An $\ell_{2,0}$-norm constrained matrix optimization via extended discrete first-order algorithms

Ryoya Oda[*] [†] **Mineaki Ohishi**[‡]
Yuya Suzuki[§] [¶] **and Hirokazu Yanagihara**[‖]

(Last Modified: October 8, 2021)

Abstract

The present paper is concerned with a constrained matrix optimization problem. The constraint is referred to as the $\ell_{2,0}$-norm of the matrix, which is defined as the number of non-zero row vectors of the matrix. We extend the discrete first-order algorithm by Bertsimas *et al.* (2016) to solve the optimization problem. The extended algorithm is useful for selecting variables in multivariate statistical models. Then, the convergence properties of the extended algorithm are established. In numerical experiments, we apply the extended algorithm to the optimization problem for the multivariate linear regression model. Furthermore, we also incorporate selecting variables using information criteria into the optimization problem.

1 Introduction

Optimization problems exist in multiple areas and are widely required. Among optimization problems, constrained matrix optimization problems are often used in estimating parameters with constraints for multivariate statistical models, such as the multivariate linear regression model [12, 13]. Consider the following constrained matrix optimization problem for a function $f : \mathbb{R}^{k\times p} \to \mathbb{R}$:

$$\min_{\boldsymbol{\Theta}} f(\boldsymbol{\Theta}) \text{ subject to } \|\boldsymbol{\Theta}\|_{2,0} \leq q, \tag{1}$$

where $\|\boldsymbol{\Theta}\|_{2,0}$ is called the $\ell_{2,0}$-norm of a matrix $\boldsymbol{\Theta} \in \mathbb{R}^{k\times p}$ and is defined as

$$\|\boldsymbol{\Theta}\|_{2,0} = \sum_{j=1}^{k} I(\boldsymbol{\theta}_j \neq \mathbf{0}_p), \tag{2}$$

in which $I(\cdot)$ denotes the indicator function, $\boldsymbol{\theta}_j$ is the j-th row vector of $\boldsymbol{\Theta}$, i.e., $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)'$, and $\mathbf{0}_p$ is a p-dimensional vector of zeros. Note that the $\ell_{2,0}$-norm is not a norm in the usual sense because the $\ell_{2,0}$-norm lacks positive scalability: $\|a\boldsymbol{\Theta}\|_{2,0} = |a|\|\boldsymbol{\Theta}\|_{2,0}$ for any $a \in \mathbb{R}$. For

[*]Corresponding author. Email: ryoya-oda@hiroshima-u.ac.jp
[†]School of Informatics and Data Science, Hiroshima University, Higashi-Hiroshima, Japan
[‡]Education and Research Center for Artificial Intelligence and Data Innovation, Hiroshima University, Hiroshima, Japan
[§]Department of Mathematics, Graduate School of Science, Hiroshima University, Higashi-Hiroshima, Japan
[¶]Current address: Network Construction Department, DOCOMO CS SHIKOKU.Inc, Takamatsu, Japan.
[‖]Graduate School of Advanced Science and Engineering, Hiroshima University, Higashi-Hiroshima, Japan

convenience, we refer to $\|\cdot\|_{2,0}$ as the $\ell_{2,0}$-norm. By definition (2), the constraint in (1) explicitly restricts the number of non-zero row vectors of $\boldsymbol{\Theta}$. From the viewpoint of statistical models, such a constraint is useful for selecting variables in estimating parameters because variables corresponding to zero-parameters can often be regarded as redundant in the model. Furthermore, since it may be desirable to find variables that affect all of the responses in multivariate statistical models (e.g., [9, 10, 14, 15]), it is important to consider the constraint in (1) because such variable selection requires the use of a vector constraint, rather than a scalar constraint.

Optimization problems with the $\ell_{2,0}$-norm constraint are non-convex optimization problems and are NP-hard because the $\ell_{2,0}$-norm is nonconvex and discontinuous. Hence, it is generally desired to obtain algorithms to achieve a sensible solution of (1) within a reasonable computation time. In multi-class classification with linear regression, Cai *et al.* [4] considered an efficient algorithm based on the general method of augmented Lagrange multipliers for solving (1) when the objective function is expressed as the $\ell_{2,1}$-norm of a matrix, where the $\ell_{2,1}$-norm is defined as $\|\boldsymbol{A}\|_{2,1} = \sum_{j=1}^{p} \left(\sum_{i=1}^{n} a_{ij}^2\right)^{1/2}$ for a matrix $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{n\times p}$. For general objective functions, Gotoh *et al.* [5] proposed a proximal algorithm for solving (1) when the objective function is represented as a difference of two convex functions. On the other hand, Bertsimas *et al.* [3] developed an algorithm for solving (1) only when $p = 1$ with smooth convex objective functions. Note that when $p = 1$, the $\ell_{2,0}$-norm is equivalent to the ℓ_0-norm of a vector, which is given by $\|\boldsymbol{a}\|_0 = \sum_{j=1}^{k} I(a_j \neq 0)$ for a vector $\boldsymbol{a} = (a_1, \ldots, a_k)' \in \mathbb{R}^k$. Their algorithm is referred to as the discrete first-order algorithm (DFA), and they derived the asymptotic convergence properties and the global convergence results of the DFA. Moreover, they confirmed that a mixed integer optimization initialized with a solution obtained from the DFA realized a near-optimal solution of (1) in numerical studies.

In the present paper, we consider the DFA proposed by Bertsimas *et al.* [3], and we extend the algorithm to solve (1) even when $p \geq 2$. Moreover, in the numerical experiments, we combine the extended DFA and information criteria to select variables for multivariate statistical models. In the framework of multivariate analysis, information criteria are often used to select variables. Examples of such information criteria include Akaike's information criterion (AIC) [1, 2] and the Bayesian information criterion (BIC) [11]. These criteria do not work or are not defined when the number of variables exceeds the sample size unless the candidate set of variables is narrow. However, it is expected that information criteria work because the candidate set becomes narrow by using solutions to problem (1) via the extended DFA.

The remainder of the present paper is organized as follows. In section 2, we present the optimization problem and extend the DFA. In section 3, we obtain several convergence properties of the extended DFA. In section 4, we conduct numerical experiments using the extended DFA and information criteria for selecting variables in the multivariate linear regression model. Technical details are provided in the Appendix.

2 Optimization problem and Algorithm

2.1 $\ell_{2,0}$-norm constrained optimization problem

Suppose that the function $g : \mathbb{R}^{k\times p} \to \mathbb{R}$ is bounded below, is convex, and has a Lipschitz continuous gradient with constant ℓ, i.e., there exists a constant $\ell > 0$ for all $\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}} \in \mathbb{R}^{k\times p}$,

$$\|\boldsymbol{D}(\boldsymbol{\Theta}) - \boldsymbol{D}(\tilde{\boldsymbol{\Theta}})\|_F \leq \ell\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|_F, \tag{3}$$

where $\boldsymbol{D}(\boldsymbol{\Theta}) \in \mathbb{R}^{k\times p}$ is a matrix that is based on partial derivatives, i.e., $\boldsymbol{D}(\boldsymbol{\Theta}) = \partial g(\boldsymbol{\Theta})/\partial\boldsymbol{\Theta}$, and $\|\cdot\|_F$ is the Frobenius norm of a matrix that is given as $\|\boldsymbol{\Theta}\|_F = \mathrm{tr}(\boldsymbol{\Theta}'\boldsymbol{\Theta})^{1/2}$ for a matrix $\boldsymbol{\Theta}$. The function g defined in (3) was used by Bertsimas *et al.* [3]. Then, we consider the following $\ell_{2,0}$-norm constrained optimization problem for the function g:

$$\min_{\boldsymbol{\Theta}} g(\boldsymbol{\Theta}) \text{ subject to } \boldsymbol{\Theta} = (\boldsymbol{\mathcal{B}}', \boldsymbol{\Xi}')', \ \|\boldsymbol{\Xi}\|_{2,0} \leq q, \tag{4}$$

where $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{k_1\times p}$ and $\boldsymbol{\Xi} \in \mathbb{R}^{k_2\times p}$ are the partitioned matrices of $\boldsymbol{\Theta} \in \mathbb{R}^{k\times p}$, and k_1 and k_2 satisfy $k_1+k_2 = k$. Problem (4) restricts the number of non-zero vectors of $\boldsymbol{\Xi}$, but $\boldsymbol{\mathcal{B}}$ is optimized without constraints. Thus, (4) includes the following optimization problem:

$$\min_{\boldsymbol{\Theta}} g(\boldsymbol{\Theta}) \text{ subject to } \|\boldsymbol{\Theta}\|_{2,0} \leq q, \tag{5}$$

because problem (4) can be regarded as (5) by letting $k_1 = 0$ or $k_2 = k$. Problem (4) is useful for estimating parameters without constraints for a part of $\boldsymbol{\Theta}$ (e.g., the parameter corresponding to the intercept term) in multivariate statistical models. An example of (4) in the following multivariate statistical model is presented.

Example (Multivariate linear regression model). Suppose that $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)' \in \mathbb{R}^{n\times p}$ is an observation matrix stacking individual p response variables and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n\times(k-1)}$ is an observation matrix stacking individual $k-1$ explanatory variables, where n is the sample size. We assume that the column vectors of $\boldsymbol{X}$ have unit ℓ_2-norm, i.e., $\|\boldsymbol{x}_j\|_2 = 1$ $(j = 1, \ldots, k-1)$, where $\|\cdot\|_2$ is the ℓ_2-norm of vector, which is defined as $\|\boldsymbol{a}\|_2 = (\boldsymbol{a}'\boldsymbol{a})^{1/2}$ for a vector $\boldsymbol{a}$. Moreover, we assume that the intercept term is included in this model. Hence, let $\boldsymbol{Z} = (\boldsymbol{1}_n, \boldsymbol{X}) \in \mathbb{R}^{n\times k}$ be the matrix including the intercept term, where $\boldsymbol{1}_n$ is an n-dimensional vector of ones. Then, the residual sum of squares is widely used in estimating parameters:

$$g(\boldsymbol{\Theta}) = g_1(\boldsymbol{\Theta}) = \frac{1}{2n}\|(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\Theta})\boldsymbol{G}^{-1/2}\|_F^2, \tag{6}$$

where $\boldsymbol{G}$ is a positive definite matrix. Note that the intercept term does not vanish, because of non-centralizing of the column vectors of $\boldsymbol{Y}$ and $\boldsymbol{X}$. When the constraint for the intercept term is not set, we can apply (6) to problem (4) when $k_1 = 1$ and $k_2 = k-1$. Moreover, we observe that $\boldsymbol{D}(\boldsymbol{\Theta}) = -n^{-1}\boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\Theta})\boldsymbol{G}^{-1}$ and that one value of ℓ is $n^{-1}\lambda_{\max}(\boldsymbol{Z}'\boldsymbol{Z})/\lambda_{\min}(\boldsymbol{G})$, where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ are the maximum and minimum eigenvalues, respectively, of a square matrix.

2.2 Extended discrete first-order algorithm

We extend the DFA proposed by Bertsimas *et al.* [3] to solve (4). First of all, for a given $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{k_2})' \in \mathbb{R}^{k_2 \times p}$, we consider the following optimization problem:

$$\min_{\boldsymbol{\Xi}} \|\boldsymbol{\Xi} - \boldsymbol{C}\|_F^2 \text{ subject to } \|\boldsymbol{\Xi}\|_{2,0} \leq q. \tag{7}$$

Let $I_q(\boldsymbol{C})$ be the set consisting of suffixes of the q largest row vectors of $\boldsymbol{C}$ in the sense of ℓ_2-norm, i.e.,

$$I_q(\boldsymbol{C}) = \{1 \leq j \leq k_2 \mid \|\boldsymbol{c}_j\|_2 \text{ are among the largest } q \text{ of all } \|\boldsymbol{c}_j\|_2 \text{ s}\}. \tag{8}$$

Then, the optimal solutions of (7) can be derived in closed form, as is shown in the following proposition. (The proof is given in Appendix A.)

Proposition 1. Let $\hat{\boldsymbol{\Xi}} = (\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_{k_2})'$ be an optimal solution to problem (7). Then, $\hat{\boldsymbol{\Xi}}$ is given by

$$\hat{\boldsymbol{\xi}}_j = \begin{cases} \boldsymbol{c}_j & (j \in I_q(\boldsymbol{C})) \\ \boldsymbol{0}_p & (otherwise) \end{cases}, \tag{9}$$

where $I_q(\boldsymbol{C})$ is defined in (8). We denote the set of optimal solutions (9) as $\mathbf{H}_q(\boldsymbol{C})$.

Note that $\mathbf{H}_q(\boldsymbol{C})$ is expressed as the set of solutions because problem (7) may have some optimal solutions. The DFA is based on projected gradient decent methods in first-order convex optimization problems (see [7, 8]). The following proposition gives an upper bound of g and its minimizer with constraints. (The proof is given in Appendix B.)

Proposition 2. Let g be the function defined in (3). Then, for any $L \geq \ell$, we have

$$g(\tilde{\boldsymbol{\Theta}}) \leq Q_L(\tilde{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) = g(\boldsymbol{\Theta}) + \frac{L}{2}\|\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F^2 + \operatorname{tr}\left\{\boldsymbol{D}'(\boldsymbol{\Theta})(\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\right\}, \tag{10}$$

for all $\boldsymbol{\Theta} = (\boldsymbol{\mathcal{B}}', \boldsymbol{\Xi}')' \in \mathbb{R}^{k \times p}$ and $\tilde{\boldsymbol{\Theta}} = (\tilde{\boldsymbol{\mathcal{B}}}', \tilde{\boldsymbol{\Xi}}')' \in \mathbb{R}^{k \times p}$ ($\boldsymbol{\mathcal{B}}, \tilde{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{k_1 \times p}$; $\boldsymbol{\Xi}, \tilde{\boldsymbol{\Xi}} \in \mathbb{R}^{k_2 \times p}$). Moreover, the optimal solution $\boldsymbol{\Theta}_\dagger = (\boldsymbol{\mathcal{B}}_\dagger', \boldsymbol{\Xi}_\dagger')'$ ($\boldsymbol{\mathcal{B}}_\dagger \in \mathbb{R}^{k_1 \times p}$, $\boldsymbol{\Xi}_\dagger \in \mathbb{R}^{k_2 \times p}$) to $\min_{\tilde{\boldsymbol{\Theta}}:\ \|\tilde{\boldsymbol{\Xi}}\|_{2,0} \leq q} Q_L(\tilde{\boldsymbol{\Theta}}, \boldsymbol{\Theta})$ is given by

$$\boldsymbol{\mathcal{B}}_\dagger = \boldsymbol{\mathcal{B}} - L^{-1}\boldsymbol{D}_1(\boldsymbol{\Theta}),\ \boldsymbol{\Xi}_\dagger \in \mathbf{H}_q\left(\boldsymbol{\Xi} - L^{-1}\boldsymbol{D}_2(\boldsymbol{\Theta})\right), \tag{11}$$

where $\mathbf{H}_q(\cdot)$ is defined in (9) and $\boldsymbol{D}_1(\boldsymbol{\Theta}) \in \mathbb{R}^{k_1 \times p}$ and $\boldsymbol{D}_2(\boldsymbol{\Theta}) \in \mathbb{R}^{k_2 \times p}$ are the partitioned matrices of $\boldsymbol{D}(\boldsymbol{\Theta})$, i.e., $\boldsymbol{D}(\boldsymbol{\Theta}) = (\boldsymbol{D}_1'(\boldsymbol{\Theta}), \boldsymbol{D}_2'(\boldsymbol{\Theta}))'$.

Using (11), we extend the DFA proposed by Bertsimas *et al.* [3] to solve (4), which is presented as Algorithm 1. We observe that Algorithm 1 for the parameter without constraints behaves like a vanilla gradient decent algorithm. Moreover, note that Algorithm 1 corresponds to the DFA proposed by Bertsimas *et al.* [3] when $p = 1$ and $k_1 = 0$.

Algorithm 1 Extended discrete first-order algorithm to solve (4)

Require: An initial value $\boldsymbol{\Theta}_1 = (\boldsymbol{\mathcal{B}}_1', \boldsymbol{\Xi}_1')'$ satisfying $\|\boldsymbol{\Xi}_1\|_{2,0} \leq q$, a constant L ($\geq \ell$), and a small value $\varepsilon > 0$.

$m = 1$.

repeat

Obtain $\boldsymbol{\Theta}_{m+1} = (\boldsymbol{\mathcal{B}}_{m+1}', \boldsymbol{\Xi}_{m+1}')'$ from (11) as follows:

$$\boldsymbol{\mathcal{B}}_{m+1} = \boldsymbol{\mathcal{B}}_m - L^{-1}\boldsymbol{D}_1(\boldsymbol{\Theta}_m),\ \boldsymbol{\Xi}_{m+1} \in \mathbf{H}_q\left(\boldsymbol{\Xi}_m - L^{-1}\boldsymbol{D}_2(\boldsymbol{\Theta}_m)\right). \tag{12}$$

Increment m by 1.

until $g(\boldsymbol{\Theta}_m) - g(\boldsymbol{\Theta}_{m+1}) < \varepsilon$ holds.

3 Convergence properties of Algorithm 1

We present several convergence properties of Algorithm 1. First, we define a notion of first-order optimality for problem (4).

Definition 1. For $L \geq \ell$, $\tilde{\boldsymbol{\Theta}} = (\tilde{\boldsymbol{\mathcal{B}}}', \tilde{\boldsymbol{\Xi}}')'$ ($\tilde{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{k_1 \times p}, \tilde{\boldsymbol{\Xi}} \in \mathbb{R}^{k_2 \times p}$) is said to be an $\ell_{2,0}$-constrained first-order stationary point of problem (4) if $\|\tilde{\boldsymbol{\Xi}}\|_{2,0} \leq q$ holds and $\tilde{\boldsymbol{\Theta}}$ satisfies the following equation:

$$\tilde{\boldsymbol{\mathcal{B}}} = \tilde{\boldsymbol{\mathcal{B}}} - L^{-1}\boldsymbol{D}_1(\tilde{\boldsymbol{\Theta}}),\ \tilde{\boldsymbol{\Xi}} \in \mathbf{H}_q\left(\tilde{\boldsymbol{\Xi}} - L^{-1}\boldsymbol{D}_2(\tilde{\boldsymbol{\Theta}})\right). \tag{13}$$

If $\tilde{\boldsymbol{\Theta}}$ is an $\ell_{2,0}$-constrained first-order stationary point, we have $\boldsymbol{D}_1(\tilde{\boldsymbol{\Theta}}) = \boldsymbol{O}_{k_1,p}$, where $\boldsymbol{O}_{k_1,p} \in \mathbb{R}^{k_1 \times p}$ is the matrix, the elements of which are zero. Moreover, letting $\tilde{\boldsymbol{\Xi}} = (\tilde{\boldsymbol{\xi}}_1, \ldots, \tilde{\boldsymbol{\xi}}_{k_2})'$, it holds that $\tilde{\boldsymbol{\xi}}_j = \tilde{\boldsymbol{\xi}}_j - L^{-1}\boldsymbol{d}_j(\tilde{\boldsymbol{\Theta}})$ for $j \in I_q(\tilde{\boldsymbol{\Xi}} - L^{-1}\boldsymbol{D}_2(\tilde{\boldsymbol{\Theta}}))$, where $\boldsymbol{d}_j(\tilde{\boldsymbol{\Theta}})$ is the j-th row vector of $\boldsymbol{D}_2(\tilde{\boldsymbol{\Theta}})$, i.e., $\boldsymbol{D}_2(\tilde{\boldsymbol{\Theta}}) = (\boldsymbol{d}_1(\tilde{\boldsymbol{\Theta}}), \ldots, \boldsymbol{d}_{k_2}(\tilde{\boldsymbol{\Theta}}))'$. Hence, we have $\boldsymbol{d}_j(\tilde{\boldsymbol{\Theta}}) = \mathbf{0}_p$ for $j \in I_q(\tilde{\boldsymbol{\Xi}} - L^{-1}\boldsymbol{D}_2(\tilde{\boldsymbol{\Theta}}))$. The following proposition is concerned with a sufficient condition for a global minimizer to the unconstrained optimization problem $\min_{\boldsymbol{\Theta}} g(\boldsymbol{\Theta})$. (The proof is given in Appendix C.)

Proposition 3. If $\tilde{\boldsymbol{\Theta}} = (\tilde{\boldsymbol{\mathcal{B}}}', \tilde{\boldsymbol{\Xi}}')'$ satisfies (13) and $\|\tilde{\boldsymbol{\Xi}}\|_{2,0} < q$, then we have $\tilde{\boldsymbol{\Theta}} \in \arg\min_{\boldsymbol{\Theta}} g(\boldsymbol{\Theta})$.

Next, we give several asymptotic convergence properties of Algorithm 1. To do so, we make several definitions for notational convenience. Let $\boldsymbol{\Theta}_m = (\boldsymbol{\mathcal{B}}_m', \boldsymbol{\Xi}_m')'$ be the m-iterated solution in (12) by Algorithm 1, and let $\boldsymbol{\Xi}_m = (\boldsymbol{\xi}_{m,1}, \ldots, \boldsymbol{\xi}_{m,k_2})'$. Moreover, let $\boldsymbol{r}_m = (r_{m,1}, \ldots, r_{m,k_2})'$ be the k_2-dimensional vector satisfying $r_{m,j} = I(\boldsymbol{\xi}_{m,j} \neq \mathbf{0}_p)$. Denote as $\alpha_{m,q} = \|\boldsymbol{\xi}_{m,(q)}\|_2$ the ℓ_2-norm of the q-th largest row vector of $\boldsymbol{\Xi}_m$ in the sense of $\|\boldsymbol{\xi}_{m,(1)}\|_2 \geq \cdots \geq \|\boldsymbol{\xi}_{m,(k_2)}\|_2$. Using $\alpha_{m,q}$, we define $\bar{\alpha}_q = \limsup_{m\to\infty} \alpha_{m,q}$ and $\underline{\alpha}_q = \liminf_{m\to\infty} \alpha_{m,q}$. Then, we present the several asymptotic convergence properties of Algorithm 1 as the following proposition. (The proof is given in Appendix D.)

Proposition 4. For problem (4), let $\boldsymbol{\Theta}_m$ be the m-iterated solution in (12) by Algorithm 1. Then, the following properties of Algorithm 1 hold:

(a) Let $L \geq \ell$. Then, we have

$$g(\boldsymbol{\Theta}_m) - g(\boldsymbol{\Theta}_{m+1}) \geq \frac{L-\ell}{2}\|\boldsymbol{\Theta}_{m+1} - \boldsymbol{\Theta}_m\|_F^2. \tag{14}$$

Moreover, $g(\boldsymbol{\Theta}_m)$ monotonically decreases for m and converges as $m \to \infty$.

(b) For any $L > \ell$, it holds that $\boldsymbol{\Theta}_{m+1} - \boldsymbol{\Theta}_m \to \boldsymbol{O}_{k,p}$ $(m \to \infty)$.

(c) Let $L > \ell$, and $\underline{\alpha}_q > 0$. Then, there exists $M > 0$ such that for all $m \geq M$, $\boldsymbol{r}_m = \boldsymbol{r}_{m+1}$. Furthermore, the sequence $\{\boldsymbol{\Theta}_m\}$ converges to an $\ell_{2,0}$-constrained first-order stationary point.

(d) Let $L > \ell$. Then, we have $\lim_{m\to\infty} \boldsymbol{D}_1(\boldsymbol{\Theta}_m) = \boldsymbol{O}_{k_1,p}$. Furthermore, if $\underline{\alpha}_q = 0$, it holds that $\liminf_{m\to\infty} \max_{j=1,\ldots,k_2} \|\boldsymbol{d}_j(\boldsymbol{\Theta}_m)\|_2 = 0$.

(e) Let $L > \ell$. If $\bar{\alpha}_q = 0$ and the sequence $\{\boldsymbol{\Theta}_m\}$ has a limit point, then $g(\boldsymbol{\Theta}_m) \to \min_{\boldsymbol{\Theta}} g(\boldsymbol{\Theta})$ $(m \to \infty)$.

The stopping rule of Algorithm 1 is based on (a) of Proposition 4. From (c), if the q-th largest vector $\boldsymbol{\xi}_{m,(q)}$ is non-zero for sufficiently large m, then the suffixes of the non-zero vectors of $\boldsymbol{\Xi}_m$ are fixed after that. Moreover, (c) ensures the global convergence to an $\ell_{2,0}$-constrained first-order stationary point of Algorithm 1. From (e), the objective function g converges to an optimal value for unconstrained optimization problem $\min_{\boldsymbol{\Theta}} g(\boldsymbol{\Theta})$ under minor assumptions.

Finally, we refer to the $\ell_{2,0}$-constrained first-order stationary point and a rate of convergence of Algorithm 1. The following proposition is concerned with some properties of the $\ell_{2,0}$-constrained first-order stationary point. (The proof is given in Appendix E.)

Proposition 5. For $L \geq \ell$, the following properties hold:

(a) If $\tilde{\boldsymbol{\Theta}} = (\tilde{\boldsymbol{\mathcal{B}}}', \tilde{\boldsymbol{\Xi}}')'$ $(\tilde{\boldsymbol{\mathcal{B}}} \in \mathbb{R}^{k_1\times p}, \tilde{\boldsymbol{\Xi}} \in \mathbb{R}^{k_2\times p})$ is an $\ell_{2,0}$-constrained first-order stationary point in Definition 1, then $\mathbf{H}_q(\tilde{\boldsymbol{\Xi}} - L^{-1}\boldsymbol{D}_2(\tilde{\boldsymbol{\Theta}}))$ has exactly one element.

(b) Global minimizers of problem (4) are $\ell_{2,0}$-constrained first-order stationary points.

The following theorem presents knowledge about the rate of convergence of Algorithm 1. (The proof is given in Appendix F.)

Theorem 1. Let $L \geq \ell$. Then, Algorithm 1 iterated M times satisfies

$$\min_{m=1,\ldots,M} \|\boldsymbol{\Theta}_{m+1} - \boldsymbol{\Theta}_m\|_F^2 \leq \frac{2\{g(\boldsymbol{\Theta}_1) - g_*\}}{M(L-\ell)},$$

where $g(\boldsymbol{\Theta}_m) \downarrow g_*$ as $m \to \infty$.

The result of Theorem 1 is an extension of Theorem 3.1 of Bertsimas *et al.* [3] and coincides with it when $p = 1$ and $k_1 = 0$.

4 Numerical Studies

We conduct numerical experiments based on Algorithm 1 for the $\ell_{2,0}$-norm constrained optimization problem (4) in terms of variable selection for the multivariate linear regression model (see Example). Denote the $n \times p$ multivariate normal distribution with mean matrix $\boldsymbol{A}$ and covariance matrix $\boldsymbol{B}$ as $N_{n\times p}(\boldsymbol{A}, \boldsymbol{B})$. The explanatory matrix $\boldsymbol{X}$, the true parameter $\boldsymbol{\beta}_*$ corresponding to the intercept term, and $\boldsymbol{\Xi}_*$ were determined as follows:

$$\boldsymbol{X} \sim N_{n\times k}(\boldsymbol{O}_{n,k}, \boldsymbol{\Psi} \otimes \boldsymbol{I}_n)\ , \boldsymbol{\Theta}_* = (\boldsymbol{\beta}_*, \boldsymbol{\Xi}_*', \boldsymbol{O}_{k-k_*,p}')',$$
$$\boldsymbol{\beta}_* \sim N_{p\times 1}(5\mathbf{1}_p, \boldsymbol{I}_p \otimes 1),\ \boldsymbol{\Xi}_* \sim N_{k_*\times p}(5\mathbf{1}_p\mathbf{1}_p', \boldsymbol{I}_p \otimes \boldsymbol{I}_{k_*}),$$

where $\boldsymbol{I}_p \in \mathbb{R}^{p\times p}$ is the identity matrix, the (a,b)-th element of $\boldsymbol{\Psi}$ is $(0.5)^{|a-b|}$, and k_* is the number of non-zero row vectors of $\boldsymbol{\Xi}_*$. Note that $\boldsymbol{X}$, $\boldsymbol{\beta}_*$, and $\boldsymbol{\Xi}_*$ were generated only once and were used throughout the simulation studies. Then, we made the column vectors of $\boldsymbol{X}$ have unit ℓ_2-norm. Using the notation in Example, the response matrix $\boldsymbol{Y}$ was generated by $\boldsymbol{Y} = (\boldsymbol{y}_{(1)}, \ldots, \boldsymbol{y}_{(p)}) \sim N_{n\times p}(\boldsymbol{Z}\boldsymbol{\Theta}_*, \boldsymbol{\Sigma}\otimes\boldsymbol{I}_n)$, where $\boldsymbol{Z} = (\boldsymbol{1}_n, \boldsymbol{X})$ and $\boldsymbol{\Sigma} = 0.4\{(1-0.8)\boldsymbol{I}_p + 0.8\boldsymbol{1}_p\boldsymbol{1}_p'\}$. Then, we made the column vectors of $\boldsymbol{Y}$ have unit ℓ_2-norm.

Since Algorithm 1 gives a solution for fixed q, we evaluate the best estimator by combining Algorithm 1 and information criteria, which are used to select variables. We performed the following steps:

Step 1. Give a value $\hat{\boldsymbol{\Theta}}_{*,k_2} = (\hat{\boldsymbol{\mathcal{B}}}'_{*,k_2}, \hat{\boldsymbol{\Xi}}'_{*,k_2})'$ $(\hat{\boldsymbol{\mathcal{B}}}_{*,k_2} \in \mathbb{R}^{k_1\times p}, \hat{\boldsymbol{\Xi}}_{*,k_2} \in \mathbb{R}^{k_2\times p})$ satisfying $\|\hat{\boldsymbol{\Xi}}_{*,k_2}\|_{2,0} \leq k_2$ and set $q = k_2 - 1$.

Step 2. Give a value $\tilde{\boldsymbol{\Theta}}_{*,q} = (\tilde{\boldsymbol{\mathcal{B}}}'_{*,q}, \tilde{\boldsymbol{\Xi}}'_{*,q})'$ $(\tilde{\boldsymbol{\mathcal{B}}}_{*,q} \in \mathbb{R}^{k_1\times p}, \tilde{\boldsymbol{\Xi}}_{*,q} \in \mathbb{R}^{k_2\times p})$ satisfying $\|\tilde{\boldsymbol{\Xi}}_{*,q}\|_{2,0} \leq q$.

Step 3. For the given q, obtain the solution $\hat{\boldsymbol{\Theta}}_{*,q} = (\hat{\boldsymbol{\mathcal{B}}}'_{*,q}, \hat{\boldsymbol{\Xi}}'_{*,q})'$ $(\hat{\boldsymbol{\mathcal{B}}}_{*,q} \in \mathbb{R}^{k_1\times p}, \hat{\boldsymbol{\Xi}}_{*,q} \in \mathbb{R}^{k_2\times p})$ by Algorithm 1 for the initial value $\boldsymbol{\Theta}_1 = \tilde{\boldsymbol{\Theta}}_{*,q}$. Then, decrement q by 1.

Step 4. Repeat Steps 2 and 3 until $q = 0$.

Step 5. Decide the best selection number as $\hat{k}_* = \arg\min_{q=0,\ldots,k_2} \mathrm{IC}(\hat{\boldsymbol{\Theta}}_{*,q})$ and obtain the best estimator by $\hat{\boldsymbol{\Theta}}_* = \hat{\boldsymbol{\Theta}}_{*,\hat{k}_*}$, where $\mathrm{IC}(\cdot)$ is an information criterion.

In the above steps and Algorithm 1, $k_1 = 1$, $\hat{\boldsymbol{\Theta}}_{*,k_2} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$ and $\varepsilon = 10^{-4}$, and the value $\tilde{\boldsymbol{\Theta}}_{*,q}$ in Step 2 is given as follows:

Step 2-1. Denote $\mathcal{A}_q = \{a_1, \ldots, a_{q+1}\}$ $(a_1 < \cdots < a_{q+1})$ as the active set of $\hat{\boldsymbol{\Xi}}_{*,q+1}$ defined by $\{k_1 + 1 \leq j \leq k \mid \hat{\boldsymbol{\theta}}_{*,q+1,j} \neq \boldsymbol{0}_p\}$, where $\hat{\boldsymbol{\theta}}_{*,q+1,j}$ is the j-th row vector of $\hat{\boldsymbol{\Theta}}_{*,q+1}$ in Step 3.

Step 2-2. Set $\bar{\mathcal{A}}_q = \{1, \ldots, k_1\} \cup \mathcal{A}_q$ and

$$(\bar{\boldsymbol{\mathcal{B}}}'_q, \bar{\boldsymbol{\Xi}}'_q)' = \left(\boldsymbol{Z}'_{\bar{\mathcal{A}}_q}\boldsymbol{Z}_{\bar{\mathcal{A}}_q}\right)^{-1}\boldsymbol{Z}'_{\bar{\mathcal{A}}_q}\boldsymbol{Y} \ (\bar{\boldsymbol{\mathcal{B}}}_q \in \mathbb{R}^{k_1\times p}, \bar{\boldsymbol{\Xi}}_q \in \mathbb{R}^{|\mathcal{A}_q|\times p}),$$

where $\boldsymbol{Z}_{\bar{\mathcal{A}}_q}$ is the $n \times |\bar{\mathcal{A}}_q|$ matrix consisting of columns of $\boldsymbol{Z}$ indexed by the elements of $\bar{\mathcal{A}}_q$. Furthermore, denote the j-th row vectors of $\bar{\boldsymbol{\mathcal{B}}}_q$ and $\bar{\boldsymbol{\Xi}}_q$ as $\bar{\boldsymbol{\beta}}_{q,j}$ and $\bar{\boldsymbol{\xi}}_{q,a_j}$, respectively.

Step 2-3. Give the j-th row vector of $\tilde{\boldsymbol{\Theta}}_{*,q}$ used in Step 2 as follows:

$$\begin{cases} \bar{\boldsymbol{\beta}}_{q,j} & (1 \leq j \leq k_1) \\ \bar{\boldsymbol{\xi}}_{q,a_j} & ((j \in \mathcal{A}_q) \wedge (j \neq a_{q,\min})) \\ \boldsymbol{0}_p & \big((j \in \{k_1+1, \ldots, k\} \cap \mathcal{A}_q^c) \vee (j = a_{q,\min})\big) \end{cases},$$

where $a_{q,\min} = \arg\min_{j\in\mathcal{A}_q} \|\bar{\boldsymbol{\xi}}_{q,a_j}\|_2$.

Furthermore, the BIC proposed by Schwarz [11] was used as an information criterion and is defined by

$$\mathrm{IC}(\boldsymbol{\Theta}) = n\log\left|n^{-1}(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\Theta})'(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\Theta})\right| + p\|\boldsymbol{\Theta}\|_{2,0}\log n.$$

For problem (4), we set $g = g_1$, $\boldsymbol{G} = (n-k)^{-1}\boldsymbol{Y}'\{\boldsymbol{I}_n - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\}\boldsymbol{Y}$ and $L = \lceil n^{-1}\lambda_{\max}(\boldsymbol{Z}'\boldsymbol{Z})/\lambda_{\min}(\boldsymbol{G})\rceil$, where $\lceil\cdot\rceil$ is the ceiling function. For these settings, the above steps were carried out for 1,000 simulation iterations.

In these numerical studies, we examine the following properties.

- The two relative mean square errors (RMSEs):

$$\mathrm{RMSE}_{\boldsymbol{\Theta}_*} = \frac{E[\|\boldsymbol{\Theta}_* - \hat{\boldsymbol{\Theta}}_*\|_F^2]}{E[\|\boldsymbol{\Theta}_* - \hat{\boldsymbol{\Theta}}_{*,k_2}\|_F^2]} \times 100\ (\%),$$

$$\mathrm{RMSE}_{\boldsymbol{Y}} = \frac{E[\|\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\Theta}}_*\|_F^2]}{E[\|\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\Theta}}_{*,k_2}\|_F^2]} \times 100\ (\%).$$

  In our numerical settings, $E[\|\boldsymbol{\Theta}_* - \hat{\boldsymbol{\Theta}}_{*,k_2}\|_F^2] = \mathrm{tr}\{(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\}\mathrm{tr}(\boldsymbol{\Sigma})$ and $E[\|\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\Theta}}_{*,k_2}\|_F^2] = (n-k)\mathrm{tr}(\boldsymbol{\Sigma})$. These RMSEs are approximated by the average value of 1,000 simulation iterations. Note that the smaller the $\mathrm{RMSE}_{\boldsymbol{\Theta}_*}$ and $\mathrm{RMSE}_{\boldsymbol{Y}}$, the better the accuracy of the estimation of $\boldsymbol{\Theta}_*$ and the prediction accuracy of $\boldsymbol{Y}$, respectively.

- The probability (%) such that the suffixes of the non-zero vectors of $\hat{\boldsymbol{\Theta}}_*$ and $\boldsymbol{\Theta}_*$ are equivalent among 1,000 simulation iterations.

- The CPU time (s) obtained as the average value of 1,000 simulation iterations.

Table 1. Properties of the estimation results for $\boldsymbol{\Theta}_*$ by the combination of the extended DFA and the BIC in the multivariate linear regression model.

n	k	$\mathrm{RMSE}_{\boldsymbol{\Theta}_*}$	$\mathrm{RMSE}_{\boldsymbol{Y}}$	Probability	CPU time
100	20	47.01	14.69	95.2	0.848
100	40	17.26	20.83	90.1	1.187
100	60	8.61	31.88	85.1	1.477
100	80	4.30	77.50	76.9	2.411
300	20	51.28	4.02	99.0	0.125
300	40	23.36	4.62	97.8	0.150
300	60	14.43	4.69	96.7	0.214
300	120	5.43	6.67	95.4	0.422
300	180	2.42	10.00	93.7	0.872
300	240	1.08	22.50	88.9	2.167
500	20	52.73	2.34	98.8	0.032
500	40	23.99	2.45	99.2	0.059
500	100	8.59	2.81	98.4	0.220
500	200	3.17	4.17	96.8	0.544
500	300	1.39	5.63	96.1	1.331
500	400	0.67	13.75	88.7	2.423

Table 1 shows the above properties when we set $p = 3$ and $k_* = 10$. From Table 1, we observe that both the RMSEs ($\mathrm{RMSE}_{\boldsymbol{\Theta}_*}$ and $\mathrm{RMSE}_{\boldsymbol{Y}}$) are smaller than 100. This means that

the estimator $\hat{\boldsymbol{\Theta}}_*$ by the combination of the extended DFA and the BIC is better than the least squares estimator $\hat{\boldsymbol{\Theta}}_{*,k_2}$ in terms of the accuracy of the estimation of $\boldsymbol{\Theta}_*$ and the prediction accuracy of $\boldsymbol{Y}$. Moreover, the probabilities are high and the CPU times are short. Therefore, we can confirm that the combination of the extended DFA and the BIC is valid for the estimation of $\boldsymbol{\Theta}_*$.

Appendix

A Proof of Proposition 1

Since the Frobenius norm is invariant to the exchange of rows, without loss of generality, the given $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{k_2})'$ in (7) can be regarded as $\|\boldsymbol{c}_1\|_2 \geq \cdots \geq \|\boldsymbol{c}_{k_2}\|_2$. Problem (7) can be rewritten as

$$\min_{\|\boldsymbol{\Xi}\|_{2,0} \leq q} \sum_{j=1}^{k_2} \|\boldsymbol{\xi}_j - \boldsymbol{c}_j\|_2^2.$$

From the above expression, we can see that the optimal solution for (7) is limited to the case in which q row vectors of $\boldsymbol{\Xi}$ become $\boldsymbol{c}_j$ and the $(k_2 - q)$ remainder becomes $\boldsymbol{0}_p$. On the other hand, let $\mathcal{S} = \{1 \leq j \leq k_2 \mid \boldsymbol{\xi}_j \neq \boldsymbol{0}_p\}$. Then, we have

$$\min_{\|\boldsymbol{\Xi}\|_{2,0} \leq q} \sum_{j=1}^{k_2} \|\boldsymbol{\xi}_j - \boldsymbol{c}_j\|_2^2 = \min_{\|\boldsymbol{\Xi}\|_{2,0} \leq q} \left\{ \sum_{j \in \mathcal{S}} \|\boldsymbol{\xi}_j - \boldsymbol{c}_j\|_2^2 + \sum_{j \notin \mathcal{S}} \|\boldsymbol{c}_j\|_2^2 \right\}.$$

Since the above problem is optimal when $\boldsymbol{\xi}_j = \boldsymbol{c}_j$ for $j \in \mathcal{S}$ and $\sum_{j \notin \mathcal{S}} \|\boldsymbol{c}_j\|_2^2$ is minimum, we can see that $\mathcal{S} = \{1, \ldots, q\}$. □

B Proof of Proposition 2

First, we show (10). Let $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta}')$ and $\tilde{\boldsymbol{\theta}} = \mathrm{vec}(\tilde{\boldsymbol{\Theta}}')$ for any $\boldsymbol{\Theta}, \tilde{\boldsymbol{\Theta}} \in \mathbb{R}^{k \times p}$. Then, by rewriting $g(\tilde{\boldsymbol{\Theta}})$ as $h(\tilde{\boldsymbol{\theta}})$, the following inequality can be derived (see, e.g., [7]):

$$h(\tilde{\boldsymbol{\theta}}) \leq h(\boldsymbol{\theta}) + \frac{L}{2}\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 + (\partial h(\boldsymbol{\theta})/\partial \boldsymbol{\theta})' (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}).$$

From the properties of the vec operator and the Frobenius norm, the above inequality can be expressed as

$$g(\tilde{\boldsymbol{\Theta}}) \leq g(\boldsymbol{\Theta}) + \frac{L}{2}\|\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F^2 + \mathrm{tr}\left\{ \boldsymbol{D}'(\boldsymbol{\Theta})(\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \right\}.$$

Next, we show (11). The following equation can be derived:

$$\begin{aligned} Q_L(\tilde{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) &= \frac{L}{2} \left\| \tilde{\boldsymbol{\Theta}} - \left(\boldsymbol{\Theta} - L^{-1} \boldsymbol{D}(\boldsymbol{\Theta}) \right) \right\|_F^2 - \frac{1}{2L} \|\boldsymbol{D}(\boldsymbol{\Theta})\|_F^2 + g(\boldsymbol{\Theta}) \\ &= \frac{L}{2} \left\| \tilde{\boldsymbol{\mathcal{B}}} - \left(\boldsymbol{\mathcal{B}} - L^{-1} \boldsymbol{D}_1(\boldsymbol{\Theta}) \right) \right\|_F^2 + \frac{L}{2} \left\| \tilde{\boldsymbol{\Xi}} - \left(\boldsymbol{\Xi} - L^{-1} \boldsymbol{D}_2(\boldsymbol{\Theta}) \right) \right\|_F^2 \\ &\quad - \frac{1}{2L} \|\boldsymbol{D}(\boldsymbol{\Theta})\|_F^2 + g(\boldsymbol{\Theta}). \end{aligned}$$

Hence, the optimal solution to $\min_{\tilde{\boldsymbol{\Theta}}:\ \|\tilde{\boldsymbol{\Xi}}\|_{2,0}\leq q} Q_L(\tilde{\boldsymbol{\Theta}}, \boldsymbol{\Theta})$, is derived as follows:

$$\begin{aligned}\min_{\tilde{\boldsymbol{\Theta}}:\ \|\tilde{\boldsymbol{\Xi}}\|_{2,0}\leq q} Q_L(\tilde{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) = \frac{L}{2} \min_{\tilde{\boldsymbol{\mathcal{B}}}} \left\| \tilde{\boldsymbol{\mathcal{B}}} - \left(\boldsymbol{\mathcal{B}} - L^{-1}\boldsymbol{D}_1(\boldsymbol{\Theta})\right)\right\|_F^2 \\ + \frac{L}{2} \min_{\|\tilde{\boldsymbol{\Xi}}\|_{2,0}\leq q} \left\| \tilde{\boldsymbol{\Xi}} - \left(\boldsymbol{\Xi} - L^{-1}\boldsymbol{D}_2(\boldsymbol{\Theta})\right)\right\|_F^2 \\ - \frac{1}{2L}\|\boldsymbol{D}(\boldsymbol{\Theta})\|_F^2 + g(\boldsymbol{\Theta}). \qquad \text{(B.1)}\end{aligned}$$

This completes the proof of (11). □

C Proof of Proposition 3

Let $\boldsymbol{\mathcal{M}} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{k_2})'$, satisfying $\boldsymbol{\mu}_j = \tilde{\boldsymbol{\xi}}_j - L^{-1}\boldsymbol{d}_j(\tilde{\boldsymbol{\Theta}})$ for $j = 1, \ldots, k_2$. First, we show that $\boldsymbol{d}_j(\tilde{\boldsymbol{\Theta}}) = \boldsymbol{0}_p$ for all $j \notin I_q(\boldsymbol{\mathcal{M}})$. From (13), it is straightforward to observe that $\boldsymbol{\mu}_i = \tilde{\boldsymbol{\xi}}_i$ for $i \in I_q(\boldsymbol{\mathcal{M}})$. On the other hand, since $\tilde{\boldsymbol{\xi}}_j = \boldsymbol{0}_p$ for $j \notin I_q(\boldsymbol{\mathcal{M}})$, we have $\boldsymbol{\mu}_j = L^{-1}\boldsymbol{d}_j(\tilde{\boldsymbol{\Theta}})$ for $j \notin I_q(\boldsymbol{\mathcal{M}})$. These imply that $\|\tilde{\boldsymbol{\xi}}_i\|_2 \geq \|L^{-1}\boldsymbol{d}_j(\tilde{\boldsymbol{\Theta}})\|_2$ for any $i \in I_q(\boldsymbol{\mathcal{M}})$ and $j \notin I_q(\boldsymbol{\mathcal{M}})$ because $\|\boldsymbol{\mu}_i\|_2 \geq \|\boldsymbol{\mu}_j\|_2$. Note that it follows from $\|\tilde{\boldsymbol{\Xi}}\|_{2,0} < q$ that $\min_{i\in I_q(\boldsymbol{\mathcal{M}})} \|\boldsymbol{\mu}_i\|_2 = \min_{i\in I_q(\boldsymbol{\mathcal{M}})} \|\tilde{\boldsymbol{\xi}}_i\|_2 = 0$. Hence, we have $\boldsymbol{d}_j(\tilde{\boldsymbol{\Theta}}) = \boldsymbol{0}_p$ for all $j \notin I_q(\boldsymbol{\mathcal{M}})$.

Next, we show that $\boldsymbol{d}_i(\tilde{\boldsymbol{\Theta}}) = \boldsymbol{0}_p$ for all $i \in I_q(\boldsymbol{\mathcal{M}})$. From (13), it is straightforward to observe that $\tilde{\boldsymbol{\xi}}_i = \tilde{\boldsymbol{\xi}}_i - L^{-1}\boldsymbol{d}_i(\tilde{\boldsymbol{\Theta}})$ for $i \in I_q(\boldsymbol{\mathcal{M}})$. Hence, we have $\boldsymbol{d}_i(\tilde{\boldsymbol{\Theta}}) = \boldsymbol{0}_p$ for all $i \in I_q(\boldsymbol{\mathcal{M}})$. Therefore, it holds that $\boldsymbol{D}_2(\tilde{\boldsymbol{\Theta}}) = \boldsymbol{O}_{k_2,p}$. Moreover, since it is straightforward to observe that $\boldsymbol{D}_1(\tilde{\boldsymbol{\Theta}}) = \boldsymbol{O}_{k_1,p}$, we have $\boldsymbol{D}(\tilde{\boldsymbol{\Theta}}) = \boldsymbol{O}_{k,p}$. This fact and the convexity of g lead to $\tilde{\boldsymbol{\Theta}} \in \arg\min_{\boldsymbol{\Theta}} g(\boldsymbol{\Theta})$. □

D Proof of Proposition 4

D.1 Proof of (a)

Let $\boldsymbol{\Theta}_\dagger \in \arg\min_{\tilde{\boldsymbol{\Theta}}:\ \|\tilde{\boldsymbol{\Xi}}\|_{2,0}\leq q} Q_L(\tilde{\boldsymbol{\Theta}}, \boldsymbol{\Theta})$, where $Q_L(\tilde{\boldsymbol{\Theta}}, \boldsymbol{\Theta})$ is as defined in (10). Then, for $\boldsymbol{\Theta} = (\boldsymbol{\mathcal{B}}', \boldsymbol{\Xi}')'$ ($\boldsymbol{\mathcal{B}} \in \mathbb{R}^{k_1\times p}, \boldsymbol{\Xi} \in \mathbb{R}^{k_2\times p}$) such that $\|\boldsymbol{\Xi}\|_{2,0} \leq q$, we have

$$\begin{aligned} g(\boldsymbol{\Theta}) &= Q_L(\boldsymbol{\Theta}, \boldsymbol{\Theta}) \\ &\geq \inf_{\tilde{\boldsymbol{\Theta}}:\ \|\tilde{\boldsymbol{\Xi}}\|_{2,0}\leq q} Q_L(\tilde{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) \\ &= Q_L(\boldsymbol{\Theta}_\dagger, \boldsymbol{\Theta}) \\ &= g(\boldsymbol{\Theta}) + \frac{L}{2}\|\boldsymbol{\Theta}_\dagger - \boldsymbol{\Theta}\|_F^2 + \mathrm{tr}\left\{\boldsymbol{D}'(\boldsymbol{\Theta})(\boldsymbol{\Theta}_\dagger - \boldsymbol{\Theta})\right\} \\ &= g(\boldsymbol{\Theta}) + \frac{L-\ell}{2}\|\boldsymbol{\Theta}_\dagger - \boldsymbol{\Theta}\|_F^2 + \frac{\ell}{2}\|\boldsymbol{\Theta}_\dagger - \boldsymbol{\Theta}\|_F^2 + \mathrm{tr}\left\{\boldsymbol{D}'(\boldsymbol{\Theta})(\boldsymbol{\Theta}_\dagger - \boldsymbol{\Theta})\right\} \\ &= \frac{L-\ell}{2}\|\boldsymbol{\Theta}_\dagger - \boldsymbol{\Theta}\|_F^2 + Q_\ell(\boldsymbol{\Theta}_\dagger, \boldsymbol{\Theta}) \\ &\geq \frac{L-\ell}{2}\|\boldsymbol{\Theta}_\dagger - \boldsymbol{\Theta}\|_F^2 + g(\boldsymbol{\Theta}_\dagger). \qquad \text{(D.1)}\end{aligned}$$

Since we can regard $\boldsymbol{\Theta}_\dagger$ as $\boldsymbol{\Theta}_{m+1}$ by letting $\boldsymbol{\Theta} = \boldsymbol{\Theta}_m$, (D.1) is expressed as

$$g(\boldsymbol{\Theta}_m) - g(\boldsymbol{\Theta}_{m+1}) \geq \frac{L-\ell}{2}\|\boldsymbol{\Theta}_{m+1} - \boldsymbol{\Theta}_m\|_F^2.$$

Hence, $g(\boldsymbol{\Theta}_m)$ monotonically decreases for m. Moreover, it is straightforward to observe that $g(\boldsymbol{\Theta}_m)$ converges as $m \to \infty$ because it is bounded below. □

D.2 Proof of (b)

Since $g(\boldsymbol{\Theta}_m)$ converges as $m \to \infty$ from (a) in Proposition 4, $g(\boldsymbol{\Theta}_m) - g(\boldsymbol{\Theta}_{m+1})$ converges to 0. Hence, $\|\boldsymbol{\Theta}_{m+1} - \boldsymbol{\Theta}_m\|_F^2$ in (14) also converges to 0 for $L > \ell$. This implies that $\boldsymbol{\Theta}_{m+1} - \boldsymbol{\Theta}_m \to \boldsymbol{O}_{k,p}$ $(m \to \infty)$. □

D.3 Proof of (c)

First, we prove that there exists $M > 0$ such that for all $m \geq M$, $\boldsymbol{r}_m = \boldsymbol{r}_{m+1}$ by contradiction. Assume that for any $M > 0$, there exists $\tilde{m} \geq M$ such that $\boldsymbol{r}_{\tilde{m}} \neq \boldsymbol{r}_{\tilde{m}+1}$. Since $\underline{\alpha}_q > 0$, we observe that $\|\boldsymbol{\Xi}_m\|_{2,0} = q$ for sufficiently large m. Hence, by considering $M > m$, we can see that there exist i, j $(i \neq j)$ such that

$$\boldsymbol{\xi}_{\tilde{m},i} = \boldsymbol{0}_p,\ \boldsymbol{\xi}_{\tilde{m},j} \neq \boldsymbol{0}_p,\ \boldsymbol{\xi}_{\tilde{m}+1,i} \neq \boldsymbol{0}_p,\ \boldsymbol{\xi}_{\tilde{m}+1,j} = \boldsymbol{0}_p,$$

for infinitely many $\tilde{m} \geq M$. Using the above equations, the following inequality can be derived:

$$\|\boldsymbol{\Xi}_{\tilde{m}} - \boldsymbol{\Xi}_{\tilde{m}+1}\|_F \geq \sqrt{\|\boldsymbol{\xi}_{\tilde{m}+1,i}\|_2^2 + \|\boldsymbol{\xi}_{\tilde{m},j}\|_2^2} \geq \frac{\|\boldsymbol{\xi}_{\tilde{m}+1,i}\|_2 + \|\boldsymbol{\xi}_{\tilde{m},j}\|_2}{\sqrt{2}}.$$

From the above, we observe that the ℓ_2-norms of the non-zero vectors $\|\boldsymbol{\xi}_{\tilde{m}+1,i}\|_2$ and $\|\boldsymbol{\xi}_{\tilde{m},j}\|_2$ converge to 0 as $\tilde{m} \to \infty$ because $\|\boldsymbol{\Theta}_{\tilde{m}} - \boldsymbol{\Theta}_{\tilde{m}+1}\|_F \to 0$ from (b) in Proposition 4. This contradicts $\underline{\alpha}_q > 0$.

Next, we show that the sequence $\{\boldsymbol{\Theta}_m\}$ converges to an $\ell_{2,0}$-constrained first-order stationary point. Since $\boldsymbol{r}_m = \boldsymbol{r}_{m+1}$ for sufficiently large m, we can set $\mathcal{L} = \{1 \leq j \leq k_2 \mid r_{m,j} = 1\}$. Note that the elements in $\mathcal{L}$ are invariant for sufficiently large m. Hence, using (B.1), for sufficiently large m we have

$$\begin{aligned}
&\min_{\tilde{\boldsymbol{\Theta}}:\ \|\tilde{\boldsymbol{\Xi}}\|_{2,0} \leq q} Q_L(\tilde{\boldsymbol{\Theta}}, \boldsymbol{\Theta}_m) \\
&= \frac{L}{2} \min_{\|\tilde{\boldsymbol{\Xi}}\|_{2,0} \leq q} \left\| \tilde{\boldsymbol{\Xi}} - \left(\boldsymbol{\Xi}_m - L^{-1}\boldsymbol{D}_2(\boldsymbol{\Theta}_m)\right) \right\|_F^2 - \frac{1}{2L}\|\boldsymbol{D}(\boldsymbol{\Theta}_m)\|_F^2 + g(\boldsymbol{\Theta}_m) \\
&= \frac{L}{2} \min_{\tilde{\boldsymbol{\xi}}_j:\ j \in \mathcal{L}} \sum_{j \in \mathcal{L}} \left\| \tilde{\boldsymbol{\xi}}_j - \left(\boldsymbol{\xi}_{m,j} - L^{-1}\boldsymbol{d}_j(\boldsymbol{\Theta}_m)\right) \right\|_2^2 + \frac{L}{2} \sum_{j \notin \mathcal{L}} \left\| \boldsymbol{\xi}_{m,j} - L^{-1}\boldsymbol{d}_j(\boldsymbol{\Theta}_m) \right\|_2^2 \\
&\quad - \frac{1}{2L}\|\boldsymbol{D}(\boldsymbol{\Theta}_m)\|_F^2 + g(\boldsymbol{\Theta}_m).
\end{aligned}$$

This implies that Algorithm 1 behaves like a vanilla gradient decent algorithm for minimizing a convex function over a closed convex. Therefore, the sequence $\{\boldsymbol{\Theta}_m\}$ converges to an $\ell_{2,0}$-constrained first-order stationary point. □

D.4 Proof of (d)

From (12) and (b), it is straightforward to observe that $\lim_{m\to\infty} \boldsymbol{D}_1(\boldsymbol{\Theta}_m) = \boldsymbol{O}_{k_1,p}$. Assume that $\underline{\alpha}_q = 0$. Let $\boldsymbol{\mathcal{M}}_m = (\boldsymbol{\mu}_{m,1}, \ldots, \boldsymbol{\mu}_{m,k_2})' = \boldsymbol{\Xi}_m - L^{-1}\boldsymbol{D}_2(\boldsymbol{\Theta}_m)$. From (12), for any $i \in I_q(\boldsymbol{\mathcal{M}}_m)$ and $j \notin I_q(\boldsymbol{\mathcal{M}}_m)$, we have $\|\boldsymbol{\mu}_{m,i}\|_2 \geq \|\boldsymbol{\mu}_{m,j}\|_2$. Hence, the following inequality can be derived:

$$\liminf_{m\to\infty} \min_{i \in I_q(\boldsymbol{\mathcal{M}}_m)} \|\boldsymbol{\mu}_{m,i}\|_2 \geq \liminf_{m\to\infty} \max_{j \notin I_q(\boldsymbol{\mathcal{M}}_m)} \|\boldsymbol{\mu}_{m,j}\|_2. \tag{D.2}$$

On the other hand, from (12) and (b), it is straightforward to observe that

$$\boldsymbol{\xi}_{m,i} - \boldsymbol{\xi}_{m+1,i} = \begin{cases} L^{-1}\boldsymbol{d}_i(\boldsymbol{\Theta}_m) & \to \mathbf{0}_p \ (m \to \infty) \quad (i \in I_q(\boldsymbol{\mathcal{M}}_m)) \\ \boldsymbol{\xi}_{m,i} & \to \mathbf{0}_p \ (m \to \infty) \quad (i \notin I_q(\boldsymbol{\mathcal{M}}_m)) \end{cases}. \tag{D.3}$$

Using (D.3) and the triangle inequality, we have

$$\begin{aligned} &\liminf_{m\to\infty} \min_{i\in I_q(\boldsymbol{\mathcal{M}}_m)} \|\boldsymbol{\mu}_{m,i}\|_2 \\ &\leq \liminf_{m\to\infty} \min_{i\in I_q(\boldsymbol{\mathcal{M}}_m)} \|\boldsymbol{\xi}_{m,i}\|_2 + \liminf_{m\to\infty} \max_{i\in I_q(\boldsymbol{\mathcal{M}}_m)} \|L^{-1}\boldsymbol{d}_i(\boldsymbol{\Theta}_m)\|_2 \\ &= \liminf_{m\to\infty} \min_{i\in I_q(\boldsymbol{\mathcal{M}}_m)} \|\boldsymbol{\xi}_{m,i}\|_2, \end{aligned} \tag{D.4}$$

$$\begin{aligned} &\liminf_{m\to\infty} \max_{j\notin I_q(\boldsymbol{\mathcal{M}}_m)} \|\boldsymbol{\mu}_{m,j}\|_2 \\ &\geq \limsup_{m\to\infty} \min_{i\notin I_q(\boldsymbol{\mathcal{M}}_m)} \|\boldsymbol{\xi}_{m,i}\|_2 + \liminf_{m\to\infty} \max_{i\notin I_q(\boldsymbol{\mathcal{M}}_m)} \|L^{-1}\boldsymbol{d}_i(\boldsymbol{\Theta}_m)\|_2 \\ &= \liminf_{m\to\infty} \max_{i\notin I_q(\boldsymbol{\mathcal{M}}_m)} \|L^{-1}\boldsymbol{d}_i(\boldsymbol{\Theta}_m)\|_2. \end{aligned} \tag{D.5}$$

By combining (D.2), (D.4), and (D.5), the following inequality can be derived:

$$\liminf_{m\to\infty} \min_{i\in I_q(\boldsymbol{\mathcal{M}}_m)} \|\boldsymbol{\xi}_{m,i}\|_2 \geq \liminf_{m\to\infty} \max_{j\notin I_q(\boldsymbol{\mathcal{M}}_m)} \|L^{-1}\boldsymbol{d}_i(\boldsymbol{\Theta}_m)\|_2. \tag{D.6}$$

Since $\underline{\alpha}_q = \liminf_{m\to\infty} \min_{i\in I_q(\boldsymbol{\mathcal{M}}_m)} \|\boldsymbol{\xi}_{m,i}\|_2 = 0$, the right-hand side of (D.6) becomes 0. This fact and (D.3) imply that $\liminf_{m\to\infty} \max_{j=1,\ldots,k_2} \|\boldsymbol{d}_j(\boldsymbol{\Theta}_m)\|_2 = 0$. □

D.5 Proof of (e)

By modifying lim inf to lim sup in (D.6), we can derive the following inequality:

$$\limsup_{m\to\infty} \min_{i\in I_q(\boldsymbol{\mathcal{M}}_m)} \|\boldsymbol{\xi}_{m,i}\|_2 \geq \limsup_{m\to\infty} \max_{j\notin I_q(\boldsymbol{\mathcal{M}}_m)} \|L^{-1}\boldsymbol{d}_i(\boldsymbol{\Theta}_m)\|_2, \tag{D.7}$$

where $\boldsymbol{\mathcal{M}}_m = (\boldsymbol{\mu}_{m,1}, \ldots, \boldsymbol{\mu}_{m,k_2})' = \boldsymbol{\Xi}_m - L^{-1}\boldsymbol{D}_2(\boldsymbol{\Theta}_m)$. Since $\bar{\alpha}_q = 0$, the right-hand side of (D.7) becomes 0. Since this fact and (D.3) imply that $\boldsymbol{D}_2(\boldsymbol{\Theta}_m) \to \boldsymbol{O}_{k_2,p}$ $(m \to \infty)$, we have $\boldsymbol{D}(\boldsymbol{\Theta}_m) \to \boldsymbol{O}_{k,p}$ $(m \to \infty)$ from (d) in Proposition 4. Let $\boldsymbol{\Theta}_\infty$ be a limit point of the sequence $\{\boldsymbol{\Theta}_m\}$. Then, there exists a subsequence $\{\tilde{m}\}$ such that $\boldsymbol{\Theta}_{\tilde{m}} \to \boldsymbol{\Theta}_\infty$ and $g(\boldsymbol{\Theta}_{\tilde{m}}) \to g(\boldsymbol{\Theta}_\infty)$. Since $\boldsymbol{D}(\boldsymbol{\Theta})$ is Lipschitz continuous, it holds that $\boldsymbol{D}(\boldsymbol{\Theta}_{\tilde{m}}) \to \boldsymbol{D}(\boldsymbol{\Theta}_\infty) = \boldsymbol{O}_{k,p}$ as $\tilde{m} \to \infty$. This implies that $\boldsymbol{\Theta}_\infty$ is a solution to $\min_{\boldsymbol{\Theta}} g(\boldsymbol{\Theta})$. Therefore, $g(\boldsymbol{\Theta}_m) \to \min_{\boldsymbol{\Theta}} g(\boldsymbol{\Theta})$ $(m \to \infty)$ holds. □

E Proof of Proposition 5

E.1 Proof of (a)

Suppose that $\boldsymbol{\Theta}_\dagger = (\boldsymbol{\mathcal{B}}_\dagger', \boldsymbol{\Xi}_\dagger')'$ satisfies $\|\boldsymbol{\Xi}_\dagger\|_{2,0} \leq q$ and the following equation:

$$\boldsymbol{\mathcal{B}}_\dagger = \tilde{\boldsymbol{\mathcal{B}}} - L^{-1}\boldsymbol{D}_1(\tilde{\boldsymbol{\Theta}}), \ \boldsymbol{\Xi}_\dagger \in \mathbf{H}_q\left(\tilde{\boldsymbol{\Xi}} - L^{-1}\boldsymbol{D}_2(\tilde{\boldsymbol{\Theta}})\right).$$

Using (D.1), we can derive the following inequality:

$$g(\tilde{\boldsymbol{\Theta}}) - g(\boldsymbol{\Theta}_\dagger) \geq \frac{L-\ell}{2}\|\boldsymbol{\Theta}_\dagger - \tilde{\boldsymbol{\Theta}}\|_F^2. \tag{E.1}$$

On the other hand, since the function g is convex and differentiable, we have

$$g(\boldsymbol{\Theta}_\dagger) - g(\tilde{\boldsymbol{\Theta}}) \geq \mathrm{tr}\left\{\boldsymbol{D}'(\tilde{\boldsymbol{\Theta}})(\boldsymbol{\Theta}_\dagger - \tilde{\boldsymbol{\Theta}})\right\}. \tag{E.2}$$

Since $\boldsymbol{\Theta}_\dagger$ and $\tilde{\boldsymbol{\Theta}}$ are $\ell_{2,0}$-constrained first-order stationary points, we observe that $\boldsymbol{D}_1(\tilde{\boldsymbol{\Theta}}) = \boldsymbol{O}_{k_1,p}$, $\boldsymbol{d}_j(\tilde{\boldsymbol{\Theta}}) = \boldsymbol{0}_p$ for $j \in I_q(\tilde{\boldsymbol{\Xi}} - L^{-1}\boldsymbol{D}_2(\tilde{\boldsymbol{\Theta}}))$ and $\boldsymbol{\xi}_{\dagger,j} = \tilde{\boldsymbol{\xi}}_j = \boldsymbol{0}_p$ for $j \notin I_q(\tilde{\boldsymbol{\Xi}} - L^{-1}\boldsymbol{D}_2(\tilde{\boldsymbol{\Theta}}))$, where $\boldsymbol{\xi}_{\dagger,j}$ is the j-th row vector of $\boldsymbol{\Xi}_\dagger$. Hence, we have $\mathrm{tr}\{\boldsymbol{D}'(\tilde{\boldsymbol{\Theta}})(\boldsymbol{\Theta}_\dagger - \tilde{\boldsymbol{\Theta}})\} = 0$. Therefore, it follows from (E.1) and (E.2) that $\|\boldsymbol{\Theta}_\dagger - \tilde{\boldsymbol{\Theta}}\|_F^2 = 0$ holds for $L > \ell$. This implies that $\mathbf{H}_q(\tilde{\boldsymbol{\Xi}} - L^{-1}\boldsymbol{D}_2(\tilde{\boldsymbol{\Theta}}))$ has exactly one element. □

E.2 Proof of (b)

For any global minimizer $\hat{\boldsymbol{\Theta}}_{\mathrm{glo}} = (\hat{\boldsymbol{\mathcal{B}}}_{\mathrm{glo}}', \hat{\boldsymbol{\Xi}}_{\mathrm{glo}}')'$ of (4), let $\tilde{\boldsymbol{\Theta}} = (\tilde{\boldsymbol{\mathcal{B}}}', \tilde{\boldsymbol{\Xi}}')'$ be the matrix satisfying $\|\tilde{\boldsymbol{\Xi}}\|_{2,0} \leq q$ and the following equation:

$$\tilde{\boldsymbol{\mathcal{B}}} = \hat{\boldsymbol{\mathcal{B}}}_{\mathrm{glo}} - L^{-1}\boldsymbol{D}_1(\hat{\boldsymbol{\Theta}}_{\mathrm{glo}}),\ \tilde{\boldsymbol{\Xi}} \in \mathbf{H}_q\left(\hat{\boldsymbol{\Xi}}_{\mathrm{glo}} - L^{-1}\boldsymbol{D}_2(\hat{\boldsymbol{\Theta}}_{\mathrm{glo}})\right).$$

By the definition of $\hat{\boldsymbol{\Theta}}_{\mathrm{glo}}$, we have $g(\tilde{\boldsymbol{\Theta}}) \geq g(\hat{\boldsymbol{\Theta}}_{\mathrm{glo}})$. Moreover, as with (E.1), the following inequality can be derived:

$$g(\hat{\boldsymbol{\Theta}}_{\mathrm{glo}}) - g(\tilde{\boldsymbol{\Theta}}) \geq \frac{L-\ell}{2}\|\tilde{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}_{\mathrm{glo}}\|_F^2.$$

Hence, $\|\tilde{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}}_{\mathrm{glo}}\|_F^2 = 0$ holds for $L > \ell$. This implies that $\hat{\boldsymbol{\Theta}}_{\mathrm{glo}}$ is an $\ell_{2,0}$-constrained first-order stationary point. □

F Proof of Theorem 1

By summing (14) for $m = 1, \ldots, M$, we have

$$\sum_{m=1}^{M}\{g(\boldsymbol{\Theta}_m) - g(\boldsymbol{\Theta}_{m+1})\} \geq \frac{L-\ell}{2}\sum_{m=1}^{M}\|\boldsymbol{\Theta}_{m+1} - \boldsymbol{\Theta}_m\|_F^2.$$

Since $g(\boldsymbol{\Theta}_m)$ monotonically decreases for m and converges to g_* as $m \to \infty$, we obtain the following inequality:

$$\begin{aligned} g(\boldsymbol{\Theta}_1) - g_* &\geq \sum_{m=1}^{M}\{g(\boldsymbol{\Theta}_m) - g(\boldsymbol{\Theta}_{m+1})\} \\ &\geq \frac{L-\ell}{2}\sum_{m=1}^{M}\|\boldsymbol{\Theta}_{m+1} - \boldsymbol{\Theta}_m\|_F^2 \\ &\geq \frac{M(L-\ell)}{2}\min_{m=1,\ldots,M}\|\boldsymbol{\Theta}_{m+1} - \boldsymbol{\Theta}_m\|_F^2. \end{aligned}$$

This completes the proof of Theorem 1. □

Acknowledgments

Ryoya Oda is supported by JSPS KAKENHI Grant Numbers JP19K21672, JP20K14363, and JP20H04151. Mineaki Ohishi is supported by JSPS KAKENHI Grant Numbers JP20H04151 and JP21K13834. Hirokazu Yanagihara is supported by JSPS KAKENHI Grant Numbers JP16H03606, JP18K03415, and JP20H04151.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B. N. Petrov & F. Csáki), pp. 267–281. Akadémiai Kiadó, Budapest.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **AC-19**, 716–723.

[3] Bertsimas, D., King, A. and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.*, **44**, 813–852.

[4] Cai, X., Nie, F. and Huang, H. (2013). Exact top-k feature selection via $\ell_{2,0}$-norm constraint. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence.*

[5] Gotoh, J., Takeda, A. and Tono, K. (2018). DC formulations and algorithms for sparse optimization problems. *Math. Program.*, **169**, 141–176.

[6] Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective.* Springer-Verlag, New York.

[7] Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization.* Kluwer Academic Publishers, Boston, MA.

[8] Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Math. Program.*, **140**, 125–161.

[9] Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.*, **39**, 1–47.

[10] Oda, R. and Yanagihara, H. (2020). A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables. *Electron. J. Statist.*, **14**, 1386–1412.

[11] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

[12] Srivastava, M. S. (2002). *Methods of Multivariate Statistics.* John Wiley & Sons, New York.

[13] Timm, N. H. (2002). *Applied Multivariate Analysis.* Springer-Verlag, New York.

[14] Yanagihara, H. (2006). Corrected version of AIC for selecting multivariate normal linear regression models in a general nonnormal case. *J. Multivariate Anal.*, **97**, 1070–1089.

[15] Yanagihara, H., Kamo, K., Imori, S. and Satoh, K. (2012). Bias-corrected AIC for selecting variables in multinomial logistic regression models. *Linear Algebra Appl.*, **436**, 4329–4341.