = |\bar{\mathcal{S}} \cap \widehat{\mathcal{S}}|_c/|\bar{\mathcal{S}}|_c$, where for a set S, $|S|_c$ denotes its cardinality. We add precision measures $\mathrm{PPV}_{\mathcal{D}}$, $\mathrm{PPV}_{\mathcal{A}}$, and $\mathrm{PPV}_{\mathcal{S}}$ for the sensitivity analysis, where $\mathrm{PPV}_{\mathcal{D}} = |\bar{\mathcal{D}} \cap \widehat{\mathcal{D}}|_c/|\widehat{\mathcal{D}}|_c$, that is, the ratio of true positives to the sum of the true and false positives. The measures $\mathrm{PPV}_{\mathcal{A}}$ and $\mathrm{PPV}_{\mathcal{S}}$ are defined accordingly. We measure the efficiency gain of a randomly selected element, say β_{ij}, using the efficiency ratio R_{ij}, defined as

$$R_{ij} = \frac{\mathrm{var}(\tilde{\beta}_{ij})}{\mathrm{var}(\hat{\beta}_{ij})}, \tag{4.1}$$

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

r	200	300	500	1,000	200	300	500	1,000	200	300	500	1,000
	$\mathbf{X} \sim N_p(0, 0.5^2\mathbf{I}_p)$				$\mathbf{X} \sim N_p(0, 0.25^2\mathbf{I}_p)$				$\mathbf{X} \sim N_p(0, \frac{1}{8}(1_p 1_p^T + \mathbf{I}_p))$			
$\text{TPR}_{\mathcal{D}}$	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_{MSE}^{all}	99.12	140.20	206.09	370.40	87.63	138.18	205.06	370.40	98.80	140.73	201.52	363.26
$R_{\text{MSE}}^{\mathcal{D}}$	4.440	3.934	3.029	2.391	3.188	3.891	2.993	2.391	4.356	3.834	3.030	2.410

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 $\hat{\beta}_{ij}$, respectively, calculated based on 200 replications. Then, R_{median} is the median of all R_{ij} for the nonzero elements in $\boldsymbol{\beta}$. The results are provided in Table 1. Both the TPR and the PPV measures show that the variable selection procedure identifies 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 effect on the results. The correlated predictors do not seem to have any obvious negative effect on the variable selection or efficiency gains. We also investigate the ratio of the MSEs. The measure R_{MSE}^{all} computes the median of $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F^2 / \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F^2$ (over 200 replications), where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Because $\widetilde{\boldsymbol{\beta}}$ is the OLS estimator using all the responses, it is not sparse, and 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 to 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 significant effect on $\widetilde{\boldsymbol{\beta}}$. Thus, the ratios R_{MSE}^{all} are very large. We also investigate $R_{\text{MSE}}^{\mathcal{D}}$, which is similar to R_{MSE}^{all}, but focuses only 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 200 replications). The ratios are still significantly greater than one, indicating that the response variable selection procedure indeed improves the estimation performance.

An additional simulation for a low-dimensional setting is given in Section S12 of the Supplementary Material. The simulation shows the consistency of the variable selection for all three categories of the responses, as well as the estimator of $\boldsymbol{\beta}$. It also demonstrates that the estimator from the two-stage selection procedure is more efficient than the estimator of $\boldsymbol{\beta}_{\mathcal{D}}$, which uses the oracle information of the true dynamic responses. In other words, the efficiency gains from the ancillary responses can be sufficient to offset the cost of selecting 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 data set from the Cancer Genome Atlas (TCGA) Research network contains expression values for various microRNA and genes on 192 patients with GBM. Following Wang (2015) and 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)). Although the post-genomic era has provided an abundance of gene expression profiling (GEP) data, microRNA expression data are not as prevalent. Hence, methods for imputing 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. Furthermore, identifying the dynamic, ancillary, and static responses might help identify functionally relevant microRNAs for GBM, and shed light on the internal dependence structure of the microRNA expressions. Because the number of predictors is larger than the sample size, before applying the response variable selection procedure, we reduce the dimension of the predictors using two procedures: a multi-response lasso (Simon, Friedman and Hastie (2013)), and a principal component analysis (PCA).

The R package *glmnet* is used to perform the predictor variable selection using a multi-response lasso, with 31 genes 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 are identified as dynamic: miR-124a and miR-219. The role of miR-124a in inhibiting the proliferation of GBM is 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; the remainder are identified as static.[1] We also reduced the dimension of the predictors using a PCA, and kept 34 principal components, which explains 80% of the total variation in 500 genes. After performing the response variable selection, the same two microRNAs (miR-124a and miR-219) are identified as dynamic. Eight microRNAs are identified as ancillary: the six aforementioned ancillary microRNAs, and two additional microRNAs, namely, miR-204 and miR-370.

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

[1] For additional validation, we explored microRNA 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.

that the OLS estimator is computed using the entire response vector $\mathbf{Y}$. To compare the estimation efficiency, we bootstrapped the residuals 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 using a multi-response lasso). Then, we computed the ratio R_{ij} in (4.1). The ratios range from 1.43 to 1.90, which implies that to achieve the same efficiency, the OLS estimator needs to be at least $1.43^2 \approx 2$ times the original sample size. To test the prediction performance, we split the data randomly into two parts of equal size. Half of the data are used as the training set, and the other half are 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 $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}^T_{\widehat{\mathcal{D}}}, \mathbf{0})^T$. Compared with 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 a PCA, the efficiency ratio R_{ij} ranges 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 with the OLS estimator $\widetilde{\boldsymbol{\beta}}$.

We now demonstrate the 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 (CNA) is associated with the development or progression of human breast tumors (Pollack et al. (2002)). CGH is a molecular cytogenetic method for detecting DNA and CNA in tumor cells, and measures the DNA copy number in several spots along a chromosome (Witten, Tibshirani and Hastie (2009)). There is a close association between the gene expression profiling data and the CGH measurements. Models that predict CNA values based on gene expression profiling data can be useful for imputing CNA for analyses of data sets in which only gene expression profiling data are available (Geng et al. (2011)). In particular, following Chen, Dong and Chan (2013), Lian, Feng and Zhao (2015), and Molstad and Rothman (2016), we use a multivariate linear regression, with CGH measurements as the response variables, and gene expression profiles as the predictor variables. Both the predictor and the response variables are standardized. Chen, Dong and Chan (2013) focus on chromosome 21, and Lian, Feng and Zhao (2015) focus on chromosome 18. We include the results for all 23 chromosomes. Each chromosome has 66 to 1942 gene expression profiles. Thus, p is larger than $n = 89$ for most chromosomes. Using a multi-response lasso to select a common

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

small set of predictors for all of the response variables is not appropriate in this setting, because, in general, gene expressions around a region are expected to be more informative of the corresponding CNA values than are expressions at more distant sites. This insight is supported by earlier analyses in Chen, Dong and Chan (2013) and Molstad and Rothman (2016). Hence, we instead applied a PCA to the predictors and, because of 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, we performed 23 response variable selection procedures, 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, and for others, all responses are chosen as static (entire $\boldsymbol{\beta}$ estimated as zero). A third group comprises a non-trivial mix of the three categories. For example, for Chromosome 9, the CGH measurements at 36 chromosomal spots, including 2644, 12628, and 35800, are chosen as dynamic, the CGH measurements at seven chromosomal spots, including 13369, 33163, and 36175, are chosen as ancillary, and the other 64 responses are chosen as static. As discussed in the introduction, removing 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.

We also compare the prediction errors of the OLS estimator $\widetilde{\boldsymbol{\beta}}$ and the proposed estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{D}}}^T, \mathbf{0})^T$. The prediction error is computed using cross-validation, averaged over 500 random splits of the data. The results are included in Table 3. For example, Chromosome 9 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 $r = 107$ and $p = 17$. The OLS estimator $\widetilde{\boldsymbol{\beta}}$ has a prediction error of 1.90. In this example, 36 responses are selected as dynamic, seven 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 zero, and the prediction error is 1.74 (an 8.42% reduction). For Chromosome 11, 96 responses are selected as

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%	

dynamic, and 83 responses are selected as ancillary (no static responses). Because we are fitting a regression with $\mathbf{X}$ and $\mathbf{Y}_{\widehat{\mathcal{A}}}$ as predictors (see Proposition 4), the sample size of 44 in the training data set is too small for the regression. Thus, we set the dynamic response coefficients as their OLS estimators, and the rest of the coefficients as zero. This still achieves an 8.14% gain in the prediction error compared with the OLS estimator. Table 3 demonstrates that the proposed response variable selection procedure can significantly improve the prediction error compared with the OLS estimator in a practical setting with $r_n > n$.

Supplementary Material

The online Supplementary Material includes proofs of all propositions and theorems, implementation details, additional simulations and discussion of future research directions.

Acknowledgments

The authors thank the Associate Editor and two reviewers for their careful review and helpful suggestions. The authors greatly appreciate Wonyul Lee and Yufeng Liu sharing the glioblastoma multiforme cancer data. The research of Zhihua Su was supported in part by Simons Foundation grant 632688.

References

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

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

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

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

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

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.

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

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association* **107**, 1533–1545.

Chen, X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics* **38**, 3696–3723.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L. et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529–541.

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

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

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

Geng, H., Iqbal, J., Chan, W. C. and Ali, H. H. (2011). Virtual CGH: An integrative approach to predict genetic abnormalities from gene expression microarray data applied in lymphoma. *BMC Medical Genomics* **4**, 32.

Ha, M. J., Stingo, F. C. and Baladandayuthapani, V. (2020). Bayesian structure learning in multi-layered genomic networks. *Journal of the American Statistical Association* **116**, 1–33.

Khare, K., Oh, S.-Y. and Rajaratnam, B. (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**, 803–825.

Kuo, T.-Y., Hsi, E., Yang, I.-P., Tsai, P.-C., Wang, J.-Y. and Juo, S.-H. H. (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 Liu, Y. (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., Wang, Y.-G. and Zhu, M. (2009). Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method. *Biostatistics* **10**, 436–445.

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

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

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

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

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

Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group Lasso estimator for linear models. Technical report 743. Department of Statistics, University of California, Berkeley.

Obozinski, G., Taskar, B. and Jordan, M. (2007). Joint covariate selection for grouped classification. Technical report 743. University of California, Berkeley.

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

Pollack, J. R., Sørlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E. et al. (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**, 12963–12968.

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

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

Shea, A., Harish, V., Afzal, Z., Chijioke, J., Kedir, H., Dusmatova, S. et al. (2016). MicroRNAs in glioblastoma multiforme pathogenesis and therapeutics. *Cancer Medicine* **5**, 1917–1946.

Silber, J., Lim, D. A., Petritsch, C., Persson, A. I., Maunakea, A. K., Yu, M. et al. (2008). MiR-124 and miR-137 inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells. *BMC Medicine* **6**, 14.

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

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

Stingo, F., Chen, Y., Vannucci, M., Barrier, M. and Mirkes, P. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. *The Annals of Applied Statistics* **4**, 2024–2028.

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

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

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

Wang, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statistica Sinica* **25**, 831–851.

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

Witten, D. M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534.

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

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

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **68**, 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 Zou, H. (2014). Sparse precision matrix estimation via Lasso penalized D-trace loss. *Biometrika* **101**, 103–120.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Kshitij Khare
Department of Statistics, University of Florida, Gainesville, FL 32611, USA.
E-mail: kdkhare@ufl.edu
Zhihua Su
Department of Statistics, University of Florida, Gainesville, FL 32611, USA.
E-mail: zhihuasu@ufl.edu

(Received April 2022; accepted October 2022)