HHS Public Access
Author manuscript
Math Program. Author manuscript; available in PMC 2020 July 01.

Published in final edited form as:
Math Program. 2019 July ; 176(1-2): 429–463. doi:10.1007/s10107-019-01386-z.

Robust Multicategory Support Matrix Machines

Chengde Qian,
School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, P. R. China.

Quoc Tran-Dinh,
Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill. quoctd@email.unc.edu

Sheng Fu,
Department of Industrial and Systems Engineering, National University of Singapore

Changliang Zou,
School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, P. R. China. nk.chlzou@gmail.com

Yufeng Liu
Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill. yfliu@email.unc.edu

Abstract

We consider the classification problem when the input features are represented as matrices rather than vectors. To preserve the intrinsic structures for classification, a successful method is the Support Matrix Machine (SMM) in [19], which optimizes an objective function with a hinge loss plus a so-called spectral elastic net penalty. However, the issues of extending SMM to multicategory classification still remain. Moreover, in practice, it is common to see the training data contaminated by outlying observations, which can affect the robustness of existing matrix classification methods. In this paper, we address these issues by introducing a robust angle-based classifier, which boils down binary and multicategory problems to a unified framework. Benefitting from the use of truncated hinge loss functions, the proposed classifier achieves certain robustness to outliers. The underlying optimization model becomes nonconvex, but admits a natural DC (difference of two convex functions) representation. We develop a new and efficient algorithm by incorporating the DC algorithm and primal-dual first-order methods together. The proposed DC algorithm adaptively chooses the accuracy of the subproblem at each iteration while guaranteeing the overall convergence of the algorithm. The use of primal-dual methods removes a natural complexity of the linear operator in the subproblems and enables us to use the proximal operator of the objective functions, and matrix-vector operations. This advantage allows us to solve large-scale problems efficiently. Theoretical and numerical results indicate that for problems with potential outliers, our method can be highly competitive among existing methods.

Keywords

Angle-based classifiers; DCA (difference of convex function) algorithm; Fisher consistency; Nonconvex optimization; Robustness; Spectral elastic net

1 Introduction

Many popular classification methods are originally developed for data with a vector of covariates, such as linear discriminant analysis, logistic regression, support vector machine (SVM), and Adaboost [12]. Recent advances in technology enable the generation of a wealth of data with complex structures, where the input features are represented by multi-linear geometric objects such as matrices or tensors, rather than by the form of vectors or scalars. The matrix-type datasets are often encountered in a wide range of real applications, e.g., the face recognition [31] and the analysis of medical images, such as the electroencephalogram data [36].

One common strategy to handle the matrix data classification is to stack a matrix into a long vector, and then employ some existing vector-based methods. This approach has several drawbacks. First, after vectorization, the dimensionality of the resulting vector typically becomes exceedingly high, which in turn leads to the curse of dimensionality, i.e. the large p and small n phenomenon. Second, vectorization of matrix-type data can destroy informative structure and correlation of data matrix, such as the neighbor information and the adjacent relation. Third, under the statistical learning framework, the regularization of vector and matrix data should be different due to their intrinsic structures. To exploit the correlation among the columns or rows of the data matrix, several methods were developed, for example, [6], [27], [24], [14]. These methods are essentially built on the low-rank assumption. Another major direction is to extend regularization techniques commonly used in vector-based classification methods to the present matrix-type data, under certain sparsity assumptions. The regularization with the nuclear norm of matrix of parameters is popular in a variety of settings; see [7] for matrix completion with a low rank constraint, and [36] for matrix regression problems based on generalized linear models. Specifically, [19] proposed the *Support Matrix Machine* (SMM) which employs a so-called spectral elastic net penalty for binary classification problems. The spectral elastic net penalty is the combination of the squared Frobenius matrix norm and the nuclear norm, in parallel to the elastic net [37]. They showed that the SMM classifier enjoys the property of grouping effect, while keeping a low-rank representation.

Our approach and contribution:

Though the SMM model is simple yet effective, two major issues still remain. The first one is how to extend it to address the problem of multicategory classification. One may reduce the multicategory problem via a sequence of binary problems, for example, using one-versus-rest or one-versus-one techniques. However, the one-versus-rest method can be inconsistent when there is no dominating class, and one-versus-one method may suffer a tie-in-vote problem [17, 18]. Another issue is that existing classifiers may not be robust against outliers, and thus they may have unstable performance in practice [30]. To address these two issues, we propose a new multicategory angle-based SMM using truncated hinge loss functions, which not only provides a natural generalization of binary SMM methods, but also achieves certain robustness to outliers. Our proposed classifier can be viewed as a robust matrix counterpart of the robust vector-based classifier in [32]. We show that the proposed classifier enjoys Fisher consistency and other attractive theoretical properties.

Because the truncated hinge loss is nonconvex and the spectral elastic net regularization is not smooth, the optimization problem involved in our classifier is highly non-trivial. We first show that this problem admits a global optimal solution by exploiting special structures of the model. Next, we show that the optimization problem has a natural DC (difference of two convex functions) decomposition. Hence, one can apply a DC algorithm (DCA) [2] to solve this problem. However, the convex subproblem is rather complicated with nonsmooth objective functions and linear operators, and cannot be solved exactly. This prevents us from solely applying DCA to solve our nonconvex problem. We instead develop a new variant, namely the *inexact proximal DCA*, to solve this problem. By using the proximal term, we obtain a strongly convex subproblem. Then, to approximately solve this subproblem, we propose to use primal-dual first-order methods proposed in [8, 28]. These methods allow us to exploit the special structures of the problem by utilizing the proximal operator of the objective terms, and matrix-vector multiplications. One drawback of this approach is to match the number of inner iterations in the primal-dual scheme and the inexactness of the proximal DCA scheme. By exploiting the problem structure, we show how to estimate this number of inner iterations at each step of the DCA scheme to obtain a unified DCA algorithm for solving the nonconvex optimization problem. We prove that by adaptively controlling the number of iterations in the primal-dual routine, we can still achieve a global convergence of our DCA variant, which converges to a stationary point. Our method can be implemented efficiently and does not require to estimate any parameter with expensive computational cost. To our limited knowledge, we are not aware of any efficient method to solve SMM-type problems in the literature except the alternating direction method of multipliers (ADMM)-based scheme [5]. In order to examine the efficiency of our method, we compare it with an ADMM-based scheme [5]. As shown in Section 5, our method outperforms ADMM in terms of computational time, and our new model has highly competitive performance among existing methods in different aspects.

Paper organization:

The rest of the article is organized as follows. In Section 2, we briefly review some related works, and then introduce our proposed model and methodology. In Section 3, we describe a new inexact proximal DCA algorithm and investigate its convergence. Some statistical learning results, including Fisher consistency, risk and robustness analysis, are presented in Section 4. Numerical studies are given in Section 5 on both synthetic and real data. Section 6 concludes our work with some remarks, and theoretical proofs are delineated in the appendix.

Notation:

For a matrix $\mathbf{A} \in \mathbb{R}^{p \times q}$ of rank $r\,(r \le \min(p, q))$, $r(r \le \min(p, q))$, $\mathbf{A} = \mathbf{U_A}\Sigma_\mathbf{A}\mathbf{V}_\mathbf{A}^\top$ represents the condensed singular value decomposition (SVD) of $\mathbf{A}$, where $\mathbf{U_A} \in \mathbb{R}^{p \times r}$ and $\mathbf{V_A} \in \mathbb{R}^{q \times r}$ satisfy $\mathbf{U}_\mathrm{A}^\top\mathbf{U_A} = \mathbf{I}_r$ and $\mathbf{V}_\mathbf{A}^\top\mathbf{V_A} = \mathbf{I}_r$, and $\Sigma_\mathbf{A}$ = diag $\{\sigma_1(\mathbf{A}), \cdots, \sigma_r(\mathbf{A})\}$ with $\sigma_1(\mathbf{A}) \ge \cdots \ge \sigma_r(\mathbf{A}) > 0$. For each $\tau > 0$, the singular value thresholding operator $\mathscr{D}_\tau(\cdot)$ is defined as follows:

$$\mathscr{D}_\tau(\mathbf{A}) = \mathbf{U_A}\mathscr{D}_\tau(\mathbf{\Sigma_A})\mathbf{V}_\mathbf{A}^\top,$$

where $\mathscr{D}_\tau(\mathbf{\Sigma_A}) = \; diag\left\{[\sigma_1(\mathbf{A}) - \tau]_+, \cdots, [\sigma_r(\mathbf{A}) - \tau]_+\right\}$ with $[a]_+ = \max\{0, a\}$. For $\mathbf{A} \in \mathbb{R}^{p \times q}$, $\left\|\mathbf{A}\right\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ denotes the Frobenius norm of $\mathbf{A}$, $\left\|\mathbf{A}\right\|_* = \sum_{i=1}^r \sigma_i(\mathbf{A})$ denotes the nuclear norm of $\mathbf{A}$, and $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$ stands for the spectral norm of $\mathbf{A}$. The inner product between two matrices is defined as $\langle \mathbf{A}, \mathbf{B}\rangle = \sqrt{tr\left(\mathbf{A}^\top\mathbf{B}\right)} = \sqrt{\sum_{i,j} a_{i,j} b_{i,j}}$. It is well-known that the nuclear norm $\|\mathbf{A}\|_*$, as a mapping from $\mathbb{R}^{p \times q}$ to $\mathbb{R}$, is not differentiable, but convex. Alternatively, one considers the subdifferential of $\|\mathbf{A}\|_*$, which is the set of subgradients and denoted by $\partial\|\mathrm{A}\|_*$. For a matrix $\mathbf{A}$, vec($\mathbf{A}$) denotes its vectorization. We use $\langle\cdot, \cdot\rangle$ to denote the inner product.

For a proper, closed and convex function $\varphi: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, dom($\boldsymbol{\varphi}$) denotes the domain of $\boldsymbol{\varphi}$, $prox_\varphi(\mathrm{x}) \triangleq \mathrm{argmin}_\mathrm{y}\left\{\varphi(\mathrm{y}) + \frac{1}{2}\right\}\left\|\mathrm{y} - \mathrm{x}\right\|^2$ denotes its proximal operator, and $\varphi^*(y) \triangleq \sup\left\{\mathrm{x}^\top\mathrm{y} - \varphi(\mathrm{x})\right\}$ denotes its Fenchel conjugate. We say that $\boldsymbol{\varphi}$ has a "friendly" proximal operator if its proximal operator $\mathrm{prox}_{\boldsymbol{\varphi}}$ can be computed efficiently by, e.g., closed-form or polynomial time algorithms. We say that $\boldsymbol{\varphi}$ is $\mu_{\boldsymbol{\varphi}}$-strongly convex if $\varphi(\cdot) - \frac{1}{2}\mu_\varphi\left\|\cdot\right\|_F^2$ is convex, where $\mu_{\boldsymbol{\varphi}} \ge 0$. Given a nonnegative real number x, we denote $\lfloor x \rfloor$ the largest integer that is less than or equal to x.

2 Methodology

Assume that the underlying joint distribution of $(\mathbf{X}, \mathscr{Y})$ is $Pr(\mathbf{X}, \mathscr{Y})$, where $\mathbf{X} \in \mathbb{R}^{p \times q}$ is the matrix of predictors and $\mathscr{Y}$ is the label. We are given a set of training samples of matrix-type data $\mathscr{T}_N = \{\mathbf{X}_i, y_i\}_{i=1}^N$ collected independently and identically distributed (i.i.d.) from Pr, where $\mathbf{X}_i \in \mathbb{R}^{p \times q}$ is the ith input sample and y_i is its corresponding class label. Here, we assume that $\mathbf{X}_i$'s are zero-centered; otherwise we can make transformation by $\mathbf{X}_i - \overline{\mathbf{X}}$, where $\overline{\mathbf{X}} = N^{-1}\sum_{i=1}^r \mathbf{X}_i$. We take the structure into consideration and handle all X_i's in the matrix form. Based on the given training set, the target of a classification problem is to estimate a classifier $\mathscr{T}_N$, by minimizing the empirical prediction error

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\hat{y}\left(\mathbf{X}_i\right) \neq y_i\right),$$

where $\mathbb{I}(\cdot)$ is the indicator function. Because $\mathbb{I}(\cdot)$ is discontinuous, in practice, we use some surrogate loss function to approximate it. As an example, in the case of the SVM, the hinge loss is adopted.

2.1 Review of the Support Matrix Machine

We take the binary problem as a special example with the encoded class labels set {+1, −1}. The optimization problem of [19]'s SMM can be expressed as

$$\min_{\mathbf{M}_1, b}\left\{\frac{1}{N}\sum_{i=1}^{N}\ell\left(y_i\left(\left\langle\mathbf{M}_1, \mathbf{X}_i\right\rangle + b\right)\right) + \lambda\left(\frac{1}{2}\left\|\mathbf{M}_1\right\|_F^2 + \tau\left\|\mathbf{M}_1\right\|_*\right)\right\}, \tag{1}$$

where $\mathbf{M}_1 \in \mathscr{R}^{p\times q}$, and $\ell(u) \triangleq [[1-u]_+ = \max\{1-u, 0\}$ is the hinge loss, $\boldsymbol{\tau} \geq 0$ controls the balance between the Frobenius norm and nuclear norm, and $\lambda > 0$ is a tuning parameter that balances the loss and regularization terms. The SMM (1) is a soft margin classifier, and it has a close connection to the ordinary SVM [4, 10]. With $\boldsymbol{\tau} = 0$, by vectorization of the coefficient matrix $\mathbf{M}_1$, SMM reduces to the standard form of the SVM.

The penalty term, $J\left(\mathrm{M}_1\right) \triangleq \frac{1}{2}\left\|\mathrm{M}_1\right\|_F^2 + \tau\left\|\mathrm{M}_1\right\|_*$, can be re-expressed as

$$J\left(\mathbf{M}_1\right) = \sum_{i=1}^{\min\{p, q\}}\frac{\sigma_i^2\left(\mathbf{M}_1\right)}{2} + \tau\sum_{i=1}^{\min\{p, q\}}\sigma_i\left(\mathbf{M}_1\right).$$

Clearly, this term is essentially of the form of the elastic net penalty for all singular values of the regression matrix $\mathbf{M}_1$, and thus is referred to as the spectral elastic net penalty. Such regularization encourages a low-rank constraint of the coefficient matrix. This can be better understood by the dual problem of (1), which is presented as follows:

$$\begin{cases}\min_{\alpha}\left\{\frac{1}{2}\left\|D_\tau\left(\sum_{i=1}^{N}\alpha_i y_i \mathbf{X}_i\right)\right\|_F^2 - \sum_{i=1}^{N}\alpha_i\right\} \\ \text{s.t. } 0 \leq \alpha_i \leq C, i = 1, \ldots, N; \sum_{i=1}^{N}\alpha_i y_i = 0,\end{cases} \tag{2}$$

where $C = (N\lambda)^{-1}$, and the optimum satisfies $\mathbf{M}_1 = \mathscr{D}_\tau\left(\sum_{i=1}^{N} a_i y_i \mathbf{X}_i\right)$. The derivation of (2) is given in the appendix. Under the low-rank assumption, small singular values of $\sum_{i=1}^{N} a_i y_i \mathbf{X}_i$ are more likely to be noisy, and hence SMM could be more efficient than the SVM by thresholding with an appropriate choice of $\boldsymbol{\tau}$. Moreover, due to the use of the trace norm,

[19] also showed that there is a stronger grouping effect in the estimation of $\mathbf{M}_1$ than the ordinary SVM.

2.2 Robust Multicategory SMM

For extensions of the binary classification method to the multicategory case, a common approach is to use K classification functions to stand for the K categories, and the prediction rule is based on which function has the largest value. Recently, [32] showed that this approach can be inefficient and suboptimal, and proposed an angle-based classification framework that needs to train K 1 classification functions $\mathbf{f} = (f_1, \cdots, f_{K-1})^\top$. The angle-based classifiers can enjoy better prediction performance and faster computation [33, 34, 26]. Hence, we adopt this strategy here. For simplicity, we focus on linear learning.

To be more specific, consider a centered simplex with K vertices $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_K)$ in $\mathbb{R}^{K-1}$, where these vertices are given by

$$\mathrm{w}_k = \begin{cases} (K-1)^{-\frac{1}{2}}\mathbf{1} & \text{if } k = 1, \\ -\dfrac{1+\sqrt{K}}{(K-1)^{\frac{3}{2}}}\mathbf{1} + \sqrt{\dfrac{K}{K-1}}\mathbf{e}_{k-1} & \text{if } k \ge 2. \end{cases}$$

Here, e_k is the unit vector of length $K-1$ with the kth entry 1 and 0 otherwise, and 1 is the vector of all ones. One can verify that each vector $\mathbf{w}_k$ has Euclidean norm 1, and the matrix $\mathbf{W}$ introduces a symmetric simplex in $\mathbb{R}^{K-1}$. Each $\mathbf{w}_k$ represents the kth class label. Let $\mathbf{M}$ be the linear transformation matrix which maps an input $\mathbf{X}$ into a $(K-1)$-variate vector $\mathbf{f}(\mathbf{X})$ $= \mathbf{M} \cdot \mathrm{vec}(\mathrm{X})$, where $\mathbf{M} = \left[\ \mathrm{vec}\left(\mathbf{M}_1\right), \cdots,\ \mathrm{vec}\left(\mathbf{M}_{K-1}\right)\right]^\top \in \mathbb{R}^{(K-1)\times pq}$ and $\mathbf{M}_j \in \mathbb{R}^{p\times q}$ for any $j \in \{1, \cdots, K-1\}$. The angle $\angle(\mathbf{f}(\mathbf{X}), \mathbf{w}_k)$ shows the confidence of the sample $\mathbf{X}$ belonging to class k. Thus the prediction rule is based on which angle is the smallest, i.e.,

$$\hat{y}(\mathbf{X}) = \underset{k \in \{1, \cdots, K\}}{\arg\min} \angle\left(\mathbf{f}(\mathbf{X}), \mathbf{w}_k\right).$$

It can also be verified that the least-angle prediction rule is equivalent to the largest inner product, i.e.,

$$\hat{y}(\mathbf{X}) = \underset{k \in \{1, \cdots, K\}}{\arg\max} \left\langle \mathbf{f}(\mathbf{X}), \mathbf{w}_k \right\rangle.$$

Here, we define $H_a(u) \triangleq [a-u]_+ = \max\{0, a-u\}$ and $G_a(u) \triangleq [a-u]_+ = \max\{0, a+u\}$. Based on the structure of matrix-type data, our proposed *Robust Multicategory Support·Matrix ·Machine* (RMSMM) solves

$$\min_{\mathbf{f} \in \mathscr{F}} \left[\frac{1}{N} \sum_{i=1}^{N} \left\{ \gamma T_{(K-1)s}\left(\left\langle \mathbf{f}(\mathbf{X}_i), \mathbf{w}_{y_i} \right\rangle\right) + (1-\gamma) \sum_{k \neq y_i} R_s\left(\left\langle \mathbf{f}(\mathbf{X}_i), \mathbf{w}_k \right\rangle\right) \right\} + \lambda J(\mathbf{M}) \right], \quad (3)$$

Where

- $\mathscr{F} \triangleq \left\{ \mathrm{f} | \mathrm{f}(\mathrm{X}) = \mathbf{M}\mathrm{vec}(\mathrm{X}), \mathbf{M} \in \mathbb{R}^{(K-1) \times pq} \right\}$;
- $\mathbf{f}(\mathbf{X}) \triangleq \left(f_1(\mathbf{X}), \cdots, f_{K-1}(\mathbf{X})\right)$ with $f_j(\mathbf{X}) = \langle \mathbf{M}_j, \mathbf{X} \rangle$ for $j = 1, \cdots, K-1$;
- $J(\mathbf{M}) \triangleq \sum_{j=1}^{k-1} \left(\frac{1}{2}\|\mathbf{M}_j\|_F^2 + \tau \|\mathbf{M}_j\|_*\right)$, where $\tau \geq 0$ is a balancing parameter;
- $T_s(u) \triangleq H_{K-1}(u) - H_s(u)$ and $R_s(u) \triangleq G_1(u) - G_s(u)$. The notation $s \leq 0$ a parameter that controls the location of truncation, and $\gamma \in [0, 1]$ is a convex combination parameter.

In (3), the loss term $\mathscr{L}(\mathrm{X}, y, \mathrm{M}) = \left\{\gamma T_{(K-1)s}\left(\left\langle \mathrm{f}(\mathrm{X}), \mathrm{w}_y \right\rangle\right) + (1-\gamma)\sum_{k \neq y} R_s\left(\left\langle \mathrm{f}(\mathrm{X}), \mathrm{w}_k \right\rangle\right)\right\}$ can be written as $\mathscr{L}_1(\mathbf{X}, \mathbf{y}, \mathbf{M}) - \mathscr{L}_2(\mathbf{X}, y, \mathbf{M})$, where

$$\mathscr{L}_1(\mathbf{X}, y, \mathbf{M}) = \gamma H_{(K-1)}\left(\left\langle \mathbf{f}(\mathbf{X}), \mathbf{w}_y \right\rangle\right) + (1-\gamma) \sum_{k \neq y} G_1\left(\left\langle \mathbf{f}(\mathbf{X}), \mathbf{w}_k \right\rangle\right), \text{ and}$$

$$\mathscr{L}_2(\mathbf{X}, y, \mathbf{M}) = \gamma H_{(K-1)s}\left(\left\langle \mathbf{f}(\mathbf{X}), \mathbf{w}_y \right\rangle\right) + (1-\gamma) \sum_{k \neq y} G_s\left(\left\langle \mathbf{f}(\mathbf{X}), \mathbf{w}_k \right\rangle\right).$$

The first term $\mathscr{L}_1$ of the above representation is a generalization of the reinforced multicategory loss function in the angle-based framework proposed by [33]. Note that $\mathscr{L}_1$ explicitly encourages $\langle \mathbf{f}(\mathbf{X}), \mathbf{w}_y \rangle$ to be small for $k \neq y_i$ In parallel to [33], we will show later that this convex combination of hinge loss functions enjoys Fisher consistency with $\gamma \in \left[0, \frac{1}{2}\right]$ and $s \leq 0$.

The use of the second term $\mathscr{L}_2$ is motivated by [30] to alleviate the effect of potential outliers, resulting in a truncated hinge loss. It can be seen that for any potential outlier $(\mathbf{X}, y)$ with a sizable $\langle \mathbf{f}(\mathbf{X}), \mathbf{w}_y \rangle$, its loss $\mathscr{L}$ is upper bounded by a constant for any $\mathbf{f}$. Thus, the impact of outliers can be alleviated by using $\mathscr{L}$. Note that when $s > 0$, $T_s(u)$ and $R_s(u)$ are constants within $[-s, s]$. In this case, the loss for some correctly classified observations is the same as that of those misclassified ones. Hence, it is more desirable to set $s \leq 0$. As recommended by [32], the choice of $s = -(K-1)^{-1}$ works well and will be used in our simulation study.

The truncated hinge loss is nonconvex, which makes the optimization problem (3) more involved than that of SMM. We next present an efficient algorithm to implement our RMSMM.

3 Optimization Algorithm

Since the optimization problem (3) admits a DC decomposition, we propose to apply DCA [2] to solve this problem. At each iteration of DCA, it requires to solve a convex subproblem, which does not have a closed form. We instead solve this convex subproblem up to a given accuracy and design an inexact variant of DCA so that it automatically adapts the accuracy of the subproblem to guarantee the overall convergence of the full algorithm.

3.1 A DC Representation of (3)

Problem (3) is nonconvex, but fortunately, it possesses a natural DC representation. Indeed, due to the relation $f(\mathbf{X}) \triangleq \mathbf{M} \cdot \operatorname{vec}(\mathbf{X})$, we can write

$$\left\langle f(\mathbf{X}), \mathbf{w} \right\rangle = \mathbf{w}^{\top}\mathbf{M} \cdot \operatorname{vec}(\mathbf{X}) = \mathbf{a}^{\top} \operatorname{vec}(\mathbf{M}),$$

where $\mathbf{a} \triangleq \operatorname{vec}(\mathbf{X}) \otimes \mathbf{w}$ with $\otimes$ denoting the Kronecker product. Let us define

$$\mathbf{a}_i \triangleq \operatorname{vec}\left(\mathbf{X}_i\right) \otimes \mathbf{w}_{y_i}, \quad \text{and} \quad \mathbf{b}_{ik} \triangleq \operatorname{vec}\left(\mathbf{X}_i\right) \otimes \mathbf{w}_k, \; i = 1, \cdots, N, k = 1, \cdots, K - 1. \tag{4}$$

Then, we can rewrite problem (3) as

$$\min_{\mathrm{M} \in \mathbb{R}^{(K-1) \times pq}} \left\{ F(\mathrm{M}) \triangleq \frac{1}{N} \sum_{i=1}^{N} \left[\gamma T_{s(K-1)}\left(\mathbf{a}_i^{\top} \operatorname{vec}(\mathrm{M})\right) + (1-\gamma) \sum_{k \neq y_i} R_s\left(\mathbf{b}_{ik}^{\top} \operatorname{vec}(\mathrm{M})\right) \right] + \lambda J(\mathrm{M}) \right\}. \tag{5}$$

Problem (5) has a DC representation as follows:

$$\min_{\mathbf{M}} \left\{ F(\mathbf{M}) \triangleq \Phi(\mathbf{M}) - \Psi(\mathbf{M}) \right\}, \tag{6}$$

where

$$\begin{cases} \Phi(\mathbf{M}) \triangleq \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \left[\gamma H_{K-1}\left(\mathbf{a}_i^{\top} \operatorname{vec}(\mathbf{M})\right) + (1-\gamma) \sum_{k \neq y_i} G_1\left(\mathbf{b}_{ik}^{\top} \operatorname{vec}(\mathbf{M})\right) \right] + \lambda J(\mathbf{M}) \\ \Psi(\mathbf{M}) \triangleq \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \left[\gamma H_{s(K-1)}\left(\mathbf{a}_i^{\top} \operatorname{vec}(\mathbf{M})\right) + (1-\gamma) \sum_{k \neq y_i} G_s\left(\mathbf{b}_{ik}^{\top} \operatorname{vec}(\mathbf{M})\right) \right]. \end{cases} \tag{7}$$

Here, both function Φ and Ψ are convex, but nonsmooth. In addition, Ψ is polyhedral. Note that we can always add any strongly convex function S to Φ and Ψ to write $F = \Phi - \Psi$ as

$$F(\mathbf{M}) = \Phi(\mathbf{M}) - \Psi(\mathbf{M}) = [\Phi(\mathbf{M}) + S(\mathbf{M})] - [\Psi(\mathbf{M}) + S(\mathbf{M})], \tag{8}$$

to obtain a new DC representation. The latter representation shows that both convex functions $\Phi + S$ and $\Psi + S$ are strongly convex. This representation also leads to a strongly convex subproblem at each iteration of DCA as we will see in the sequel. However, the choice of S is crucial, and also affects the performance of the algorithm. In our implementation, we simply add a convex quadratic function which leads to a proximal DCA.

Note that $\mathrm{dom}(\Phi) \cap \mathrm{dom}(\Psi) \neq \emptyset$. Since problem (6) is nonconvex, any point $\mathbf{M}^* \in \mathbb{R}^{(K-1) \times pq}$ satisfies

$$0 \in \partial F(\mathbf{M}^*) \equiv \partial\Phi(\mathbf{M}^*) - \partial\Psi(\mathbf{M}^*) \tag{9}$$

is called a stationary point of (6). If $\mathbf{M}^*$ satisfies $\partial\Phi(\mathbf{M}^*) \cap \partial\Phi(\mathbf{M}^*) \neq \emptyset$, then we say that $\mathbf{M}^*$ is a critical point (6). We show in the following theorem that (6) has a global optimal solution.

Theorem 1—*If* $\lambda > 0$*, then problem* (6) *has at least one global optimal solution* $\mathbf{M}^*$.

Proof We first write the objective function F of (5) into the sum $F(\mathrm{M}) = \bar{F}(\mathrm{M}) + \frac{\lambda}{2}\left\|\mathrm{M}\right\|_F^2$, where $\bar{F}$ is a function combining the sum of $Ts(K-1)$, R_s, and the nuclear norm $\sum_{i=1}^{K-1}\tau\left\|\mathrm{M}_j\right\|_{*i}$ in J.

Next, we show that $\bar{F}$ is Lipschitz continuous. Indeed, using the fact that $[a]+ = \max\{0, a\} = \frac{1}{2}(a + |a|)$ we can show that $T_s(u) = H_{K-1}(u) - H_s(u) = [K-1-u]_+ - [s-u]_+$ and $R_s(u) = G_1(u) - G_s(u) = [1+u]_+ - [s+u]_+$ are both Lipschitz continuous. In addition, we have $\left\|\mathbf{M}_j\right\|_F \leq \left\|\mathbf{M}_j\right\|_* \leq [\min\{p, q\}]^{1/2}\left\|\mathbf{M}_j\right\|_F$ for $j = 1, \cdots K-1$. Hence, $\sum_{i=1}^{K-1}\tau\left\|\mathrm{M}_j\right\|_*$ is also Lipschitz continuous. As a consequence, $\bar{F}$ defined above is Lipschitz continuous. That is, there exists $\bar{L} \in [0, +\infty)$ such that $\left|\bar{F}(\mathbf{M}) - \bar{F}(\widehat{\mathbf{M}})\right| \leq \bar{L}\left\|\mathbf{M} - \widehat{\mathbf{M}}\right\|_F$ for all $\mathrm{M}, \widehat{M} \in \mathbb{R}^{(K-1) \times pq}$.

Using a fixed point $\mathrm{M}^0 \in \mathbb{R}^{(K-1) \times pq}$, we can bound F as

$$F(\mathrm{M}) \geq F\left(\mathrm{M}^0\right) - L_{\bar{F}}\left\|\mathrm{M} - \mathrm{M}^0\right\|_F + \frac{\lambda}{2}\left\|\mathrm{M}\right\|_F^2 \to +\infty, \text{ as } \left\|\mathrm{M}\right\|_E \to +\infty.$$

Hence, F is coercive, i.e., $\boldsymbol{F}(\mathbf{M}) \to +\infty$ as $\|\mathbf{M}\|_F \to +\infty$. Consequently, its sublevel set $\mathscr{L}(\beta) = \{\mathrm{M} \mid F(\mathrm{M}) \leq \beta\}$ is closed and bounded for any $\beta \in \mathbb{R}$. By the well-known Weierstrass theorem, (6) has at least one global optimal solution $\mathbf{M}^*$. □

3.2 Inexact Proximal DCA Scheme

Let us start with the standard DCA scheme [2] and propose an inexact proximal DCA scheme to solve (6). The proximal DCA is equivalent to DCA applying to the DC decomposition (8) mentioned above, but often uses an adaptive strongly convex term S.

3.2.1 The Standard DCA Scheme and Its Proximal Variant—The DCA method for solving (6) is very simple. At each iteration $t \geq 0$, given $\mathbf{M}^t$, we compute a subgradient r$(\mathbf{M}^t) \in \partial\Psi(\mathbf{M}^t)$ and form the subproblem:

$$\min_{\mathrm{M}}\left\{\tilde{F}_t(\mathbf{M}) \triangleq \Phi(\mathbf{M}) - \left\langle \nabla\Psi\left(\mathbf{M}^t\right), \mathbf{M}\right\rangle\right\}, \tag{10}$$

to compute the next iteration $\mathbf{M}^{t+1}$ as an exact solution of (10). The subproblem (10) is convex. However, it is fully nonsmooth and does not have a closed form solution.

In the proximal DC variant, we instead apply DCA to the DC decomposition (8) with $S(\mathrm{M}) \triangleq \frac{\rho}{2}\left\|\mathrm{M}\right\|_F^2$, which leads to the following scheme:

$$\mathrm{M}^{t+1} \triangleq \arg\min_{\mathrm{M}}\left\{\tilde{F}_t(\mathrm{M}) \triangleq \Phi(\mathrm{M}) - \left\langle \nabla\Psi(\mathrm{M}^t, \mathrm{M}\right\rangle + \left|\frac{\rho_t}{2}\right\|\mathrm{M} - \mathrm{M}^t\right\|_F^2\right\}, \tag{11}$$

where $\rho_t > 0$ is a given proximal parameter. Clearly, $\mathbf{M}^{t+1}$ is well-defined and unique.

3.2.2 Inexact Proximal DCA Scheme—Clearly the subproblem (11) in the proximal DCA scheme (11) does not have a closed form solution. We can only obtain an approximate solution of this problem. This certainly affects the convergence of (11). We instead propose an inexact variant of (11) by approximately solving

$$\mathbf{M}^{t+1} :\approx \arg\min_{\mathbf{M}}\left\{\tilde{F}_t(\mathbf{M}) \triangleq \Phi(\mathbf{M}) - \left\langle \nabla\Psi\left(\mathbf{M}^t\right), \mathbf{M}\right\rangle + \frac{\rho_t}{2}\left\|\mathbf{M} - \mathbf{M}^t\right\|_F^2\right\}, \tag{12}$$

where :≈ stands for the approximation between the approximate solution $\mathbf{M}^{t+1}$ and the true solution $\overline{M}^{t+1}$ of the subproblem (12), and is characterized via the objective residual as

$$\tilde{F}_t\left(\mathbf{M}^{t+1}\right) - \tilde{F}_t\left(\overline{\mathbf{M}}^{t+1}\right) \leq \frac{\delta_t^2}{2}. \tag{13}$$

We note that this condition is implementable if we apply first-order methods in convex optimization to approximately solving (12).

Clearly, by strong convexity, we have

$$\frac{\rho_t}{2}\left\|\mathbf{M}^{t+1} - \overline{\mathbf{M}}^{t+1}\right\|_F^2 \leq \tilde{F}_t\left(\mathbf{M}^{t+1}\right) - \tilde{F}_t\left(\overline{\mathbf{M}}^{t+1}\right) \leq \frac{\delta_t^2}{2}.$$

This leads to $\left\|\mathbf{M}^{t+1} - \overline{\mathbf{M}}^{t+1}\right\|_F \leq \delta_t/\sqrt{\rho_t}$, which shows the difference between the approximate solution $\mathbf{M}^{t+1}$ and the true one $\overline{M}^{t+1}$.

Under the inexact criterion (13), we can still prove the following descent property of the inexact proximal DCA scheme (12).

Lemma 1: Let Ψ be μ_Ψ-strongly convex with $\mu_\Psi \geq 0$. Let $\{\mathbf{M}^t\}$ be the sequence generated by the inexact DCA scheme (12) under the inexact criterion (13). Then

$$F\left(\mathbf{M}^{t+1}\right) \leq F\left(\mathbf{M}^t\right) - \frac{(\rho_t + \mu_\Psi)}{2}\left\|\mathbf{M}^{t+1} - \mathbf{M}^t\right\|_F^2 + \frac{\delta_t^2}{2}. \tag{14}$$

Proof Using the optimality condition of (12), we have

$$\nabla\Phi\left(\overline{M}^{t+1}\right) - \nabla\Psi\left(\mathbf{M}^t\right) + \rho_t\left(\overline{M}^{t+1} - \mathrm{M}^t\right) = 0, \quad \text{where } \nabla\Phi\left(\overline{M}^{t+1}\right) \in \partial\Phi\left(\overline{M}^{t+1}\right).$$

From the μ_Φ- and μ_Ψ-strong convexity of Φ and Ψ, respectively, we have

$$\Phi\left(\overline{M}^{t+1}\right) \leq \Phi\left(\mathrm{M}^t\right) + \left\langle \nabla\Phi\left(\overline{M}^{t+1}\right), \overline{M}^{t+1} - \mathrm{M}^t\right\rangle - \frac{\mu_\infty}{2}\left\|\overline{M}^{t+1} - \mathrm{M}^t\right\|_F^2, \; -\Psi\left(\mathrm{M}^{t+1}\right) \leq -\Psi\left(\mathrm{M}^t\right) - \left\langle \nabla\Psi\left(\mathrm{M}^t\right), \mathrm{M}^{t+1} - \mathrm{M}^t\right\rangle - \frac{\mu_w}{2}\left\|\mathrm{M}^{t+1} - \mathrm{M}^t\right\|_F^2$$
$$= -\Psi\left(\mathbf{M}^t\right) - \left\langle \nabla\Psi\left(\mathbf{M}^t\right), \overline{\mathbf{M}}^{t+1} - \mathbf{M}^t\right\rangle + \left\langle \nabla\Psi\left(\mathbf{M}^t\right), \overline{\mathbf{M}}^{t+1} - \mathbf{M}^{t+1}\right\rangle - \frac{\mu_\Psi}{2}\left\|\mathbf{M}^{t+1} - \mathbf{M}^t\right\|_F^2.$$

Summing up the last two inequalities and using the above optimality condition, we obtain

$$\Phi\left(\overline{\mathbf{M}}^{t+1}\right) - \Psi\left(\mathbf{M}^{t+1}\right) \leq F\left(\mathbf{M}^t\right) - \rho_t\left\|\overline{\mathbf{M}}^{t+1} - \mathbf{M}^t\right\|_F^2 + \left\langle \nabla\Psi\left(\mathbf{M}^t\right), \overline{\mathbf{M}}^{t+1} - \mathbf{M}^{t+1}\right\rangle - \frac{\mu_\infty}{2}\left\|\overline{\mathbf{M}}^{t+1} - \mathbf{M}^t\right\|_F^2 - \frac{\mu_\mathbf{w}}{2}\left\|\mathbf{M}^{t+1} - \mathbf{M}^t\right\|_F^2.$$

Here, $\boldsymbol{F}(\mathbf{M}) = \boldsymbol{\Phi}(\mathbf{M}) - \boldsymbol{\Psi}(\mathbf{M})$. Next, using (13), we have

$$\Phi\left(\mathrm{M}^{t+1}\right) \leq \Phi\left(\overline{\mathrm{M}}^{t+1}\right) - \left\langle \nabla\Psi\left(\mathbf{M}^t\right), \overline{\mathrm{M}}^{t+1} - \mathrm{M}^{t+1}\right\rangle + \frac{\delta_t^2}{2} + \frac{\rho_t}{2}\left\|\overline{\mathbf{M}}^{t+1} - \mathbf{M}^t\right\|_F^2 - \frac{\rho_t}{2}\left\|\mathbf{M}^{t+1} - \mathbf{M}^t\right\|_F^2.$$

Summing up the last two inequalities and using $F = \Phi - \Psi$ again, we obtain

$$F\left(\mathbf{M}^{t+1}\right) \leq F\left(\mathbf{M}^t\right) - \frac{1}{2}\left[\left(\rho_t + \mu_\Phi\right)\left\|\overline{\mathbf{M}}^{t+1} - \mathbf{M}^t\right\|_F^2 + \left(\rho_t + \mu_\Psi\right)\left\|\mathbf{M}^{t+1} - \mathbf{M}^t\right\|_F^2\right] + \frac{\delta_t^2}{2}.$$

This implies (14) by neglecting the term $-\frac{1}{2}\left(\rho_t + \mu_\Phi\right)\left\|\overline{\mathbf{M}}^{t+1} - \mathbf{M}^t\right\|_F^2$

3.3 Solution of The Convex Subproblem

By rescaling the objective function by a factor of $\frac{1}{\lambda}$, we can rewrite the strongly convex subproblem (12) at the iteration t of the inexact proximal DCA scheme as follows:

$$\min_{\mathrm{M}}\left\{\widetilde{F}_t(\mathbf{M}) \triangleq P_t(\mathscr{A}(\mathbf{M})) + Q_t(\mathbf{M})\right\}, \tag{15}$$

where

$$P_t(\mathscr{A}(\mathbf{M})) \triangleq \frac{1}{\lambda N}\sum_{i=1}^{N}\left[\gamma H_{K-1}\left(\mathbf{a}_i^\top \text{ vec }(\mathbf{M})\right) + (1-\gamma)\sum_{k \neq y_i} G_1\left(\mathbf{b}_{ik}^\top \text{ vec }(\mathbf{M})\right)\right] - \frac{1}{\lambda}\left\langle \nabla \Psi\left(\mathbf{M}^t\right), \mathbf{M}\right\rangle,$$

and

$$Q_t(\mathbf{M}) \triangleq J(\mathbf{M}) + \frac{\rho_t}{2}\left\|\mathbf{M} - \mathbf{M}^t\right\|_F^2 = \sum_{j=1}^{K-1}\left[\frac{1}{2}\left\|\mathbf{M}_j\right\|_F^2 + \tau\left\|\mathbf{M}_j\right\|_* + \frac{\rho_t}{2}\left\|\mathbf{M}_j - \mathbf{M}_j^t\right\|_F^2\right].$$

Here, $\mathscr{A}$ is a linear operator concatenating all vectors a_i and b_{ik}, and the subgradient $\nabla \Psi(\mathbf{M}^t)$ in P_t, and P_t is a nonsmooth convex function, but has a "friendly" proximal operator that can be computed in linear time (see Subsection 3.5 for more details). Due to the strong convexity of J, (15) is strongly convex even for $\rho_t = 0$. However, one can adaptively choose $\rho_t \geq 0$ such that we have a "good" strong convexity parameter. If we do not add a regularization term $\frac{1}{2}\left\|\mathbf{MI}_j\right\|_F^2$, then (15) is strongly convex if $\rho_t > 0$. Since $\mu_\Psi = 0$ in (6), to strictly get a descent property in Lemma 1, we require $\rho_t > 0$. The following lemma will be used in the sequel, whose proof is given in the appendix.

Lemma 2—The objective function $P_t(\cdot)$ of (15) is Lipschitz continuous, i.e., there exists $L_0 \in (0, +\infty)$ such that $\left|P_t(\mathbf{u}) - P_t(\hat{\mathbf{u}})\right|) \leq L_0\left\|\mathbf{u} - \hat{\mathbf{u}}\right\|_F$ for all u, $\hat{\mathbf{u}}$, where L_0 is independent of t. Consequently, the domain $\text{dom}\left(P_t^*\right)$ of the conjugate P_t^* is bounded uniformly in t, i.e., its diameter $D_{P^*} \triangleq 2\sup\left\{\left\|\mathbf{v}\right\| \mid \mathbf{v} \in \text{ dom }\left(P_t^*\right)\right\}$ is $\cdot$ finite and independent of t.

Denote by

$$\mathscr{L}(\beta) \triangleq \left\{\mathrm{M} \in \mathbb{R}^{(K-1) \times pq} \middle| F(\mathrm{M}) \leq \beta\right\}, \tag{16}$$

the sublevel set of (5). As we proved in Theorem 1, the sublevel set $\mathscr{L}(\beta)$ is closed and bounded for any $\beta \in \mathbb{R}$. We define

$$D_{\mathscr{L}} \triangleq 2\sup\left\{\left\|\mathbf{M}\right\|_F \middle| F(\mathbf{M}) \leq F\left(\mathbf{M}^0\right)\right\} \tag{17}$$

the diameter of this sublevel set, which is finite, i.e., $D_{\mathscr{L}} \in (0, +\infty)$.

3.3.1 Primal-dual Schemes for Solving (15)—Problem (15) can be written into a minimax saddle-point problem using the Fenchel conjugate of P_t. It is natural to apply primal-dual first-order methods to solve this problem. We propose in this subsection two different primal-dual schemes to solve (15).

Our first algorithm is the common Chambolle-Pock primal-dual method proposed in [8]. This method is described as follows. Starting from $\widehat{\mathbf{M}}_0^t = \widetilde{\mathbf{M}}_0^t = \mathbf{M}^t$, and $\mathbf{Y}_0^t = \mathbf{Y}^t$ as an initial dual variable with $\mathbf{Y}^0 = \mathbf{0}$, set $\mathrm{M}_0^t = 0$, and at each inner iteration $l \geq 0$, we perform

$$\begin{cases} \mathbf{Y}_{l+1}^t = \ prox_{\sigma_l^t P_t^*}\left(\mathbf{Y}_l^t + \sigma_l^t \mathscr{A}\left(\widehat{\mathbf{M}}_l^t\right)\right), \\ \widetilde{\mathbf{M}}_{l+1}^t = \ prox_{\omega_l^t Q_t}\left(\widetilde{\mathbf{M}}_l^t - \omega_l^t \mathscr{A}^*\left(\mathbf{Y}_{l+1}^t\right)\right), \\ \theta_l^t = \dfrac{1}{\sqrt{1 + 2\left(1 + \rho_t\right)\omega_l^t}}, \ \omega_{l+1}^t = \theta_l^t \omega_l^t, \ \sigma_{l+1}^t = \dfrac{\sigma_l^t}{\theta_l^t}, \\ \widehat{\mathbf{M}}_{l+1}^t = \widetilde{\mathbf{M}}_t^{l+1} + \theta_l^t\left(\widetilde{\mathbf{M}}_{l+1}^t - \widetilde{\mathbf{M}}_l^t\right), \\ \mathbf{M}_{l+1}^t = \left(1 - s_l^t\right)\mathbf{M}_l^t + s_l^t\,\widetilde{\mathbf{M}}_{l+1}^t, \ \text{with } s_l^t = \dfrac{\sigma_l^t}{\sum_{j=1}^l \sigma_j^t}. \end{cases} \tag{18}$$

Here, we use the index t for the DCA scheme as the outer iteration counter, and the index l for the inner iteration counter. The initial stepsizes are set to be $\sigma_0^t = \omega_0^t = c\left\|\mathscr{A}\right\|^{-1}$, where $\left\|\mathscr{A}\right\|$ is the operator norm of $\mathscr{A}$, and $c = 0.999$; $\mathscr{A}^*$ is the adjoint operator of $\mathscr{A}$ (i.e., when $\mathscr{A}$ is a matrix, $\mathscr{A}^*$ is the transpose of $\mathscr{A}$), $prox_{\sigma P_t^*}$ is the proximal operator of the Fenchel conjugate P_t^* of P_t, and $\mathrm{prox}_{\omega Q_t}$ is the proximal operator of $\omega \cdot Q_t$. Alternatively, we can also apply [28, Algorithm 2] to solve (15). Originally, [28, Algorithm 2] works directly on the primal space, and has a convergence guarantee on the primal sequence $\left\{\mathbf{M}_l^t\right\}$ that is independent of the dual variable $\left\{\mathbf{M}_l^t\right\}$ as we can see in Lemma 3 below. Let us describe this scheme here to solve (15). Starting from $\mathbf{M}_0^t = \mathbf{M}^t$, $\widetilde{\mathbf{M}}_0^t = \mathbf{M}^t$ and $\mathbf{Y}_0^t = \mathbf{Y}^t$, at each inner iteration $l \geq 0$, we update

$$\begin{cases} \mathbf{Y}_{l+1}^t = \ prox_{\sigma_l^t P_t^*}\left(\mathbf{Y}_0^t + \sigma_l^t \mathscr{A}\left(\widehat{\mathbf{M}}_l^t\right)\right) \\ \widetilde{\mathbf{M}}_{l+1}^t = \ prox_{Q_t/\omega_l^t \beta_l^t}\left(\widetilde{\mathbf{M}}_l^t - \dfrac{1}{\omega_l^t \beta_l^t}\mathscr{A}^*\left(\mathbf{Y}_{l+1}^t\right)\right) \\ \mathbf{M}_{l+1}^t = \left(1 - \omega_l^t\right)\mathbf{M}_l^t + \omega_l^t \widetilde{\mathbf{M}}_{l+1}^t \\ \omega_{l+1}^t = \dfrac{\omega_i^t}{2}\left(\sqrt{\left(\omega_l^t\right)^2 + 4} - \omega_l^t\right), \ \sigma_{l+1}^t = \dfrac{\sigma_i^t}{1 - \omega_{l+1}^t}, \ \beta_{l+1}^t = \left\|\mathscr{A}\right\|^2 \sigma_{l+1}^t, \\ \widehat{\mathbf{M}}_{l+1}^t = \mathbf{M}_{l+1}^t + \dfrac{\omega_{l+1}^t\left(1 - \omega_l^t\right)}{\omega_l^t}\left(\mathbf{M}_{l+1}^t - \mathbf{M}_l^t\right). \end{cases} \tag{19}$$

Here, the initial values $\omega_0^t = 1$ and $\sigma_0^t = \frac{1}{2}\left\|\mathscr{A}\right\|^{-2}\left(1 + \rho_t\right)$ are given.

Note that both schemes (18) and (19) look quite similar at first glance, but they are fundamentally different. First, the dual step $\mathbf{Y}_l^t$ in (19) fixes $\mathbf{Y}_0^t$ for all iterations l, while it is recursive with $\mathbf{Y}_l^t$ in (18). Second, (18) has an extra averaging step at the last line, while (19) has a linear coupling step at the last line, where it works similarly as the accelerated gradient method of Nesterov [23]. Finally, the way of updating parameters in both schemes are really different.

In terms of complexity, (18) and (19) essentially have the same per-iteration complexity with one proximal operator $\mathrm{prox}_{sP_t^*}$, one proximal operator prox_{rQ_t}, one matrix-vector multiplication $\mathscr{A}(\mathrm{M})$, and one adjoint operation $\mathscr{A}^*(\mathbf{Y})$.

The following lemma provides us conditions to design a stopping criterion for the inner loop (i.e., the l-iterative loop), whose proof is given in the appendix.

Lemma 3: Let $\overline{M}^{t+1}$ be the unique solution of (15) at the outer iteration t. Then, the sequence $\left\{\mathbf{M}_l^t\right\}_{l \geq 0}$ generated by (18) satisfies

$$\tilde{F}_t\left(\mathbf{M}_l^t\right) - \tilde{F}_t\left(\overline{\mathbf{M}}^{t+1}\right) \leq \frac{\left(1 + \rho_t + \|\mathscr{A}\|\right)\|\mathscr{A}\|}{\left(1 + \rho_t\right)l^2}\left(\left\|\mathbf{M}_0^t - \overline{\mathbf{M}}^{t+1}\right\|_F^2 + \left\|\mathbf{Y}_0^t - \mathbf{Y}^{t+1}\right\|_F^2\right), \tag{20}$$

where $\overline{Y}^{t+1}$ is the corresponding exact dual solution of (15). Alternatively, the sequence $\left\{\mathbf{M}_l^t\right\}_{l \geq 0}$ generated by (19) satisfies

$$\begin{aligned} \tilde{F}_t\left(\mathbf{M}_l^t\right) - \tilde{F}_t\left(\overline{\mathbf{M}}^{t+1}\right) &\leq \frac{4L_0\|\mathscr{A}\|}{(l+1)^2}\left[\frac{2L_0\|\mathscr{A}\|}{1+\rho_t} + \sqrt{3}\left\|\mathbf{M}_0^t - \overline{\mathbf{M}}^{t+1}\right\|_F\right] \\ &+ \frac{3\left(\rho_t + 1\right)\left\|\mathbf{M}_0^t - \overline{\mathbf{M}}^{t+1}\right\|_F^2}{(l+1)^2}, \end{aligned} \tag{21}$$

where L_0 is given in Lemma 2

One advantage of (19) over (18) is that the right-hand side bound (21) does not depend on the dual variables $\mathbf{Y}_0^t$ and $\overline{Y}^{t+1}$ as in (20).

3.3.2 The Upper Bound of the Inner Iterations—Our next step is to specify the maximum number of inner iterations $l_{\max}(t)$ to guarantee the condition (13) at each outer iteration t.

First, from both schemes (18) and (19), one can see that $\left\{\mathbf{Y}_l^t\right\} \subset \mathrm{dom}\left(P_t^*\right)$. Hence, by Lemma 2, we can bound $\left\|\mathbf{Y}_0^t - \overline{\mathbf{Y}}^{t+1}\right\|_F \leq D_{P^*}$. On the other hand, by Theorem 1, the sublevel set

$\mathscr{L}\left(F\left(\mathbf{M}^0\right)\right)$ defined by (16) is bounded. We can also bound $\left\|\mathbf{M}_0^t - \overline{\mathbf{M}}^{t+1}\right\|_F \le D_{\mathscr{L}}$, where and $D_{\mathscr{L}}$ is given by (17). Using these upper bounds (20), we can show that

$$\tilde{F}_t\left(\mathbf{M}_l^t\right) - \tilde{F}_t\left(\overline{\mathbf{M}}^{t+1}\right) \le \frac{\left(1+\rho_t+\|\mathscr{A}\|\right)\|\mathscr{A}\|}{\left(1+\rho_t\right)l^2}\left(D_{\mathscr{L}}^2 + D_{P*}^2\right).$$

Let $\bar{K}_t \triangleq \left(1+\rho_t\right)^{-1}\left(1+\rho_t+\|\mathscr{A}\|\right)\|\mathscr{A}\|$ be a constant. In order to guarantee (13), we require to choose the number of iterations l at most

$$l_{\max}(t) \triangleq \left\lceil \frac{1}{\delta_t}\sqrt{\bar{K}_t\left(D_{\mathscr{L}}^2 + D_{P*}^2\right)}\right\rceil + 1 \;\text{ with }\; \delta_t = \frac{1}{(t+1)^{\alpha}}\sqrt{D_{\mathscr{L}}^2 + D_{P*}^2}. \tag{22}$$

Here, $\alpha > 1$ is a given constant specified by the user. With such a choice of δ_t, we have $l_{\max}(t) = \left\lfloor\sqrt{\bar{K}_t}(t+1)^{\alpha}\right\rfloor + 1$, which is independent of $D_{\mathscr{L}}$ and D_{P*}.

If we apply (19) to solve (15), then we have the bound (21). Let $\hat{K}_t \triangleq \frac{8L_0^2\|\mathscr{A}\|^2}{1+\rho_t} + 4\sqrt{3}L_0\left\|\mathscr{A}\right\|D_{\mathscr{L}} + 3\left(\rho_t+1\right)D_{\mathscr{L}}^2$. Since $\left\|\mathbf{M}_0^t - \overline{\mathbf{M}}^{t+1}\right\|_F \le D_{\mathscr{L}}$, in order to achieve $\tilde{F}_t\left(\mathbf{M}_l^t\right) - \tilde{F}_t\left(\overline{\mathbf{M}}^{t+1}\right) \le \delta_t^2/2$, we require $(l+1)^{-2}\hat{K}_t \le \delta_t^2/2$ which implies $l+1 \ge \sqrt{2\hat{K}_t/\delta_t}$ Hence, we can choose

$$l_{\max}(t) \triangleq \left\lceil \frac{\sqrt{2\hat{K}_t}}{\delta_t}\right\rceil + 1, \;\text{ with }\; \delta_t = \frac{C_0\sqrt{2\hat{K}_t}}{(t+1)^{\alpha}} \text{ and } C_0 \in (0,1), \tag{23}$$

to terminate the primal-dual scheme (19). With such a choice of δ_t, we can exactly evaluate $l_{\max}(t) = \left\lfloor C_0^{-1}(t+1)^{\alpha} + 1\right\rfloor$, which is also independent of $D_{\mathscr{L}}$.

Remark 1: By the choice of δ_t as in (22) or (23), the maximum number of inner iterations $l_{\max}(t)$ is independent of the two constants $D_{\mathscr{L}}$ and D_{P*}. These constants only show up when we prove the convergence of Algorithm 1 in Theorem 2, but they do not need to be evaluated in Algorithm 1 below. Hence, in the implementation of Algorithm 1, we simply use $l_{\max}(t) = \left\lfloor\sqrt{\bar{K}_t}(t+1)^{\alpha}\right\rfloor + 1$ for (18), or $l_{\max}(t) = \left\lfloor C_0^{-1}(t+1)^{\alpha} + 1\right\rfloor$ for (19) to specify the maximum number of inner iterations, where $\alpha > 1$ is a given number, e.g., $\alpha = 1.1$.

Algorithm 1

(Inexact Proximal DC Algorithm with primal-dual iterations)

1: **Initialization:**

2: Input an accuracy $\varepsilon > 0$. Choose an initial point $\mathrm{M}^0 \in \mathbb{R}^{(K-1)\times pq}$, and choose $\mathrm{Y}^0 \triangleq 0$.

3: Choose two parameters $0 < \underline{\rho} < \bar{\rho} < +\infty$, and $\sigma_0 = \sigma_0 = \omega_0 = 0.999\|\mathscr{A}\|$.

4: **For** $t = 0$ to T, **perform**

5: Evaluate a subgradient $\nabla\Psi\left(\mathbf{M}^t\right) \in \partial\Psi\left(\mathbf{M}^t\right)$ and choose $\rho_t \in [\underline{\rho}, \overline{\rho}]$.

6: *Initialization of inner loop:* Initialize $\mathbf{M}_0^t$, $\widehat{\mathbf{M}}_0^t$, $\widetilde{\mathbf{M}}_0^t$, $\mathbf{Y}_0^t$, σ_0^t and ω_0^t Compute $l_{\max}(t)$.

7: *Inner loop:* For $l = 0, 1, \cdots, l_{\max}(t)$, perform either (18) or (19).

8: *Terminate the inner loop:* If $l \geq l_{\max}(t)$, then set $\mathbf{M}^{t+1} = \mathbf{M}_{l_{\max}(t)}^t$ and $\mathbf{Y}^{t+1} = \mathbf{Y}_{l_{\max}(t)}^t$.

9: *Stopping criterion:* If $\left\|\mathrm{M}^{t+1} - \mathrm{M}^t\right\|_F \leq \varepsilon \ \max\left\{1, \left\|\mathrm{M}^t\right\|_F\right\}$, then terminate and return $\mathbf{M}^{t+1}$.

10: **End for**

3.4 The Overall Algorithm and Its Convergence Guarantee

We now combine the inexact proximal DCA scheme (12), and the primal-dual scheme (18) (or (19)) to complete the full algorithm for solving (5) as in Algorithm 1.

In the sequel, we will explicitly specify the evaluation of a subgradient $\nabla\Psi(\mathbf{M}^t)$ of Ψ, the choice of ρ_t, and the evaluation of $\mathrm{prox}_{sP_t^*}$ and prox_{rQ_t}. The number of maximum iterations T of the outer loop is not necessary to specify. However, we use T as a safeguard value to prevent the algorithm from an infinite loop. Practically, we can set T to be a relatively large value, e.g., $T = 10^3$. Nevertheless, the stopping criterion at Step 9 will terminate Algorithm 1 earlier. For large-scale problems, we can evaluate the operation norm $\|\mathscr{A}\|$> of $\mathscr{A}$ by a power method.

We state the overall convergence of Algorithm 1 in the following theorem.

Theorem 2 (Overall convergence)—Let $\{\mathbf{M}^t\}\cdot$ be the sequence generated by Algorithm 1 using (18) (respectively, (19)) for approximately solving (12) up to $l_{\max}(t)$ inner iterations as in (22) (respectively, (23)). Then, we have

$$\sum_{t=0}^{\infty}\left\|\mathbf{M}^{t+1} - \mathbf{M}^t\right\|_F^2 < +\infty \ \textit{and it implies} \ \lim_{t\to\infty}\left\|\mathbf{M}^{t+1} - \mathbf{M}^t\right\|_F = 0.$$

Moreover, the sequence $\cdot\{\mathbf{M}^t\}$ is bounded. Any cluster point $\mathbf{M}^*$ of $\{\mathbf{M}^t\}$ is a stationary point of (5). Consequently, the whole sequence $\{\mathbf{M}^t\}$ converges to a stationary point of (5).

Proof Since we apply (19) to solve the subproblem (12), with the choice of $_t$ as in (23), we can derive from Lemma 1 that

$$\sum_{t=0}^{T}\rho_t\left\|\mathbf{M}^{t+1} - \mathbf{M}^t\right\|_F^2 \leq 2\left(F\left(\mathbf{M}^0\right) - F\left(\mathbf{M}^{T+1}\right)\right) + \sum_{t=0}^{T}\delta_t.$$

By Theorem 1, we have $F(\mathrm{M}^{T+1}) \geq F(\mathrm{M}^*) > -\infty$, the global optimal value of (5). Hence, using the fact that $\rho_t \geq \underline{\rho} > 0$, we obtain

$$\underline{\rho} \sum_{t=0}^{\infty} \left\| \mathbf{M}^{t+1} - \mathbf{M}^t \right\|_F^2 \leq 2\left(F\left(\mathbf{M}^0\right) - F(\mathbf{M}^*)\right) + \sum_{t=0}^{\infty} \delta_t < +\infty .$$

Here, $\sum_{t=0}^{\infty} \delta_t < +\infty$ due to the choice of δ_t. This is exactly the first estimate in Theorem 2. The second limit in Theorem 2 is a direct consequence of the first one.

By Theorem 1 again, the sublevel set $\sum_{t=0}^{\infty} \delta_t < +\infty$ defined by (16) is bounded, and $F(\mathbf{M}^{t+1}) \leq F(\mathbf{M}^t)$ by Lemma 1, we have $\left\{\mathbf{M}^t\right\} \subset \mathscr{L}\left(F\left(\mathbf{M}^0\right)\right)$, which is bounded. For any cluster point $\mathbf{M}^*$ of $\{\mathbf{M}^t\}$, there exists a subsequence $\left\{\mathbf{M}^{t_s}\right\}$ that converges to $\mathbf{M}^*$. Now, we prove that $\mathbf{M}^*$ is a stationary point of (5). Using the optimality condition of (12), we have

$$0 \in \partial \Phi\left(\overline{M}^{t+1}\right) - \nabla \Psi\left(\mathbf{M}^t\right) + \rho_t\left(\overline{\mathbf{M}}^{t+1} - \mathbf{M}^t\right). \tag{24}$$

Note that lim $\lim_{t \to \infty} \left\| \overline{\mathbf{M}}^{t+1} - \mathbf{M}^{t+1} \right\|_F = 0$ due to the choice of δ_t. Here, we can pass this limit to a subsequence if necessary. Using this limit and the fact that $\lim_{t\to\infty} \|\mathbf{M}^{t+1} - \mathbf{M}^t\|_F = 0$, we can show $\lim_{t \to \infty} \left\| \overline{\mathbf{M}}^{t+1} - \mathbf{M}^t \right\|_F = 0$. In summary, we have $\lim_{t \to \infty} \overline{\mathbf{M}}^{t+1} = \lim_{t \to \infty} \mathbf{M}^t = \mathbf{M}*$. Using the definition of Φ and Ψ, we can see that the subgradient $\nabla \Psi(\mathbf{M}^t)$ of Ψ is uniformly bounded and independent of t. The subgradient $\nabla \Phi\left(\overline{\mathbf{M}}^{t+1}\right)$ can be represented as $\nabla \Phi\left(\overline{M}^{t+1}\right) = \bar{S}^{t+1} + \lambda \overline{M}^{t+1}$., where $\bar{S}^{t+1}$ is uniformly bounded and independent of t. By taking subsequence if necessary, both $\nabla \Phi\left(\overline{\mathbf{M}}^{t+1}\right)$ and $\nabla \Psi(\mathbf{M}^t)$ converge to $\nabla \Psi(\mathbf{M}^*)$ and $\nabla \Psi(\mathbf{M}^*)$, respectively. By [25, Theorem 24.4], we have $\nabla \Phi(\mathbf{M}^*) \in \partial \Phi(\mathbf{M}^*)$ and $\nabla \Psi(\mathbf{M}^*) \in \partial \Psi(\mathbf{M}^*)$. Using this fact, $\lim_{t \to \infty} \overline{\mathbf{M}}^{t+1} = \lim_{t \to \infty} \mathbf{M}^t = \mathbf{M}*$, and the boundedness of ρ_t, we can show that $0 \in \partial \Phi(\mathbf{M}^*) - \partial \Psi(\mathbf{M}^*)$ Hence, $\mathbf{M}^*$ a stationary point of (5). By the boundedness of $\{\mathbf{M}^1\}$ and $\lim_{t\to\infty} \|\mathbf{M}^{t+1} - \mathbf{M}^t\|_F = 0$, one can then use routine techniques to show that the whole sequence $\{\mathbf{M}^t\}$ converges to $\mathbf{M}^*$. □

While the convergence result given in Theorem 2 is rather standard and similar to those in [2], its analysis for the inexact proximal DCA seems to be new to the best of our knowledge. Note that the convex subproblem in DCA-type methods is often general and may not have closed-form solutions. It is natural to incorporate inexactness in an adaptive manner to guarantee the convergence of the overall algorithm.

3.5 Implementation Details and Comparison with ADMM

In Algorithm 1, we need to compute the proximal operator $\mathrm{prox}_{\sigma_l^t P_t^*}$ of the Fenchel conjugate P_t^* of P_t and $\mathrm{prox}_{\omega_l^t Q_t}$ of Q_t. In addition, in order to compare our method with other optimization methods, we specify the well-known ADMM to solve (12) as our comparison candidate.

3.5.1 Evaluation of Subgradient $\nabla\Psi(\mathbf{M}^t)$ *and The Choice of* ρ_t—Using the definition of Ψ from (7), we have

$$\nabla\Psi\left(\mathbf{M}^t\right) = \frac{1}{N}\sum_{i=1}^{N}\left[\gamma\nabla H_{s(K-1)}\left(\mathbf{a}_i^\top \operatorname{vec}\mathbf{M}^t\right)\mathbf{a}_i + (1-\gamma)\sum_{k \neq y_i}\nabla G_s\left(\mathbf{b}_{ik}^\top \operatorname{vec}\mathbf{M}^t\right)\mathbf{b}_{ik}\right],$$

where $\nabla H_{s(K-1)}(u) = \frac{1}{2}\cdot\ \mathrm{sign}\ (s(K-1)-u) - \frac{1}{2}$ and $\nabla G_s(v) = \frac{1}{2}\cdot\ \mathrm{sign}\ (s+v) + \frac{1}{2}$. Here, $\mathrm{sign}(\cdot)$ is the common sign function.

To choose ρ_t, we first choose a range $[\underline{\rho}, \bar{\rho}]$ in $(0, +\infty)$. For instance, we can choose $\underline{\rho} = 10^{-5}$ and $\bar{\rho} = 10^5$, and $\{\rho_t\}$ is any sequence in $[\underline{\rho}, \bar{\rho}]$. We can also fix ρ_t for all t as $\rho_t = \bar{\rho} > 0$, e.g., $\rho_t = 10^{-3}$. From our experience, we observe that if ρ_t is small, the strong convexity of (15) is $1+\rho_t$, which is also small. Hence, the number of inner iterations $l_{\max}(t)$ is large. However, the number of outer iterations t may be small. In the opposite case, if ρ_t is large, then we need a small number $l_{\max}(t)$. Nevertheless, due to a short step $\mathbf{M}^{t+1} - \mathbf{M}^t$, the number of outer iterations may increase. Therefore, trading-off the value of ρ_t is crucial and affects the performance of Algorithm 1.

3.5.2 Evaluation of Proximal Operators—To compute the proximal operator of P_t^* in (18), we can use Moreau's identity [3]:

$$prox_{\sigma P_t^*}(z) = \begin{cases} z_j - \sigma\, prox_{1/\sigma P_t}\left(z_j/\sigma\right) = z_j - \sigma\left[\mathcal{S}_{1/\sigma}\left(z_j + \mu_j\right) - \mu_j\right], & j = 1, \cdots, 2N, \\ z_j - \sigma\, prox_{1/\sigma P_t}\left(z_j/\sigma\right) = (1-\sigma)z_{2N+1} + 1, & \end{cases}$$

where $\mathcal{S}_r(v) = \ sign\ (u) \odot \max\{|v| - r, 0\}$ is the well-known soft-thresholding operator.

To compute the proximal operator of Q_t, we note that (here, $\tau_j = \tau$)

$$Q_t(\mathbf{M}) \triangleq \sum_{j=1}^{K-1}\left[\frac{1}{2}\left\|\mathbf{M}_j\right\|_F^2 + \tau_j\left\|\mathbf{M}_j\right\|_* + \frac{\rho_t}{2}\left\|\mathbf{M}_j - \mathbf{M}_j^t\right\|_F^2\right].$$

Hence, we have

$$prox_{\omega Q_t}(\mathbf{M}) = \left(prox_{\omega Q_{t_j}}\left(\mathbf{M}_j\right)\right)_{j=1}^{K-1},$$

Where $Q_{t_j}\left(\mathrm{M}_j\right) \triangleq \frac{1}{2}\left\|\mathrm{M}_j\right\|_F^2 + \tau_j\left\|\mathrm{M}_j\right\|_* + \frac{\rho_t}{2}\left\|\mathrm{M}_j - \mathrm{M}_j^t\right\|_F^2$, and

$$prox_{\omega Q_{t_j}}\left(\mathbf{M}_j\right) \triangleq \arg\min_{\widehat{\mathbf{M}}_j}\left\{\omega\tau_j\left\|\widehat{\mathbf{M}}_j\right\|_* + \frac{1+\omega\left(1+\rho_t\right)}{2}\left\|\widehat{\mathbf{M}}_j - \frac{\omega\rho_t\mathbf{M}_j^t + \mathbf{M}_j}{1+\omega\left(\rho_t+1\right)}\right\|_F^2\right\}.$$

This operator can be computed in a closed form using SVD of $\left(\omega\rho_t\mathbf{M}_j^t + \mathbf{M}_j\right)/\left[1+\omega\left(\rho_t+1\right)\right] = \mathbf{U}_j\mathbf{\Sigma}_j\mathbf{V}_j^\top$ as $prox_{\omega Q_{t_j}}\left(\mathbf{M}_j\right) = \mathbf{U}_j\mathscr{S}_r\left(\mathbf{\Sigma}_j\right)\mathbf{V}_j^\top$, where $\mathscr{S}_r$ is the soft-thresholding operator defined above with $r = \omega\tau_j/[1 + \omega(1 + \rho_t)]$.

3.5.3 ADMM Method for Solving (15)—

In Algorithm 1, we can apply ADMM to solve the subproblem (15) instead of primal-dual methods. We split the nuclear norm in Q_t of (15) by introducing an auxiliary variable $\mathbf{S}$ and rewrite (15) as

$$\begin{cases}\min\limits_{\mathbf{M},\mathbf{S}}\left\{\left[P_t(\mathscr{A}(\mathrm{M})) + \sum\limits_{j=1}^{K-1}\left[\frac{1}{2}\left\|\mathrm{M}_j\right\|_F^2 + \frac{\rho_t}{2}\left\|\mathrm{M}_j - \mathrm{M}_j^t\right\|_F^t\right]\right]_{B_t(\mathrm{M})} + \sum\limits_{j=1}^{K-1}\tau_j\left\|\mathrm{S}_j\right\|_*\right\} \\ \text{s.t } \mathrm{S}-\mathrm{M} = 0.\end{cases} \tag{25}$$

We define the corresponding augmented Lagrangian function of (25) as

$$\mathscr{L}_\beta(\mathbf{M},\mathbf{S},\mathbf{\Lambda}) \triangleq P_t(\mathscr{A}(\mathbf{M})) + \sum_{j=1}^{K-1}\left[\frac{1}{2}\left\|\mathbf{M}_j\right\|_F^2 + \frac{\rho_t}{2}\left\|\mathbf{M}_j - \mathbf{M}_j^t\right\|_F^2\right] + \sum_{j=1}^{K-1}\tau_j\left\|\mathbf{S}_j\right\|_* + trace\left(\Lambda^\top(\mathbf{S}-\mathbf{M})\right) + \frac{\beta}{2}\left\|\mathbf{S} - \mathbf{M}\right\|_F^2,$$

where $\beta > 0$ is a penalty parameter. Starting from an initial point $\mathbf{M}_0^t = \mathbf{M}^t$, $\mathbf{S}_0^t = \mathbf{M}^t$, our ADMM scheme for solving (25) updates at the inner iteration l according to the following steps:

$$\begin{cases}\mathrm{M}_{l+1}^t \triangleq \arg\min\limits_{\mathrm{M}}\left\{B_t(\mathrm{M}) + trace\left(\left(\Lambda^t\right)_l^\top\left(\mathbf{S}_l^t - \mathrm{M}\right)\right) + \frac{\beta}{2}\left\|\mathrm{S}_l^t - \mathrm{M}\right\|_F^2\right\} \\ \mathbf{S}_{l+1}^t \triangleq \arg\min\limits_{\mathbf{S}}\left\{\sum\limits_{j=1}^{K-1}\tau_j\left\|\mathbf{S}_j\right\|_* + trace\left(\left(\Lambda_l^t\right)^\top\left(\mathbf{S} - \mathbf{M}_{l+1}^t\right)\right) + \frac{\beta}{2}\left\|\mathbf{S} - \mathbf{M}_{l+1}^t\right\|_F^2\right\} \\ \Lambda_{l+1}^t \triangleq \Lambda_l^t + \beta\left(\mathbf{S}_{l+1}^t - \mathbf{M}_{l+1}^t\right).\end{cases} \tag{26}$$

In this scheme, the auxiliary sequence $\{\mathbf{S}_l^t\}$ can be computed into a closed form using SVD as we have done in Subsection 3.5.2. The sequence $\{\mathbf{M}_l^t\}$ requires to solve a general convex problem. However, this problem has a special structure so that its dual formulation becomes a boxed constrained convex quadratic program, which is very similar to (2). Hence, we solve this problem by coordinate descent methods, see, e.g., [29]. In summary, if we apply ADMM to solve (15), then our inexact proximal DCA has three loops: DCA outer iterations, ADMM inner iterations, and coordinate descent iterations for computing $\{\mathbf{M}_l^t\}$.

Remark 2 (*Convergence of the ADMM scheme* (26))

Note that (15) is strongly convex, and both subproblems in $\mathbf{M}_{l+1}^t$ and $\mathbf{S}_{l+1}^t$ of (26) are strongly convex, and therefore, uniquely solvable. Consequently, this scheme converges theoretically as proved e.g., in [5, Appendix A]. Together with asymptotic convergence guarantees, the convergence rates of ADMM, where (26) is a special case, have been studied in e.g., [11, 13, 21]. We omit the details here.

4 Statistical Properties

In this section, we explore some statistical properties of our proposed classifier RMSMM (3). In the first part, we establish the Fisher consistency result for the RMSMM, and study the finite sample bound on the misclassification rate. In the second part, we analyze the robustness property of RMSMM via the breakdown point theory.

4.1 Classification Consistency

Fisher's consistency is a fundamental property of classification methods. For an observed matrix-type data with fixed $\mathbf{X}$, and denote by $P_k(\mathbf{X}) = Pr(\mathcal{Y} = k \mid \mathbf{X})$ the class conditional probability of class $k \in \{1, 2, \cdots, K\}$. One can verify that the best prediction rule, namely, the Bayes rule, which minimizes the misclassification error rate, is $\hat{y}_{\text{Bayes}}(\mathbf{X}) = \arg\max_k P_k(\mathbf{X})$.

For a classifier, denote by $\phi(\mathbf{f}(\mathbf{X}), y)$ its surrogate loss function for classification using $\mathbf{f}$ as the classification function, and $\hat{y}_f$ the corresponding prediction rule. Assume the conditional loss is $L(\mathbf{X}) = E[\phi(\mathbf{f}(\mathbf{X}), y) \mid \mathbf{X}]$, where the expectation is taken with respect to the marginal distribution of $(\mathcal{Y} \mid \mathbf{X})$. We denote the theoretical minimizer of the conditional loss as $\mathbf{f}^*(\mathbf{X}) = \arg\min_{\mathbf{f}} L(\mathbf{X})$. When $\hat{y}_{\mathbf{f}^*}(\mathbf{X}) = \hat{y}_{\text{Bayes}}(\mathbf{X})$, we say the classifier is Fisher consistent. Let us denote by $\mathcal{L}(\mathbf{X}, y, \mathbf{M})$ the loss function in (3). Then, we have the following result.

Theorem 3—The classifier with the loss $\mathcal{L}(\mathbf{X}, y, \mathbf{M})$ is Fisher consistent when $\gamma \in \left[0, \frac{1}{2}\right]$ and $s \le 0$.

This result can be viewed as a generalization of Theorem 1 in [34] which is devised for vector-type observations. By this theorem, we know that our classifier RMSMM can achieve the best classification accuracy, given a sufficiently large matrix-type training dataset and a rich family $\mathscr{F}$. The following theorem provides an upper bound of the prediction error using the training dataset.

The proof of both Theorems 3 and 4 can be found in the appendix.

Theorem 4—Suppose that the conditional distribution of $\mathbf{X}$ given $\mathscr{Y} = k$ is the same as the distribution of $\mathbf{C}_k + \mathbf{E}$, where $\mathbf{C}_k \in \mathbb{R}^{p \times q}$ is a constant matrix and the entries of $\mathbf{E}$ are i.i.d. random variables with mean zero and finite fourth moment. Let $\widehat{\mathbf{M}} = \left[\mathrm{vec}\left(\widehat{\mathbf{M}}_1\right), \cdots, \mathrm{vec}\left(\widehat{\mathbf{M}}_{K-1}\right)\right]^\top \in \mathbb{R}(K-1) \times pq$ denote the solution of (5). Then, with probability at least $1 - \delta$, the misclassification rate of the classifier $\hat{y}$ corresponding to $\widehat{\mathbf{M}}$ can be bounded as

$$E[\mathbb{I}\{\mathscr{Y} \neq \hat{y}(\mathbf{X})\}] \leq \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left\{y_i \neq \hat{y}\left(\mathbf{X}_i\right)\right\} + \sqrt{\frac{\log\left(\delta^{-1}\right)}{N}} + \frac{cr(\sqrt{p} + \sqrt{q})}{\sqrt{N}}, \tag{27}$$

Where $r = \sum_{j=1}^{K-1} \left\|\widehat{\mathbf{M}}_j\right\|_*$, and c is a constant specified in the proof.

Theorem 4 measures the gap between the expectation error and the empirical error, which allows us to get a better understanding of the utility of the nuclear norm. For each category, the decision matrix contains $p \times q$ parameters, and therefore, if we only impose the Frobenius constraints [34] we would expect at best to obtain rates of the order $\sqrt{pq}$. By taking the low rank structure of the decision matrices into account, we use the nuclear norm penalty to control the singular values of the decision matrices. For the i-th singular vectors of the k-th decision matrix, there are $p + q + 1$ free parameters in total [22], one for the singular value σ_{ki} and the others for the orthogonal vectors with dimensions p and q. Its contribution to the gap will be $c\sigma_{ki}(\sqrt{p} + \sqrt{q})$. Hence, with the low-rank structure of the decision matrices, the nuclear-norm-penalized estimator achieves a substantially faster rate.

The rate in Theorem 4 can be further improved if we additionally impose some low-rank constraint on the noise term of $\mathbf{X}_i$. For example, consider $\mathbf{E} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{\Lambda} \in \mathbb{R}^{r_x \times r_x}$ is low-rank noise with all entries i.i.d. with mean zero and the finite fourth moment, $\mathbf{U}$ and $\mathbf{V}$ are orthogonal projection matrices independent of $\mathbf{\Lambda}$. It can be verified that the term $\sqrt{p} + \sqrt{q}$ in the rate above can be replaced by $2\sqrt{r_x}$. Finally, as a side remark, consider a special case with $q = 1$, i.e., the features are vectors rather than matrices. In such a situation, the nuclear norm reduces to the quadratic norm, and the last term of the upper bound in (27) will become $cr(\sqrt{p} + 1)/\sqrt{N}$, which is equivalent to existing results, for example, see [34].

4.2 Breakdown Point Analysis

Robustness theory has been developed to evaluate instability of statistical procedures since the 1960s [15]. The breakdown point theory focuses on the smallest fraction of contaminated data that can cause an estimator totally diverging from the original model. Here we consider the breakdown point analysis for multicategory classification models.

Let $\mathcal{T}_n$ be the original n observations, and $\widetilde{\mathcal{T}}_{n,m} = \mathcal{T}_{n-m} \cup \mathcal{V}_m$ be the contaminated sample with m observations of $\mathcal{T}_n$ contaminated, and $\widetilde{\mathbf{M}} = \widehat{\mathbf{M}}(\widetilde{\mathcal{T}}_{n,m})$ be the parameters estimated from the contaminated sample. We extend the sample angular breakdown point in [35] to the multicategory classification problem as

$$\epsilon^{\star}(\widehat{\mathbf{M}}, \mathcal{T}_n) = \min\left\{\frac{m}{n} \,\middle|\, \exists k,\ s.t.\ \mathbf{w}_k^{\top}\widehat{\mathbf{M}}\widetilde{\mathbf{M}}^{\top}\mathbf{w}_k \le 0\right\},$$

where $\widehat{\mathbf{M}} = \widehat{\mathbf{M}}(\mathcal{T}_n)$ is the estimated decision matrix from the original sample. Since the angle-based classifiers make the decision by comparing the angles between the $(K-1)$-dimensional classification function $\mathbf{f}$ and the K vertices of the simplex $\{\mathbf{w}_k\}_{k=1}^{K}$, it is reasonable to quantify the divergence between classifiers via the angles between the decision vectors $\mathbf{w}_k^{\top}\widetilde{\mathbf{M}}$ and the original counterpart, $\mathbf{w}_k^{\top}\widehat{M}$. When there exists one category k so that the angle between the two decision vectors is larger than $\pi/2$, the two classifiers would behave totally different at this category. Consequently, the classifier with contaminated samples would “break down”.

The following theorem compares the sample breakdown points of the proposed RMSMM and the multicategory SMM (MSMM) which generalizes [19]’s SMM using angle-based methods, say $\gamma = 1/2$ and $s = -\infty$ in Eq. (3).

Theorem 5—Assume that $\widehat{\mathbf{M}} \neq 0$. Then the breakdown point of MSMM is 1/n, while the breakdown point of RMSMM is not smaller than $\frac{\epsilon_1}{2(K-1)(1-s)}$, where

$$\epsilon_1 = \min_{\mathrm{M} \in \Delta^-} F(\mathrm{M}) - \min_{\mathrm{M} \in \Delta^+} F(\mathrm{M}) > 0.$$

By this theorem, only one contaminated observation will make the MSMM classifier break down. In other words, this estimator may not work well in the presence of few outliers. In contrast, the breakdown point of our proposed RMSMM, benefitting from the use of truncated hinge loss functions, has a fixed lower bound. Thus, the RMSMM has high outlier-resistance compared to its counterpart without truncation. The robustness property will be carefully examined via numerical comparisons in the next section.

5 Numerical Experiments

In this section, we investigate the performance of our proposed robust angle-based SMM using simulated and real datasets. Our configuration of the algorithm is as follows. For the primal-dual method described in Algorithm 1, we use $\mathbf{M}^0 = \mathbf{0}$ and $\rho_t = 0.01$ for every t. We set the stop criterion as $\|\mathbf{M}^{t+1} - \mathbf{M}^t\|_F \le 10^{-4} \max\{1, \|\mathbf{M}^t\|_F\}$. All the simulation results are obtained based on 100 replications.

5.1 Simulation Results

We generate simulated datasets by the following two scenarios. In the first scenario, the dimensions of input matrices are 50 × 50. For the kth category, to make the matrices low-rank, we randomly generate two 50 × 5 matrices, $\mathbf{U}_k$ and $\mathbf{V}_k$, which are standard orthonormal. More precisely, we first generate two 50 × 5 matrices with all the entries i.i.d. from the standard normal distribution and obtain $\mathbf{U}_k$ and $\mathbf{V}_k$ by the Gram-Schmidt process. The center of each class is then specified by $\mathbf{C}_k = \mathbf{U}_k\mathbf{V}_k^\top; k = 1, \cdots, K$. The observations in each class are generated by $\mathbf{C}_k + \mathbf{E}$; $k = 1,\cdots, K$, where $\mathbf{E}$ is a 50 × 50 normal random matrix with all entries i.i.d from $\mathcal{N}\left(0, \sigma^2\right)$. For the contaminated observations, we generate them by $3\mathbf{C}_1 + \mathbf{E}$ for $\mathcal{Y} \in \{1, \cdots, K\}$.

In the second scenario, the dimensions of input matrices are fixed as 80 × 100. We follow the settings in [36] to generate the true array signals by $\mathbf{C}_k = \mathbf{C}_{k,1}\mathbf{C}_{k,2}^\top; k = 1, \cdots, K$, where each entry of $\mathbf{C}_k$ is 0 or 1 and $\mathbf{C}_{k,i} \in \mathbb{R}^{p_i \times r}$, $p_1 = 80$ and $p_2 = 100$. To control the rank and the percentage of nonzero entries, we set $r = 10$ and generate $\mathbf{C}_{k,i}$ by setting each row to contain only one entry one and others zero, and the probabilities of entries being one are equal. All the entries of the noise matrix $\mathbf{E}$ are i.i.d. from $\sigma \cdot t(3)$, where $t(3)$ denotes the Student's t-distribution with three degrees of freedom. The outliers are generated by the same method as in the first scenario.

We use 10^3 observations for training, 10^4 observations for tuning and 10^4 observations for testing. The contamination ratio in the training sample ρ, is chosen as 0%, 10%, and 20%. For training the truncated model, we use the solutions of the ordinary SMM as an initial point. Following the suggestion by [33], we choose $\gamma = 1/2$ as it can provide stable classification performance. The truncation parameter, s, is fixed at $-1/(K-1)$. The other hyper-parameters, C and τ, are selected via a grid search on the tuning set.

We first consider the binary classification problem, say $K = 2$. We compare our RMSMM with the SMM in [19]. We also include a naive benchmark, the standard SVM method which is applied to the stacked-up vectors. Fig. 1 presents the classification error rates of RMSMM, SMM, and SVM on the simulated data with Scenario (I) and $K = 2$. Three noise magnitudes are considered: $\sigma = 0.5$, 0.7 and 0.9. Both two "support-matrix-based" methods, RMSMM and SMM, perform much better than the SVM. It has been observed that RMSMM generally outperforms SMM when there exits outliers, and its advantage becomes more pronounced for larger ρ. All methods are affected by different values of σ, but the comparison conclusion still holds for various σ.

Next we consider the multicategory case. Fig. 2 depicts the boxplots of the classification error rates for RMSMM and other competitors under Scenario (I) with $K = 3$ and 5. Three benchmarks are considered: the multicategory SMM using angle-based methods, MSMM; the angle-based multicategory SVM classifier [32] and its robust version RMSVM classifier [34]. In the case of $\rho = 0$, the RMSMM and its non-robust counterpart MSMM perform almost identically, which demonstrates that the truncation parameter, s, can adapt to the data structure and make the efficiency loss of RMSMM relative to MSMM minimal when there is no outlier. When $\rho = 0.1$ or $\rho = 0.2$, the advantage of RMSMM is clear: the means and standard variations of its classification error rates are generally smaller. From this figure, we can also observe that the use of the nuclear norm is prominent: the two SMM-based classifiers perform much better than the two SVM-based ones. Similar comparison conclusions can be drawn from Fig. 3, which reports the classification error rates of RMSMM and the other three methods under Scenario (II) with $\sigma = 3$, 4, and 5.

Finally, we present some comparison results of the ADMM and primal-dual algorithms for solving the RMSMM optimization problem (5). Fig. 4 reports the classification error rates and the corresponding computational time (in seconds) of the RMSMM using the two different primal-dual algorithms: (18) and (19) under Scenario (I) with $\sigma = 0.7$ and Scenario (II) with $\sigma = 4$ when $K = 3$. The bottom two panels record the total run time including the selection of tuning parameters. The tuning parameters λ and τ in the RMSMM are selected via a grid search. To be more specific, $\lambda \in [0.1, 10^4]$ and for each choice of λ, τ is tuned to make the decision matrix change from full-rank to rank one. One can see that the two algorithms perform very similarly in terms of classification rates, but the proposed primal-dual algorithm is significantly faster and the advantage is more remarkable as ρ increases. This is further confirmed by Fig. 5 which depicts the decay curves of the RMSMM objective function values versus the computational time until the two algorithms reach the desired accuracy. We consider the case under Scenario (II) with $K = 3$ and $\sigma = 4$ for a given combination of tuning parameters. In particular, we fix a combination of (λ, τ) and record the objective function values for each iteration. Clearly, the primal-dual algorithm is generally more stable and converges much faster than ADMM.

5.2 A Real-data Example

We apply the RMSMM model (5) to the Daily and Sports Activities Dataset [1] which can be found on the UCI Machine Learning Repository. The dataset comprises motion sensor data of 19 daily sport activities, each performed by 8 subjects (4 females, 4 males, between the ages of 20 and 30) in their own style for 5 minutes. The dataset was collected by several sensors. The input matrices are of dimension 125×45, where each column contains 125 samples of data acquired by a sensor over a period of 5 seconds at 25 Hz sampling frequency, and each row contains the data acquired from all of 45 sensor axes at a particular sampling instant.

To show the efficient performance of the proposed RMSMM model, we only select the first 10 categories of the dataset for simplicity. Thus the total number of instances is $N = 10 \times 8 \times 60 = 4,800$. It is a 10-category and balanced classification problem with 480 instances in each category. We equally and randomly divide the data into three parts for training, tuning, and testing, and the sample size of each part is 1, 600.

We choose $s = -K + 1$, and select the other parameters by a grid search. We report the classification accuracy of RMSMM, MSMM, RMSVM, and MSVM in Fig. 6-(left). The two matrix-based methods achieve lower classification rates than the other two vector-based classifiers, due to the benefit of the nuclear norm. This improvement can be more clear in Fig. 7, which presents the heatmap of the decision matrices of RMSMM and RMSVM; the former has a more sparse structure than the latter.

To demonstrate the effect of potential outliers on classification accuracy, we artificially contaminate the dataset with outliers by randomly relabeling 10% of the training set into another class. From Fig. 6-(right), we observe that the performances of all the methods are deteriorated by this manipulation, while the RMSMM performs the best. Both two robust classifiers, RMSMM and RMSVM, are less affected by the outliers, than the other two non-robust methods. All these numerical examples shown above suggest that the RMSMM is a practical and robust classier for a multicategory classification problem when the input features are represented as matrices.

6 Concluding Remarks

In this paper, we consider how to devise a robust multicategory classifier when the input features are represented as matrices. Our method is constructed in the angle-based classification framework, embedding a truncated hinge loss function into the support matrix machine. Although the corresponding optimization problem is nonconvex, it admits a natural DC (difference of two convex functions) representation. Hence, it is natural to apply DCA algorithms to solve this problem. Unfortunately, the convex subproblem in DCA is rather complex and does not have a closed form solution. Therefore, we develop an inexact proximal DCA variant to solve the underlying optimization problem. To approximately solve the convex subproblem, we propose to use primal-dual first-order methods. We combine both inexact proximal DCA and primal-dual methods to obtain a new proximal DCA scheme. We prove that our optimization model admits a global optimal solution, and the sequence generated by our DCA variant globally converges to a stationary point.

In terms of statistical learning perspective, we prove Fisher's consistency and prediction error bounds. Numerical results demonstrate that our new classifiers are quite efficient and much more robust than existing methods in the presence of outlying observations. We conclude the article with two remarks. First, our unified framework is demonstrated using the linear classifier. Though it is well recognized that linear learning is an effective solution in many real applications, it may be sub-efficient especially for problems with complex feature structures. Thus it is of interest to thoroughly study nonlinear learning under the proposed framework. Second, our numerical results show that the proposed procedure works well under large-dimensional scenarios. Theoretical investigation to the necessary condition on which the statistical theoretical guarantee of RMSMM holds is another interesting topic for future study.

Acknowledgments

The authors are grateful to the editor and the reviewers for their insightful comments that have significantly improved the article. Qian and Zou were supported in part by NNSF of China Grants 11690015 11622104 and 11431006, and NSF of Tianjin 18JCJQJC46000. Tran-Dinh was supported in part by US NSF-grant, DMS-1619884. Liu was supported in part by US NSF grants IIS1632951 and DMS-1821231, and NIH grant R01GM126550.

A Appendix:: Proofs of Technical Results

In this appendix, we provide all the remaining proofs of the results presented in the main text.

A.1 Proof of Lemma 2: Lipschitz continuity and boundedness

Since $[a]_+ = \max\{0, a\} = (a + |a|)/2$, the function P_t defined in (15) can be rewritten as $P_t(\mathscr{A}(\mathbf{z})) = \left\|\widehat{\mathscr{A}}\mathbf{z} + \mu\right\|_1 + \mathbf{d}_t^\top \mathbf{Z}$ for some matrix $\widehat{\mathscr{A}}$ and vectors μ and $\mathbf{d}_t$. Here, $\mathrm{d}_t \triangleq \mathrm{d} - \lambda^{-1}\mathrm{vec}\left(\nabla\Psi\left(\mathbf{M}^t\right)\right)$. However, ψ is also Lipschitz continuous due to its definition. This implies that $\nabla\psi(\mathbf{M}^t)$ is uniformly bounded, i.e., there exists a constant $C_0 \in (0, +\infty)$ such that $\|\nabla\psi(\mathbf{M}^t)\|_F \le C_0$ for all $\mathbf{M}^t \in \mathbb{R}^{(K-1)\times pq}$. As a consequence, P_t is Lipschitz continuous with the uniform constant L_0 that is independent of t, i.e., $\left|P_t(\mathbf{u}) - P_t(\hat{\mathbf{u}})\right| \le L_0\left\|\mathbf{u} - \hat{\mathbf{u}}\right\|_F$ for all $\mathbf{u}, \hat{\mathbf{u}}$. The boundedness of $dom\left(P_t^*\right)$ of the conjugate P_t^* follows from [3, Corollary 17.19].

A.2 The proof of Lemma 3: The convergence of the primal-dual methods

Let $\mathscr{G}(\mathbf{M}, \mathbf{Y}) = Q_t(\mathbf{M}) + \left\langle\mathscr{A}(\mathbf{M}), \mathbf{Y}\right\rangle - P_t^*(\mathbf{Y})$, where P_t^* is the Fenchel conjugate of P_t. Applying [9, Theorem 4] with $f = 0$, for any $\mathbf{M}$ and $\mathbf{Y}$, we have

$$\mathscr{G}\left(\mathbf{M}_l^t, \mathbf{Y}\right) - \mathscr{G}\left(\mathbf{M}, \overline{\mathbf{Y}}_l^t\right) \le \frac{1}{T_l}\left(\frac{\left\|\mathbf{M}_0^t - \mathbf{M}\right\|_F^2}{2\omega_0^t} + \frac{\left\|\mathbf{Y}_0^t - \mathbf{Y}\right\|_F^2}{2\sigma_0^t}\right), \qquad 28)$$

where $T_l = \sum_{i=1}^{l}\frac{\sigma_{i-1}^t}{\sigma_0^t}$, and $\overline{\mathbf{Y}}_l^t = \frac{1}{T_l}\sum_{j=1}^{l}\frac{\sigma_{j-1}^t}{\sigma_0^t}\mathbf{Y}_j^t$.

By the update rule in (18), we have $\omega_{l+1}^t\sigma_{l+1}^t = \omega_l^t\sigma_l^t$. Hence, by induction, we have $\omega_l^t\sigma_l^t = \omega_0^t\sigma_0^t = \left\|\mathscr{A}\right\|^{-2}$. On the other hand, by [8, Lemma 2], with the choice of $\lambda = \left\|\mathscr{A}\right\|^{-1}\left(1 + \rho_t\right)$, we have

$$\frac{\|\mathscr{A}\|}{1+\rho_t} + \frac{\|\mathscr{A}\|l}{\left\|\mathscr{A}\right\| + \left(1+\rho_t\right)} \le \frac{1}{\left(1+\rho_t\right)\omega_l^t} \le \frac{\|\mathscr{A}\|}{1+\rho_t} + l\,.$$

Using this estimate and $\sigma_l^t = \left\|\mathscr{A}\right\|^{-2}\omega_l^{-t}$, we have

$$T_l = \sum_{i=1}^{l}\frac{\sigma_{i-1}^t}{\sigma_0^t} = \frac{1}{\|\mathscr{A}\|}\sum_{i=1}^{l}\frac{1}{\omega_{i-1}^t} \ge \sum_{i=1}^{l}\left(\frac{i-1}{1+c} + 1\right) = \frac{l(l-1)}{2(1+c)} + l \ge \frac{l^2}{2(1+c)},$$

where $c = \left\|\mathscr{A}\right\|\left(1+\rho_t\right)^{-1}$. Hence, we can estimate T_l as $T_l \ge \frac{1}{2}\left(1 + \rho_t + \left\|\mathscr{A}\right\|\right)^{-1}\left(1+\rho_t\right)l^2$. Using this estimate of T_l, T_l, $\sigma_0^t = \omega_0^t = \left\|\mathscr{A}\right\|$, and $\tilde{F}_t\left(\mathbf{M}_l^t\right) - \tilde{F}_t\left(\overline{\mathbf{M}}^{t+1}\right) \le \mathscr{G}\left(\mathbf{M}_l^t, \overline{\mathbf{Y}}^{t+1}\right) - \mathscr{G}\left(\overline{\mathbf{M}}^{t+1}, \overline{\mathbf{Y}}_l^t\right)$, we obtain from (28) that

$$\tilde{F}_t\left(\mathbf{M}_l^t\right) - \tilde{F}_t\left(\overline{\mathbf{M}}^{t+1}\right) \le \frac{\left(1+\rho_t+\|\mathscr{A}\|\right)\|\mathscr{A}\|}{\left(1+\rho_t\right)l^2}\left(\left\|\mathbf{M}_0^t-\overline{\mathbf{M}}^{t+1}\right\|_F^2+\left\|\mathbf{Y}_0^t-\overline{\mathbf{Y}}^{t+1}\right\|_F^2\right).$$

This is exactly (20).

Next, we prove (21). By introducing $\mathbf{Y} = \mathscr{A}(\mathbf{M})$, we can reformulate the strongly convex subproblem (15) into the following constrained convex problem:

$$\tilde{F}_t\left(\overline{M}^{t+1}\right) = \min_{\mathrm{M},\mathrm{Y}}\left\{\tilde{F}_t(\mathrm{M},\mathrm{Y}) = P_t(\mathrm{Y}) + Q_t(\mathrm{M}) \,\middle|\, \mathscr{A}(\mathrm{M}) - \mathrm{Y} = 0\right\}. \tag{29}$$

Note that Q_t is strongly convex with the strong convexity parameter $1 + \rho_t$. We can apply [28, Algorithm 2] to solve (29). If we define

$$\Delta_{\sigma_i^t}\left(\mathrm{M}_{l+1}^t\right) = P_t\left(\mathrm{Y}_{l+1}^t\right) + Q_t\left(\mathrm{M}_{l+1}^t\right) + \frac{\sigma_l^t}{2}\left\|\mathscr{A}\left(\mathrm{M}_{l+1}^t\right) - \mathrm{Y}_{l+1}^t\right\|_F^2 - \tilde{F}_t\left(\overline{M}^{t+1}\right).$$

then, from the proof of [28, Theorem 2], we can show that

$$\Delta_{\sigma_l^t}\left(\mathbf{M}_{l+1}^t\right) \le \frac{2\left[\sigma_0^t\|\mathscr{A}\|^2 + 1 + \rho_t\right]\left\|\mathbf{M}_0^t - \overline{\mathbf{M}}^{t+1}\right\|_F^2}{(l+2)^2}. \tag{30}$$

By Lemma 2, P_t is Lipschitz continuous with the Lipschitz constant L_0. Then we have

$$\begin{aligned}
&\tilde{F}_t\left(\mathbf{M}_{l+1}^t\right) - \tilde{F}_t\left(\overline{\mathbf{M}}^{t+1}\right)\\
&= P_t\left(\mathscr{A}\left(\mathbf{M}_{l+1}^t\right)\right) + Q_t\left(\mathbf{M}_{l+1}^t\right) - \tilde{F}_t\left(\overline{\mathbf{M}}^{t+1}\right) \le P_t\left(\mathbf{Y}_{l+1}^t\right) + Q_t\left(\mathbf{M}_{l+1}^t\right) - \tilde{F}_t\left(\overline{\mathbf{M}}^{t+1}\right) + L_0\left\|\mathscr{A}\left(\mathbf{M}_{l+1}^t\right)\right.\\
&\left.- \mathbf{Y}_{l+1}^t\right\|_F.
\end{aligned}$$

Combining (30) and this estimate, we obtain

$$\begin{aligned}
0 \le\ & \tilde{F}_t\left(\mathbf{M}_{l+1}^t\right) - \tilde{F}_t\left(\overline{\mathbf{M}}^{t+1}\right)\\
\le\ & \frac{2\left[\sigma_0^t\|\mathscr{A}\|^2 + 1 + \rho_t\right]\left\|\mathbf{M}_0^t - \overline{\mathbf{M}}^{t+1}\right\|_F^2}{(l+2)^2} + L_0\left\|\mathscr{A}\left(\mathbf{M}_{l+1}^t\right) - \mathbf{Y}_{l+1}^t\right\|_F - \frac{\sigma_l^t}{2}\left\|\mathscr{A}\left(\mathbf{M}_{l+1}^t\right) - \mathbf{Y}_{l+1}^t\right\|_F^2.
\end{aligned}$$

Similar to the proof of [28, Corollary 1], by using $\sigma_0^t = \frac{1+\rho_t}{2\|\mathscr{A}\|^2}$, the last inequality leads to

$$\left\|\mathscr{A}\left(\mathbf{M}_{l+1}^{t}\right)-\mathbf{Y}_{l+1}^{t}\right\|_{F} \leq \frac{4\|\mathscr{A}\|}{(l+1)^{2}}\left[\frac{2L_{0}\|\mathscr{A}\|}{1+\rho_{t}}+\sqrt{3}\left\|\mathbf{M}_{0}^{t}-\overline{\mathbf{M}}^{t+1}\right\|_{F}\right].$$

Combining the two last estimates, we obtain

$$\tilde{F}\left(\mathbf{M}_{l}^{t}\right)-\tilde{F}\left(\overline{\mathbf{M}}^{t+1}\right) \leq \frac{4\|\mathscr{A}\|L_{0}}{(l+1)^{2}}\left[\frac{2L_{0}\|\mathscr{A}\|}{1+\rho_{t}}+\sqrt{3}\left\|\mathbf{M}_{0}^{t}-\overline{\mathbf{M}}^{t+1}\right\|_{F}\right]+\frac{3\left(\rho_{t}+1\right)\left\|\mathbf{M}_{0}^{t}-\overline{\mathbf{M}}^{t+1}\right\|_{F}^{2}}{(l+1)^{2}},$$

which is exactly (21). □

A.3 Proof of statistical properties

We provide the proof of Theorems 3 and 4 in this section.

A.3.1 Proof of Theorem 3: Fisher's consistency

In our RMSMM (3), one can abstract the truncated hinge loss function as

$$\phi(\mathbf{f}(\mathbf{X}), y)=\gamma T_{(K-1)s}\left(\left\langle\mathbf{f}(\mathbf{X}), \mathbf{w}_{y}\right\rangle\right)+(1-\gamma) \sum_{k \neq y} R_{s}\left(\left\langle\mathbf{f}(\mathbf{X}), \mathbf{w}_{k}\right\rangle\right).$$

Then, the conditional loss can be rewritten as

$$L(\mathbf{X}) \triangleq \sum_{k=1}^{K}\left[\gamma P_{k} T_{(K-1)s}\left(\left\langle\mathbf{f}(\mathbf{X}), \mathbf{w}_{k}\right\rangle\right)+\left(1-P_{k}\right) R_{s}\left(\left\langle\mathbf{f}(\mathbf{X}), \mathbf{w}_{k}\right\rangle\right)\right].$$

[34, Theorem 1] showed that for a vector data $\mathbf{x}$, the robust classifier based on the loss function $\phi(\mathbf{f}(\mathbf{x}), y)$ is Fisher consistent with $\gamma \in \left[0, \frac{1}{2}\right]$ and $s \leq 0$. By vectorizing the matrix data $\mathbf{X}$ to a new vector $\mathbf{x} = \mathrm{vec}(\mathbf{X})$, then all settings here are the same as those of Theorem 1 in [34]. In this case, Fisher consistency results can naturally be transferred to matrix-type data. □

A.3.2 Proof of Theorem 4: Misclassification rates

First, we introduce the Rademacher complexity. Let $\mathscr{G} = \{g: \mathbf{X} \times \mathscr{Y} \to \mathbb{R}\}$ be a class of loss functions. Given the sample $\mathscr{T} = \left\{\left(\mathbf{X}_i, y_i\right)\right\}_{i=1}^{N}$, we define the empirical Rademacher complexity of $\mathscr{G}$ as

$$\hat{R}_{N}(\mathscr{G})=E_{\boldsymbol{\sigma}}\left\{\sup_{g \in \mathscr{G}} \frac{1}{N} \sum_{i=1}^{N} \sigma_{i} g\left(\mathbf{X}_{i}, y_{i}\right)\right\},$$

where $\sigma = \left\{\sigma_i\right\}_{i=1}^{N}$ are i.i.d. random variables with Pr(σ_1 = 1) = Pr(σ_1 = −1) = 1/2. The Rademacher complexity of $\mathscr{G}$ is defined as

$$R_N(\mathscr{G}) = E_{\boldsymbol{\sigma}, \mathscr{T}}\left\{\sup_{g \in \mathscr{G}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i g\left(\mathbf{X}_i, y_i\right)\right\}.$$

For our model, let

$$H = \left\{h(\mathrm{X}, y) = \min_{k \neq y}\left(\left(\mathrm{f}(\mathrm{X}), \mathrm{w}_y - \mathrm{w}_k\right)\right) \middle| f \in \mathscr{F}, \sum_j \left\|\mathrm{M}_j\right\|_* \leq r\right\},$$

and

$$\mathbb{I}_\kappa(x) = \begin{cases} 1 & x < 0, \\ 1 - \frac{1}{\kappa}x & 0 \leq x \leq \kappa, \\ 0 & otherwise\,. \end{cases}$$

To prove Theorem 4, we first recall the following lemma which provides a bound on $E\left[\mathbb{I}_\kappa\{h(\mathbf{X}, y)\}\right]$ by the empirical error and the Rademacher complexity.

Lemma 4—For any h ∈ H, with probability at least 1 – δ, we have

$$E\left[\mathbb{I}_\kappa\{h(\mathbf{X}, y)\}\right] \leq \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_\kappa\left\{h\left(\mathbf{X}_i, y_i\right)\right\} + 2R_N\left(\mathbb{I}_\kappa \circ H\right) + \left\{\frac{\log\left(\delta^{-1}\right)}{N}\right\}^{1/2}.$$

The proof of Lemma 4 can be found in [34].

Now, we need to derive the upper bound of the Rademacher complexity used in Lemma 4. Since $\mathbb{I}_\kappa$ is $\frac{1}{\kappa}$-Lipschitz, we have

$$\begin{aligned} R_N\left(\mathbb{I}_\kappa \circ H\right) &\leq \frac{1}{\kappa}E_{\boldsymbol{\sigma}, \mathscr{T}}\left\{\sup_{\Sigma\left\|\mathbf{M}_j\right\|+\ \leq r} \frac{1}{N}\sum_{i=1}^{N} \sigma_i \sum_{j=1}^{K-1} tr\left(\mathbf{M}_j^\top \widetilde{\mathbf{X}}_i\right)\right\} \\ &= \frac{r}{\kappa N}E_{\boldsymbol{\sigma}, \mathscr{T}}\left\{\left\|\sum_{i=1}^{N} \sigma_i \widetilde{\mathbf{X}}_i\right\|_2\right\}, \end{aligned}$$

where $\widetilde{\mathbf{X}}_i$ denotes $\mathrm{X}_i - \overline{X}$ and $\overline{\mathbf{X}} = N^{-1}\sum_{i=1}^{N} \mathbf{X}_i$. The first inequality is due to Lemma 4.2 in [20], and the absolute values of the entries in $\mathbf{w}_y - \mathbf{w}_k$ are all bounded by 1.

Firstly, by the assumption, we can write $\mathbf{X} = E(\mathbf{X}) + \mathbf{E}$, where $E(\mathbf{X}) = \sum_{k=1}^{K} Pr(\mathscr{Y} = k)\mathbf{C}_k$ and the variance and the fourth moment of the entries are σ^2 and μ_4^4. Accordingly, $\widetilde{\mathbf{X}}_i = \mathbf{E}_i - \overline{\mathbf{E}}$, where $\overline{\mathbf{E}} = N^{-1}\sum_{i=1}^{N}\mathbf{E}_i$. Since $\left\{\left(\mathbf{X}_i, y_i\right)\right\}_{i=1}^{N}$ are the i.i.d. copies of $(\mathbf{X}, \mathscr{Y})$, we have

$$\left\|\sum_{i=1}^{N}\sigma_i\widetilde{\mathbf{X}}_i\right\|_2 \le \left|\sum_{i=1}^{N}\sigma_i\right|\left\|\overline{\mathbf{E}}\right\|_2 + \left\|\sum_{i=1}^{N}\sigma_i\mathbf{E}_i\right\|_2 .$$

Because $E\left[\left(\sum_{i=1}^{N}\sigma_i\mathbf{E}_i\right)^2\right] = N\sigma^2$ and $E\left[\left(\sum_{i=1}^{N}\sigma_i\mathbf{E}_i\right)^4\right] = N\mu_4^4 + 3N(N-1)\sigma^4$, by Theorem 2 in [16] we have

$$\begin{aligned} E_{\boldsymbol{\sigma}, \mathscr{T}}\left(\left\|\sum_{i=1}^{N}\sigma_i\mathbf{E}_i\right\|_2\right) &\le c\sigma N^{1/2}\left\{p^{1/2} + q^{1/2} + (pq)^{1/4}\left[N\mu_4^4 + 3N(N-1)\sigma^4\right]^{1/4} / \left(\sigma N^{1/2}\right)\right\} \\ &\le c\sigma\left(1 + \frac{3^{1/4}}{2}\right)N^{1/2}\left\{p^{1/2} + q^{1/2}\right\} + O\left(N^{1/4}\left(p^{1/2} + q^{1/2}\right)\right), \end{aligned}$$

where c is a constant which does not depend on $\mathscr{T}$. By similar arguments, it is easy to see that

$$\begin{aligned} E_{\boldsymbol{\sigma}, \mathscr{T}}\left(\left|\sum_{i=1}^{N}\sigma_i\right|\left\|\overline{\mathbf{E}}\right\|_2\right) &\le \sqrt{E_{\boldsymbol{\sigma}}\left\{\left(\sum_{i=1}^{N}\sigma_i\right)^2\right\}}E_{\mathscr{T}}\left(\|\overline{\mathbf{E}}\|_2\right) \\ &= N^{1/2}E_T\left(\|\overline{\mathbf{E}}\|_2\right) = O\left(p^{1/2} + q^{1/2}\right). \end{aligned}$$

Accordingly, we obtain the upper bound of the Rademacher complexity as

$$R_N\left(\mathbb{I}_\kappa \circ H\right) \le \frac{r}{\kappa\sqrt{N}}\left\{c\sigma\left(1 + \frac{3^{1/4}}{2}\right)\left(p^{1/2} + q^{1/2}\right)\right\}.$$

The proof is completed by using Lemma 4 with this bound and the fact that the continuous indicator function $\mathbb{I}_\kappa$ is an upper bound of the indicator function for any $\boldsymbol{\kappa}$. □

A.3.3 Proof of Theorem 5: Breakdown Point Analysis

Let $F(\mathbf{M}, \mathscr{T})$ denote the loss function (3) with the sample $\mathscr{T}$, and

$$\Delta^+ \triangleq \left\{\mathbf{M} \,\middle|\, \forall k,\ s.t.\ \mathbf{w}_k^\top\widehat{\mathbf{M}}\mathbf{M}^\top\mathbf{w}_k \rangle 0\right\} \text{ and } \Delta^- \triangleq \left\{\mathbf{M} \,\middle|\, \exists k,\ s.t.\ \mathbf{w}_k^\top\widehat{\mathbf{M}}\mathbf{M}^\top\mathbf{w}_k \le 0\right\}.$$

For the MSMM classifier, we can choose the contaminated observation as $(\mathbf{X}^o, k)$ with $\text{vec}\left(\mathbf{X}^o\right)^\top = -c\mathbf{w}_k^\top\widehat{\mathbf{M}}$. For any $\mathbf{M} \in \Delta^+$, $\mathbf{w}_k^\top\widehat{\mathbf{M}}\mathbf{M}^\top\mathbf{w}_k > 0$, then

$\mathbf{w}_k^\top \mathbf{M}\text{vec}\left(\mathbf{X}^o\right) = -c\mathbf{w}_k^\top \widehat{\mathbf{M}}\mathbf{M}^\top \mathbf{w}_k \to -\infty$ as $c \to \infty$. In this situation, the loss term corresponding to this contaminated observation will tend to infinity. Hence, we have $\widetilde{\mathbf{M}} \in \Delta^-$ and the classifier breaks down.

For the RMSMM, since $\widehat{\mathbf{M}} \neq 0, \widehat{\mathbf{M}}$ is an interior point of Δ^+, the claim

$$\epsilon_1 = \min_{\mathbf{M} \in \Delta^-} F\left(\mathbf{M}, \mathcal{T}_n\right) - \min_{\mathbf{M} \in \Delta^+} F\left(\mathbf{M}, \mathcal{T}_n\right) > 0$$

is true. Note that the loss function

$$l(\mathbf{X}, \mathcal{Y}, \mathbf{M}) = \gamma T_{s(K-1)}\left(\mathbf{w}_y^\top \mathbf{M}\text{vec}\,(\mathbf{X})\right) + (1-\gamma)\sum_{k \neq \mathcal{Y}} R_s\left(\mathbf{w}_k^\top \mathbf{M}\text{vec}\,(\mathbf{X})\right)$$

is bounded by $(K-1)(1-s)$. For any $m \leq n\epsilon_1/[2(1+\delta)(K-1)(1-s)]$ with $\delta > 0$ being any positive constant, any corresponding $n-m$ clean subset $\mathcal{T}_{n-m} \subset \mathcal{T}_n$, and any $\mathbf{M} \in \mathbb{R}^{p \times q}$, we have

$$0 \leq F\left(\mathbf{M}, \mathcal{T}_n\right) - \frac{n-m}{n} F\left(\mathbf{M}, \mathcal{T}_{n-m}\right) = \frac{1}{n}\sum_{i \in \mathcal{T}_n \backslash \mathcal{T}_{n-m}} l\left(\mathbf{X}_i, y_i, \mathbf{M}\right) \leq \frac{m(K-1)(1-s)}{n} < \frac{\epsilon_1}{2+2\delta}.$$

Therefore,

$$\left|\min_{\mathrm{M} \in \Delta^-} F\left(\mathrm{M}, \mathcal{T}_n\right) - \min_{\mathrm{M} \in \Delta^+} F\left(\mathrm{M}, \mathcal{T}_n\right) - \min_{\mathrm{M} \in \Delta^-} F\left(\mathrm{M}, \tilde{T}_{n,m}\right) + \min_{\mathrm{M} \in \Delta^+} F\left(\mathrm{M}, \tilde{T}_{n,m}\right)\right| \leq \frac{\epsilon_1}{1+\delta},$$

and

$$\min_{\mathrm{M} \in \Delta^-} F\left(\mathrm{M}, \tilde{T}_{n,m}\right) - \min_{\mathrm{M} \in \Delta^+} F\left(\mathrm{M}, \tilde{T}_{n,m}\right) > \frac{\epsilon_1 \delta}{1+\delta} > 0.$$

The last inequality reveals that $\widetilde{\mathbf{M}} \in \Delta^+$ and thus the classifier would not break down when $m \leq n\epsilon_1/[2(1+\delta)(K-1)(1-s)]$ observations are contaminated. Finally, the proof is complete by setting $\delta \to 0$. □

A.4 Derivation of Eq. (2): The dual problem

Lemma 5

For a $p \times q$ real matrix $\mathbf{A}$, the subdifferential of the nuclear norm $\|\cdot\|_*$ is given as

$$\partial\left\|\mathbf{A}\right\|_* = \left\{\mathbf{U}_\mathbf{A}\mathbf{V}_\mathbf{A}^\top + \mathbf{Z} \middle| \mathbf{Z} \in \mathbb{R}^{p \times q}, \mathbf{U}_\mathbf{A}^\top \mathbf{Z} = \mathbf{0}, \mathbf{Z}\mathbf{V}_\mathbf{A} = 0, \left\|\mathbf{Z}\right\|_2 \leq 1\right\},$$

where $\mathbf{U_A \Sigma_A V_A}^\top$ *is the SVD of* $\mathbf{A}$*, and* ∂ *stands for the operator of subgradients.*

Lemma 6

Suppose that $\mathbf{X} \in \mathbb{R}^{p \times q}$*,* $\partial G(\mathbf{X}) = \rho\mathbf{X} - \mathbf{P} + \tau\partial\|\mathbf{X}\|_*$*, where* $\mathbf{P} \in \mathbb{R}^{p \times q}$ *is a constant matrix w.r.t.* $\mathbf{X}$*. Let the SVD of* $\mathbf{P}$ *be*

$$\mathbf{P} = \mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^\top + \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^\top,$$

where Σ_0 *contains the singular values of* $\mathbf{P}$ *which are greater than* τ*, and* Σ_1 *contains the rest. Then, we have* $\mathbf{0} \in \partial G(\mathbf{X}^*)$*, where* $\mathbf{X}^* = \rho^{-1}\mathscr{D}_\tau(\mathbf{P}) = \rho^{-1}\mathbf{U}_0(\mathbf{\Sigma}_0 - \tau\mathbf{I})\mathbf{V}_0^\top$.

Lemma 6 can be verified by using Lemma 5 with $\mathbf{Z} = \tau^{-1}\mathbf{U}_1\Sigma_1\mathbf{V}_0^\top$.

Now we derive the dual problem (2) of (1). As in the classical SVM, by setting $C = (N\lambda)^{-1}$, we can rewrite (1) into the following form:

$$\begin{cases} \min\limits_{\mathrm{M}, b, \xi} \left\{ \frac{1}{2} tr\left(\mathbf{M}^\top\mathbf{M}\right) + \tau\left\|\mathbf{M}\right\|_* + C\sum\limits_{i=1}^N \xi_i \right\} \\ \text{s.t } \xi_i \geq 0,\ y_i\left[tr\left(\mathbf{M}^\top\mathbf{X}_i\right) + b\right] \geq 1 - \xi_i,\ i = 1, \cdots, N. \end{cases}$$

The corresponding Lagrange function of this problem can be written as

$$\begin{aligned} L_P(\mathrm{M}, b, \xi, \alpha, \mu) = {} & \frac{1}{2} tr\left(\mathbf{M}^\top\mathbf{M}\right) + \tau\left\|\mathbf{M}\right\|_* + C\sum_{i=1}^N \xi_i \\ & - \sum_{i=1}^N \alpha_i\left[y_i\left\{tr\left(\mathbf{M}^\top\mathbf{X}_i\right) + b\right\} - 1 + \xi_i\right] - \sum_{i=1}^N \mu_i\xi_i, \end{aligned} \tag{31}$$

where $\alpha_i \geq 0$ and $\mu_i \geq 0$ are corresponding Lagrange multipliers. By setting the derivatives w.r.t. b and ξ_i of this Lagrange function to zero, we get

$$\begin{cases} \sum\limits_{i=1}^N \alpha_i y_i = 0, \\ C - \alpha_i - \mu_i = 0,\ \ i = 1, \cdots, N. \end{cases}$$

Based on Lemma 6 and setting the derivative w.r.t. $\mathbf{M}$ to zero, we have $\mathbf{M} = \mathscr{D}_\tau\left(\sum_{i=1}^N \alpha_i y_i \mathbf{X}_i\right)$. Substituting these conditions into (31), we obtain

$$\begin{cases} \min_{\alpha} \left\{ \frac{1}{2} \left\| \mathscr{D}_{\tau} \left(\sum_{i=1}^{N} \alpha_i y_i \mathbf{X}_i \right) \right\|_F^2 - \sum_{i=1}^{N} \alpha_i \right\} \\ \text{s.t } 0 \le \alpha_i \le C; i = 1, \ldots, N, \sum_{i=1}^{N} \alpha_i y_i = 0. \end{cases}$$

This gives us the dual problem (2) of (1).

References

1. Altun K and Barshan B. Human activity recognition using inertial/magnetic sensor units In International Workshop on Human Behavior Understanding, pages 38–51. Springer, 2010.
2. Le Thi HA and Pham Dinh T. Solving a class of linearly constrained indefinite quadratic problems by dc algorithms. Journal of Global Optimization, 11(3):253–285, 1997.
3. Bauschke H and Combettes PL. Convex Analysis and Monotone Operators Theory in Hilbert spaces. Springer International Publishing, 2nd edition, 2017.
4. Boser BE, Guyon IM, and Vapnik VN. A training algorithm for optimal margin classifiers In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages 144–152. ACM, 1992.
5. Boyd S. Alternating direction method of multipliers In Talk at NIPS workshop on optimization and machine learning, 2011.
6. Cai D, He X, Wen J-R, Han J, and Ma W-Y. Support tensor machines for text categorization. Technical report, 2006.
7. Cai J-F, Candès EJ, and Shen Z. A singular value thresholding algorithm for matrix completion. SIAM J. Optim, 20(4):1956–1982, 2010.
8. Chambolle A and Pock T. A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis, 40(1):120–145, 2011.
9. Chambolle A and Pock T. On the ergodic convergence rates of a first-order primal–dual algorithm. Math. Program, 159(1–2):253–287, 2016.
10. Cortes C and Vapnik V. Support-vector networks. Machine Learning, 20(3):273–297, 1995.
11. Davis D and Yin W. Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions. Math. Oper. Res, 2014.
12. Friedman J, Hastie T, and Tibshirani R. The Elements of Statistical Learning Springer Series in Statistics. Springer-Verlag New York, 2nd edition, 2001.
13. He BS and Yuan XM. On the O(1/n) convergence rate of the Douglas-Rachford alternating direction method. SIAM J. Numer. Anal, 50:700–709, 2012.
14. Hou C, Nie F, Zhang C, Yi D, and Wu Y. Multiple rank multi-linear SVM for matrix data classification. Pattern Recognition, 47(1):454–469, 2014.
15. Huber PJ and Ronchetti E. Robust Statistics. Wiley series in probability and statistics John Wiley & Sons, Inc., Hoboken, New Jersey, 2nd edition, 2009.
16. Latala R. Some estimates of norms of random matrices. Proceedings of the American Mathematical Society, 133(5):1273–1282, 2005.
17. Lee Y, Lin Y, and Wahba G. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. Journal of the American Statistical Association, 99(465):67–81, 2004.
18. Liu Y. Fisher consistency of multicategory support vector machines In Artificial Intelligence and Statistics, pages 291–298, 2007.
19. Luo L, Xie Y, Zhang Z, and Li W-J. Support matrix machines. In Proceedings of the 32nd International Conference on Machine Learning, number 1, pages 938–947, Lille, France, 2015.
20. Mohri M, Rostamizadeh A, and Talwalkar A. Foundations of Machine Learning Adaptive computation and machine learning series. MIT Press, Cambridge, MA, 2012.

21. Monteiro RDC and Svaiter BF. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. SIAM J. Optim, 23(1):475–507, 2013.
22. Negahban S and Wainwright MJ. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. The Annals of Statistics, 39(2):1069–1097, 2011.
23. Nesterov Y. Introductory Lectures on Convex Optimization: A Basic Course, volume 87 of Applied Optimization. Kluwer Academic Publishers, 2004.
24. Pirsiavash H, Ramanan D, and Fowlkes C. Bilinear classifiers for visual recognition In Advances in Neural Information Processing Systems, pages 1482–1490, 2009.
25. Rockafellar RT. Convex Analysis, volume 28 of Princeton Mathematics Series. Princeton University Press, New Jersey, 1970.
26. Sun H, Craig B, and Zhang L. Angle-based multicategory distance-weighted svm. Journal of Machine Learning Research, 18(85):1–21, 2017.
27. Tao D, Li X, Wu X, Hu W, and Maybank SJ. Supervised tensor learning. Knowledge and Information Systems, 13(1):1–42, 2007.
28. Tran-Dinh Q. Proximal alternating penalty algorithms for constrained convex optimization. Comput. Optim. Appl, 72(1):1–43, 2019.
29. Wright SJ. Coordinate descent algorithms. Math. Program, 151(1):3–34, 2015.
30. Wu Y and Liu Y. Robust truncated hinge loss support vector machines. Journal of the American Statistical Association, 102(479):974–983, 2007.
31. Yang J, Zhang D, Frangi AF, and Yang J-Y. Two-dimensional pca: a new approach to appearance-based face representation and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(1):131–137, 2004. [PubMed: 15382693]
32. Zhang C and Liu Y. Multicategory angle-based large-margin classification. Biometrika, 101(3): 625–640, 2014. [PubMed: 26538663]
33. Zhang C, Liu Y, Wang J, and Zhu H. Reinforced angle-based multicategory support vector machines. Journal of Computational and Graphical Statistics, 25(3):806–825, 2016. [PubMed: 27891045]
34. Zhang C, Pham M, Fu S, and Liu Y. Robust multicategory support vector machines using difference convex algorithm. Math. Program, 169(1):277–305, 2018. [PubMed: 29736090]
35. Zhao J, Yu G, Liu Y, et al. Assessing robustness of classification using an angular breakdown point. The Annals of Statistics, 46(6B):3362–3389, 2018. [PubMed: 30294050]
36. Zhou H and Li L. Regularized matrix regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(2):463–483, 2014. [PubMed: 24648830]
37. Zou H and Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005.

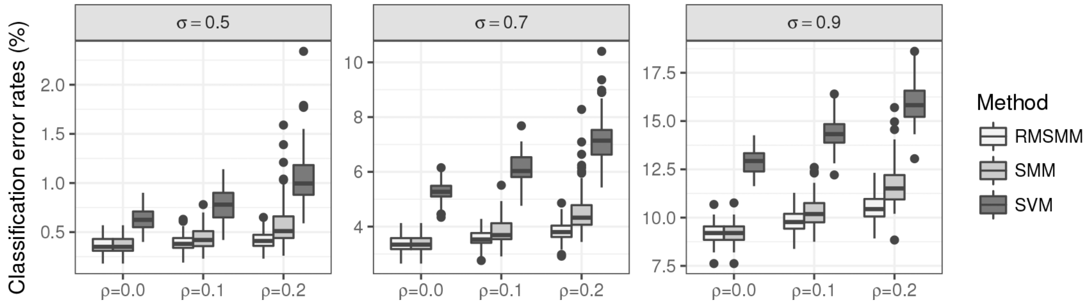


Fig. 1.

Classification error rates for RMSMM, SMM, and SVM on the simulated data with Scenario (I) and $K = 2$. Here, ρ stands for the percentage of data that are contaminated. SMM: [19]'s support matrix machine; SVM: the standard SVM applied to the stacked-up vectors.

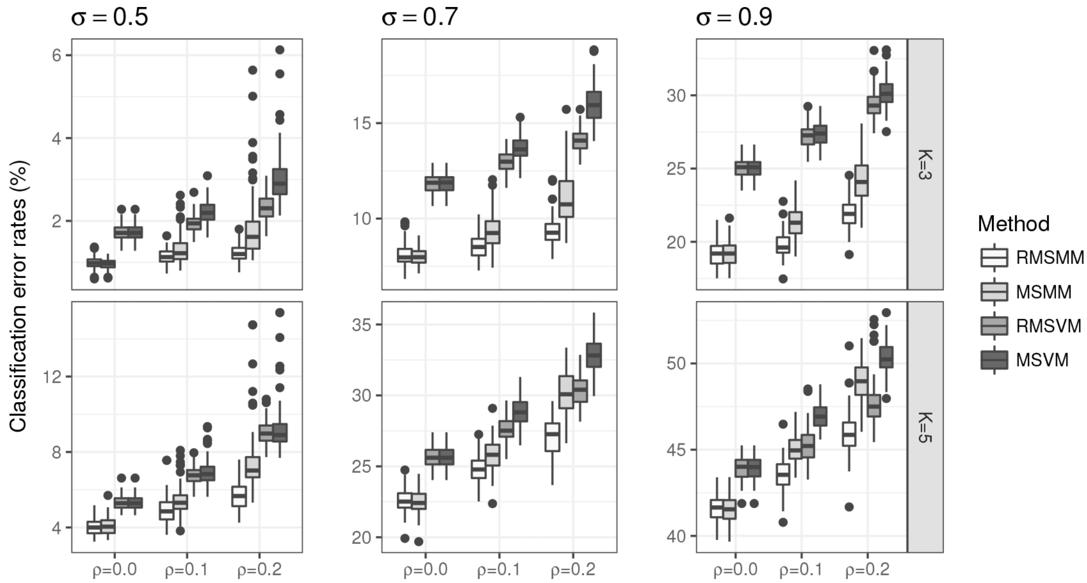


Fig. 2.

Classification error rates for RMSMM, SMM, and SVM on the simulated data with Scenario (I). The top three panels: the case with $K = 3$; the bottom three panels: the case with $K = 5$. MSMM: multicategory generalization of SMM using angle-based methods; MSVM: the angle-based multicategory SVM [32]; RMSVM: the robust angle-based multicategory SVM [34].

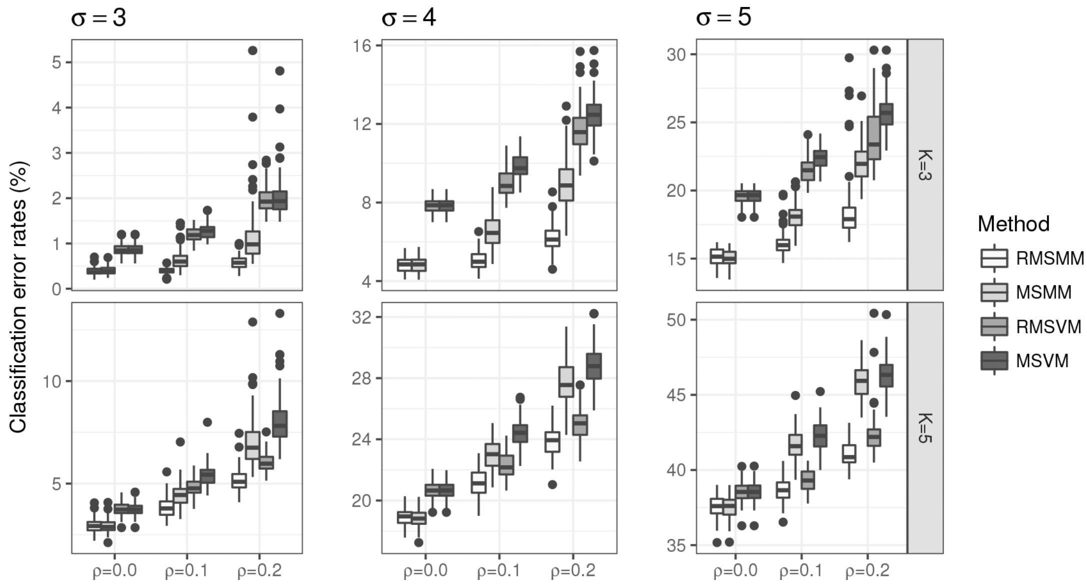


Fig. 3.

Classification error rates for RMSMM, SMM, and SVM on the simulated data with Scenario (II). The top three panels: the case of $K = 3$; the bottom three panels: the case of $K = 5$.

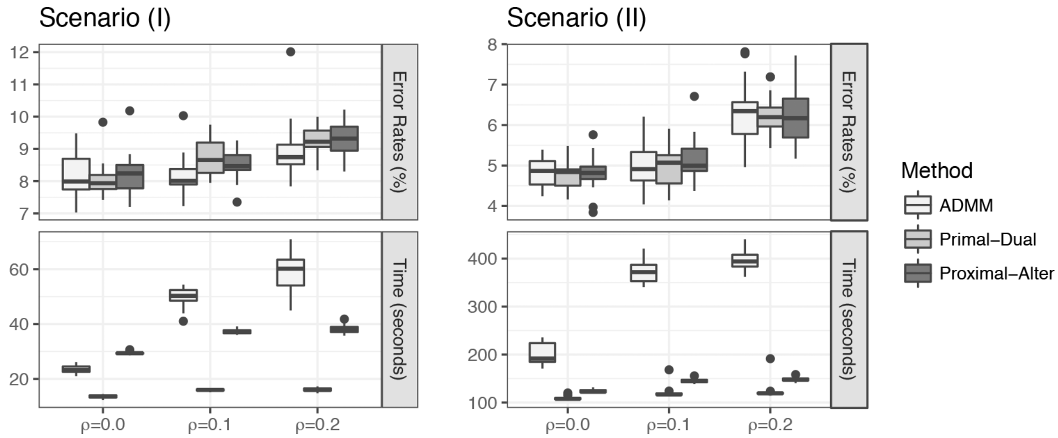


Fig. 4.

Comparison between the ADMM and primal-dual algorithms: Primal-Dual stands for (18), and Proximal-Alter stands for (19) for solving the RMSMM optimization problem (5). The top two panels: classification error rates under Scenario (I) with $\sigma = 0.7$ and Scenario (II) with $\sigma = 4$ when $K = 3$; The bottom two panels: the corresponding computational time (in seconds).

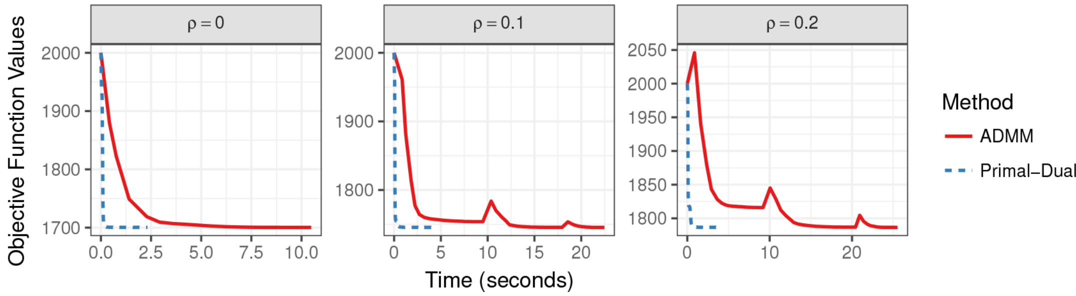


Fig. 5.

The decrease of the RMSMM objective values with respect to the computational time under Scenario (II) with $K = 3$ and $\sigma = 4$.

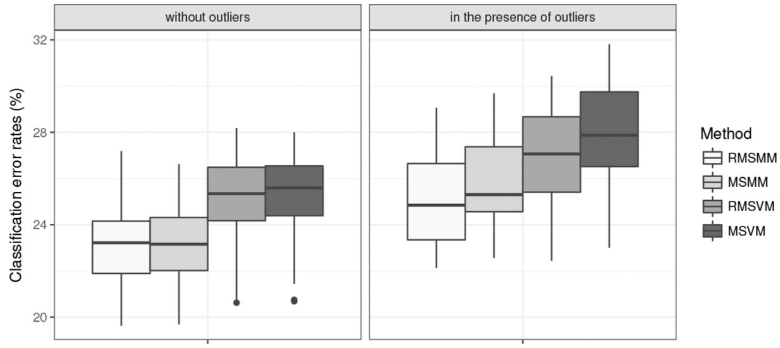
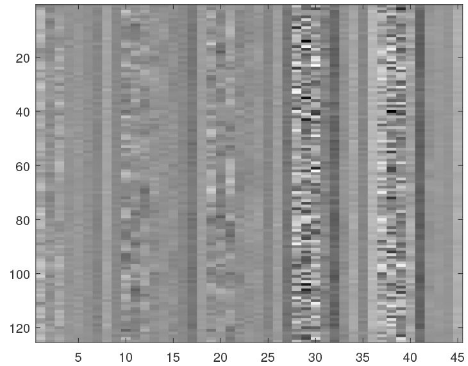


Fig. 6.

Classification error rates for RMSMM, MSMM, RMSVM, and MSVM on the Daily and Sports Activities Dataset. The left and right panels present the results when the data are clean or contaminated, respectively.



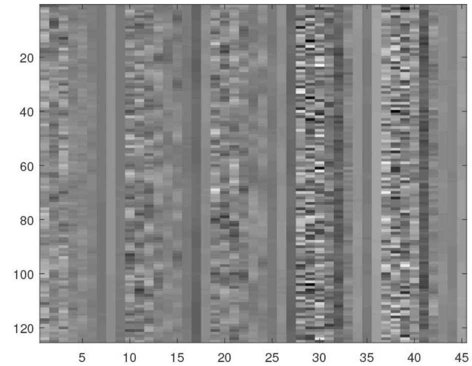


Fig. 7.
Heatmaps of the first decision matrices of RMSMM (left panel) and RMSVM (right panel)