Sparse multivariate regression with missing values and its application to the prediction of material properties

Keisuke Teramoto¹ | Kei Hirose²,³

¹Department of Mathematics, Hiroshima University, Higashi-Hiroshima, Japan
²Institute of Mathematics for Industry, Kyushu University, Fukuoka, Japan
³RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

Correspondence
Keisuke Teramoto, Department of Mathematics, Hiroshima University, Kagamiyama 1-3-1, Higashi-Hiroshima 739-8526, Japan.
Email: kteramoto@hiroshima-u.ac.jp

Abstract
In the field of materials science and engineering, statistical analysis and machine learning techniques have recently been used to predict multiple material properties from an experimental design. These material properties correspond to response variables in the multivariate regression model. In this study, we conduct a penalized maximum likelihood procedure to estimate model parameters, including the regression coefficients and covariance matrix of response variables. In particular, we employ $\ell_1$-regularization to achieve a sparse estimation of the regression coefficients and inverse covariance matrix of response variables. In some cases, there may be a relatively large number of missing values in the response variables, owing to the difficulty of collecting data on material properties. We therefore propose a method that incorporates a correlation structure among the response variables into a statistical model to improve the prediction accuracy under the situation with missing values. The expectation maximization algorithm is also constructed, which enables application to a dataset with missing values in the responses. We apply our proposed procedure to real data consisting of 22 material properties.

Keywords
graphical lasso, missing data, multivariate regression, sparse estimation

1 | INTRODUCTION

The importance of data analysis applications using statistics and machine learning for materials science and engineering has been steadily increasing (cf. References 1-10). Owing to the recent development of machine learning methods, data-centric informatics applied to a sufficiently large amount of data is useful for identifying materials that have desirable properties, such as durability and flexibility. Desirable materials are often difficult to identify using only experiments and physical simulations.

The data structure in that field often has two features that introduce new challenges. The first feature is that we often predict multiple properties of the materials; that is, we must construct a regression model with multiple responses (multivariate regression). For example, in a study on adhesion structure, the yield strength, ultimate tensile strength, fracture, and Young’s modulus must be predicted, among many other material properties that are not described here. However, it would be difficult to find a material that satisfies multiple desired properties simultaneously because there are typically trade-offs among these properties.¹¹,¹² These trade-offs are expressed as a correlation matrix among the response
variables. This study assumes the correlation structure among response variables and employs a likelihood procedure to estimate regression coefficients and a covariance matrix of response variables. The estimated covariance matrix assists engineers with interpreting the relationship between properties and improves the prediction accuracy.\(^{13,14}\)

When the number of responses is relatively large, it may be difficult to estimate all correlation pairs because the number of parameters is proportional to the square of the number of variables. In such cases, regularization methods have typically been employed to achieve a stable estimation of the covariance matrix of response variables. In particular, the least absolute shrinkage and selection operator (LASSO)-type sparse estimation (cf. Reference 15) is applied for simultaneous variable selection and model estimation, which enables an interpretation of the relationship among material properties. A wide variety of regularization methods that induce a sparse structure have been proposed, such as elastic net,\(^{16}\) group LASSO,\(^{17}\) graphical LASSO,\(^{18}\) generalized LASSO,\(^{19}\) and overlapping group LASSO.\(^{20}\) Rothman et al.\(^{13}\) introduced the multivariate regression with covariance estimation (MRCE) method, in which sparse regression coefficients and a sparse inverse covariance matrix of the response variables are simultaneously estimated. This is a generalization of the LASSO regression to the sparse multivariate regression analysis. The model parameter is estimated using the penalized maximum likelihood procedure with the LASSO. In this study, we perform data analysis based on the MRCE method.

The second feature that introduces challenges is that the data values in the response variables are often missing because it is difficult to observe all the properties of the materials, owing to large-scale experiments. When the ratio of missing data is small, we can exclude the corresponding observations and conduct a data analysis. This method is referred to as the complete-case analysis.\(^{21}\) However, responses tend to have many missing values. In fact, a dataset used in this study comprised 50% or more missing values (see Figure 1). If we conduct a complete-case analysis, the number of observations becomes extremely small, which results in low prediction accuracy. The full information maximum likelihood (FIML) approach (see Reference 22) provides a means to handle a large number of missing values; Hirose et al.\(^{22}\) showed that the FIML approach provides a good estimator when the ratio of missing data is 90%. FIML produces a consistent estimator, even when the number of missing values is large under the missing at random (MAR) assumption (cf. Reference 21). Moreover, the FIML approach enables missing value interpolation, which may assist with the understanding of some hidden structures/relations. For multivariate data without response variables, Städler and Bühlmann\(^{24}\) proposed the graphical LASSO with missing values MissGLASSO method for data with missing values. An \(l_1\)-regularized likelihood method is used to estimate the sparse inverse covariance matrix. Moreover, they proposed an efficient EM algorithm (see Reference 25) for optimization with provable numerical convergence properties. Städler and Bühlmann\(^{24}\) extended MissGLASSO to multiple (not multivariate) regression analysis. However, they only assumed the case where the exploratory variables have missing values; MissGLASSO cannot be directly applied to multivariate regression analysis when there are missing values in the response variables.

As mentioned above, the MRCE simultaneously estimates the regression coefficients and covariance matrix of the response variables. However, it is applicable only to complete multivariate data; thus, we cannot apply this method directly for data with missing values. Therefore, we need a suitable extension of the MRCE to apply to data with missing values. Notably, MissGLASSO can be applied to data with missing values. The aim of this study is to propose a multivariate regression model with missing values by combining these two methods.

In this study, we establish a new algorithm called the sparse multivariate regression with missing data (SMRM) algorithm to estimate the inverse covariance matrix and interpolate the data with missing values (see Section 3). To estimate the multivariate regression coefficients and the covariance structure, we need to solve a particular \(l_1\)-regularized likelihood type optimization problem with two regularization parameters; one is related to the correlation structure of the responses, and the other is related to the regression coefficients matrix. Here, we note that multiple regularization parameters for regression coefficients are assumed because the error variances vary among the response variables. For this optimization problem, we employ the EM algorithm. As in the case of the MRCE method, the coordinate descent algorithm and graphical LASSO algorithm are applied in the maximization (M) step of the EM algorithm. Using sparse estimation, the SMRM algorithm can conduct stable estimation, even for a dataset with a relatively large number of missing values. In addition, we can improve the prediction accuracy by using the correlation structure among the response variables. We estimate the sparse inverse covariance matrix to introduce our method instead of the covariance matrix itself because spurious correlations among responses may be excluded.\(^{26}\) Note that there is an interesting application of the MRCE method by She et al.\(^{27}\) that deals with time series data, particularly, long-scale linear dynamical systems. Although She et al.\(^{27}\) and our study use similar mathematical techniques, She et al.\(^{27}\) do not deal with missing data. That is a significant difference. In the last section, we apply the SMRM algorithm to real data and investigate the influences of regularization parameters. Furthermore, we compare the prediction accuracy obtained by our method to that of the LASSO.
2 | PRELIMINARIES

2.1 | Conditional distribution

In this section, we briefly review some notions and facts from multivariate regression analysis. For a more detailed explanation, refer to Reference 28.

Let \( x_j = (x_{1j}, \ldots, x_{pj})^T \) \((1 \leq j \leq p)\) be the predictor variables, \( y_l = (y_{1l}, \ldots, y_{ql})^T \) \((1 \leq l \leq q)\) response variables. (We consider \( x_j \) and \( y_l \) as column vectors.) Then, we set matrices \( X, \tilde{X}, \) and \( Y \) as

\[
X = (x_1, \ldots, x_p) = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^p \\ x_1^2 & x_1^3 & \cdots & x_1^p \\ \vdots & \vdots & \ddots & \vdots \\ x_1^n & x_1^{n+1} & \cdots & x_1^p \\ x_2^1 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ x_p^1 & x_p^2 & \cdots & x_p^p \end{pmatrix}, \quad \tilde{X} = (1, x_1, \ldots, x_p) = \begin{pmatrix} 1 & x_1^1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2^1 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_p^1 & x_p^2 & \cdots & x_p^p \end{pmatrix},
\]

\[
Y = (y_1, y_2, \ldots, y_q) = \begin{pmatrix} y_1^1 \\ y_2^1 \\ \vdots \\ y_p^1 \\ y_1^2 \\ y_2^2 \\ \vdots \\ y_p^2 \\ \vdots \\ \vdots \\ y_1^q \\ y_2^q \\ \vdots \\ y_p^q \end{pmatrix}.
\]

(1)

Let \( \mathbf{x}^i = (x^i_1, x^i_2, \ldots, x^i_p), \mathbf{y}^i = (1, x^i_1, x^i_2, \ldots, x^i_p) \), and \( y^i = (y^i_1, y^i_2, \ldots, y^i_q) \) \((1 \leq i \leq n)\) be the \( i \)th row vectors of \( X, \tilde{X}, \) and \( Y, \) as in (1), respectively. (We consider \( \mathbf{x}^i, \mathbf{y}^i \) as row vectors.) We then consider the multivariate linear regression model of the form

\[
\mathbf{y}^i = \mathbf{x}^i \beta + \mathbf{e}^i,
\]

where \( \beta \) is a \( q \times p \) matrix of regression coefficients and \( \mathbf{e}^i \) is a \( q \times 1 \) vector of random errors.
\[ Y = \tilde{X}B + E = (1_n, X) \left( \begin{array}{c} b_0^T \\ B \end{array} \right) + E, \]

where \( b_0 = (b_0^0, b_0^1, \ldots, b_0^q)^T \in \mathbb{R}^q \) is a vector of the regression intercept, \( B \) is a regression coefficient matrix of the form

\[ B = \begin{pmatrix} b_1^1 & b_1^2 & \cdots & b_1^q \\ b_2^1 & b_2^2 & \cdots & b_2^q \\ \vdots & \vdots & \ddots & \vdots \\ b_q^1 & b_q^2 & \cdots & b_q^q \end{pmatrix}, \]

and \( E \) is the error matrix given by

\[ E = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_q) = \begin{pmatrix} \varepsilon_1^1 & \varepsilon_1^2 & \cdots & \varepsilon_1^q \\ \varepsilon_2^1 & \varepsilon_2^2 & \cdots & \varepsilon_2^q \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_q^1 & \varepsilon_q^2 & \cdots & \varepsilon_q^q \end{pmatrix}. \]

We denote the \( i \)th row vector of \( E \) (\( 1 \leq i \leq n \)) as \( \varepsilon^i = (\varepsilon_1^i, \varepsilon_2^i, \ldots, \varepsilon_q^i) \). We assume that the \( n \) subjects are independent. We then obtain the following:

\begin{itemize}
  \item \( \varepsilon_1 \sim N(0_q, \sigma_1^2 I_n) \) (\( 1 \leq l \leq q \)),
  \item \( (\varepsilon^1)^T \sim_{i.i.d.} N(0_q, \Sigma) \) (\( 1 \leq i \leq n \)),
\end{itemize}

where \( 0_q \) is the \( q \times q \) zero vector, \( I_n \) is the \( n \times n \) identity matrix, and \( \Sigma \) is the covariance matrix of the form

\[ \Sigma = \begin{pmatrix} \sigma_1^1 & \sigma_1^2 & \cdots & \sigma_1^q \\ \sigma_2^1 & \sigma_2^2 & \cdots & \sigma_2^q \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_q^1 & \sigma_q^2 & \cdots & \sigma_q^q \end{pmatrix}, \]

with \( \sigma_l^l = \sigma_l^{l'} \) (\( 1 \leq l, l' \leq q \)). We assume the independence condition in the following. Under these assumptions, we note that \( y^i | \tilde{x}^i \sim N(\mu_i, \Sigma) \).

We now consider a partition \( y^i = (y^{i,1}, y^{i,2}) \) for each \( i \in \{1, \ldots, n\} \). \((y^{i,2})^T(y^{i,1})^T\) follows a linear regression on \((y^{i,1})^T\) with a mean of \( \mu_{i,2} + \Sigma_{i,21} \Sigma_{i,11}^{-1} (y^{i,1}_1 - \mu_{i,1}) \) and covariance of \( \Sigma_{i,22} - \Sigma_{i,21} \Sigma_{i,11}^{-1} \Sigma_{i,12} \). Here, we divide \( \mu_i \) and \( \Sigma \) into

\[ \mu_i = \begin{pmatrix} \mu_{i,1} \\ \mu_{i,2} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{i,11} & \Sigma_{i,12} \\ \Sigma_{i,21} & \Sigma_{i,22} \end{pmatrix} \]

for each \( i \). (For example, if we divide \( y^i \) into \( y^{i,1} = (y^i_1, \ldots, y^i_l) \) and \( y^{i,2} = (y^i_{l+1}, \ldots, y^i_q) \), then \( \mu_{i,1}, \mu_{i,2}, \Sigma_{i,11}, \Sigma_{i,12}, \Sigma_{i,21}, \) and \( \Sigma_{i,22} \) are a \( l \times 1 \) matrix, \( (q-l) \times 1 \) matrix, \( l \times l \) matrix, \( (q-l) \times l \) matrix, \( (q-l) \times (q-l) \) matrix, and \( (q-l) \times (q-l) \) matrix, respectively.) Thus, it can be observed that

\[ (y^{i,2})^T(y^{i,1})^T \sim N(\mu_{i,2} + \Sigma_{i,21} \Sigma_{i,11}^{-1} (y^{i,1}_1 - \mu_{i,1}), \Sigma_{i,22} - \Sigma_{i,21} \Sigma_{i,11}^{-1} \Sigma_{i,12}). \]

Let \( K \) be a \( q \times q \) matrix that satisfies \( KS = I_q \). We call \( K \) the precision matrix or the inverse covariance matrix. For \( i \), if we divide \( (y^i)^T \) into \((y^{i,1}, y^{i,2})^T\), then it holds that

\[ (y^{i,2})^T(y^{i,1})^T \sim N((\mu_{i,2} + \Sigma_{i,21} \Sigma_{i,11}^{-1} (y^{i,1}_1 - \mu_{i,1}), \Sigma_{i,22} - \Sigma_{i,21} \Sigma_{i,11}^{-1} \Sigma_{i,12})). \]
\[
\begin{pmatrix}
K_{i,11} & K_{i,12} \\
K_{i,21} & K_{i,22}
\end{pmatrix}
\begin{pmatrix}
\Sigma_{i,11} & \Sigma_{i,12} \\
\Sigma_{i,21} & \Sigma_{i,22}
\end{pmatrix}
= 
\begin{pmatrix}
I & 0 \\
0 & I
\end{pmatrix}.
\]

By (3) and (4), we obtain
\[
(y^{i,2})^T | (y^{i,1})^T \sim N(\mu_{i,2} - K_{i,22}^{-1} K_{i,21} (y^{i,1})^T - \mu_{i,1}), K_{i,22}^{-1}).
\]

This relation is the key part of our algorithm.

### 2.2 The LASSO

Here, we briefly review the LASSO. (For further details, refer to Reference 15.) This method is used in Section 4 to evaluate the prediction accuracy obtained by our proposed method according to real data. Let \( x_j = (x_1^j, \ldots, x_n^j)^T \) be predictor variables \((1 \leq j \leq p)\) and \( y = (y^1, \ldots, y^n)^T \) be the response variables. We set a matrix \( \tilde{X} \) using \( x_j \), as in (1). We then consider the linear regression model
\[
y = \beta_0 1_n + X \beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n),
\]
where \( \beta_0, \beta \in \mathbb{R} \) are parameters of the regression. In this case, the LASSO optimizes the following form
\[
\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - \beta_0 1_n - X \beta\|_2^2 + \lambda \|\beta\|_1 \right\}
= \left( \|\beta\|_1 = \sum_{j=1}^p |\beta_j| \right).
\]

where \( \lambda > 0 \). By solving the optimization problem, as in (6), we obtain estimators of \( \beta \) and \( \beta_0 \). Generally, the regularization parameter \( \lambda \) of the LASSO is chosen to minimize the predicted errors of each response. Such a regularization parameter is typically called the “best” regularization parameter.

### 3 INTERPOLATION FOR DATA WITH MISSING VALUES

#### 3.1 Responses with missing values

Let \( x_j \in \mathbb{R}^n \) \((1 \leq j \leq p)\) be the predictor varieties and \( y_l \in \mathbb{R}^n \) the response varieties \((1 \leq l \leq q)\). We assume that relation (2) holds. We consider the case in which the matrix of responses \( Y \), as in (1), has missing values. Then, we divide the \( i \)th row vector \( y^i \) of \( Y \) into
\[
y^i = (y^{i,\text{obs}}, y^{i,\text{mis}}),
\]
where \( y^{i,\text{obs}} \) is a vector that consists of the observed values, and \( y^{i,\text{mis}} \) consists of missing values. By (5), it follows that
\[
(y^{i,\text{mis}})^T | (y^{i,\text{obs}})^T \sim N(\mu_{i,\text{mis}} - K_{i,\text{mis,obs}}^{-1} K_{i,\text{mis,obs}}(y^{i,\text{obs}})^T - \mu_{i,\text{obs}}), K_{i,\text{mis,mis}}^{-1}).
\]

where we divide the mean vector \( \mu_i = \tilde{\beta}^T \bar{x}^i \) and the precision matrix \( K \) into
\[
\mu_i = (\mu_{i,\text{obs}}; \mu_{i,\text{mis}}), \quad K = \begin{pmatrix} K_{i,\text{obs,obs}} & K_{i,\text{obs,mis}} \\ K_{i,\text{mis,obs}} & K_{i,\text{mis,mis}} \end{pmatrix}
\]
for each \( i \). For the remainder of this article, we assume that for the matrix \( Y \) given by
\[ Y = (y_1, \ldots, y_q) = \begin{pmatrix} y_1^T \\ \vdots \\ y_n^T \end{pmatrix} = \begin{pmatrix} y_1^1 & y_1^2 & \cdots & y_1^q \\ \vdots & \vdots & \ddots & \vdots \\ y_n^1 & y_n^2 & \cdots & y_n^q \end{pmatrix}, \]  

(8)

there are no columns with entries that are missing values.

### 3.2 Algorithm to interpolate the data with missing values

Let us now derive an algorithm that performs multivariate regression and interpolates data with missing values. We assume the same conditions as in the previous subsection. For each \( i \), it follows that \((y_i^T|\hat{x}_i)^T \sim N(\mu_i = \tilde{B}^T \hat{x}_i, \Sigma)\); hence, the likelihood function \( L((y_i^T|\hat{x}_i)^T) \) is

\[
L((y_i^T|\hat{x}_i)^T) = (2\pi)^{\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left( -\frac{1}{2}((y_i^T - \mu_i)^T \Sigma^{-1}(y_i^T - \mu_i)) \right).
\]

where \(|\Sigma|\) is the determinant of \( \Sigma \). Thus, the log-likelihood function can be expressed as

\[
\sum_{i=1}^{n} \log L((y_i^T|\hat{x}_i)^T) = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} ((y_i^T - \mu_i)^T \Sigma^{-1}(y_i^T - \mu_i)) = -\frac{np}{2} \log(2\pi) + \frac{n}{2} \log |K| - \frac{1}{2} \sum_{i=1}^{n} ((y_i^T - \mu_i)^T \Sigma^{-1}(y_i^T - \mu_i)).
\]

Using this function, we set

\[
l(\tilde{B}, K; Y) = \frac{n}{2} \log |K| - \frac{n}{2} \sum_{i=1}^{n} \mu_i^T K \mu_i + \frac{1}{2} \sum_{i=1}^{n} (y_i^T)^T K (y_i^T) - \frac{1}{2} \text{tr} (K Y^T Y),
\]

(9)

where \( Y \) is a matrix given by (8).

We set the following \( l_1 \)-regularization of the function \( l \):

\[
- l(\tilde{B}, K; Y) + \lambda_1 \sum_{i \notin p} |k_i^T| + 2 \sum_{j=1}^{p} \sum_{j=1}^{q} \lambda_{j,2} |b_i^j| = \frac{n}{2} \log |K| - \frac{n}{2} \sum_{i=1}^{n} \mu_i^T K \mu_i + \frac{1}{2} \sum_{i=1}^{n} (y_i^T)^T K (y_i^T) - \frac{1}{2} \text{tr} (K Y^T Y) + \lambda_1 \sum_{i \notin p} |k_i^T| + 2 \sum_{j=1}^{p} \sum_{j=1}^{q} \lambda_{j,2} |b_i^j|,
\]

(10)

where \( \lambda_1 \geq 0 \) and \( \lambda_{j,2} \geq 0 \) are regularization parameters. We consider the following conditional mean of \( -l(\tilde{B}, K; Y) + \lambda_1 \sum_{i \notin p} |k_i^T| + 2 \sum_{j=1}^{p} \sum_{j=1}^{q} \lambda_{j,2} |b_i^j| \) as in (10):

\[
Q(\tilde{B}, K|\tilde{B}', K') = -\mathbb{E}[l(\tilde{B}, K; Y)|Y_{\text{obs}}, \tilde{B}', K'] + \lambda_1 \sum_{i \notin p} |k_i^T| + 2 \sum_{j=1}^{p} \sum_{j=1}^{q} \lambda_{j,2} |b_i^j|.
\]

(11)

In this case, we derive an algorithm to impute the data \( y^{\text{mis}} \) by solving the optimization problem for \( Q \) by applying the EM-algorithm. We call this procedure the sparse multivariate regression for responses with missing data (SMRM) algorithm. First, we provide initial values \( \tilde{B}^{(0)} \) and \( K^{(0)} \). Then, we compute the E-steps and M-steps as follows (cf. Reference 24).

**E-step:** For each \( i \), we denote the mean vector and precision matrix in the \( m \)-step (\( m = 0, 1, 2, \ldots \)) as \( \mu_i^{(m)} = \tilde{B}^{(m)T} (\hat{x}_i)^T \) and \( K^{(m)} \), respectively. We set

\[
c_i^{(m)} = \mu_i^{(m)} - (K_i^{(m)}|_{\text{mis mis}})^{-1} K_i^{(m)}|_{\text{mis obs}} ((y_i^{\text{obs}})^T - \mu_i^{(m)}|_{\text{obs}})
\]

(12)
for each $i$. We consider $\mathbf{c}^{(m)}$ as a column vector and use vectors $\mathbf{c}^{(m)}$ to impute the missing values $(\mathbf{y}^{\text{mis}(m)})^T$ in the $m$-step for each $i$. Then, the conditional means $\mathbb{E}[\mathbf{y}^i_j|\mathbf{y}^{\text{obs}}^i, \mu_i^{(m)}, K^{(m)}]$ and $\mathbb{E}[\mathbf{y}^j_{\ell'}|\mathbf{y}^{\text{obs}}_{\ell'}, \mu_{\ell'}^{(m)}, K^{(m)}]$ can be calculated as

\[
\mathbb{E}[\mathbf{y}^i_j|\mathbf{y}^{\text{obs}}^i, \mu_i^{(m)}, K^{(m)}] = \begin{cases} 
\mathbf{y}^i_j & \text{if } y^i_j \text{ is observed,} \\
\mathbf{c}^{(m)}_j & \text{if } y^i_j \text{ is missing,}
\end{cases}
\]

(13)

\[
\mathbb{E}[\mathbf{y}^j_{\ell'}|\mathbf{y}^{\text{obs}}_{\ell'}, \mu_{\ell'}^{(m)}, K^{(m)}] = \begin{cases} 
\mathbf{y}^j_{\ell'} & \text{if both } y^i_j \text{ and } y^j_{\ell'} \text{ are observed,} \\
\mathbf{c}^{(m)}_{j, \ell'} & \text{if } y^j_{\ell'} \text{ is observed and } y^i_j \text{ is missing,} \\
(\mu_i^{(m)})^{-1}_{y^i_j, y^j_{\ell'}} + \mathbf{c}^{(m)}_j \mathbf{c}^{(m)}_{\ell'} & \text{if both } y^i_j \text{ and } y^j_{\ell'} \text{ are missing.}
\end{cases}
\]

(14)

Using this method, we compute the function $Q(\mathbf{B}, K) = (\mathbf{B}, K)$, as in (11).

**M-step:** We compute the updates $(\mathbf{B}^{(m+1)}, K^{(m+1)})$ as the minimizer of $Q(\mathbf{B}, K) = (\mathbf{B}, K)$. To do this, we define the following function:

\[
g(\mathbf{B}, K) = \text{tr} \left[ \frac{1}{n} (\mathbf{Y} - \mathbf{B}^T K) (\mathbf{Y} - \mathbf{B}^T K) \right] - \log |K|.
\]

Then, our aim corresponds to solving the following optimization problem:

\[
(\hat{\mathbf{B}}, \hat{K}) = \arg\min_{\mathbf{B}, K} \left\{ g(\mathbf{B}, K) + \lambda_1 \sum_{i \neq \ell} |k_{i, \ell}| + 2 \sum_{j=1}^{p} \sum_{l=1}^{q} \lambda_{2, j} |b_{i, j}^l| \right\}.
\]

(15)

The algorithms used to numerically solve the above problem are the MRCE and MissGLASSO algorithms. By applying the MRCE-type and MissGLASSO-type algorithms, we solve the above problem and provide updates.

In summary, the algorithm works as follows:

**Algorithm** (SMRM). For fixed $\lambda_1$ and $\lambda_{2, j}$, initialize $\mathbf{B}^{(0)}$ and $K^{(0)}$.

**Step 1:** Impute $\mathbf{y}^{(m)}$ using $\mathbf{c}^{(m)}$ given by (12) for each $i = 1, \ldots, n$.

**Step 2:** Compute $(\hat{\mathbf{B}}^{(m+1)}, K^{(m+1)}) = (\mathbf{B}^{(m+1)}, K^{(m+1)})$ by solving (15) using MRCE and MissGLASSO. (See also Remark 1.)

**Step 3:** If $\sum_{j \neq \ell} |b_{i, j}^{(m+1)} - b_{i, j}^{(m)}| < \varepsilon$ for a given sufficiently small $\varepsilon > 0$, then stop. Otherwise, go to Step 1.

**Remark 1.** To compute the mean of $g$, we remark the following process:

1. The case of optimizing $\mathbf{B}$ for fixed $K$: to compute the mean of $g$, we use $\mathbf{Y}$ as a matrix with a row vector that consists of the complement vector $(\mathbf{y}^{(m)})^T = (\mathbf{y}^{\text{obs}}, (\mathbf{c}^{(m)})^T)^T$ given by the rule $\mathbb{E}[\mathbf{y}^i_j|\mathbf{y}^{\text{obs}}^i, \mu_i^{(m)}, K^{(m)}]$, as in (13).
2. The case of optimizing $K$ for fixed $\hat{\mathbf{B}}$: to compute the mean of $\mathbf{Y}^T \mathbf{Y}$, we use the rule of computation for $\mathbb{E}[\mathbf{y}^j_{\ell'}|\mathbf{y}^{\text{obs}}_{\ell'}, \mu_{\ell'}^{(m)}, K^{(m)}]$, as in (14).

**Remark 2.** For complete data, the SMRM algorithm performs similarly to the LASSO for sufficiently large $\lambda_1$. Thus, we may consider the SMRM algorithm as a generalization of the LASSO for multivariate regression analysis.

### 4 Applying the Algorithm to Real Data

In this section, we apply the SMRM algorithm to real data provided by Toray Industries, Inc. The data consist of the physical/mechanical properties of particular polymer compounds. To maintain confidentiality, we cannot display the full dataset; however, we describe the size and components of the data. The sample size of the data is $n = 114$, and the number of predictors and responses are $p = 26$ and $q = 22$, respectively. The predictor variables consist of the compounding ratios of the source materials. The response variables consist of the mechanical characteristics...
created by the source materials, such as Young’s modulus, tensile strength, elongation at break, flexural modulus, flexural strength, and the Charpy impact strength of polymer compounds. The responses have missing values, of which the rates range from 5% to 80% for each observation. In particular, the total ratio of missing values in the responses is 59.7% which is typical in the materials science field owing to the development process of focusing on the specific properties (see Figure 1).

During data analysis, we apply the SMRM algorithm to the data and compare the prediction accuracy of our proposed method with that of the LASSO. In this study, we use R version 4.0.2.

### 4.1 Procedure

Let \( x_j = (x_j^1, \ldots, x_j^p)^T \) \((1 \leq j \leq p)\) be predictor varieties, and let \( y_l = (y_l^1, \ldots, y_l^q)^T \) \((1 \leq l \leq q)\) be response varieties. Then, we divide the original data into training data:test data = 8:2; that is, we partition \( x_j \) and \( y_l \) into \( x_j^T = (x_{j, \text{train}}^T, x_{j, \text{test}}^T) \) and \( y_l^T = (y_{l, \text{train}}^T, y_{l, \text{test}}^T) \), respectively, with a partition rate of 8:2 for each \( j \) and \( l \) (see Figure 1). Although \( y_{l, \text{train}} \) and \( y_{l, \text{test}} \) may have missing values, we assume that \( x_j, x_{j, \text{train}}, \) and \( x_{j, \text{test}} \) are complete data. We perform the following analysis to compare the SMRM algorithm and the LASSO:

#### Step 1:
For each \( l \), we set \( y_{l, \text{train}}, \text{obs} \) and \( X_{l, \text{train}}, \text{obs} \), where \( y_{l, \text{train}}, \text{obs} \) is a vector, of which the elements are observed values in \( y_{l, \text{train}} \), and \( X_{l, \text{train}}, \text{obs} \) is the matrix corresponding to \( y_{l, \text{train}}, \text{obs} \). Then, we apply the LASSO to the dataset \((y_{l, \text{train}}, \text{obs}, X_{l, \text{train}}, \text{obs})\) for each \( l \). Regularization parameters, such as \( \lambda_{l, \text{train}} \) \((1 \leq l \leq q)\), are chosen via cross-validation. The prediction values, \( \hat{y}_{l, \text{test}}, \text{obs} \), are then computed. For each \( l \), we calculate the mean squared errors for the LASSO estimation, \( \text{MSE}_{l, \text{lasso}} \), for each \( l \) using

\[
\text{MSE}_{l, \text{lasso}} = \frac{||y_{l, \text{test}}, \text{obs} - \hat{y}_{l, \text{test}}, \text{obs}||^2}{\text{length}(y_{l, \text{test}}, \text{obs})},
\]

where \( ||w||^2 = w^T w \) for \( w \in \mathbb{R}^d \).

#### Step 2:
We assume the multivariate linear regression model \( Y_{\text{train}} = \tilde{X}_{\text{train}} \tilde{\beta}_{\text{train}} + \epsilon_{\text{train}} \) for the training data, where \( \epsilon_{l, \text{train}} \sim N(0, \Sigma_{\text{train}}) \) \((1 \leq l \leq q)\). Then, we apply the SMRM method to the training data \((\tilde{X}_{\text{train}}, Y_{\text{train}})\) for the appropriate pair \((\lambda_1, \lambda_2)\), where \( \lambda_2 = (\lambda_{2,j})_{1 \leq j \leq p, 1 \leq l \leq q} \) is a matrix with elements that are defined based on the regularization parameters \( \lambda_{l, \text{train}} \). Then, we obtain the estimator \( \hat{\beta}_{\text{train}} \) of parameter \( \tilde{\beta}_{\text{train}} \). Using \( \hat{\beta}_{\text{train}} \), we can compute the matrix of the prediction value of \( Y_{\text{test}} \), \( \hat{y}_{\text{SMRM}} \). We remark that \( \hat{y}_{\text{SMRM}} \) is complete data, whereas \( Y_{\text{test}} \) is data that have missing values. We calculate \( \text{MSE}_{l, \text{SMRM}} \) using

\[
\text{MSE}_{l, \text{SMRM}} = \frac{||y_{l, \text{test}}, \text{obs} - \hat{y}_{l, \text{test}}, \text{obs}||^2}{\text{length}(y_{l, \text{test}}, \text{obs})}
\]

for each \( l \), similarly to the LASSO.

#### Step 3:
We set \( \text{MSE}_{\text{lasso}} \) and \( \text{MSE}_{\text{SMRM}} \) using

\[
\text{MSE}_{\text{lasso}} = \sum_{l=1}^{q} \text{MSE}_{l, \text{lasso}}, \quad \text{MSE}_{\text{SMRM}} = \sum_{l=1}^{q} \text{MSE}_{l, \text{SMRM}}.
\]

The above MSEs are primarily affected by response variables with large variances. Thus, we define the MSEs that are not affected by the variance of the response variables as follows:

\[
\overline{\text{MSE}}_{\text{lasso}} = \sum_{l=1}^{q} (\text{MSE}_{l, \text{lasso}})^{-1} \text{MSE}_{l, \text{lasso}} (= q), \quad \overline{\text{MSE}}_{\text{SMRM}} = \sum_{l=1}^{q} (\text{MSE}_{l, \text{lasso}})^{-1} \text{MSE}_{l, \text{SMRM}}.
\]

Then, we compare \( \overline{\text{MSE}}_{\text{lasso}} \) and \( \overline{\text{MSE}}_{\text{SMRM}} \).
Remark 3. Because the SMRM algorithm is based on the multivariate normal distribution, the predicted $\hat{Y}_{SMRM}$ contains negative values. However, the physical property $y_i$ cannot assume negative values in a real-world situation. To avoid this, we first set $\log(Y)$ and consider it as the response matrix. Then, applying $\exp(\log(\hat{Y}))$ to the predicted matrix $\log(\hat{Y})$, we have $\hat{Y}$.

In Step 2, we use the $\lambda_2$ matrix for the SMRM algorithm. If the responses are standardized, we define the $\lambda_2$ matrix as $\lambda_2 = r\lambda$, where $r \in R \setminus \{0\}$ and

$$
\lambda = \begin{pmatrix}
\lambda_{1, \text{train}} & \cdots & \lambda_{q, \text{train}} \\
\vdots & \ddots & \vdots \\
\lambda_{1, \text{train}} & \cdots & \lambda_{q, \text{train}}
\end{pmatrix}_{p}.
$$

(19)

However, when the responses are not standardized, the variance of the response variable affects the regularization parameter; $\lambda_2$ must be different among the response variables. Thus, we perform the following procedure to reduce the effect of the variances:

Step 1: For each $l \in \{1, \ldots, q\}$, we estimate $y_{l, \text{train}, \text{obs}}$ using the LASSO with the regularization parameter $\lambda_{l, \text{train}}$, which is chosen via cross-validation.

Step 2: For each $l$, we calculate the MSEs, $t_l = ||y_{l, \text{train}, \text{obs}} - \hat{y}_{l, \text{train}, \text{obs}}^{\text{lasso}}||^2 / \text{length}(y_{l, \text{train}, \text{obs}})$, for the training data, where $\hat{y}_{l, \text{train}, \text{obs}}^{\text{lasso}}$ is the estimator for $y_{l, \text{train}, \text{obs}}$ by the LASSO in Step 1.

Step 3: Using $t_l$, which was obtained in Step 2, we define a vector $a$ as

$$
a = \begin{pmatrix}
a_1 \\
\vdots \\
a_q
\end{pmatrix},
$$

(20)

Step 4: We define a matrix as

$$
\lambda = \begin{pmatrix}
\lambda_{1, \text{train}} a_1 & \cdots & \lambda_{q, \text{train}} a_q \\
\vdots & \ddots & \vdots \\
\lambda_{1, \text{train}} a_1 & \cdots & \lambda_{q, \text{train}} a_q
\end{pmatrix}_{p}.
$$

(21)

Step 5: We set $\lambda_2 = r\lambda$ ($r \in R \setminus \{0\}$) and apply the SMRM algorithm using this matrix.

4.2 Comparison between the SMRM algorithm and the LASSO

Following the procedure explained in the previous subsection, we compare our method (the SMRM algorithm) to the LASSO using the data provided by Toray Industries, Inc. The regularization parameters $\lambda_{l, \text{train}}$ ($1 \leq l \leq 22$) for the training data and MSEs (MSE$^{\text{lasso}}$) for the test data by the LASSO are summarized in the first and second rows of Table 1.

The second row of Table 1 shows that the mechanical characteristics A, G, H, K, L, M, N, and V have large MSE values. These values significantly act on MSE$^{\text{lasso}}$ and also act on MSE$^{\text{SMRM}}$. Therefore, it is better to use MSE, as in (18), to compare the prediction accuracy between the LASSO and the SMRM algorithm by avoiding the dependence of the variance of responses.

We subsequently apply the SMRM algorithm. Because the responses of the data are not standardized, we calculate the vector $a$, as in (20) (this is a $22 \times 1$ matrix), to obtain the $\lambda_2$ matrix. The vector $a$ is listed in the third row of Table 1. Using the regularization parameters $\lambda_{l, \text{train}}$ ($1 \leq l \leq 22$), as in the first row of Table 1 and $a$, we obtain the $\lambda$ matrix, as given in (21). The row vector of $\lambda$ is shown in the fourth row of Table 1.

The values in the third row of Table 1 correspond to the reciprocals of the MSEs for each mechanical characteristic of the training data obtained by the LASSO. These MSE values vary significantly. The values in the fourth row of Table 1 can be considered as modified regularization parameters obtained by the LASSO in Step 1. A comparison of the first and
fourth rows of Table 1 shows that the values in the fourth row vary only slightly. Thus, one may make a stable estimation using the SMRM algorithm with $\lambda_2$ constructed by (21) instead of the matrix given by (19).

We consider the following cases of the pair $(\lambda_1, \lambda_2)$ for the SMRM algorithm:

- For $\lambda_1$, we consider $6.5 \times 10^{-3} \leq \lambda_1 \leq 1$ divided into 200 points of equal length under the log scale.
- We consider $\lambda_2$ as $r \lambda$ for $r = 3, 2, 1, 0.75, 0.5, 0.225, 0.2, 0.175, 0.1$.

As we apply the SMRM algorithm, we use the warm start method, which is outlined as follows. Let $\{\lambda_1^{(s)}\}_{s=1}^{200}$ be a sequence of $\lambda_1$, of which the initial value is $\lambda_1^{(1)} = 1$, and the end is $\lambda_1^{(200)} = 6.5 \times 10^{-3}$. For a fixed $\lambda_2$, we start with $(\lambda_1^{(1)}, \lambda_2) = (1, 2)$. Then, we obtain $\hat{B}$ and $\hat{K}^{(1)}$ via the SMRM algorithm. Next, we apply the SMRM algorithm to the pair of regularization parameters $(\lambda_1^{(2)}, \lambda_2)$ with the initial values $\hat{B} = B^{(1)}$ and $K = K^{(1)}$, which were obtained in the previous step. Inductively, we practice a similar analysis until $\lambda_1^{(200)} = 6.5 \times 10^{-3}$.

Our proposed algorithm incorporates the EM algorithm, which requires iteration. In each iteration, we need to conduct an E-step and an M-step. In the E-step, the maximum complexity is $O(nq^3)$, where $n$ is the sample size and $q$ is the size of the responses. Subsequently, we apply the M-step for each $\lambda_1^{(s)}$ using the MRCE algorithm. The total complexity of this algorithm is $O(nq^3T_{\lambda_1})$, where $T$ is the number of iterations of the EM algorithm and $n\lambda_t$ is the number of candidates of the $\lambda_1^{(s)}$. We also evaluate the complexity numerically by running the SMRM algorithm on this dataset. The runtimes of the SMRM algorithm for this setting are listed in Table 2. We list the runtimes for ten executions. The specifications of the personal computer (PC) used is as follows: OS, Windows 10 Pro version 21H1; processor, Intel(R) Core(TM) i7-1065G7 CPU @ 1.30 GHz & 1.50 GHz; RAM, 32 GB. The average analysis time is 2143.58 s. In our analysis, the maximum number of iterations is 2000 for each $\lambda_1^{(s)}$ ($1 \leq s \leq 200$). Thus, the maximum number of analyses in our setting is $9 \times 200 \times 2000 = 3,600,000$. However, for a relatively large $\lambda_1^{(s)}$, the algorithm can find the estimator of $B$ and $K$ easily, and hence we can obtain the result with fewer steps than 3,600,000. Thus, our algorithm may not have a critical issue of computational timing in practical usage.

In our observation, the values of $\text{MSE}^{\text{SMRM}}$ improve gradually whenever $\lambda_1^{(s)}$ is updated for a fixed $r \leq 1$ (see Figure 2 and Figure A1 in the Appendix). However, when $\lambda_1$ becomes smaller than a particular number, the SMRM algorithm is not stable, and $\text{MSE}^{\text{SMRM}}$ deteriorates. The following reasons can be considered:

- When the regularization parameter $\lambda_1$ for the SMRM algorithm is large, the correlation structure is not considered. In this case, imputation of missing values may not be effective because missing completion is achieved by taking

| TABLE 1 List of $\lambda_{1,\text{Train}}$, $\text{MSE}_{1}^{\text{lasso}}$ elements of $a$ and elements of $\lambda$ for each mechanical characteristic |
|---------------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Mech. char. | A | B | C | D | E | F | G | H | I | J | K |
| $\lambda_{1,\text{Train}}$ | 0.0014 | 0.0321 | 0.9544 | 1.7748 | 0.0380 | 0.0292 | 0.2526 | 0.8888 | 0.0127 | 0.0738 | 0.5167 |
| $\text{MSE}_{1}^{\text{lasso}}$ | 445.3167 | 0.6529 | 0.4304 | 88.7628 | 4.6077 | 0.1835 | 5873.6704 | 40,911.2809 | 0.0712 | 0.0368 | 374.3443 |
| Elements of $a$ | 159.71 | 541.05 | 18.02 | 8.02 | 214.25 | 12.76 | 21.94 | 23.55 | 2575.31 | 24.63 | 8.00 |
| Elements of $\lambda$ | 0.22 | 17.34 | 17.20 | 14.24 | 80.39 | 0.37 | 5.54 | 20.93 | 32.77 | 1.82 | 4.13 |

| TABLE 2 List of runtimes of the SMRM algorithm for our setting |
|-----------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Runtime (s) | 2157.56 | 2162.05 | 2274.28 | 2012.64 | 2106.67 | 2107.73 | 2131.73 | 2140.82 | 2233.16 | 2072.25 |
advantage of the correlation structure of the responses. Hence, the prediction accuracy may be worse than that of the LASSO.

- When $\lambda_1$ is appropriately small, missing values are well imputed using the correlation structure among the responses (or among the residuals). The prediction accuracy improves, owing to the contribution of the correlation structure.

- When $\lambda_1$ is too small, the estimated model overfits the data. Hence, the prediction error increases again.

In the case of $r > 1$, $\text{MSE}^{\text{SMRM}}$ deteriorates independent of $\lambda_1$. When $r > 1$, elements in $\lambda_2$ assume large values. For these data, the elements of the row vectors of the estimator $\hat{B}$ tend to be zeros. This is the reason why $\text{MSE}^{\text{SMRM}}$ is worse than the case of $r \leq 1$.

For the estimator $\hat{K}$ of the precision matrix $K$, we note that $\hat{K}$ is a diagonal matrix for sufficiently large $\lambda_1$. In these steps, although the SMRM algorithm performs stably, we can see that the prediction accuracy of the SMRM algorithm is worse than that of the LASSO in contrast to $\lambda_2$ (see Figure A1).

We observed the influence of $r$ on the prediction accuracy. As $r$ gradually decreases, the prediction accuracy for the SMRM algorithm improves. In particular, we observe that $r = 0.2$, that is, $\lambda_2 = 0.2\lambda_1$ with $\log(\lambda_1) = -4.96$ provide the best prediction accuracy (see Figure 2). When $r = 0.1$, the prediction accuracy decreases compared with the case where $r = 0.2$ (see Figure A1).

Furthermore, comparing $\text{MSE}^{\text{SMRM}}$ with $\lambda_2 = 0.2\lambda$ and $\text{MSE}^{\text{lasso}}$, it can be observed that $\text{MSE}^{\text{SMRM}} < \text{MSE}^{\text{lasso}}$ holds; that is, the SMRM algorithm is superior to the LASSO for a suitably small $\lambda_1$. Because $\lambda_1$ affects the correlation structure among responses, the prediction accuracy may be improved by estimating the responses multivariately with the appropriate correlation structure, instead of individually. In fact, the ratios of the maximum and minimum eigenvalues of $\hat{K}$ increase when $\lambda_1$ gets small. Figure 3 shows the ratios for $r = 0.2$. It can be seen that the ratio looks unchanged around $-3 \leq \log(\lambda_1) \leq 0$. This implies that $\hat{K}$ is estimated as a diagonal one for such $\lambda_1$. Moreover, one can see that the ratio takes a large value for a small $\lambda_1$. This can be considered as $\hat{K}$ being no diagonal matrix, that is, the correlation structure is estimated among the responses.

Next, we consider the MSEs in the case of $\lambda_2 = 0.2\lambda$ for mechanical characteristics individually (see Figure A2). In the individual analysis, we consider the mechanical characteristics $C$ and $D$, that is, the third and fourth mechanical characteristics (see Figure 4). For $-4.35 \leq \log(\lambda_1) \leq 0$, $\text{MSE}^{\text{SMRM}} > \text{MSE}^{\text{lasso}}$ (1) holds, where $l = 3, 4$. This implies that the prediction accuracy obtained by the SMRM algorithm is lower than that of the LASSO. However, for $\log(\lambda_1) \leq -4.38$, we find that $\text{MSE}^{\text{SMRM}} < \text{MSE}^{\text{lasso}}$ ($l = 3, 4$) holds. Thus, it can be observed that the prediction accuracy is improved by using the correlation structure among the responses. Although $\log(\lambda_1) = -4.96$ exhibits the best $\text{MSE}^{\text{SMRM}}$ prediction accuracy, as mentioned above, $\text{MSE}^{\text{SMRM}}_3$ and $\text{MSE}^{\text{SMRM}}_4$ provide the best values for $\log(\lambda_1) = -4.99$ (one after “the best” with respect to $\text{MSE}$).

![Figure 2](image-url)  

**FIGURE 2** $\text{MSE}^{\text{SMRM}}$ (blue) and $\text{MSE}^{\text{lasso}}$ (red). The x-axis represents $\log(\lambda_1)$ and the y-axis $\text{MSE}$. The dotted vertical line (pink) indicates $\log(\lambda_1) = -4.96$, which provides the best case.
FIGURE 3  Relation between ratios $\rho$ of the maximum and minimum eigenvalues of $\hat{R}$ and $\lambda_1$ for $r = 0.2$. The x-axis represents $\log(\lambda_1)$ and the y-axis represents the ratio $\rho$.

FIGURE 4  $\overline{\text{MSE}_{\text{SMRM}}}$ (blue) and $\overline{\text{MSE}_{\text{lasso}}}$ (red) of mechanical characteristics C (left) and D (right). The x-axis represents $\log(\lambda_1)$. The dotted vertical line (pink) represents $\log(\lambda_1) = -4.96$, which indicates the best case for MSE.

To consider the reason for these differences, we use heat maps of correlation structures (see Figure 5). Here, “the best,” “better1,” and “better2” are named with respect to $\overline{\text{MSE}_{\text{SMRM}}}$. (When we observe $\overline{\text{MSE}_{\text{lasso}}}$ individually, note that “better2” represents the case for the best prediction accuracy of C and D.) We first note that C and D have a particular physical/mechanical relationship with each other.* It is noticeable that C and D have a strong positive correlation. Therefore, it appears that if a positive correlation between C and D gradually increases, the prediction accuracy improves.

On the other hand, the SMRM algorithm estimates the correlation structure among C, D, and O (15th mechanical characteristic). Because C (resp. D) and O represent different mechanical properties, it is difficult to emphasize their correlation structure via experiments. Therefore this may be considered as a hidden relation among mechanical properties, and it seems that the sparse multivariate regression method using a precision matrix contributes to identifying such relations. If a positive correlation among C, D, and O is suitably set, the prediction accuracy of C and D will improve. For the
FIGURE 5 Heat maps of correlation structures among responses for $\lambda_2 = 0.2\lambda_1$. From top left to bottom right, we present heat maps for $\log(\lambda_1) = -4.93, -4.96, -4.99, -5.04$

mechanical characteristic O, we notice that the prediction accuracy is better when the positive correlation of O with C and D increases (see Figure 6).

By the above observations, a suitable positive correlation structure among C, D, and O for this dataset affects the prediction accuracy of these mechanical characteristics. Furthermore, we can identify unexpected relations among the responses, similar to the above characteristics, using our method.

For other mechanical characteristics, see the Appendix.

Remark 4. Normalized (scaled) data can be applied to the SMRM algorithm. However, we should remark that the covariance matrix constructed without the missing values may not be positive definite. Actually, the covariance matrix of our data is a non-positive definite matrix. In addition, if we do not have sufficient data that are not missing values, we might not get precise normalization of the data. Moreover, we need to change parameters ($\lambda_1, \lambda_2$) suitably for scaled data so that the algorithm converges. In fact, although we tried to apply the SMRM algorithm to scaled data, the convergence of the algorithm tended to be slow when $\lambda_1$ got progressively smaller. Therefore it is difficult to normalize data with missing values and apply the SMRM algorithm to such data.
TERAMOTO and HIROSE

FIGURE 6  $\text{MSE}^{\text{SMRM}}$ (blue) and $\text{MSE}^{\text{lasso}}$ (red) of the mechanical characteristic O. The x-axis represents $\log(\lambda_1)$. The dotted vertical line (pink) represents $\log(\lambda_1) = -4.96$

5 | CONCLUSION

In this study, we proposed a novel method called the SMRM algorithm with the intention of applying it to materials science. As the data structure of materials science often contains two features—(1) material properties are multivariate, and (2) they have often missing values—it seems that the MRCE and MissGLASSO algorithms should work effectively. Unfortunately, these methods cannot be applied directly in our setting. However, by modifying them and establishing a suitable framework, we constructed the proposed algorithm (Section 3). Owing to the regularization for the correlation structure, we can improve the prediction accuracy and may find the unexpected relation among the response variables. Actually, in the real data analysis, we found the unexpected relation among the response variables of the data (Section 4). Further, we verified that our proposed method is superior to the LASSO for the data. For these reasons, we expect that our proposed method has a possibility to contribute to the progress of materials science and related areas.

Our proposed procedure performed worse than the LASSO for some material properties. The poor performance is possibly due to there being a large number of missing values. As the ratio of missing values increases, the prediction accuracy of our proposed method becomes poor. As future work, it would be interesting to investigate the influence of the ratio of missing values on our proposed method. Further, applying the SMRM to a dataset where both $X$ and $Y$ have missing values would be an interesting future prospect. In this case, we need to assume a suitable joint distribution for $X$ and $Y$. Moreover, we would need to prepare new parameters to estimate $X$. Therefore to handle this case, we will improve the SMRM algorithm to resolve these issues. However, such work is beyond the scope of the present article. Thus, it will be considered as a future research topic.

ACKNOWLEDGMENTS

The authors are grateful to Keiichiro Nomura, Sadayuki Kobayashi, and Kohei Koyanagi for fruitful discussions and valuable advice. They also express their gratitude to Professors Keiji Tanaka, Satoru Yamamoto, Shigeru Kuchii, and Shigeru Taniguchi for their valuable comments. They would also like to thank the anonymous referees for reading the manuscript carefully and providing constructive comments. This work was supported by JST-Mirai Program Grant Number JPMJMI18A2, Japan.

DATA AVAILABILITY STATEMENT

Embargo on data due to commercial restrictions.

ENDNOTES

∗This is indicated by Nomura, Kobayashi, and Koyanagi, who provided these data.
†This is suggested by the referee.
ORCID
Keisuke Teramoto  https://orcid.org/0000-0002-6990-4898

REFERENCES
1. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-555.
2. Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. NPJ Comput Mater. 2017;3(1):1-13.
3. Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. NPJ Comput Mater. 2018;4(1):1-8.
4. Pilania G, Wang C, Jiang X, Rajasekaran S, Ramprasad R. Accelerating materials property predictions using machine learning. Sci Rep. 2013;3(1):1-6.
5. Chen C-T, Gu GX. Machine learning for composites materials. MRS Commun. 2019;9(2):556-566.
6. Wei J, Chu X, Sun X-Y, et al. Machine learning in materials science. InfoMat. 2019;1(3):338-358.
7. Brook RJ, Arnold GC. Applied Regression Analysis and Experimental Design. CRC Press; 1985.
8. Vanschoren J, Blockeel H, Pfahringer B, Holmes G. Experiment databases. Mach Learn. 2012;87(2):127-158.
9. Brook R, Arnold G. Applied Regression Analysis and Experimental Design. CRC Press; 1985.
10. Da Ros S, Schwab M, Pinto JC. Parameter Estimation and Statistical Methods. Elsevier; 2017.
11. Han SZ, Choi E-A, Lim SH, Kim S, Lee J. Alloy design strategies to increase strength and its trade-offs together. Prog Mater Sci. 2021;117:100720.
12. Han SZ, Choi E-A, Lim SH, Kim S, Lee J. Alloy design strategies to increase strength and its trade-offs together. Prog Mater Sci. 2021;117:100720.
13. Rothman AJ, Levina E, Zhu J. Sparse multivariate regression with covariance estimation. J Comput Graph Stat. 2010;19:947-962.
14. Wang J. Joint estimation of sparse multivariate regression and conditional graphical models. Stat Sin. 2015;25(3):831-851.
15. Tibshirani R. Regression shrinkage and selection via the lasso. Ann Stat. 1996;58(1):267-288.
16. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B. 2005;67(2):301-320.
17. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9:432-441.
18. Roth V. The generalized LASSO. IEEE Trans Neural Netw. 2004;15(1):16-28.
19. Yuan L, Liu J, Ye J. Efficient methods for overlapping group lasso. IEEE Trans Pattern Anal Mach Intell. 2013;35(9):2104-2116.
20. Graham JW. Missing Data. Analysis and Design. Statistics for Social and Behavioral Sciences. Springer; 2012.
21. Enders CK, Bandalos DL. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. Struct Equ Model. 2001;8(3):430-457.
22. Hirose K, Kim S, Kano Y, Imada M, Yoshida M, Matsuo M. Full information maximum likelihood estimation in factor analysis with a large number of missing values. J Stat Comput Simul. 2016;86(1):91-104.
23. Städler N, Bühlmann P. Missing values: sparse inverse covariance estimation and an extension to sparse regression. Stat Comput. 2012;22(2):19-235.
24. Buck SF. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. J R Stat Soc Ser B. 1960;22:302-306.
25. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. Ann Stat. 2006;34(3):1436-1462.
26. She Y, He Y, Li S, Wu W. Joint association graph screening and decomposition for large-scale linear dynamical systems. IEEE Trans Signal Process. 2015;63:389-401.
27. Chatfield C, Collins A. Introduction to Multivariate Analysis. Chapman & Hall; 1980.
28. Lauritzen SL. Graphical Models. Oxford Statistical Science Series. Vol 17. The Clarendon Press, Oxford University Press, Oxford Science Publications; 1996.

How to cite this article: Teramoto K, Hirose K. Sparse multivariate regression with missing values and its application to the prediction of material properties. Int J Numer Methods Eng. 2022;123(2):530-546. doi: 10.1002/nme.6867

APPENDIX. FIGURES OF THE PREDICTION ACCURACY OF REAL DATA

This appendix presents figures obtained by real data analysis (see Section 4). First, figures for $\text{MSE}^{\text{SMRM}}$ where $r = 3, 2, 1, 0.75, 0.5, 0.225, 0.2, 0.175, 0.1, \lambda_2 = r \lambda_1$, and $\text{MSE}^{\text{lasso}}$ are shown in Figure A1. When $r > 1$, $\text{MSE}^{\text{SMRM}}$ is quite inferior to $\text{MSE}^{\text{lasso}}$. One can recognize that taking $r$ progressively smaller, $\text{MSE}^{\text{SMRM}}$ is improved in stages and $\text{MSE}^{\text{SMRM}}$ achieves the best value at $r = 0.2$. Subsequently, $\text{MSE}^{\text{SMRM}}$ gets worse again.
We next list $\hat{\text{MSE}}^{\text{SMRM}}$ with $\lambda_2 = 0.2 \lambda$ and $\hat{\text{MSE}}^{\text{lasso}}$ for each mechanical characteristic in Figure A2. By the modification, $\hat{\text{MSE}}^{\text{lasso}}_l$ takes 1 for each mechanical characteristic. One can observe that $\hat{\text{MSE}}^{\text{SMRM}}_l$ for C, D, N, O, Q, and T (i.e., \(l = 3, 4, 15, 17, 20\)) affect the total $\hat{\text{MSE}}^{\text{SMRM}}$.

**FIGURE A1** $\hat{\text{MSE}}^{\text{SMRM}}$ (blue) and $\hat{\text{MSE}}^{\text{lasso}}$ (red)
Figure A2: $\overline{\text{MSE}}_{\text{SMRM}}$ (blue) and $\overline{\text{MSE}}_{\text{lasso}}$ (red) for each mechanical characteristic. The x-axis represents $\log(\lambda_1)$ and the y-axis $\overline{\text{MSE}}$. 