Support Matrix Machine via Joint $\ell_{2,1}$ and Nuclear Norm Minimization Under Matrix Completion Framework for Classification of Corrupted Data

Imran Razzak, *Senior Member, IEEE*, Mohamed Reda Bouadjenek, Raghib Abu Saris, and Weiping Ding

***Abstract*— Traditional support vector machines (SVMs) are fragile in the presence of outliers; even a single corrupt data point can arbitrarily alter the quality of the approximation. If even a small fraction of columns is corrupted, then classification performance will inevitably deteriorate. This article considers the problem of high-dimensional data classification, where a number of the columns are arbitrarily corrupted. An efficient Support Matrix Machine that simultaneously performs matrix Recovery (SSMRe) is proposed, i.e. feature selection and classification through joint minimization of $\ell_{2,1}$ (the nuclear norm of L). The data are assumed to consist of a low-rank clean matrix plus a sparse noisy matrix. SSMRe works under incoherence and ambiguity conditions and is able to recover an intrinsic matrix of higher rank in the presence of data densely corrupted. The objective function is a spectral extension of the conventional elastic net; it combines the property of matrix recovery along with low rank and joint sparsity to deal with complex high-dimensional noisy data. Furthermore, SSMRe leverages structural information, as well as the intrinsic structure of data, avoiding the inevitable upper bound. Experimental results on different real-time applications, supported by the theoretical analysis and statistical testing, show significant gain for BCI, face recognition, and person identification datasets, especially in the presence of outliers, while preserving a reasonable number of support vectors.**

***Index Terms*— Joint matrix recovery, outliers, support matrix machine, support vector machines (SVMs).**

I. Introduction

MORE often than not, real-world data are multidimensional and imperfect. These attributes pose serious challenges, especially in datasets of limited size. Traditional support vector machines (SVMs) [1], [2], [3] are fragile in the presence of outliers; even a single corrupt point can arbitrarily affect classification performance. Persistent or nonprobabilistic data corruption stems from failures in sensor inputs, or from malicious data tampering.

In addition to data corruption, some of the available data may not conform to the presumed low-dimensional model, i.e., if most of the columns are in low-dimensional space, the corresponding matrix is low rank and a small number of columns are outliers that correspond to a column-sparse matrix [4], [5], [6].

Vector-based methods have been successively applied for classification and with good results. State-of-the-art vector-based methods are linear discriminant analysis (LDA) [7], [8], [9], [10], SVMs [1], [2], [3], and k-nearest neighbor (KNN) [11], [12], [13]. However, the data have to be reshaped into vectors for further classification and this can in turn destroy the structural information embedded within. An alternative solution to avoid this problem is to concatenate the matrix into a vector for classification. However, this results in an increase in dimensionality that leads to over-fitting. Recently, some efforts have been made to cast matrices into vectors using common spatial patterns [14], [15], [16], [17], [18], [19]. However, these methods ignore the topological structure embedded in matrix data, and this structural information is of great interest, considering it helps to improve classification.

Recently, several efforts have been made to classify matrices directly without conversion to their respective vector representation, in so doing exploiting the correlation between the columns or rows of the matrix. Rank-k SVM [20] and Rank-k Logistic Regression [4], [5] model the regression matrix as a sum of the k rank-one orthogonal matrix. Song et al. [5] presented 2-D large margin nearest neighbor to improve the KNN classification for matrix data by adopting left projection and right projection matrix to define the matrix-based Mahalanobis distance. In another work, authors have presented rank-k 2-D multinomial logistic regression for multiclass matrix classification problem by modeling each category through left and right projection matrices with rank k. Pirsiavash et al. [21] presented a bilinear classifier by applying the hinge loss for model fitting through factorization of the regression matrix into a low-rank matrix. Zhang et al. [22] devised low-rank linearization to transform the nonlinear SVM to a corresponding linear SVM, through a kernel map computed from the low-rank approximation of matrices. One major disadvantage

Manuscript received 28 June 2021; revised 4 October 2022 and 12 May 2023; accepted 4 July 2023. *(Corresponding authors: Imran Razzak; Weiping Ding.)*

Imran Razzak is with the School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 1466, Australia (e-mail: imran.razzak@unsw.edu.au).

Mohamed Reda Bouadjenek is with the School of Information Technology, Deakin University, Geelong, VIC 3216, Australia (e-mail: reda.bouadjenk@deakin.edu.au).

Raghib Abu Saris is with the College of Public Health and Helaht Informatics, King Saud bin Abdulaziz University for Health Sciences, Riyadh 11426, Saudi Arabia (e-mail: sarisr@kaust-hs.edu.sa).

Weiping Ding is with the School of Information Science and Technology, Nantong University, Nantong 226019, China (e-mail: dwp9988@163.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3293888.

Digital Object Identifier 10.1109/TNNLS.2023.3293888

of these methods is that each new feature in a low-dimensional subspace is the linear combination of all the original features in high-dimensional space. Thus, such treatments usually affect classification performance due to the inclusion of redundant features. Furthermore, it is often difficult to interpret new features when treated in this way.

To tackle the challenge of robust feature selection, the sparsity regularization in dimensionality reduction has recently been investigated for feature selection i.e., ℓ_1 [23], [24], ℓ_q [25], $\ell_{2,0}$ [26], and $\ell_{2,1}$ [27], [28]. The Frobenius norm has also been applied to introduce sparsity in a regression matrix [4], [5], [29]. These approaches work well and consider the correlation between columns and rows under the low-rank assumptions and provide satisfactory performance [30]. However, these approaches consider all the entities of the matrix as explanatory factors, whereas in the real world, some features might be redundant (even useless) for certain classification tasks; in other words only a small set of useful features are used when classifying unseen data. For example, due to the low-ranked nature of gene or human facial images, obtaining relevant features by removing irrelevant and redundant features, reducing computational costs without significant loss of information nor negatively degrading learning performance.

Besides intra-sample outliers, traditional classifiers, whether vector-based or matrix-based support machines, are fragile in the presence of outliers [31] and are not suitable as classifiers at all in the presence of corrupt data. Recent matrix classifiers (such as rank-k SVM and bilinear classifiers) incorporate the low-rank property and introduce certain constraints on the regression matrix to leverage the correlation. However, it requires predetermination of rank of the regression matrix which is complex and requires tuning. Recently, low-rank matrix completion methods have proven the importance of exact matrix recovery from partial observations. For instance, suppose we are given a partially observed matrix, and we know that the full matrix can be decomposed as $X = L + S$, where matrix L is low rank and S is sparse and consists of only few nonzero columns. Here, both matrices L and S have arbitrary magnitude, the rank of matrix L as well as position and number of corrupted columns of the matrix S are unknown. Can we classify this type of corrupted data efficiently?

The short answer is yes. We can classify these data by combining the low-rank matrix completion with the support matrix machine, even where a fraction of columns is corrupted. The solution is dependent on the efficient recovery of matrix L on noncorrupted columns, and the selection of structural and intrinsic features. To solve this, we propose joint minimization of matrix recovery and hinge loss which helps to account intra-sample outliers.

Another challenge is the dimensionality and loss of structural information from reshaping data into vectors for classification [32], [33], [34]. Reshaping into vectors can destroy the structural information embedded within the data, as well as increasing its dimensionality. To address the aforementioned, we simultaneously optimize nuclear norm, $\ell_{2,1}$ norm (the nuclear norm of L), and hinge loss on matrix data. We provide convex optimization formulation of a proposed objective function and identify the sufficient conditions under which it classifies corrupted data efficiently through a low-rank feature recovery process. Results show, under certain natural conditions, the optimum of this convex program yields the best classification performance through low-rank feature recovery, even when a fraction of columns is corrupted. Compared to the state-of-the-art featured selection methods, we can describe the theoretical and empirical key contributions of this work as follows:

1) a novel classifier is proposed effectively combining the hinge loss function for model fitting, low-rank matrix recovery, and elastic net penalty for regularization on a regression matrix, and simultaneous matrix recovery is performed followed by clean feature extraction and classification;
2) we present a method, called **S**upport **M**atrix **M**achine, by simultaneously performing matrix **Re**covery (abbreviated SMMRe), that is able to classify data with denser corruption ($L \leq (C_r n / \log(n))$ and $S \leq C_s n$, where C_s and C_r are numerical constants) through exact recovery of the intrinsic matrix of higher rank based on incoherence conditions;
3) since convex optimization cannot perform an exact recovery of a corrupted matrix, an oracle problem for matrix recovery is used. As a result, convex optimization-based SMMRe performs correct matrix recovery as well as identifying outliers, which improve classification performance;
4) the goals above are achieved by employing a regularizing term [a combination of low-rank and $\ell_{2,1}$ (the nuclear norm of L)] which promotes structural sparsity and matrix recovery as well as selecting features across all data points with joint sparsity. The low-rank matrix recovery helps to recover unobserved entities as well as avoid the inevitable upper bound for the number of selected features occurring in $\ell_{2,1}$-norm SVM;
5) since the optimization is convex (but nonsmooth), one of the major challenges is how nonsmooth optimization can be efficiently solved. To this end, we devised an efficient algorithm to solve the proposed objective function.

Rest of the article is organized as follows. Section II introduces the basic notations and preliminaries used in this article followed by motivation of this work in section III. In Section IV, we present the related work on support matrix machines and in Section V, we present the proposed problem formulation followed by the proposed objective function, its optimization, and theoretical justification in Section VI. Section IX reports the experimental results on different datasets. Finally, Section X conclude this article.

II. Notations and Preliminaries

We start by establishing the notations and preliminaries used throughout this article. Following standard conventions, scalar, vector, and matrix are represented by lowercase letters (e.g., x), lowercase bold letters (e.g., **x**), and uppercase letter (e.g., X), respectively. Let Ω and Ω' represent both sets of matrix entries and the linear space of matrices supported on these entries;

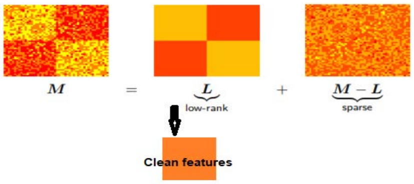
M
=
$\underbrace{L}_{\text{low-rank}}$
+
$\underbrace{M-L}_{\text{sparse}}$
Clean features

Fig. 1. Motivation for joint low-rank plus matrix recovery-based classification for missing plus corrupted data.

similarly, A_o and A_S denote both the set of column indices and the linear space of matrices supported on these columns. We let I_p denoted by $p \times p$ matrix. For a linear subspace $\mathbb{S}$, we let $\mathbb{P}(\mathbb{S})$ denotes the orthogonal projection onto linear subspace $\mathbb{S}$. For a matrix $\mathbf{X} \in \mathbb{R}^{p\times q}$, $\ell_{2,1}$, norm of matrix is denoted as $||X||_{2,1} = \sum_{i=0}^{q} ||x^i||_2 = \sum_{i=0}^{q}(\sum_{j=0}^{p} ||X_{i,j}^2||)^{1/2}$. It is rotational invariant for rows for any rotational matrix R i.e., $||X_R||_{2,1} = ||X||_{2,1}$. The $\ell_{2,1}$ norm can be generalized to r, p-norm.

As we know, the nuclear norm $||X||_* = \sum_{i=1}^{r} \sigma_i$ of a matrix X as a function from $\mathbb{R}^{p\cdot q}$ to $\mathbb{R}^1$ cannot be differentiated. Alternatively, we have to consider the sub-differential of $||X||_*$ that is denoted by $\partial||A||_*$. It is a set of sub-gradients. For a matrix X of dimension $p \times q$ of rank r

$$\partial||A||_* = \big\{U_X V_X^T + Z : Z \in \mathbb{R}^{p\times q},\\ U_X^T Z = 0, ZV_X = 0, ||Z||_2 \leq 1\big\}. \quad (1)$$

We further introduce the singular value thresholding (SVD) operator to approximate the matrix with minimum nuclear norm. It simply applies a soft-thresholding rule to the singular values of X, effectively shrinking these toward zero [35], [36], [37].

For any $\tau \geq 0$, the SVD operator is defined as follows:

$$\mathbb{D}_\tau[X] = U_X S_\tau[\Sigma_X] V_X^T$$

where $S_\tau[\Sigma] = \text{diag}([\sigma_1(X) - \tau]_+, \ldots, [\sigma_r(X - \tau]_+)$ and $[z]_+ = \max(z, 0)$.

III. Motivation

In this article, our concern is classification problem on a set of corrupted data matrix. Input data are high in dimension and noisy; hence, we focus our attention on regularizers that have the ability to recover the corrupted data and promote structural sparsity to find robust solutions against outliers. Moreover, our target is to endow the feature space that does not penalize the features individually as in the case of the ℓ_1 norm. Recently, low-rank matrix recovery has shown tremendous performance for the recovery of unobserved noisy data [38]. Inspired by this performance, we intend to combine the matrix recovery into support matrix machines through simultaneous optimization. As a result, iteratively SMMRe is not only able to recover the unobserved entities, but also combines the property of low rank and sparsity together. Figs. 1 and 2 illustrate the proposed framework. Fig. 2 shows that SMMRe first recover the clean matrix followed by classification.

IV. Related Work

In this section, we provide a brief description and formalization of the matrix classification problem. Practically, it has been noticed that the selection of features and model designs is far more important than the choice of loss [39]. Hence, in this coherence, we focused the regularization term in promoting the structural sparsity and leveraging the intrinsic structure of data.

We have given a set of training samples $T = \{X, y_i\}_{i=1}^n$, where $X_i \in \mathbb{R}^{p\times q}$ is the the ith input sample matrix and $y_i \in \{1, -1\}$ is its corresponding class label. Generally, the data needs to be transformed into corresponding vector. To fit a classifier, matrix X is needed to be stacked into vector. Let $x_i = \text{vec}(X_i^T) = ([X_i]_{11}, [X_i]_{12}, \ldots, [X_i]_{1q}, [X_i]_{2,1}, [X_i]_{22}, \ldots, [X_i]_{pq})^T \in \mathbb{R}^{pq}$.

The classical soft margin SVM is defined as

$$\arg\min \frac{1}{2}\text{tr}(w^T w) + C\sum_{i=1}^{n} 1 - y_i[\text{tr}(W^T x_i) + b]_+ \quad (2)$$

where $1 - y_i[\text{tr}(W^T X_i) + b]_+$ is the hinge loss, $W \in \mathbb{R}^{pq}$ is the vector of regression coefficients, $b \in \mathbb{R}^{pq}$ is an offset term, and C is a regularization parameter.

In (2), we need to reshape the matrix into vectors which result in losing the correlation among columns or rows in the matrix. An alternative solution for this problem is to concatenate the matrix into a vectors for classification. However, it results in increase in dimensionality that leads to model overfitting. Recently, some efforts have been made to suppress the matrix into vectors using common spatial patterns [14], [15], [16], [17], [18], [19]. However, these methods ignore the topological structure (the relationship between neighboring data points) embedded in the matrix data, whereas considering structural information is of great interest and helps to improve the classification.

By directly transforming the (2) for matrix, we get

$$\arg\min \frac{1}{2}\text{tr}(W^T W) + C\sum_{i=1}^{n} 1 - y_i[\text{tr}(W^T X_i) + b]_+. \quad (3)$$

It is known that $\text{tr}(WW^T) = \text{vec}(W)\text{vec}(W^T)$ and $\text{tr}(W^T X_i) = \text{vec}(W)^T\text{vec}(X_i)$, thus the above objective function cannot capture the intrinsic structure of each input matrix efficiently, due to the loss of structural information during the reshaping process. To take the advantage of intrinsic structural information within each matrix, one intuitive way is to capture the correlation within each matrix through low-rank constraints on the regression parameter.

As the hinge loss enjoys the large margin principle, it also embodies sparseness and robustness, which are two desirable properties for a good classifier. Motivated by this, Luo et al. [40] presented the sparse matrix machine shown in (4) . The objective function in (4) consists of hinge loss plus nuclear norm and Frobenius norm as a regularizer

$$\arg\min_{W,b} \frac{1}{2}\text{tr}(W^T W) + \tau||W||_* + C\sum_{i=1}^{n}\{1 - y_i[\text{tr}(W^T X_i) + b]\}_+. \quad (4)$$

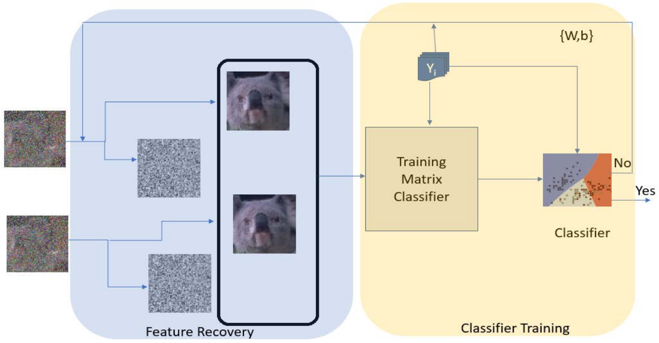

Fig. 2. Proposed joint low-rank plus matrix recovery-based classification framework (matrix recovery + classifier).

The spectral elastic net regularization $(1/2)\mathrm{tr}(W^T W) + \tau||W||_*$ captures the correlation within each matrix. In addition, the nuclear norm in the regularizer is used to control the rank of W that is NP-hard problem. In this scenario, it provides the best approximation of rank of the matrix W. The objective function shown in (4) is capable of capturing the latent structure within each matrix and further performs the classification based on all entities of each matrix which effect the classification performance, thus, making the model complicated. To overcome this challenge, Zheng et Al. [41] presented sparse support matrix machine that consists of loss plus nuclear norm and ℓ_1 as regularizer term

$$\arg\min \gamma||W||_1 + \tau||W||_* + C\sum_{i=1}^{n}\{1 - y_i[\mathrm{tr}(W^T X_i) + b]\}_+. \tag{5}$$

The classification function in (7) incorporates the loss and constraints on the regression matrix which is a linear combination of ℓ_1 norm and nuclear norm. ℓ_1 norm encourages matrix W to be sparse by serving as a convex surrogate for nonzero entries. The regularizer term in (7) is combination of ℓ_1 norm and nuclear norm which provides structural sparsity. A common features of approach based on Frobenius norm [40] and ℓ_1 norm [41] are that they treat both indices (row and column) in the same way. However, they have different meanings i.e., i and j run through data points and spatial dimensions, respectively. This subtle distinction makes it easy to get loss for the matrix, whereas, $\ell_{2,1}$ norm captures this subtle distinction and provides structural sparsity. Furthermore, studies have shown that $\ell_{2,1}$ is sparser than ℓ_1-regularization as it finds the joint solutions and encourages multiple predictors to share similar sparsity patterns.

V. Problem Formulation

Suppose we are given data X with dimension $p \times q$ to classify, a fraction of these columns span r-dimensional subspace while rest of the columns are arbitrarily corrupted. We are given only a partial set of observations and our goal is to classify such type of data based on the partial set of observations. The data matrix can be decomposed as $X = L + S$. L is the column-sparse matrix that corresponds to corrupted columns, thus at most αn columns are nonzeros. L corresponds to noncorrupted matrix, thus $\mathrm{rank}(L) = r$ and $(1-\alpha)n$ columns of matrix L are nonzeros, corresponding to the outliers. Better performance cannot be guaranteed in all cases because there could be completely unobserved rows or columns resulting in no hope of selecting features belonging to the missing data; in such case, missing value cannot be recovered. Notice that we have fraction of observed. Suppose $\Omega \subset [p] \times [q]$ are observed entities, and $P(\Omega)$ is the orthogonal projection onto the linear subspace of matrices supported on Ω i.e., $P_\Omega(M) = M_{i,j}$ if $i, j \in \Omega$ and $P_\Omega(M) = 0$ if $i, j \notin \Omega$. We intend to classify the corrupted data efficiently through matrix recovery framework. Thus, we propose to optimize matrix recovery and classification with an additional objective of low-rank feature representation. We assume that the matrix L satisfies the incoherence conditions ($\max|Ue_i|^2 \le \mu(r/p)$ and $\max|Ve_j|^2 \le \mu(r/(1-\alpha)n)$), where e is the unit matrix.

VI. Proposed SMMRe

In this section, we introduce the proposed SMMRe, which, as a matter of fact is a novel classifier. SMMRe simultaneously recovers the corrupted matrix while removing the redundant information. The classifier also selects the discriminant patterns and considers the strong correlation of rows and columns in the matrix. It is well known that hinge loss enjoys the large margin as it provides tight and convex upper bound on the indicator function which penalizes misclassifications. It embodies sparseness and robustness as it acts like a regularizer which induces joint sparsity (in term of support vectors, SVM is sparse as compared to least-squares SVM). In this regard, we adopt the loss function and propose a robust approach that efficiently performs matrix recovery, cleans feature extraction from recovered matrix, imposes sparseness, and preserves the structural information. The proposed objective function is joint optimization of low-rank matrix recovery, hinge loss for model fitting plus the regularization on regression matrix. To this end, we have the objective function

$$\begin{aligned}\arg \min_{W,b,\{L_i,S_i\}_{i=1}^n} & \sum_{i=1}^{n}(\alpha_1||L_i||_* + \alpha_2||S_i||_{2,1}) + \tau||W||_* \\ & + \sum_{i=1}^{n}\{1 - y_i[\mathrm{tr}(W^T L_i) + b]\}_+ \\ \text{such that } \forall i, \quad & X_i = L_i + S_i \end{aligned} \tag{6}$$

where $L_i \in \mathbb{R}^{p\times q}$, $S_i \in \mathbb{R}^{p\times q}$, and $W \in \mathbb{R}^{\times q}$ are the low-rank matrix corresponding to noncorrupted columns, sparse matrix corresponding to corrupted columns, and regression matrix, respectively. α_1, α_2, and τ are positive scalars that penalize the sparse matrix, nuclear norm of low rank and nuclear norm of regression matrix, respectively.

The above (6) is a combination of four terms, hinge loss function, matrix recovery ($\ell_{2,1}$, nuclear norm of L), and nuclear norm of W. In results, the objective function not only inherits the properties of matrix recovery and identifies the corrupted column with high probability but also holds the properties of low rank and sparsity together which helps to deal with outliers and corrupted data. Moreover, the regularizer terms in (6) are able to encode the prior knowledge and guide the selection of features by modeling the structure of the feature space.

A. Matrix Recovery and Training

The objective function in (7) consists of four terms, all of which are convex. The $\ell_{2,1}$-norm and nuclear norm are convex as both satisfy the triangle and homogeneity properties whereas the other term is a linear function thus it is also convex. The optimization problem for the SMMRe is convex, nonsmooth, and nondifferentiable; however, the combination of hinge loss, $\ell_{2,1}$-norm and nuclear norm makes the problem nontrivial to be solved directly. To decouple the hinge loss and nuclear norm with respect to W in SMMRe, we have introduced an auxiliary variable, and applied Lagrange multiplier. The above equation can be written as

$$\begin{aligned}\min_{W,b,\{L_i,S_i\}_{i=1}^n} &\sum_{i=1}^{n}(\alpha_1||L_i||_* + \alpha_2||S_i||_{2,1}) + \tau||W||_* \\ &+ C\sum_{i=1}^{n} h(W,b,L_1) \\ \text{s.t. } \forall i, \quad & X_i = L_i + S_i \text{ and} \\ & W = Z \text{ where } Z \text{ is auxiliary variable.}\end{aligned} \tag{7}$$

Now the constrained problem in (7) can be efficiently solved using augmented Lagrangian multiplier algorithm (ALM). The key of ALM method is to search for a saddle point of the augmented Lagrangian function instead of solving the original constrained optimization problem. The augmented Lagrangian function is given as follows:

$$\begin{aligned}&\mathcal{L}(W,Z,b,L_i,S_i,V,M)\\ &\quad= \sum_{i=1}^{n} h(W,b,L_i) + \tau_1||Z||_*\\ &\qquad + \operatorname{tr}[V^T(Z-W)] + \frac{\mu_1}{2}||Z-W||_F^2\\ &\qquad + \sum_{i=1}^{n}\Big\{\alpha_1||L_i||_* + \alpha_2||S_i||_{2,1} + \operatorname{tr}\\ &\qquad\qquad \times\big[M_i^T(X_i - L_i - S_i) + \frac{\mu_2}{2}||X_i - L_i - S_i||_F^2\Big\}\end{aligned} \tag{8}$$

where $h(W,b,L_i) = 1 - y_i[\operatorname{tr}(W^TL_i)+b]_+$, M, $V \in \mathbb{R}^{pq}$ are the Lagrange multiplier. μ_1 and μ_2 are the positive penalty parameters. α_1, α_2, and τ control the trade-off between hinge loss and regularization terms i.e., α_1, α_2 controls the recovery process and clean feature selection whereas τ captures the correlation of data matrix. Updating Lagrange multipliers as

$$(W^k, Z^k, b^k) = \min_{W,Z,b} \mathcal{L}(W,Z,b,L_i^{k-1},V^{k-1}) \tag{9}$$

$$(L^k, S^k) = \min_{L_i,S_i} \mathcal{L}\big(W^k, b^k, L_i, S_i, M_i^{k-1}\big) \tag{10}$$

$$V^k = V^{k-1} + \mu_1(Z^k - W^k) \tag{11}$$

$$M_i^k = M_i^{k-1} + \mu_2\big(X_i - L_i^k - S_i^k\big). \tag{12}$$

Notice that, the (9) estimates the model parameter for matrix classification, (10) perform the matrix recovery and clean feature selection simultaneously. Thus, it validates the core objective of clean feature extraction through matrix recovery. As (9) is difficult to solve directly, thus, we solved (described in Theorem 1) it by minimizing $\mathcal{L}$ against W, Z, and b.

To compute Z, minimizing (8) ($\mathcal{L}(W,Z,b,L_i,S_i,V,M)$) with respect to Z, we get

$$f(Z) = \tau_1||Z||_* + \operatorname{tr}(V^TZ) + \frac{\mu}{2}||Z-F||_F^2. \tag{13}$$

Z can be updated based on the following theorem.

Theorem 1: For any positive scalars α and μ_1, consider $f(Z)$ denotes $\tau_1||Z||_* + \operatorname{tr}(V^TZ) + (\mu/2)||Z-F||_F^2$.

We have $\partial f(Z) = 0$.

Minimizing $f(Z)$ with respect to Z, we reach the following optimal solution:

$$Z = \frac{1}{\mu_1}\mathbb{D}_\xi(\mu_1 W - V). \tag{14}$$

$\mathbb{D}_\xi$ can be computed as

$$\mathbb{D}_\xi = U\mathcal{S}_\tau(\Sigma)V^T$$

where $\mathcal{S}_\tau$ is the entry-wise soft thresholding operator.

Proof: The (13) consist of quadratic terms, thus $f(Z)$ is convex. There exist an optimal minimizer Z' such that $Z = (1/\mu_1)\mathbb{D}_\xi(\mu_1 W - V)$. Z' minimizes $f(Z)$ only if sub-gradient of $f(Z')$ is 0. We can write

$$0 \in \partial||Z'||_* + V + \mu_1(Z' - W) \tag{15}$$

where $\partial||Z||_*$ is the set of sub-gradients of nuclear norm.

Consider Z is an arbitrary matrix, we can write

$$\begin{aligned}\partial||Z||_* &= UV^T + M \text{ s.t. } M \in \mathbb{R}^{p\times q}\\ U^TM &= 0, \quad MV = 0, \quad ||M||_F \le 1.\end{aligned} \tag{16}$$

To prove $Z = (1/\mu_1)\mathbb{D}_\xi(\mu_1 W - V)$ satisfies (15), we decompose $\mu_1 W - V$ into following components:

$$\mu_1 W - V = U_0\Sigma_0V_0^T + U_1\Sigma_1V_1^T.$$

From the above equation, we can write

$$\begin{aligned}\mu_1(W - Z') - V &= \mu_1 W - V - \mu_1 Z'\\ &= \tau(U_0V_0^T + \frac{1}{\tau}U_1\Sigma_1V_1^T.\end{aligned} \tag{17}$$

Comparing (16), we can define $M = (1/\tau)U_1\Sigma_1V_1^T$. Thus, it can be verified that $U_0M_0 = 0$ and $MV_0 = 0$ and $||M||_F \le 0$. Thus, we have $\mu_1(W - Z') - V \in \tau\partial||Z'||_*$, Hence proved. ■

Similar to compute W and b, we can rewrite the (8) as

$$\min_{W,b}\sum_{i=1}^{n} h(W,b,L_i) - \operatorname{tr}(V^TW) + \frac{\mu_2}{2}||Z-W||_F^2. \tag{18}$$

W is computed as

$$\begin{aligned}W &= \frac{1}{\mu}\left(\mu Z + V + \sum_{i=1}^{n} a_i y_i L_i\right)\\ a &= \max_{\alpha} -\frac{1}{2}\alpha^Y K\alpha + q^T\alpha\\ K &= \frac{1}{\alpha_1} y_i y_j \operatorname{tr}(L_i^T, L_j)\end{aligned}$$

TABLE I

ALGORITHMIC PROCEDURE OF PROPOSED SPARSE SUPPORT MATRIX MACHINE UNDER MATRIX RECOVERY FRAMEWORK (SMMRE)

```
Input: : Labeled Training dataset: $[X_i, y_i]$ where $X_j \in R^{m\times n}$ for $j = 1, ..., N$, low-rank co-efficient $\tau$, sparsity coefficient $\alpha_1$ $\alpha_2$ $\alpha_3$,; smoothing parameter $\alpha$, weights $w_1$ and $w_2$

Output: Matrices $W$, $L$, $S$ and bias $b$

Step-I: Initialize the matrix $W, = 0$, $L_i = X_i$, $S_i = X_i - L_i$, $M = 0$, $V = 0$
While not converge do
    Step-II:Compute $Z = \frac{1}{\mu_1} D_\xi(\mu_1 W - V)$
    Step-III: Compute $W = \frac{1}{\mu}(\mu Z + V + \sum_{i=1}^{n} a_i y_i L_i)$
    Step-IV: Compute $\min_S = \alpha_3 ||S_i||_{2,1} - tr(M^T S_i) + \frac{\alpha_2}{2} ||X_i - S_i) + M_i)$

    Step-VI: Update $b = \frac{1}{n}\sum_{i=1}^{n}(y_i - tr(W^T L_i))$
    Step-VII: Update $M = M^{k-1} + \alpha_2(X_i - L_i - S_i)$

end while
Step-VII: Return W, L, S and b
```

TABLE II

SUMMARY OF DATASET

Dataset	subject	Dimension	Train	Test
Caltch Face	435	320×280	218	217
BCI-III IVa	5	120×300	140	140
BCI-VI 2a	54	240×150	72	72
BCI-VI 2b	9	150×24	200	160

and

$$q = 1 - \frac{1}{\alpha_1} y_i \operatorname{tr}(\alpha_1 Z + V)^T L_i)$$
$$b = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \operatorname{tr}(W^T L_i)\big).$$

Finally, to compute Lagrange multipliers, differentiating (8), we get

$$\min_{L_i} h(W, b, L_i) + \alpha_1 ||L_i||_* - \operatorname{tr}\big(M_i^T L_i\big) + ||X_i - L_i - S_i||_F^2 \tag{19}$$

$$L_i = \mathbb{D}_\xi(y_i W + \alpha_1(X_i - S_i) + M_i)$$
$$\min S = \alpha_2 ||S_i||_{2,1} - \operatorname{tr}(M^T S_i) + \frac{\alpha_1}{2}||X_i - S_i) + M_i). \tag{20}$$

The above equation can be computed using column wise soft thresholding.

Now updating the Lagrange multipliers and coefficient

$$M^K = M^{k-1} + \alpha_2(X_i - L_i - S_i)$$
$$V^K = V^{k-1} + \alpha_1(Z - W) \tag{21}$$
$$\alpha_1 = p\alpha_1$$
$$\alpha_2 = p\alpha_2. \tag{22}$$

The above convex optimization cannot recover the matrix correctly. To overcome this challenge, we used an oracle problem that is defined by the structure we are interested in recovering. Thus, oracle-based convex optimization-based SMMRe algorithm is able to recover the corrupted columns correctly as well as can identify the outliers.

For the algorithm to succeed, it is sufficient for the recovered pair (L', S') to have the right column space and correct column of noncorrupted matrix L. Similarly, it requires right column support for sparse matrix S. To identify such a solution, we consider the oracle problem. α denotes the space of matrices supported on the set of all entries in the noncorrupted columns plus the observed entries in the corrupted columns. We are required to minimize $\min ||L||_* + ||S||_{2,1}$ subject to $\mathbb{P}_\alpha(L + C) = \mathbb{P}_\alpha X$, $\mathbb{P}(L) = L$ and $\mathbb{P}_I(S) = S$. Consider (L, S) is the solution for the oracle problem as we know it is feasible due to the feasibility of true pair (L', S'). Now, we must satisfy the conditions, (L', S'), as an optimal solution to Algorithm 1 and it must have correct column space and column support. Q is a dual certificate as long as it satisfies the following conditions: (I) $Q' \in \Omega$; (II) $\mathbb{P}_\alpha(Q') - UV^T = 0$; (III) $\mathbb{P}_\alpha(Q') < 1$; (IV) $|P_I(Q')|_{\infty,2}$; and (V) $P_I(Q') \in \lambda H$ s.t $H \in \mathbb{R}^{pq}|P_I(H) = 0)$. The next step is to consider any feasible perturbation, $(L' + \Delta_L, S' + \Delta_S)$. For a given Q', if it satisfies the above conditions, it shows that $(L' + \Delta_L, S' + \Delta_S)$ is suboptimal solution

$$\begin{aligned}\sum_{i=1}^{n}\xi + \sum_{i=1}^{n}(\alpha_1||L_i||_* + \alpha_2||S_i||_{2,1}) + \tau||W||_* &\leq \sum_{i=1}^{n}\xi\\ &+ \sum_{i=1}^{n}(\alpha_1||L_i + \Delta_L||_* + \alpha_2||S_i + \Delta_S||_{2,1}) + \tau||W||_*.\end{aligned} \tag{23}$$

The next step is the construction of dual certificate that satisfies the following conditions: (I) $Q' \in \Omega$; (II) $\mathbb{P}_\alpha(Q') - UV^T = \mathbb{P}_\alpha\mathbb{R}^{-1}(\mathbb{B})$ s.t. $\mathbb{B} = ((m/2pn)\lambda)^{1/2}$; (III) $\mathbb{P}_\alpha(Q') \leq 0.5$; (IV) $|P_I(Q')|_{\infty,2}$; and (V) $P_I(Q') \in (\lambda/H)$ s.t $H \in \mathbb{R}^{pq}|P_I(H) = 0)$. Ignoring the requirement of $Q' \in \Omega$ is a more manageable problem that allows to consider the fully observed problem of separating the low-rank matrix from a column-sparse matrix. The final step is the sampling i.e., compute Q from Q' that is performed by modified batched sampling-with replacement scheme [42].

B. Convergence of SMMRe

We consider the convergence of SMMRe algorithm described in Table I. Alternative direction method of minimization (ADMM) is an optimization algorithm that has recently become very popular due to its capabilities to solve large-scale and/or distributed problems. We have used ADMM optimizer to converge to an optimal primal-dual optimal solution. Equation (8) depicts the augmented Lagrangian $\mathcal{L}$ of (7), where μ_1 and μ_2 are fixed positive parameters. Furthermore, Table I is a step-by-step algorithm of the ADMM method depicted in (9) and (10).

To be specific, $\mathcal{L}$ is first minimized with respect to (W, Z, b) holding (L, S, V, M) fixed at $(L^{k-1}, S^{k-1}, V^{k-1}, M^{k-1})$, and then $\mathcal{L}$ is minimized with respect to (L, S) holding (W, Z, b, V, M) fixed at $(W^k, Z^k, b^k, V^{k-1}, M^{k-1})$.

TABLE III

CLASSIFICATION PERFORMANCE (ACCURACY) OF DIFFERENT ALGORITHMS ON DATASET BCI 2B

Sub.	BCI-Win	SVM [43]	SSVM [44]	RGLM [45]	LSVM [46]	TSVM [47]	SSM [40]	Rank KNN [4]	RSSM [48]	MSMM [49]	SMMRe [Proposed]
S1	0.60	0.68	0.73	0.69	0.69	0.68	0.68	0.74	0.73	0.74	0.736
S2	0.40	0.50	0.53	0.51	0.51	0.51	0.52	0.62	0.56	0.55	**0.64**
S3	0.21	0.52	0.54	0.53	0.51	0.53	0.53	0.61	0.56	0.56	**0.622**
S4	0.95	0.91	0.91	0.92	0.87	0.93	0.93	0.96	0.97	0.94	**0.975**
S5	0.86	0.8	0.83	0.82	0.80	0.84	0.83	0.90	0.88	0.87	**0.890**
S6	0.61	0.73	0.82	0.76	0.79	0.74	0.75	0.84	0.79	0.82	**0.861**
S7	0.56	0.69	0.76	0.75	0.72	0.71	0.72	0.76	0.78	0.77	**0.798**
S8	0.85	0.82	0.91	0.87	0.85	0.86	0.83	0.91	0.92	0.92	**0.932**
S9	0.74	0.74	0.84	0.77	0.78	0.76	0.76	0.84	0.83	0.86	**0.886**
Avg.	0.67	0.71	0.76	0.74	0.72	0.73	0.73	0.85	0.848	0.868	**0.878**

TABLE IV

COMPARATIVE EVALUATION BASED ON AVERAGE CLASSIFICATION ACCURACY ON BCI 2A

Motor Imagery	SVM [43]	SSVM [44]	RGLM [45]	LSVM [46]	TSVM [47]	SSM [40]	Rank KNN [4]	RSSM [48]	MSMM [49]	SMMRe [Proposed]
LvsR	0.80	0.82	0.82	0.81	0.81	0.81	0.84	0.83	0.83	**0.858**
LvsF	0.87	0.89	.89	0.88	0.89	0.88	0.92	0.90	0.90	**0.904**
LvsT	0.86	0.88	.88	0.88	0.88	0.88	0.93	0.91	0.90	**0.933**
RvsF	0.88	0.87	.87	0.88	0.89	0.89	0.90	0.89	**0.90**	0.896
RvsT	0.87	0.87	0.86	0.89	0.89	0.88	0.91	0.90	0.90	**0.912**
FvsT	0.80	0.80	0.81	0.80	0.81	0.81	0.88	0.84	0.84	**0.882**

TABLE V

COMPARATIVE EVALUATION BASED ON AVERAGE CLASSIFICATION ACCURACY ON BCI III–IVA

Subject	SVM [43]	SSVM [44]	RGLM [45]	LSVM [46]	TSVM [47]	SSM [40]	Rank KNN [4]	RSSM [48]	MSMM [49]	SMMRe [Proposed]
aa	0.73	0.74	0.71	0.72	0.75	0.74	0.78	0.76	0.77	**0.79**
a1	0.98	1	0.98	0.99	1	1	1	1	1	1
av	0.68	0.67	0.66	0.68	0.68	0.67	0.71	0.7	0.7	**0.72**
aw	0.7	0.75	0.71	0.71	0.72	0.74	0.82	**0.83**	0.81	0.82
ay	0.7	0.71	0.7	0.69	0.7	0.69	**0.79**	0.76	0.76	**0.78**
avg	0.76	0.77	0.75	0.76	0.77	0.76	0.80	0.81	0.81	**0.82**

Fig. 3. Sample images from Caltech face dataset (first two rows) and INRIA person dataset (last two rows). The human detection is challenging due to similar appearance of persons and human statues.

Note, (W^k, Z^k, b^k) are the new values. Finally, the dual variables V and M are updated in a gradient ascent way. Consequently, a proof can be established in a way similar to the one detailed in Mota et al. [43]. This shows that all the sequences produced by ADMM converge.

C. Classification

Since the test data are rich in outliers, thus, we also need to consider the noise reduction to be classified by SMMRe. For a given set of input test data $[X_t]_{t=1}^n$, X_t can be decomposed into low-rank matrix and sparse noise by optimizing the following:

$$\arg\min_{[L_t, S_t]_{t=1}^n} ||L_t||_* + \gamma ||S_t||_{2,1} \quad \text{s.t.} \quad X_t = L_t + S_t \tag{24}$$

where L_t and S_t is the low rank and sparse matrix, respectively. γ is the positive scalar that adds penalty for sparse noise. We can notice that equation (24) is similar to RPCA [38]. Once the noisy input matrix is decomposed into noisy and clean, we have used only clean matrix for testing. We can predict the label using learned parameters as

$$Y_t = \text{sgn}(\text{tr}(W^T L_t) + b). \tag{25}$$

VII. DATASET

We evaluated the proposed approach on the most fundamental applications of classification. We have applied SMMRe on important datasets (Caltech face dataset and INRIA dataset) and BCI competition (III–IVa and BCI IV–IIa).

A. Caltech Face Dataset

It is gender recognition dataset of 435 individuals that consist of images containing various facial expressions of size 592 × 896 captured under different illumination conditions and backgrounds shown in Fig. 3. We have divided the dataset into training dataset (147 male and 71 female) and test dataset (131 male and 86 female). Images are converted to gray scale and the face in the image has been cropped using Viola-Jones face detector. We have re-sized the face to 320 × 280 and used the pixel values as an input matrix without any advanced feature extraction techniques. Fig. 3 shows sample images of Caltech face dataset. Notice that, the images share similar features in terms of face outlines and structure; however, gender can be differentiated from small detail such as persons' eyes, hair, etc.

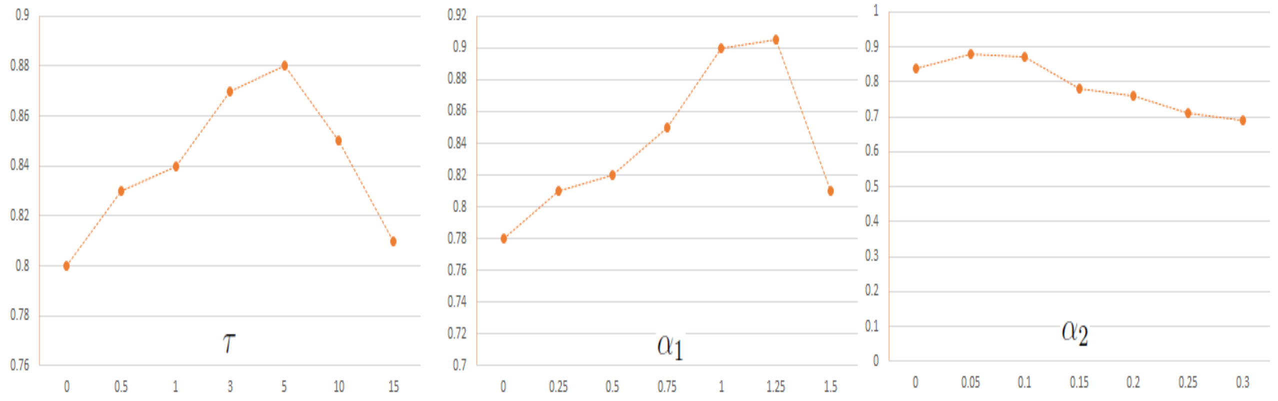

Fig. 4. Effect of different parameters (τ, α_1, and α_2) values.

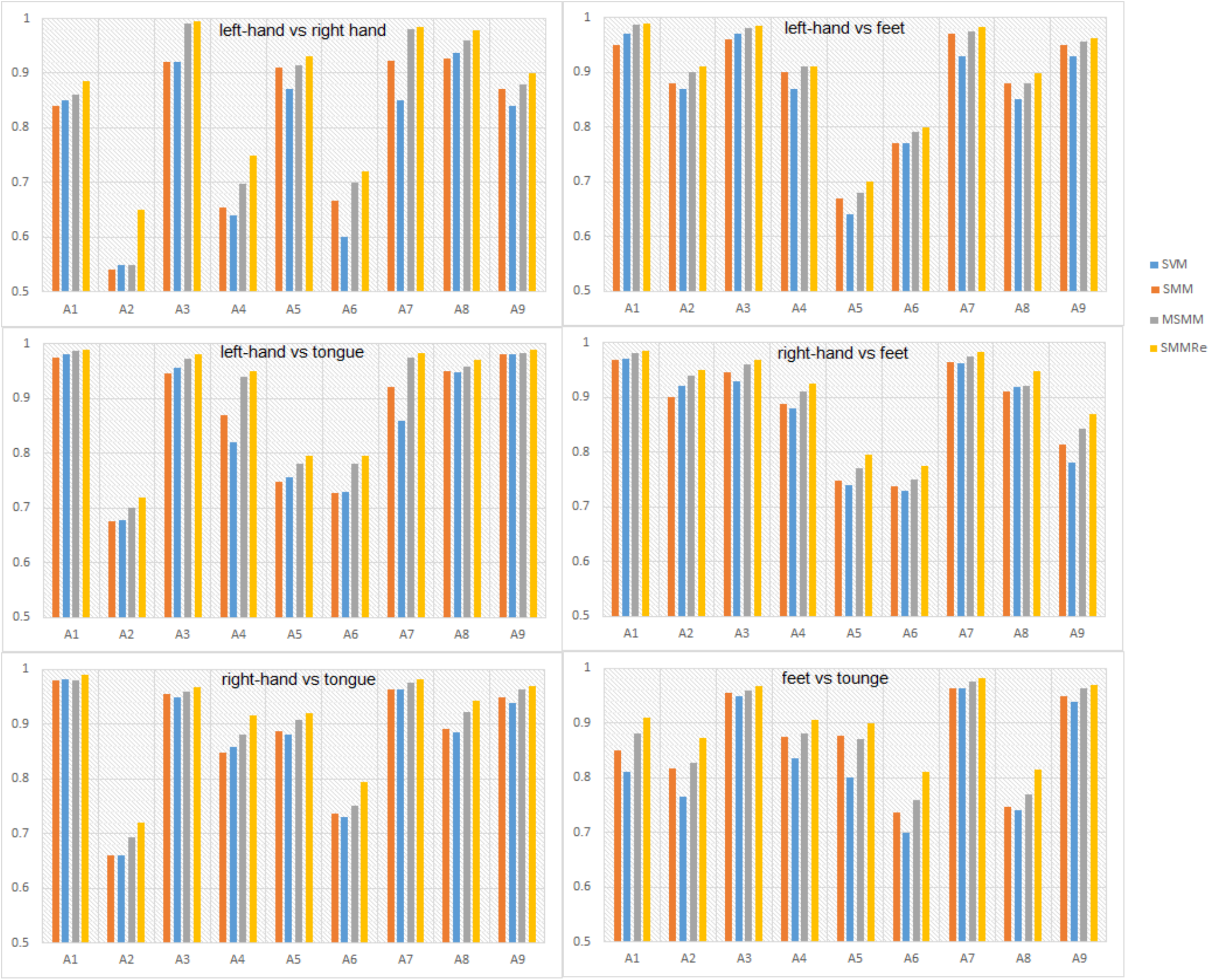

Fig. 5. Comparative evaluation of SVM, SMM, MSMM, and SMMRe on IVa: top left to bottom right (left-hand versus right hand, left-hand versus feet, left-hand versus tongue, right-hand versus feet, right-hand versus tongue, feet versus tounge).

B. INRIA Person Dataset

It is collected to detect the existence of a person in an image or video. INRIA person dataset is divided in two formats: original images with corresponding annotation files and positive images in normalized 64 × 128 pixel format. It consists of 2416 images with people and 1218 people-free images for training, and 1126 images with people and 453 people-free samples for testing. Person detection is a challenging task due to similar background and arbitrary appearance of human in the image. Fig. 3 shows sample image of dataset. In this experiment, we have converted each image into gray scale with dimensions (160 × 96). For person detection, we have used gray-scale image as it is without feature extraction to show the structural correlation of pixels; thus, we have converted the input image into gray level of size 160 × 96.

C. BCI Competition

We further evaluate the SMMRe on the application of electroencephalogram (EEG) data classification. EEG signals consist of 2-D matrices that have high correlation among the rows and columns within each sample, which could be effectively captured by matrix classification methods [41]. In this experiment, three EEG data observations from BCI competition-IV, namely BCI III–IVa[1], BCI IV–IIa[2], and BCI IV–IIb[3], are used to evaluate the performance of proposed approach. Table II describes the detail of the datasets. Both datasets contain a small number of samples with redundant data, a property that makes EEG classification challenging.

VIII. Parameter Selection

There are three key parameter that need to be carefully selected for optimal performance. We consider the influence of parameters ($\tau, \alpha_1,$ and α_2) on SMMRe performance. τ is the penalty applied to the nuclear norm of the regression matrix that controls sparseness. α_1 is the penalty term on the nuclear norm that controls the recovery process. α_2 is the penalty on the $\ell_{2,1}$ norm to overcome the affect of outliers in the feature matrix and, as a result, it helps extract robust features from the cleaned matrix. To select the optimal range of parameter, we first fix two parameters α_1 and α_2 and performed several experiment to find optimal range of τ for each dataset. Once we found the optimal range of τ, we selected the optimal range of α_1 and α_2. We observe that the objective function degenerates to a traditional support matrix machine for $\tau, \alpha_1, \alpha_2 = 0$ that show that SMMRe is a special case of support matrix machines. Similarly, fixing $\alpha_1 = 0$ degenerates the model to SMM. To study the influence of the parameter, we fix α_1 and α_2 and find the best optimum value of τ to control sparseness. Once we have a sparseness control, we repeated the process for the other two terms. Figs. 6, 8, and 9 shows the effect of different parameter setting of τ, α_1, and α_2. Fig. 4 shows the behavior of proposed on different value of parameters τ, α_1, and α_2 on subject S9 of BCI 2b dataset.

IX. Result and Discussion

To evaluate the SMMRe performance, we compare it to the state-of-the-art vector-based methods such as SVM [44], sparse SVM (SSVM) [45], LSVM [47], BSVM [51], TSVM [48] as well as with the state-of-the-art matrix-based classifiers, i.e., Rank KNN [4], SSMM [41], RSMM [49], SMM [40], MSMM [50], and regularized matrix regression (RGLM) [46] on benchmark face recognition, person identification, and EEG datasets.

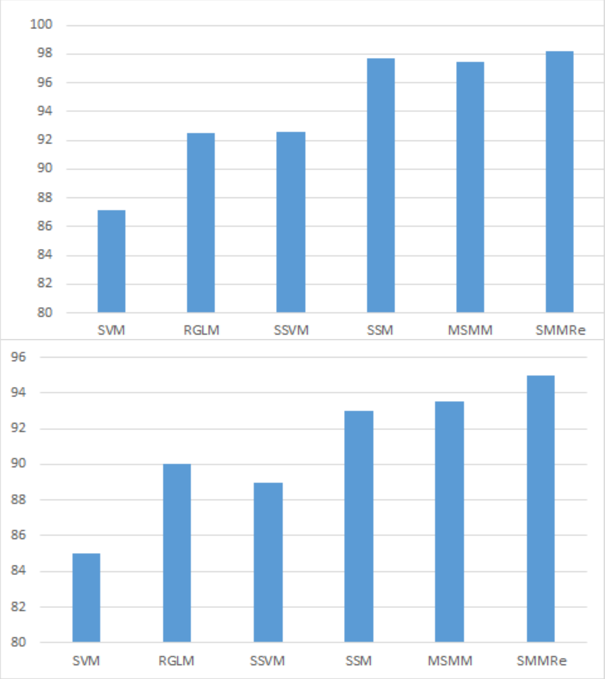

Fig. 6. Comparative evaluation (accuracy) based on average classification accuracy on real (top) corrupted (bottom) INRIA person dataset.

Surprisingly, we can simultaneously perform matrix recovery, low-rank feature extraction, identification of noncorrupted columns and their position and classification based on set of fraction of observed entries. Fig. 9 shows the effect of different parameters values on the classification. Tables III–V and Fig. 5 show the classification results on EEG datasets (BCI 2a, BCI 2b, and IVa). From Fig. 5, we can notice that SMMRe considerably performed better against challenging conditions (A2 and A5 in left-hand versus right-hand, A2, A5, and A6 in left-hand versus tongue, A5, A6 in left-hand versus feet and right-hand versus feet) in comparison to others.

Results showed that support matrix machines based on matrix recovery outperform the state-of-the-art methods. Similar results can be noticed in Figs. 6 and 7 for person identification on INRIA and Caltech face datasets, respectively. Furthermore, we can observe that classifiers based on the matrix data provided better results as compared to those methods based on data as a vector, which shows that vector-based methods ignore the structural information thus, they are very sensitive to the curse of dimensionality. However, matrix-based approaches leverage the structural information of the data which is greatly beneficial to the improvement of the classification performance. The other main reason is the low-rank property as discriminant features exist in sparse structure and images are low rank.

In comparison to matrix-based methods, SMMRe outperforms both sparse (i.e., SSVM) and low-rank methods (i.e., BSVM, SMM, and SSMM) which validate the claim that SMMRe promotes the structural sparsity and shares similar sparsity patterns across multiple predictors. To further validate

[1] http://www.bbci.de/competition/III/#download
[2] http://www.bbci.de/competition/iv/#dataset2a
[3] http://www.bbci.de/competition/iv/#dataset2b

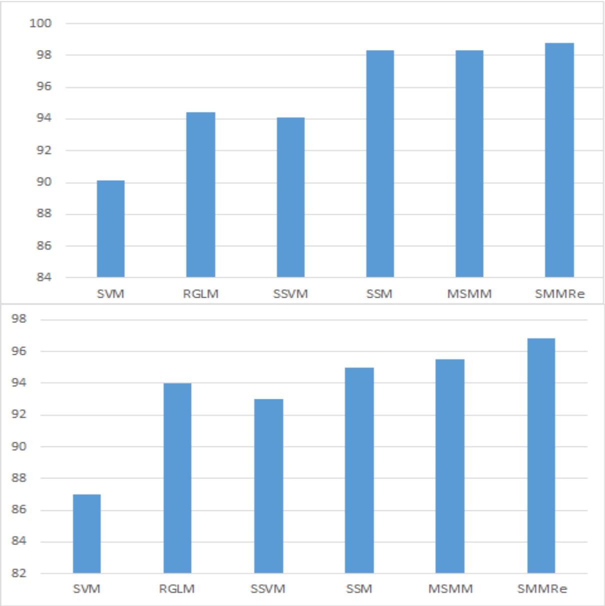

Fig. 7. Comparative evaluation (accuracy) based on average classification accuracy on real (top) contaminated (bottom) Caltech face dataset.

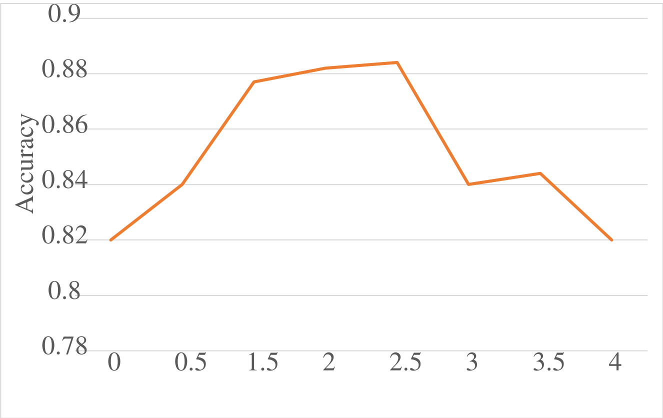

Fig. 8. Effect of different parameters τ.

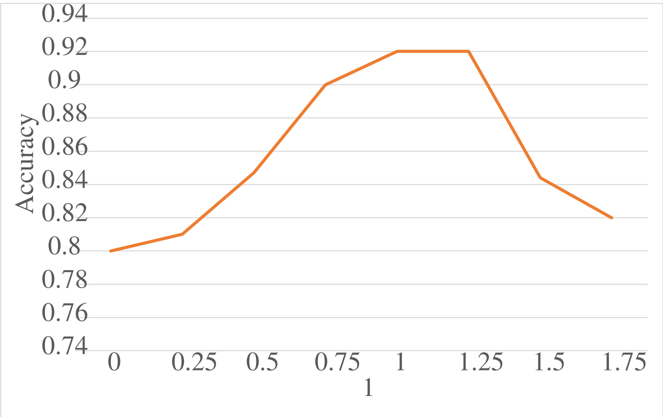

Fig. 9. Effect of different parameters α_1 values.

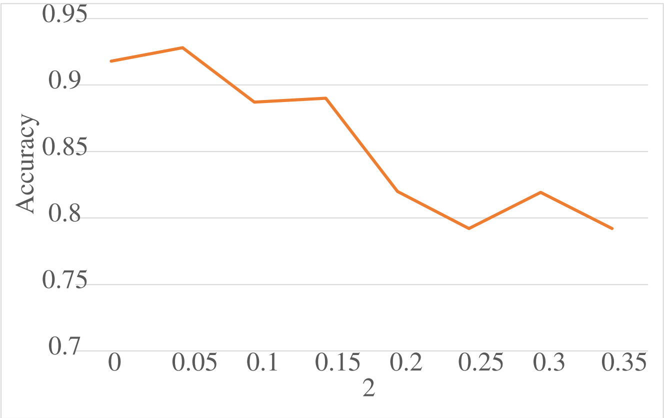

Fig. 10. Effect of different parameters α_2 values.

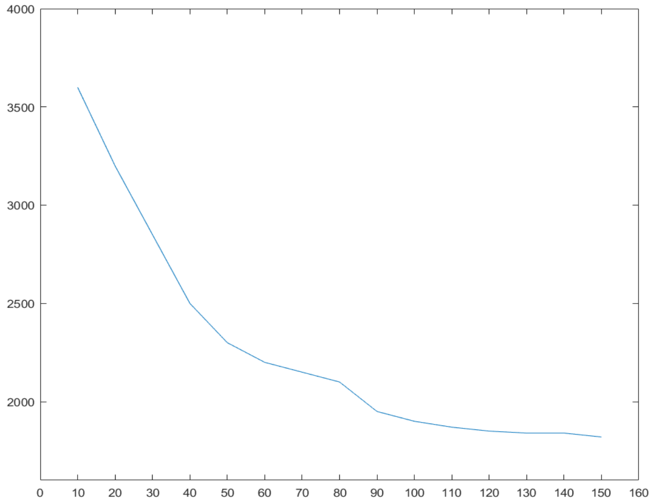

Fig. 11. Convergence curve of SMMRe [objective function value (y-axis) versus iteration (x-axis)].

the robustness against outliers, we have contaminated both Caltech face dataset and INRIA dataset with random noise, specifically we have randomly selected 20% images to add noise in each dataset. We corrupted both datasets via the addition of random noise as well as block occlusions. Random noise is salt and pepper noise spread randomly at 30%, 50% on random selection of images from dataset. Similarly, block occlusion is added by placing blocks of different sizes at random locations with variable size 5 × 5, 10 × 10, and 10 × 15. For evaluation on contaminated datasets, we have selected 60% and 70% and 80% samples per individual for each dataset as training dataset and add blocks of variable sizes. Figs. 6(b) and 7(b) show the comparative evaluation on INRIA and Caltech face dataset, respectively. Note that SMMRe considerably performed better against outliers or challenging conditions in comparison to others. This is due to the matrix recovery through identification of noncorrupted columns, low-rank robust feature extraction, and classification. It shows that SMMRe is robust even from partially observed matrix which validate our claim that SMMRe is able to classify data with denser corruptions through exact recovery of intrinsic matrix of higher rank based on the incoherence conditions.

We also consider the influence of parameters (τ, α_1, and α_2) on the performance of SMMRe. τ is the penalty on nuclear norm of regression matrix that controls the sparseness. α_1 is the penalty term on nuclear norm that controls the recovery process. α_2 is the penalty on $\ell_{2,1}$ norm to overcome the affect of outliers in feature matrix; as a result, it helps to extract robust features from cleaned matrix. The objective function degenerates to traditional support matrix machine for $\tau, \alpha_1, \alpha_2 = 0$, that shows that SMMRe is the special case of support matrix machines. Similarly, fixing $\alpha_1 = 0$, degenerate the model to SMM. To study the influence of parameter, we fix α_1 and α_2 and find the best optimum value of τ to control the sparseness. Once we have sparseness control, we repeated the process for other two terms. Figs. 8–10 show the effect of different parameter setting of τ, α_1, and α_2. Fig. 11, illustrate the convergence of SMMRe.

X. Conclusion

In this article, we have integrated matrix recovery and support matrix machines for the classification of dense corrupted data. The method proposed—SMMRe—is simultaneously able to perform matrix recovery, low-rank feature representation, and classification, and thus able to classify data with denser corruptions through the exact recovery of the intrinsic matrix of higher rank based on incoherence conditions. The regularization term promotes low-rank matrix recovery and structural sparsity as well as sharing a similar sparsity pattern across multiple predictors. Furthermore, it also leverages structural information and avoids the inevitable upper-bound that simultaneously promotes a good fit to the data. A comprehensive experimental study on four publicly available datasets for image and EEG classification was carried out to validate the proposed SMMRe approach. The experimental results showed the gain in performance in most cases with an average increase to 0.878 from 0.864, 0.906 from 0.893, and 0.82 from 0.81 for BCI 2b, BCI 2a, and BCI III–IVa datasets, respectively. Similar trends can be noticed for person identification and face recognition task. This shows the effectiveness of the SMMRe approach for solving classification problems, even when a large fraction of columns is corrupted, while keeping reasonable the number of support vectors. Our observation showed that proposed approach is robust against outliers, it has the property of low rank and joint sparsity to select features across all the classes. The proposed method is not restricted to recovery of low-rank matrices. In future, we plan to explore different incoherence and ambiguity conditions for highly noisy data.

References

[1] Y. Zhang et al., "Multi-kernel extreme learning machine for EEG classification in brain-computer interfaces," *Expert Syst. Appl.*, vol. 96, pp. 302–310, Apr. 2018.

[2] W.-C. Hsu, L.-F. Lin, C.-W. Chou, Y.-T. Hsiao, and Y.-H. Liu, "EEG classification of imaginary lower limb stepping movements based on fuzzy support vector machine with kernel-induced membership function," *Int. J. Fuzzy Syst.*, vol. 19, no. 2, pp. 566–579, Apr. 2017.

[3] D. S. de Lucena, S. R. Moreno, V. C. Mariani, and L. D. S. Coelho, "Support vector machine optimized by artificial bee colony applied to EEG pattern recognition," in *Proc. 1st Int. Conf. Brain Function Assessment Learn.*, vol. 10512. Patras, Greece: Springer, Sep. 2017, p. 213.

[4] K. Song, F. Nie, J. Han, and X. Li, "Rank-*k* 2-D multinomial logistic regression for matrix data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3524–3537, Aug. 2018.

[5] K. Song, F. Nie, and J. Han, "Two dimensional large margin nearest neighbor for matrix classification," in *Proc. IJCAI*, Aug. 2017, pp. 2751–2757.

[6] I. Razzak, R. A. Saris, M. Blumenstein, and G. Xu, "Integrating joint feature selection into subspace learning: A formulation of 2DPCA for outliers robust feature selection," *Neural Netw.*, vol. 121, pp. 441–451, Jan. 2020.

[7] F. Qi, Y. Li, and W. Wu, "RSTFC: A novel algorithm for spatio-temporal filtering and classification of single-trial EEG," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 12, pp. 3070–3082, Dec. 2015.

[8] A. Subasi and M. Ismail Gursoy, "EEG signal classification using PCA, ICA, LDA and support vector machines," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8659–8666, Dec. 2010.

[9] Y. Zhang, Y. Wang, J. Jin, and X. Wang, "Sparse Bayesian learning for obtaining sparsity of EEG frequency bands based feature vectors in motor imagery classification," *Int. J. Neural Syst.*, vol. 27, no. 2, Mar. 2017, Art. no. 1650032.

[10] C. Yan et al., "Self-weighted robust LDA for multiclass classification with edge classes," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 1, pp. 1–19, Feb. 2021.

[11] A. Datta and R. Chatterjee, "Comparative study of different ensemble compositions in EEG signal classification problem," in *Emerging Technologies in Data Mining and Information Security*. Singapore: Springer, 2019, pp. 145–154.

[12] M. Li, H. Xu, X. Liu, and S. Lu, "Emotion recognition from multichannel EEG signals using K-nearest neighbor classification," *Technol. Health Care*, vol. 26, no. S1, pp. 509–519, 2018.

[13] T. Nezam, R. Boostani, V. Abootalebi, and K. Rastegar, "A novel classification strategy to distinguish five levels of pain using the EEG signal features," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 131–140, Jan. 2021.

[14] S. Lemm, B. Blankertz, G. Curio, and K.-R. M´uller, "Spatio-spectral filters for improving the classification of single trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 9, pp. 1541–1548, Sep. 2005.

[15] R. Tomioka, G. Dornhege, G. Nolte, B. Blankertz, K. Aihara, and K.-R. Müller, "Spectrally weighted common spatial pattern algorithm for single trial EEG classification," Dept. Math. Eng., Univ. Tokyo, Tokyo, Japan, Tech. Rep. 40, 2006.

[16] X. Liao, D. Yao, D. Wu, and C. Li, "Combining spatial filters for the classification of single-trial EEG in a finger movement task," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 5, pp. 821–831, May 2007.

[17] H. Zhang, H. Yang, and C. Guan, "Bayesian learning for spatial filtering in an EEG-based brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1049–1060, Jul. 2013.

[18] T.-E. Kam, H.-I. Suk, and S.-W. Lee, "Non-homogeneous spatial filter optimization for ElectroEncephaloGram (EEG)-based motor imagery classification," *Neurocomputing*, vol. 108, pp. 58–68, May 2013.

[19] A. S. Aghaei, M. S. Mahanta, and K. N. Plataniotis, "Separable common spatio-spectral patterns for motor imagery BCI systems," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 1, pp. 15–29, Jan. 2016.

[20] L. Wolf, H. Jhuang, and T. Hazan, "Modeling appearances with low-rank SVM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–6.

[21] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Bilinear classifiers for visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1482–1490.

[22] L. Lan, Z. Wang, S. Zhe, W. Cheng, J. Wang, and K. Zhang, "Scaling up kernel SVM on limited resources: A low-rank linearization approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 369–378, Feb. 2019.

[23] H. Lian and Z. Fan, "Divide-and-conquer for debiased l_1-norm support vector machine in ultra-high dimensions," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6691–6716, 2018.

[24] J. Spilka, J. Frecon, R. Leonarduzzi, N. Pustelnik, P. Abry, and M. Doret, "Sparse support vector machine for intrapartum fetal heart rate classification," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 3, pp. 664–671, May 2017.

[25] Y.-H. Shao, C.-N. Li, M.-Z. Liu, Z. Wang, and N.-Y. Deng, "Sparse L_q-norm least squares support vector machine with feature selection," *Pattern Recognit.*, vol. 78, pp. 167–181, Jun. 2018.

[26] T. Pang, F. Nie, J. Han, and X. Li, "Efficient feature selection via $\ell_{2,0}$-norm constrained sparse regression," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 5, pp. 880–893, Jun. 2018.

[27] X. Cai, F. Nie, H. Huang, and C. Ding, "Multi-class $\ell_{2,1}$-norm support vector machine," in *Proc. IEEE 11th Int. Conf. Data Mining*, Dec. 2011, pp. 91–100.

[28] X. Chang and Y. Yang, "Semisupervised feature analysis by mining correlations among multiple tasks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2294–2305, Oct. 2017.

[29] T. Kobayashi and N. Otsu, "Efficient optimization for low-rank integrated bilinear classifiers," in *Computer Vision–ECCV*. Berlin, Germany: Springer, 2012, pp. 474–487.

[30] Q. Zheng, F. Zhu, J. Qin, and P.-A. Heng, "Multiclass support matrix machine for single trial EEG classification," *Neurocomputing*, vol. 275, pp. 869–880, Jan. 2018.

[31] Z. Zhai, B. Gu, X. Li, and H. Huang, "Safe sample screening for robust support vector machine," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 4, 2020, pp. 6981–6988.

[32] A. Rosales-Perez, S. Garcia, H. Terashima-Marin, C. A. Coello Coello, and F. Herrera, "MC^{2}ESVM: Multiclass classification based on cooperative evolution of support vector machines," *IEEE Comput. Intell. Mag.*, vol. 13, no. 2, pp. 18–29, May 2018.

[33] A. P. Castaño, "Support vector machines," in *Practical Artificial Intelligence*. U.K.: Springer, 2018, pp. 315–365.

[34] A. Rosales-Pérez, A. E. Gutierrez-Rodríguez, S. García, H. Terashima-Marín, C. A. C. Coello, and F. Herrera, "Cooperative multi-objective evolutionary support vector machines for multiclass problems," in *Proc. Genetic Evol. Comput. Conf.* Stroudsburg, PA, USA: ACM, Jul. 2018, pp. 513–520.

[35] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, Jan. 2010.

[36] T. Oh, Y. Matsushita, Y. Tai, and I. S. Kweon, "Fast randomized singular value thresholding for low-rank optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 376–391, Feb. 2018.

[37] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Dec. 2009.

[38] H. Zhang, Z. Lin, C. Zhang, and E. Y. Chang, "Exact recoverability of robust pca via outlier pursuit with tight recovery bounds," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3143–3149.

[39] A. F. Martins, N. A. Smith, P. M. Aguiar, and M. A. Figueiredo, "Structured sparsity in structured prediction," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA: ACM, 2011, pp. 1500–1511.

[40] L. Luo, Y. Xie, Z. Zhang, and W.-J. Li, "Support matrix machines," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn. (ICML)*, vol. 37, Jul. 2015, pp. 938–947. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045118.3045219

[41] Q. Zheng, F. Zhu, J. Qin, B. Chen, and P.-A. Heng, "Sparse support matrix machine," *Pattern Recognit.*, vol. 76, pp. 715–726, Apr. 2018.

[42] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, Jan. 2010.

[43] J. F. Mota, J. M. Xavier, P. M. Aguiar, and M. Puschel, "Distributed basis pursuit," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1942–1956, 2011.

[44] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[45] J. Zhu, S. Rosset, R. Tibshirani, and T. J. Hastie, "1-norm support vector machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 49–56.

[46] H. Zhou and L. Li, "Regularized matrix regression," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 76, no. 2, pp. 463–483, Mar. 2014.

[47] O. L. Mangasarian and D. R. Musicant, "Lagrangian support vector machines," *J. Mach. Learn. Res.*, vol. 1, pp. 161–177, Sep. 2001.

[48] B. Richhariya and M. Tanveer, "A reduced universum twin support vector machine for class imbalance learning," *Pattern Recognit.*, vol. 102, Jun. 2020, Art. no. 107150.

[49] Q. Zheng, F. Zhu, and P. Heng, "Robust support matrix machine for single trial EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 551–562, Mar. 2018.

[50] I. Razzak, M. Blumenstein, and G. Xu, "Multiclass support matrix machines by maximizing the inter-class margin for single trial EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1117–1127, Jun. 2019.

[51] T. Joachims, T. Finley, and C.-N.-J. Yu, "Cutting-plane training of structural SVMs," *Mach. Learn.*, vol. 77, no. 1, pp. 27–59, Oct. 2009.