Hindawi
Journal of Mathematics
Volume 2024, Article ID 3692883, 8 pages
https://doi.org/10.1155/2024/3692883

Research Article

The Alternating Direction Method of Multipliers for Sufficient Dimension Reduction

Sheng Ma, Qin Jiang, and Zaiqiang Ku

Department of Mathematics, Huanggang Normal University, Huanggang, Hubei, China

Correspondence should be addressed to Qin Jiang; jiangqin999@126.com

Received 4 November 2023; Revised 17 April 2024; Accepted 24 April 2024; Published 13 May 2024

Academic Editor: Barbara Martinucci

The minimum average variance estimation (MAVE) method has proven to be an effective approach to sufficient dimension reduction. In this study, we apply the computationally efficient optimization algorithm named alternating direction method of multipliers (ADMM) to a particular approach (MAVE or minimum average variance estimation) to the problem of sufficient dimension reduction (SDR). Under some assumptions, we prove that the iterative sequence generated by ADMM converges to some point of the associated augmented Lagrangian function. Moreover, that point is stationary. It also presents some numerical simulations on synthetic data to demonstrate the computational efficiency of the algorithm.

1. Introduction

It is well known that dimension reduction is a highly efficient way in visualization and statistical analysis of high-dimensional data. It is assumed that the predictor vector $\mathbf{X} \in \mathbb{R}^p$ affects the response variable $Y \in \mathbb{R}$ only through a few linear combinations $\beta_1^T\mathbf{X}, \beta_2^T\mathbf{X}, \cdots, \beta_d^T\mathbf{X}$, with $d < p$. Thus, it implies that all the information of $\mathbf{X}$ about Y is summarized by $\mathbf{B}^T\mathbf{X}$, with $\mathbf{B} = (\beta_1, \beta_2, \cdots, \beta_d)$. The column vectors of $\mathbf{B}$ span the subspace $\mathbf{S}(\mathbf{B})$. We call this subspace $\mathbf{S}(\mathbf{B})$ the sufficient dimension reduction (SDR) subspace (see Li [1], Cook [2]). To estimate $\mathbf{S}(\mathbf{B})$, we do dimension reduction with the smallest possible value of d.

In the literature of dimension reduction, two most popular problems have been widely studied. One question is how does the conditional distribution of $Y \mid \mathbf{X}$ vary with $\mathbf{X}$. Specifically, it aims at seeking a few linear combinations $\mathbf{B}^T\mathbf{X}$ such that

$$P(Y \le y \mid \mathbf{X}) = P\left(Y \le y \mid \mathbf{B}^T\mathbf{X}\right), \quad y \in \mathbb{R}. \tag{1}$$

Then, we can find the subspace $\mathbf{S}(\mathbf{B})$ satisfying (1). If $\cap\mathbf{S}(\mathbf{B})$ is a SDR subspace, we call it the central subspace (CS) and denote it by $\mathbf{S}_{Y|\mathbf{X}}$. In many real applications, $E(Y \mid \mathbf{X})$ is of most interest to the researchers, so the objective of the problem is to seek out a matrix $\mathbf{B}_{p\times d}$, satisfying

$$E(Y \mid \mathbf{X}) = E\left(Y \mid \mathbf{B}^T\mathbf{X}\right). \tag{2}$$

We call the intersection of all SDR subspaces, the central mean space (CMS), and denote it by $\mathbf{S}_{\mathrm{E}(Y|\mathbf{X})}$, if it is still a SDR subspace. Cook [2] gave more discussions about the CS and the CMS, including Cook and Li [3].

There were various methods to recover $\mathbf{S}_{Y|\mathbf{X}}$ or $\mathbf{S}_{\mathrm{E}(Y|\mathbf{X})}$ in the references. For example, Li and Duan [4] studied the ordinary least squares. Later, Li [1] studied sliced inverse regression. Cook and Weisberg [5] put forward sliced inverse variance estimation. Then, Li [6] studied principal Hessian directions. Then, Xia et al. [7] also considered minimum average variance estimation (MAVE). Li and Wang [8] considered directional regression, too. Then, density MAVE was of interest to Xia [9]. Sliced regression-based MAVE was considered by Wang and Xia [10]. Efficient semiparametric estimation was studied by Ma and Zhu [11–13]. Most of these approaches used the principle of inverse regression with the well-known limitation—the need for a linearity condition on the covariates. MAVE and some other semiparametric methods do not have strong

hypotheses on the design of covariates. Moreover, they have wider applicability. Furthermore, in many situations, they yield more accurate estimators of the reduced dimensional space, see references Ma and Zhu [11], Xia et al. [7], and Xia [9]. We know MAVE is not robust to outliers in the response because of the use of least squares and is more computationally intensive due to the use of kernel smoothing. The MAVE-type methods estimate the column space of $\mathbf{B} \in \mathbb{R}^{p\times d}$ through minimizing the loss criterion described later in (14) with the orthogonality constraint $\mathbf{B}^T\mathbf{B} = I_d$. Thus, the implementation of the MAVE-type methods involves nonconvex optimization problems based on the nonconvexity with the orthogonality constraint. Therefore, heuristic algorithms are often used to have no convergence guarantees in the literature.

Recently, many researchers adopted ADMM algorithm to high-dimensional statistics and machine learning, see Yin et al. [14], Xue et al. [15], Zhang and Zou [16], Gu et al. [17], and Kapla et al. [18], and many other topics, such as matrix completion, tensor completion, and sparse recovery, see Li et al. [19, 20], Liu et al. [21], and Shi et al. [22]. The main advantages of the ADMM algorithm are its flexibility at simplifying a diversity of optimization problems and its good convergence property, see Boyd et al. [23]. In this study, we demonstrate that the ADMM algorithm can be adapted to solve the optimization problem of the aforementioned MAVE-type methods. Moreover, we prove that the proposed algorithm will converge to some point that is stationary, through Wang et al. [24]'s theory for ADMM on nonconvex problems. To our knowledge, for MAVE, the proposed ADMM algorithm is the first one with the convergence property. Details refer to Theorem 2 in the study. In addition, Zhang et al. [25] proposed a robust estimation through regularization with case-specific parameters to achieve robust estimation and outlier detection simultaneously.

In the rest of this article, we present the proposed ADMM algorithm and prove its convergence properties in Section 2. Then, numerical simulations are conducted to illustrate the proposed algorithm in Section 3. Finally, a brief discussion is concluded in Section 4. Some technical details are relegated to Appendix.

2. The Proposed Algorithm Based on ADMM

The following ADMM algorithm is applicable to all the MAVE-type approaches. For simplicity, we only present ADMM algorithm to estimate the CMS problem. Xia et al. [7] considered the following model:

$$y = g\left(\mathbf{B}^T x\right) + \varepsilon, \tag{3}$$

for dimension reduction. Here, the matrix $\mathbf{B} \in \mathbb{R}^{p\times d}$ satisfies $\mathbf{B}^T\mathbf{B} = I_d$ with some $d < p$. The function g is smooth, but it is not known, with $E(\varepsilon \mid x) = 0$. Here, we focus on the estimation of $\mathbf{B}$ by assuming that d is known.

Consider the simple case $g(t) = t$. Let $\mathbf{B}_0$ be the solution of

$$\min_{\mathbf{B}} \mathrm{E}\left\{y - \mathrm{E}\left(y \mid \mathbf{B}^T\mathbf{X}\right)\right\}^2. \tag{4}$$

It is known that the conditional variance, for given $\mathbf{B}^T\mathbf{X}$, should be

$$\sigma^2\left(\mathbf{B}^T\mathbf{X}\right) = E\left[\left\{y - \mathrm{E}\left(y \mid \mathbf{B}^T\mathbf{X}\right)\right\}^2 \middle| \mathbf{B}^T\mathbf{X}\right], \tag{5}$$

for $\mathbf{B} = (\beta_1, \beta_2, \cdots, \beta_d)$. Thus, from the conditional expectation, we get

$$E\left\{y - \mathrm{E}\left(y \mid \mathbf{B}^T\mathbf{X}\right)\right\}^2 = E\left(\sigma^2\left(\mathbf{B}^T\mathbf{X}\right)\right). \tag{6}$$

So, we know the expression (6) is equivalent to

$$\min_{\mathbf{B}} E\left(\sigma^2\left(\mathbf{B}^T\mathbf{X}\right)\right), \quad \mathbf{B}^T\mathbf{B} = I_d. \tag{7}$$

This is called MAVE.

For a sample $\{(\mathbf{X}_i, y_i)\}$, $i = 1, \cdots, n$, let

$$g_{\mathbf{B}}(v_1, v_2, \cdots, v_d) = \mathrm{E}\left[\left\{y - \mathrm{E}\left(y \mid \beta_1^T\mathbf{X} = v_1, \beta_2^T\mathbf{X} = v_2, \cdots, \beta_d^T\mathbf{X} = v_d\right)\right\}^2 \mid \mathbf{B}^T\mathbf{X}\right]. \tag{8}$$

For any given $\mathbf{X}_0$, we know $\mathrm{E}(y_i \mid \mathbf{B}^T\mathbf{X}_i)'$ local linear expansion, at $\mathbf{X}_0$, can be expressed as follows:

$$\mathrm{E}\left(y_i \mid \mathbf{B}^T\mathbf{X}_i\right) \approx a + b^T\mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_0\right). \tag{9}$$

Here, $a = g_{\mathbf{B}}(\mathbf{B}^T\mathbf{X}_0)$ and $b^T = (b_{(1)}, b_{(2)}, \cdots, b_{(d)})$. In addition,

$$b_{(k)} = \frac{\partial g_{\mathbf{B}}(v_1, v_2, \cdots, v_d)}{\partial v_k} \mid_{v_1=\beta_1^T\mathbf{X}_0, v_2=\beta_2^T\mathbf{X}_0, \cdots, v_d=\beta_d^T\mathbf{X}_0}, \quad k = 1, 2, \cdots, d. \tag{10}$$

Obviously, the residuals we want to find are the following expressions:

$$y_i - g_{\mathbf{B}}\left(\mathbf{B}^T\mathbf{X}_i\right) \approx y_i - \left\{a + b^T\mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_0\right)\right\}. \tag{11}$$

Through the spirit of the local estimation that is linear smoothing, it is natural to estimate $\sigma^2(\mathbf{B}^T\mathbf{X})$, using the following approximation:

$$\begin{aligned}\sum_{i=1}^{n}\left\{y_i - \mathrm{E}\left(y_i \mid \mathbf{B}^T\mathbf{X}_i\right)\right\}^2\omega_{i0} &\approx \sum_{i=1}^{n}\left\{y_i - \left(a + b^T\mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_0\right)\right)\right\}^2\omega_{i0}.\end{aligned} \tag{12}$$

Here, ω_{i0} are the weights satisfying $\sum_{i=1}^{n}\omega_{i0} = 1$. Two choices of the weights ω_{i0} were given in Xia et al. [7]. The estimation of a is the minimum point of expression (6). Of course, the

same is true for b. Then, we know the estimation of $\sigma^2(\mathbf{B}^T\mathbf{X}_0)$ is the minimum value of the expression (6). That is,

$$\widehat{\sigma}_{\mathrm{B}}^2\left(\mathbf{B}^T\mathbf{X}_0\right) = \min_{a,b}\left(\sum_{i=1}^{n}\left\{y_i - \left(a + b^T\mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_0\right)\right)\right\}^2\omega_{i0}\right). \tag{13}$$

Due to the expressions (4), (7), and (13), the MAVE method estimates $\mathbf{B}$ by solving the following minimization problem:

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\left\{y_i - a_j - b_j{}^T\mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_j\right)\right\}^2\omega_{\mathrm{ij}} \quad \text{subject to } \mathbf{B}^T\mathbf{B} = I_d, \tag{14}$$

where $b_j{}^T = (b_{j1}, b_{j2}, \cdots, b_{jd})$. Xia et al. [7] solved the minimizer of (14) by an iterative algorithm for $\mathbf{B}$ and $(a_j, b_j)_{,j=1,\cdots,n}$. Although the iterative algorithm seems feasible, it is difficult to establish its convergence property due to the nonconvexity of the orthogonality constraint $\mathbf{B}^T\mathbf{B} = I_d$. Zhu [26] studied optimization problems on Stiefel manifold and mentioned that problem (14) can be solved.

Machine learning, computer vision, and statistics are all research hotspots. In the fields of them, there are many optimization problems that are structured convex, such as Zanni et al. [27]. For solving various convex or nonconvex problems that arise in the fields of machine learning, computer vision, and statistics, the alternating direction method with multipliers (ADMM) has been a powerful and successful method, since Gabay and Mercier [28] introduced the ADMM, and then, its convergence properties for convex objective functions have been extensively studied. Many researchers successfully applied ADMM to solve these problems. For example, Boyd et al. [23] gave the recent survey paper. Liu et al. [29] and Cascarano et al. [30] also studied them.

Recently, many researchers successfully used some variants of ADMM to solve some previous nonconvex problems. For example, Hong et al. [31] discussed the convergence properties of variants of ADMM and then applied them to nonconvex problems. Moreover, they established the iteration complexity of ADMM. For the multiblock separable optimization problems, Guo et al. [32] studied the case of linear constraints and no convexity of the related component functions. For the iteration sequence generated by ADMM, they drew a conclusion that each clustering point of it is a critical point. For the multiblock proximal ADMM's two linearized variants, Jiang et al. [33] studied their iteration complexity. Especially, for minimizing an objective function, lack of convexity, and possible smoothness and constrained by coupled linear identities, Wang et al. [24] studied the convergence of ADMM. To solve the optimization problems with separability and nonconvexity, Jia et al. [34] considered the convergence rate of the ADMM.

In this study, we are interested in the following ADMM algorithm to optimize problem (14) under the framework of Wang et al. [24]. General ADMM flow is referred to the original paper, such as Gabay and Mercier [28], for better understanding.

Let $\mathbf{S} = \left\{\mathbf{B} \in \mathbb{R}^{p\times d}\colon \mathbf{B}^T\mathbf{B} = I_d\right\}$. Then, we can reformulate the problem of minimizing (8) as that of minimizing

$$f(\mathbf{B}, \mathbf{a}, \mathbf{b}, \mathbf{A}) = \mathrm{I}_{\mathbf{S}}(\mathbf{A}) + h(\mathbf{B}, \mathbf{a}, \mathbf{b}), \quad \text{subject to} \quad \mathbf{A} - \mathbf{B} = 0, \tag{15}$$

where the function $\mathrm{I}_{\mathbf{S}}$ is simply the indicator function of the set $\mathbf{S}$, i.e., $\mathrm{I}_{\mathbf{S}}(\mathbf{A}) = 0$ if $\mathbf{A} \in \mathbf{S}$ or ∞ if $\mathbf{A} \notin \mathbf{S}$ and

$$h(\mathbf{B}, \mathbf{a}, \mathbf{b}) = \sum_{j=1}^{n}\sum_{i=1}^{n}\left\{y_i - a_j - b_j{}^T\mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_j\right)\right\}^2\omega_{\mathrm{ij}}, \tag{16}$$

with $\mathbf{a} = \left\{a_j, j = 1, \cdots, n\right\}, \mathbf{b} = \left\{b_j, j = 1, \cdots, n\right\}$. The function h is proper, differentiable, and nonconvex.

The inner product of $\mathbf{A}$ and $\mathbf{B}$ is, as usual, written to be $\langle\mathbf{A}, \mathbf{B}\rangle = \mathrm{Tr}(\mathbf{A}^T\mathbf{B})$. Here, $\mathrm{Tr}(\cdot)$ is the trace operator. Then, the following function,

$$L_\rho(\mathbf{B}, \mathbf{a}, \mathbf{b}, \mathbf{A}, \Theta) = f(\mathbf{B}, \mathbf{a}, \mathbf{b}, \mathbf{A}) + \langle\Theta, \mathbf{A} - \mathbf{B}\rangle + \frac{\rho}{2}\|\mathbf{A} - \mathbf{B}\|_{\mathrm{F}}^2, \tag{17}$$

is the Lagrangian function of (15), where ρ is the penalty parameter, $\Theta \in \mathbb{R}^{p\times d}$ is the Lagrange multiplier, and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm.

We can compute the estimations of $(\mathbf{B}, \mathbf{a}, \mathbf{b}, \mathbf{A}, \Theta)$ through the following ADMM algorithm.

For a given value of $\mathbf{A}^{(m)}$, $\mathbf{B}^{(m)}$, and $\Theta^{(m)}$ at step m, the iterative process is as follows:

$$\left(\mathbf{a}^{(m+1)}, \mathbf{b}^{(m+1)}\right) = \arg\min_{\mathbf{a},\mathbf{b}} L_\rho\left(\mathbf{B}^{(m)}, \mathbf{a}, \mathbf{b}, \mathbf{A}^{(m)}, \Theta^{(m)}\right), \tag{18}$$

$$\mathbf{B}^{(m+1)} = \arg\min_{\mathbf{B}} L_\rho\left(\mathbf{B}, \mathbf{a}^{(m+1)}, \mathbf{b}^{(m+1)}, \mathbf{A}^{(m)}, \Theta^{(m)}\right), \tag{19}$$

$$\mathbf{A}^{(m+1)} = \arg\min_{\mathbf{A}} L_\rho\left(\mathbf{B}^{(m+1)}, \mathbf{a}^{(m+1)}, \mathbf{b}^{(m+1)}, \mathbf{A}, \Theta^{(m)}\right), \tag{20}$$

$$\Theta^{(m+1)} = \Theta^{(m)} + \rho\left(\mathbf{B}^{(m+1)} - \mathbf{A}^{(m+1)}\right). \tag{21}$$

Obviously, the function,

$$\left(\mathbf{a}^{(m+1)}, \mathbf{b}^{(m+1)}\right) = \arg\min_{\mathbf{a},\mathbf{b}} h\left(\mathbf{B}^{(m)}, \mathbf{a}, \mathbf{b}\right), \tag{22}$$

is equivalent to (10). Then, by the definition of $h(\mathbf{B}^{(m)}, \mathbf{a}, \mathbf{b})$ and simple algebraic manipulation, we obtain the explicit form

$$\begin{pmatrix} a_j^{(m+1)} \\ b_j^{(m+1)} \end{pmatrix} = \left\{ \sum_{i=1}^{n} \omega_{\mathrm{ij}} \begin{pmatrix} 1 \\ \mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_j\right) \end{pmatrix} \begin{pmatrix} 1 \\ \mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_j\right) \end{pmatrix}^T \right\}^{-1} \sum_{j=1}^{n} \omega_{\mathrm{ij}} \begin{pmatrix} 1 \\ \mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_j\right) \end{pmatrix} y_j, \tag{23}$$

for $j = 1, \cdots, n$.

In (19), after throwing away the terms independent of $\mathbf{B}$, one has

$$\mathbf{B}^{(m+1)} = \underset{\mathbf{B}}{\operatorname{argmin}}\, h\left(\mathbf{B}, \mathbf{a}^{(m+1)}, \mathbf{b}^{(m+1)}\right) + \frac{\rho}{2}\left\|\mathbf{B} - \mathbf{A}^{(m)} - \frac{1}{\rho}\Theta^{(m)}\right\|_{\mathrm{F}}^2. \tag{24}$$

Since the right term of above expression is the quadratic function of $\mathbf{B}$, $\mathbf{B}^{(m+1)}$ has an explicit form. Specifically, we can obtain $\mathbf{B}^{(m+1)}$ by the gradient of the right term of above expression. By some algebraic manipulation, we have

$$\begin{aligned} 0 &= \nabla\left\{h\left(\mathbf{B}, \mathbf{a}^{(m+1)}, \mathbf{b}^{(m+1)}\right) + \frac{\rho}{2}\left\|\mathbf{B} - \mathbf{A}^{(m)} - \frac{1}{\rho}\Theta^{(m)}\right\|_{\mathrm{F}}^2\right\} \\ &= -\sum_{j=1}^{n}\sum_{i=1}^{n}\left\{y_i - a_j\right\}\omega_{ij}\left(\mathbf{X}_i - \mathbf{X}_j\right)\left(b_j^{(m+1)}\right)^T \\ &\quad + \sum_{j=1}^{n}\sum_{i=1}^{n}\left\{\left(b_j^{(m+1)}\right)^T\mathbf{B}^T\left(\mathbf{X}_i - \mathbf{X}_j\right)\right\}\omega_{ij}\left(\mathbf{X}_i - \mathbf{X}_j\right)\left(b_j^{(m+1)}\right)^T - \rho\mathbf{A}^{(m)} - \Theta^{(m)} + \rho\mathbf{B}. \end{aligned} \tag{25}$$

Then, we obtain $\mathbf{B}^{(m+1)}$ by

$$\mathrm{vec}\left(\mathbf{B}^{(m+1)}\right) = H_{(m+1)}^{-1}\mathrm{vec}\left(G_{(m+1)}\right)H_{(m+1)}^{-1}G(m+1), \tag{26}$$

where $\mathrm{vec}(\mathbf{A})$ denotes vectorization for any matrix $\mathbf{A}$ and

$$\begin{aligned} G_{(m+1)} &= \sum_{j=1}^{n}\sum_{i=1}^{n}\left\{y_i - a_j\right\}\omega_{ij}\left(\mathbf{X}_i - \mathbf{X}_j\right)\left(b_j^{(m+1)}\right)^T + \rho\mathbf{A}^{(m)} + \Theta^{(m)}, \\ H_{(m+1)} &= \sum_{j=1}^{n}\sum_{i=1}^{n}\left\{b_j^{(m+1)}\left(b_j^{(m+1)}\right)^T\right\} \otimes \left\{\left(\mathbf{X}_i - \mathbf{X}_j\right)\left(\mathbf{X}_i - \mathbf{X}_j\right)^T\right\}\omega_{ij} + \rho\mathrm{I}_{pd}. \end{aligned} \tag{27}$$

The function,

$$\mathbf{A}^{(m+1)} = \underset{\mathbf{A}}{\operatorname{argmin}}\, \mathrm{I}_{\mathbf{S}}(\mathbf{A}) + \frac{\rho}{2}\left\|\mathbf{B}^{(m+1)} - \frac{1}{\rho}\Theta^{(m)} - \mathbf{A}\right\|_{\mathrm{F}}^2, \tag{28}$$

is equivalent to (12). The solution of the function above is given in the following lemma.

Lemma 1. *For any matrix $\mathbf{C} \in \mathbb{R}^{\mathrm{p}\times\mathrm{d}}$ with rank d, its projection, $\mathbf{P}_{\mathbf{S}}(\mathbf{C})$, onto $\mathbf{S}$ is defined as the solution of*

$$\underset{\mathbf{B}\in\mathbf{S}}{\operatorname{argmin}}\left\|\mathbf{C} - \mathbf{B}\right| \left|_{\mathrm{F}}^2. \right. \tag{29}$$

Then, we have $\mathbf{P}_{\mathbf{S}}(\mathbf{C}) = \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1/2}$.

In Appendix, we give the Proof of Lemma 1. We apply Lemma 1 to the update step of $\mathbf{A}$ and obtain

$$\begin{aligned} \mathbf{A}^{(m+1)} &= \mathbf{P}_{\mathbf{S}}\left(\mathbf{B}^{(m+1)} - \frac{1}{\rho}\Theta^{(m)}\right) \\ &= \left(\rho\mathbf{B}^{(m+1)} - \Theta^{(m)}\right)\left[\left(\rho\mathbf{B}^{(m+1)} - \Theta^{(m)}\right)^T\left(\rho\mathbf{B}^{(m+1)} - \Theta^{(m)}\right)\right]^{-1/2}. \end{aligned} \tag{30}$$

Finally, the update of Θ is given in (21). The stopping rule is $\left\|\mathbf{P}_{\mathbf{B}^{(m+1)}} - \mathbf{P}_{\mathbf{B}^{(m)}}\right\|_F \leq \varepsilon$, for some small value ε, where $\mathbf{P}_{\mathbf{C}} = \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T$ for any matrix $\mathbf{C}$. We summarize the above procedure in Algorithm 1.

We next derive the convergence property of the proposed ADMM algorithm.

Theorem 2. *For any sufficiently large ρ, there is one limit point at least, for the sequence $\{\mathbf{a}^{(m)}, \mathbf{b}^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)}, \Theta^{(m)}\}$ that is obtained by Algorithm 1. Moreover, each limit point is a stationary point of the associated augmented Lagrangian function $L_\rho(\mathbf{B}, \mathbf{a}, \mathbf{b}, \mathbf{A}, \Theta)$.*

The above result is based on the ADMM theory on nonconvex problems established by Wang et al. [24]. The proof of our theorem is relegated to Appendix.

3. Numerical Simulations

Numerical simulation studies are done to examine the numerical performances of the algorithm. Matlab is used for the numerical tests. The estimation accuracy is assessed by two different measures: the norm distance and the trace correlation $\mathrm{Tr}(\mathbf{P}_{\mathbf{B}}\mathbf{P}_{\widehat{\mathbf{B}}})/d$. In addition, the norm distance refers to the canonical distance between the projection matrix $\mathbf{P}_{\mathbf{B}}$ and $\mathbf{P}_{\widehat{\mathbf{B}}} = \widehat{\mathbf{B}}(\widehat{\mathbf{B}}^T\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^T$. That is, the norm distance is defined by $\left\|\mathbf{P}_{\mathbf{B}} - \mathbf{P}_{\widehat{\mathbf{B}}}\right\|_F$. The performance of the ADMM algorithm is studied in the first two examples, for single models $d = 1$ and multiple index models $d > 1$. The third example aims to explore how different penalty parameters ρ affect estimation accuracy.

Example 1. In this example, two single index models are considered.

(i) The vector $\mathbf{X} \in \mathbb{R}^2$ is independent uniformly distributed [0, 1], and we generate the data also by the model [35]

$$Y = 4\left\{\frac{(x_1 + x_2 - 1)}{\sqrt{2}}\right\}^2 + 4 + 0.5\varepsilon; \tag{31}$$

(ii) The vector $\mathbf{X} \in \mathbb{R}^3$ is independent uniformly distributed [0, 1] and we generate the data by the model

$$Y = \sin\left\{\frac{\pi(x_1 + x_2 + x_3)/\sqrt{3} - A}{(B - A)}\right\} + 0.2\varepsilon; \tag{32}$$

where $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$. This model was investigated by Carroll et al. [36].

The error ε has the standard normal distribution. The true parameters of models (18) and (19) are $\beta = (1, 1)^T/\sqrt{2}$ and $\beta = (1, 1, 1)^T/\sqrt{3}$.

Example 2. In this example, two multiple index models are studied.

(i) $X \sim N(0, I_7)$ and we generate the data, by using the model [1]

Require: Initial values $m = 0$, $\mathbf{B}^{(0)}$, $\mathbf{A}^{(0)}$, and $\Theta^{(0)}$;
While stopping criterion is not satisfied do
  Compute $\mathbf{a}^{(m+1)}, \mathbf{b}^{(m+1)}$ by (23);
  Compute $\mathbf{B}^{(m+1)}$ by (26);
  Compute $\mathbf{A}^{(m+1)}$ by (30);
  Compute $\Theta^{(m+1)}$ by (21);
  $m \longleftarrow m + 1$;
end while
return $\mathbf{a}^{(m+1)}, \mathbf{b}^{(m+1)}$, $\mathbf{A}^{(m+1)}$, $\mathbf{B}^{(m+1)}$, $\Theta^{(m+1)}$.

Algorithm 1: ADMM for the problem of minimizing (9).

$$Y = \frac{x_1}{0.5 + \{1.5 + x_2\}^2} + 0.5\varepsilon, \tag{33}$$

where the resulting parameters are $\beta_1 = (1, 0, \cdots, 0)^T$, $\beta_2 = (0, 1, \cdots, 0)^T$, and $\varepsilon \sim N(0, 1)$.

(ii) $X \sim N(0, I_6)$ and we generate the data, by using the model [7]

$$Y = \left(\beta_1^T\mathbf{X}\right)^2 + \beta_2^T\mathbf{X} + 0.1\varepsilon, \tag{34}$$

where $\beta_1 = (1, 1, 1, 0, 0, 0)^T/\sqrt{3}$, $\beta_2 = (0, 0, 0, 1, 1, 1)^T/\sqrt{3}$, and $\varepsilon \sim N(0, 1)$.

We consider the sample size $n = 100$, 200, and 400. In each case, the simulation is repeated 1000 times. The penalty parameter is $\rho = 1$. The initial values $\mathbf{A}^{(0)} = \mathbf{B}^{(0)} = \Theta^{(0)}$ are obtained by Xia et al.'s (2002) OPG method.

The simulation results are shown in the following Tables 1 and 2.

Tables 1 and 2 report the two different measures of estimation accuracy, their corresponding standard errors (in parentheses), and the running times. We run the desired numerical simulation on an Intel(R) Core(TM) i7-5600U CPU at 2.60 GHz with 8 GB memory. The results suggest that the ADMM algorithm can provide slightly more accurate estimates than the MAVE method, according to the two different measures: the norm distance and the trace correlation. We also see that the ADMM algorithm is slightly faster than the MAVE method of Xia et al. [7] in most cases, according to the running times.

Example 3. This example examines the performance of the ADMM algorithm for different penalty parameters.

The data are generated from model (19) and model (20), and ρ is, respectively, taken as 0.01, 0.5, 5, 10, 20, and 50.

For Example 3, 1000 replications are simulated with $n = 200$. We display the results in Table 3. For simplicity, one only presents the mean and standard deviations of the Frobenius norm distance and the running times. Obviously, the mean and standard deviations of the Frobenius norm distance and the running times have very little difference when ρ is taken as 0.01, 0.5, 5, 10, 20, and 50, respectively. These results reveal that the ADMM algorithm is almost insensitive to the choice of penalty parameter. Furthermore, we can find that the results based on the penalty parameters $\rho = 10$ and 20 are slightly faster than the other results.

Table 1: Simulation results for Example 1: the mean and standard deviations (in parentheses) of two different measures and the running times.

Model	Method	Sample size	Frobenius norm	Trace correlation	Runtime (second)
Model (18)	ADMM	100	0.1338 (0.0996)	0.9861 (0.0196)	62.3846
		200	0.0941 (0.0710)	0.9931 (0.0095)	152.3062
		400	0.0678 (0.0509)	0.9964 (0.0050)	213.2868
	Xia et al. [7]	100	0.1321 (0.0973)	0.9865 (0.0187)	64.2370
		200	0.0950 (0.0698)	0.9931 (0.0094)	161.7568
		400	0.0697 (0.0514)	0.9963 (0.0050)	216.0695
Model (19)	ADMM	100	0.0676 (0.0338)	0.9971 (0.0028)	76.1789
		200	0.0444 (0.0241)	0.9987 (0.0014)	184.7710
		400	0.0318 (0.0174)	0.9993 (0.0007)	242.7430
	Xia et al. [7]	100	0.0698 (0.0348)	0.9970 (0.0030)	76.3474
		200	0.0483 (0.0252)	0.9985 (0.0015)	185.15548
		400	0.0322 (0.0170)	0.9993 (0.0007)	245.4477

Table 2: Simulation results for Example 2: the mean and standard deviations of two different measures and the running times.

Model	Method	Sample size	Frobenius norm	Trace correlation	Runtime (second)
Model (20)	ADMM	100	0.5964 (0.2257)	0.8984 (0.0865)	292.5071
		200	0.3297 (0.0977)	0.9704 (0.0189)	584.7095
		400	0.2045 (0.0567)	0.9887 (0.0067)	743.9010
	Xia et al. [7]	100	0.6020 (0.2320)	0.8960 (0.0900)	253.4045
		200	0.3307 (0.1035)	0.9700 (0.0201)	660.8530
		400	0.2086 (0.0533)	0.9884 (0.0061)	760.7147
Model (21)	ADMM	100	0.0684 (0.0209)	0.9987 (0.0008)	161.0862
		200	0.0356 (0.0100)	0.9997 (0.0002)	361.0665
		400	0.0203 (0.0055)	0.9999 (0.0001)	510.9169
	Xia et al. [7]	100	0.0693 (0.0214)	0.9987 (0.0009)	165.3294
		200	0.0353 (0.0103)	0.9997 (0.0002)	481.7147
		400	0.0204 (0.0055)	0.9999 (0.0001)	557.5861

Table 3: Simulation results for Example 3: the ADMM algorithm for the different penalty parameters.

	Model	ρ					
		0.01	0.5	5	10	20	50
Model (19)	Frobenius norm's mean	0.0451	0.0462	0.0467	0.0466	0.0472	0.0459
	Frobenius norm's std	0.0240	0.0245	0.0236	0.0242	0.0238	0.0237
	Time (s)	183.5471	184.7868	184.2315	182.5929	182.2877	184.0295
Model (20)	Frobenius norm's mean	0.3303	0.3350	0.3348	0.3408	0.3548	0.3714
	Frobenius norm's std	0.1037	0.1028	0.1077	0.1199	0.1419	0.1835
	Time (s)	566.8275	626.9739	573.4154	541.7302	529.0828	551.4599

4. Concluding Remarks

Xia and his students have proposed an efficient direct algorithm to achieve MAVE and contributed it into an R package "MAVE." The detailed description can be seen in the website https://CRAN.R-project.org/package=MAVE.

In this study, we develop an ADMM algorithm to solve the MAVE-type methods and establish its convergence properties. The computational efficiency of the algorithm has been demonstrated by numerical experiments. It is noteworthy that the proposed ADMM algorithm is applicable to all the MAVE-type approaches. For example, for survival data, Xia et al. [37] proposed the hMAVE method for survival data, which can be regarded as a censored vision of Xia et al. [7]'s MAVE method. In this setting, our ADMM algorithm can also be used. It has been proved that the ADMM algorithm is quite flexible in handling many large-scale statistical problems, see Boyd et al. [23]. Therefore, it is desirable to deal with sufficient dimension reduction for high or ultrahigh-dimensional data by our ADMM algorithm.

Appendix

We first give the Proof of Lemma 1.

Proof of Lemma 1. We use the method of Lagrange multipliers to show this result. Consider the Lagrangian function

$$L(\mathbf{B}, \mathbf{V}) = \|\mathbf{C} - \mathbf{B}\|_{\mathrm{F}}^2 + \mathrm{Tr}\left(\mathbf{V}\left\{\mathbf{B}^T\mathbf{B} = I_d\right\}\right), \tag{A.1}$$

where $\mathbf{V} \in \mathbb{R}^{d\times d}$ are the Lagrange multipliers with $\mathbf{V} = \mathbf{V}^T$. Then, we have

$$0 = \frac{\partial L(\mathbf{B}, \mathbf{V})}{\partial \mathbf{B}} = \frac{\partial \mathbf{Tr}\left(\{\mathbf{C} - \mathbf{B}\}^T \{\mathbf{C} - \mathbf{B}\}\right)}{\partial \mathbf{B}} + \frac{\partial \mathrm{Tr}\left(\mathbf{V}^T \left\{\mathbf{B}^T \mathbf{B} = I_d\right\}\right)}{\partial \mathbf{B}} \tag{A.2}$$

$$= -2(\mathbf{C} - \mathbf{B}) + 2\mathbf{B}\mathbf{V}.$$

$$0 = \frac{\partial L(\mathbf{B}, \mathbf{V})}{\partial \mathbf{V}} = \mathbf{B}^T \mathbf{B} - I_d. \tag{A.3}$$

It follows from (A.3) that

$$\mathbf{C} = \mathbf{B}(\mathbf{V} + I_d). \tag{A.4}$$

Combining (A.4) with (A.3), we obtain

$$\mathbf{C}^T \mathbf{C} = (\mathbf{V} + I_d)\mathbf{B}^T \mathbf{B}(\mathbf{V} + I_d) = (\mathbf{V} + I_d)^2 I_d. \tag{A.5}$$

Combining (A.4) with (A.5), the solution of (16) is

$$\mathbf{B} = \mathbf{P}_{\mathbf{S}}(\mathbf{C}) = \mathbf{C}\left(\mathbf{C}^T \mathbf{C}\right)^{-1/2}. \tag{A.6}$$

Furthermore, it is easy to show that $\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1/2} \in \mathbf{S}$.

For readers' convenience, we add the Assumptions A1–A5 of Wang et al. [24].

The model in the first scenario is

$$\text{minimize}_{x=(x_0, x_1, \cdots x_p), y} \varphi(x, y) = f(x) + h(y), \quad \text{subject to } Ax + By = b. \tag{A.7}$$

(A1) Denote $\mathbf{F}: = \{(x, y) \in \mathbb{R}^{n+q}: Ax + By = 0\}$. Obviously, φ is coercive on $\mathbf{F}$. In addition, for any continuous φ, A1 holds trivially since $\mathbf{F}$ is bounded.

(A2) $\mathrm{Im}(A) \subseteq \mathrm{Im}(B)$. Here, $\mathrm{Im}(\cdot)$ denotes the image of a matrix.

(A3) For any fixed $x, \mathrm{argmin}_y\{\varphi(x, y): By = u\}$ is Lipschitz continuous on u. Moreover, it has a unique minimizer.

(A4) Set

$$f(X) = g(X) + \sum_{i=0}^{p} f_i(x_i). \tag{A.8}$$

Here, the function $g(X)$ is differentiable and Lipschitz. The function f_0 is lower semicontinuous. Moreover, $f_i(x_i)$ has restricted prox-regularity, which can be read in Definition 2 of Wang et al. [24].

(A5) The function $h(y)$ is also Lipschitz and differentiable. □

Proof of Theorem 2. Based on Theorem 2 of Wang et al. [24], Assumptions A1–A5 are verified to prove this theorem.

Since the feasible set $\mathbf{S}$ is bounded and $h(\mathbf{B}, \mathbf{a}, \mathbf{b})$ in (9) is a continuous function on $\mathbf{S}$, Assumption A1 holds.

Note that, in our settings, the coefficient matrices of the linear equality constraints in (9) are identity matrices. Thus, Assumptions A2 and A3 hold.

The assumption A4 holds because $\mathbf{I}_{\mathbf{S}}(\cdot)$ (see, $f_0(\cdot)$ in A4 of Wang et al. [24]) is lower semicontinuous.

The assumption A5 holds because $h(\mathbf{B}, \mathbf{a}, \mathbf{b})$ (see, $h(\cdot)$ in (7) of Wang et al. [24]) is continuous. □

Data Availability

No underlying data were collected or produced in this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] K.-C. Li, "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, vol. 86, pp. 316–342, 1991.

[2] R.-D. Cook, *Regression Graphics: Ideas for Studying Regressions through Graphics*, John Wiley & Sons, Hoboken, NJ, USA, 1998.

[3] R.-D. Cook and B. Li, "Dimension reduction for conditional mean in regression," *Annals of Statistics*, vol. 30, no. 2, pp. 455–474, 2002.

[4] K.-C. Li and N. Duan, "Regression analysis under link violation," *Annals of Statistics*, vol. 17, no. 3, pp. 1009–1052, 1989.

[5] R.-D. Cook and S. Weisberg, "Sliced inverse regression for dimension reduction: comment," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 328–332, 1991.

[6] K.-C. Li, "On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma," *Journal of the American Statistical Association*, vol. 87, no. 420, pp. 1025–1039, 1992.

[7] Y. Xia, H. Tong, W.-K. Li, and L.-X. Zhu, "An adaptive estimation of dimension reduction space," *Journal of the Royal Statistical Society-Series B: Statistical Methodology*, vol. 64, no. 3, pp. 363–410, 2002.

[8] B. Li and S.-L. Wang, "On directional regression for dimension reduction," *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 997–1008, 2007.

[9] Y. Xia, "A constructive approach to the estimation of dimension reduction directions," *Annals of Statistics*, vol. 35, no. 6, pp. 2654–2690, 2007.

[10] H. Wang and Y. Xia, "Sliced regression for dimension reduction," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 811–821, 2008.

[11] Y.-Y. Ma and L.-P. Zhu, "A semiparametric approach to dimension reduction," *Journal of the American Statistical Association*, vol. 107, no. 497, pp. 168–179, 2012.

[12] Y.-Y. Ma and L.-P. Zhu, "Efficient estimation in sufficient dimension reduction," *Annals of Statistics*, vol. 41, no. 1, pp. 250–268, 2013.

[13] Y.-Y. Ma and L.-P. Zhu, "On estimation efficiency of the central mean subspace," *Journal of the Royal Statistical Society-Series B: Statistical Methodology*, vol. 76, no. 5, pp. 885–901, 2014.

[14] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for ℓ_1-Minimization with applications to compressed sensing," *SIAM Journal on Imaging Sciences*, vol. 1, pp. 143–168, 2008.

[15] L. Xue, S. Ma, and H. Zou, "Positive-definite ℓ_1-penalized estimation of large covariance matrices," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1480–1491, 2012.

[16] T. Zhang and H. Zou, "Sparse precision matrix estimation via lasso penalized d-trace loss," *Biometrika*, vol. 101, no. 1, pp. 103–120, 2014.

[17] Y.-W. Gu, J. Fan, L.-C. Kong, S.-Q. Ma, and H. Zou, "ADMM for high-dimensional sparse penalized quantile regression," *Technometrics*, vol. 60, no. 3, pp. 319–331, 2018.

[18] D. Kapla, L. Fertl, and E. Bura, "Fusing sufficient dimension reduction with neural networks," *Computational Statistics & Data Analysis*, vol. 168, 2022.

[19] X.-P. Li and H.-C. So, "Robust low-rank tensor completion based on tensor ring rank via $\ell_{p,\epsilon}$," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3685–3698, 2021.

[20] X.-P. Li, Z.-L. Shi, C.-S. Leung, and H.-C. So, "Sparse index tracking with K-sparsity or ϵ-deviation constraint via ℓ_0-norm minimization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10930–10943, 2023.

[21] Q. Liu, X.-P. Li, H. Cao, and Y.-T. Wu, "From simulated to visual data: a robust low-rank tensor completion approach using ℓ_p-regression for outlier resistance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3462–3474, 2022.

[22] Z.-L. Shi, X.-P. Li, C.-S. Leung, and H.-C. So, "Cardinality constrained portfolio optimization via alternating direction method of multipliers," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 2, pp. 2901–2909, 2024.

[23] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, pp. 1–122, 2010.

[24] Y. Wang, W. Yin, and J.-S. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, 2019.

[25] J. Zhang, Q. Wang, and D. Mays, "Robust mave through nonconvex penalized regression," *Computational Statistics & Data Analysis*, vol. 160, pp. 107247–107247, 2021.

[26] X. Zhu, "A Riemannian conjugate gradient method for optimization on the Stiefel manifold," *Computational Optimization and Applications*, vol. 67, no. 1, pp. 73–110, 2017.

[27] L. Zanni, A. Benfenati, M. Bertero, and V. Ruggiero, "Numerical methods for parameter estimation in Poisson data inversion," *Journal of Mathematical Imaging and Vision*, vol. 52, no. 3, pp. 397–413, 2015.

[28] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.

[29] Y.-Y. Liu, F.-H. Shang, H.-Y. Liu, L. Kong, L.-C. Jiao, and Z.-C. Lin, "Accelerated variance reduction stochastic admm for large-scale machine learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4242–4255, 2021.

[30] P. Cascarano, A. Sebastiani, M.-C. Comes, G. Franchini, and F. Porta, "Combining weighted total variation and deep image prior for natural and medical image restoration via admm," in *Proceedings of the 2021 21st International Conference on Computational Science and Its Applications (ICCSA)*, pp. 39–46, Cagliari, Italy, September 2021.

[31] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.

[32] K. Guo, D. Han, D. Z.-W. Wang, and T.-T. Wu, "Convergence of ADMM for multi-block nonconvex separable optimization models," *Frontiers of Mathematics in China*, vol. 12, no. 5, pp. 1139–1162, 2017.

[33] B. Jiang, T.-Y. Lin, S.-Q. Ma, and S.-Z. Zhang, "Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis," *Computational Optimization and Applications*, vol. 72, no. 1, pp. 115–157, 2019.

[34] Z.-H. Jia, X. Gao, X.-J. Cai, and D. Han, "Local linear convergence of the alternating direction method of multipliers for nonconvex separable optimization problems," *Journal of Optimization Theory and Applications*, vol. 188, pp. 1–25, 2021.

[35] W. Hardle, P. Hall, and H. Ichimura, "Optimal smoothing in single-index models," *Annals of Statistics*, vol. 21, no. 1, pp. 157–178, 1993.

[36] R.-J. Carroll, J.-Q. Fan, I. Gijbels, and M.-P. Wand, "Generalized partially linear single-index models," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 477–489, 1997.

[37] Y. Xia, D. Zhang, and J. Xu, "Dimension reduction and semiparametric estimation of survival models," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 278–290, 2010.