A NON-CONVEX PROXIMAL APPROACH FOR CENTROID-BASED CLASSIFICATION

Mewe-Hezoudah Kahanam[†] *Laurent Le-Brusquet*[⋆] *Ségolène Martin*[†] *Jean-Christophe Pesquet*[†]

[†] Université Paris-Saclay, Inria, CentraleSupélec, Centre de Vision Numérique
[⋆] Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes

ABSTRACT

In this paper, we propose a novel variational approach for supervised classification based on transform learning. Our approach consists of formulating an optimization problem on both the transform matrix and the centroids of the classes in a low-dimensional transformed space. The loss function is based on the distance to the centroids, which can be chosen in a flexible manner. To avoid trivial solutions or highly correlated clusters, our model incorporates a penalty term on the centroids, which encourages them to be separated. The resulting non-convex and non-smooth minimization problem is then solved by a primal-dual alternating minimization strategy. We assess the performance of our method on a bunch of supervised classification problems and compare it to state-of-the-art methods.

Index Terms— Supervised classification, centroid-based classification, non-convex optimization, transform learning

1. INTRODUCTION

Given labeled multidimensional observations split into k classes of data having similar characteristics, the problem of supervised classification consists in learning a classifier which associates each data with a cluster. However, classification of high-dimensional data remains limited by both computational issues and the predictive power of traditional methods, vectors tending to become indiscernible as the dimension grows [1, 2]. When facing such situations, a common strategy is to map those data to a lower-dimensional subspace without losing essential information regarding the discriminative characteristics of the original variables. A popular approach is to perform a Principal Component Analysis (PCA) prior to classification [3]. However, this approach is not always relevant in practice since it only accounts for the second-order correlations within the data and does not take into account all the available information [4]. A generalization of the PCA approach consists in searching for a low-dimensional transform that enjoys some optimality properties. For instance, one may search for a matrix that maximizes the separation between classes in the transformed space, such as in *Linear Discriminant Analysis* [5]. Ideally, the clusters in the transformed space should be as far as possible from each other, and the data points compactly distributed within each cluster.

A simple way to tackle the supervised classification problem is to formulate it as a least squares optimization problem on the transform matrix W [6, 7]:

$$\underset{W}{\text{minimize}}\ \|Y - XW\|_{\mathrm{F}}^2, \tag{1}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, $X \in \mathbb{R}^{m\times d}$ is a matrix whose rows correspond to the m samples belonging to the feature space $\mathbb{R}^d$, and $Y \in \{0,1\}^{m\times k}$ denote the matrix of one-hot encoded labels, where $k \geqslant 2$ is the number of classes. Each row vector $(Y_i)_{1\leq i\leq m}$ of Y contains only one non-zero coefficient, which indicates the cluster of the sample X_i. Many authors have proposed to add an ℓ_1 norm sparse regularizer on the transformation matrix W in Problem (1), therefore retrieving the standard LASSO loss function [8, 9, 10]. However, in [11], it was observed that using the Frobenius norm in (1) makes the approach sensitive to outliers and it was suggested to replace it by an $\ell_{1,2}$ norm. In the same spirit, an ℓ_1 data-fit function was used in [12], leading to the minimization of the sum of two non-smooth convex functions.

A more flexible criterion proposed in [13] and inspired from the k-means algorithm, optimizes jointly the transform matrix W and the centroid matrix M through a data-fidelity term of the general form $(W, M) \mapsto f(YM - XW)$, where f is a function defined on $\mathbb{R}^{m\times\ell}$. Yet, as highlighted by the authors, such an approach requires adding adequate regularization and penalty terms to break separability and avoid the trivial undesirable solution $(M, W) = (0, 0)$.

In this paper, we present a new transform-based variational approach for supervised classification. Our contribution is twofold. The first one is to propose a suitable criterion on both the transform matrix and the centroids. Our formulation differs from the one in [13] as it adds a term to the loss favoring centroid separation. Although the introduction of this term makes the approach more effective, it leads to a loss of the convexity property, thus making the optimization problem more challenging. Secondly, we propose an alternating proximal minimization algorithm to solve this problem. Sub-iterations for computing the involved proximity operator are performed using a primal-dual formulation.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

The paper is organized as follows. Section 2 presents the centroid-based approach for classification and the difficulties it raises. In Section 3 we formulate a novel criterion for separating the centroids in the transformed space. The algorithm we implement to address this problem is described in Section 4. Finally, comparisons in terms of accuracy of our approach with state-of-the-art methods are conducted in Section 5, before drawing some conclusions in Section 6.

2. CENTROID-BASED CLASSIFICATION

Let $(X_i)_{1\leq i\leq m}$ be vectors in $\mathbb{R}^d$ containing the rows of data matrix $X\in\mathbb{R}^{m\times d}$, corresponding to data samples. Centroid based classification consists in mapping X to a lower dimensional space $\mathbb{R}^\ell$ using a linear transform $W\in\mathbb{R}^{d\times\ell}$. We also introduce k *centroid* vectors, each of which can be thought of as the center of the samples $W^\top X_i$ belonging to the same cluster. The matrix of centroids $(M_j)_{1\leq j\leq k}$ is denoted by $M=[M_1,\dots M_k]^\top\in\mathbb{R}^{k\times\ell}$.

The centroid-based classification approach amounts to minimizing the following loss function with respect to the matrix variables W and M:

$$(\forall M\in\mathbb{R}^{k\times\ell})\ (\forall W\in\mathbb{R}^{d\times\ell})\quad \mathcal{L}(M,W)=f(YM-XW)+g(W)+h(M),\quad (2)$$

where $f\colon\mathbb{R}^{m\times\ell}\longrightarrow]-\infty,+\infty]$, $g\colon\mathbb{R}^{d\times\ell}\longrightarrow]-\infty,+\infty]$ and $h\colon\mathbb{R}^{k\times\ell}\longrightarrow]-\infty,+\infty]$ satisfy the following assumption:

Assumption 1. *Functions f, g, and h are convex, proper, lower semi-continuous. Moreover, f is row-wise separable, i.e.*

$$(\forall Z=[Z_1,\dots,Z_m]^\top\in\mathbb{R}^{m\times\ell})\quad f(Z)=\sum_{i=1}^m\varphi(Z_i),\quad (3)$$

for some function $\varphi\colon\mathbb{R}^\ell\longrightarrow]-\infty,+\infty]$.

In formulation (2), f corresponds to the data-fit function which plays a crucial role in learning from the training set (X,Y), while g (resp. h) is a regularization function on W (resp. M).

Example 1.

- A classical choice for f corresponds to $f=\|\cdot\|_{\mathrm{F}}^2$ [14, 15]. Note that, in this case, the data-fit term in (2) is reminiscent of the standard k-means approach since

$$f(YM-XW)=\sum_{j=1}^k\sum_{i\in C_j}\|M_j-W^\top X_i\|_2^2,\quad (4)$$

  where $(C_j)_{1\leq j\leq k}$ denote the classes. An alternative choice proposed in [12] is $f=\|\cdot\|_1$ and was shown to be more robust to outliers.

- A sparsity-promoting regularization is often employed for W [16, 17, 18], corresponding to $g=\alpha\|\cdot\|_1$, where $\alpha>0$. It may also be convenient to use an elastic net regularisation $g=\alpha\|\cdot\|_1+\frac{\beta}{2}\|\cdot\|_{\mathrm{F}}^2$ with $\beta>0$ to guarantee the existence of a solution to the minimization problem.

- Choices for function h are discussed in the rest of the paper.

Given the transform matrix W and the centroid matrix M obtained by minimizing (2), a new sample X_{m+1} can be assigned to a class j^*, where j^* satisfies

$$j^*\in\underset{j\in\{1,\dots,k\}}{\mathrm{Argmin}}\ \varphi(M_j-W^\top X_{m+1}).\quad (5)$$

(If such a j^* is not unique, one arbitrarily assigns to X_{m+1} the class corresponding to the smallest j^* satisfying (5).)

While model (2) has the great advantage of leading to a convex optimization problem, one can observe that the trivial solution $(M,W)=(0,0)$ is obtained when f, g are chosen as in Example 1 and h is a nonnegative function vanishing at 0. Therefore, function h should be cleverly chosen so to avoid this trivial solution. For instance, Barlaud et al. [13] circumvented this issue by introducing a regularization h of the form

$$(\forall M\in\mathbb{R}^{k\times k})\quad h(M)=\frac{\rho}{2}\|M-\mathrm{I}_k\|_{\mathrm{F}}^2,\quad (6)$$

where $\rho>0$ is a multiplicative constant. Such a choice for function h penalizes the alignment of the centroids. Although it avoids retrieving the trivial solution, it forces the data to be mapped to vectors of the same dimension $\ell=k$ as the number of classes, which may not be a suitable choice to fully separate the clusters. In the following, we propose a new expression for h which favors the separation of the clusters in a transformed space of arbitrary dimension $\ell\leq m$.

3. A NON-CONVEX FORMULATION TO CENTROID-BASED CLASSIFICATION

3.1. Separation of the centroids

We propose to resort to the following modified loss function:

$$(\forall M\in\mathbb{R}^{k\times\ell})\ (\forall W\in\mathbb{R}^{d\times\ell})\quad \mathcal{L}(M,W)=f(YM-XW)+g(W)-\gamma h(M),\quad (7)$$

where f, g and h satisfy Assumption 1, and $\gamma>0$ is a penalty parameter. We opt for a particular choice of function h which pushes the centroids to be distant from each others, namely

$$(\forall M\in\mathbb{R}^{k\times\ell})\quad h(M)=\sum_{1\leqslant i<j\leqslant k}\|M_j-M_i\|_1.\quad (8)$$

Note that for this choice of function h, Model (7) is nonconvex. In order to handle this term more easily, it is reexpressed as

$$(\forall M \in \mathbb{R}^{k\times\ell}) \quad h(M) = \|AM\|_1, \tag{9}$$

where A is a matrix of dimension $q\times k$, where $q = \ell(\ell-1)/2$.

3.2. Bounding the problem

The loss function defined in (7) may have no minimizer. Let us show this, in the particular case when f, g and h are homogeneous functions with the same scaling factor, in the sense that there exists a function $a\colon \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ such that $a(0) = 0$, $\lim_{\eta\to+\infty} a(\eta) = +\infty$ and, for every $\eta \in \mathbb{R}_+$,

$$\begin{cases} (\forall Z \in \mathbb{R}^{m\times\ell}) & f(\eta Z) = a(\eta)f(Z), \\ (\forall W \in \mathbb{R}^{d\times\ell}) & g(\eta W) = a(\eta)g(W), \\ (\forall M \in \mathbb{R}^{k\times\ell}) & h(\eta M) = a(\eta)h(M). \end{cases} \tag{10}$$

Then, if there exists a pair (M, W) such that

$$f(MY - WX) + g(M) - \gamma h(W) < 0, \tag{11}$$

Criterion (7) is unbounded from below. Otherwise, $(M, W) = (0, 0)$ is a trivial solution.

These considerations lead us to bound the centroid matrix M. To do so, we constrain each of the centroids $(M_j)_{1\le j\le k}$ to lie in a closed ball of radius $\delta > 0$. Taking into account this constraint on the centroids, the minimization problem to be solved reads

$$\begin{aligned} \underset{(M,W)}{\text{minimize}} \quad & f(YM - XW) + g(W) \\ & - \gamma h(M) + \iota_C(M), \end{aligned} \tag{12}$$

where

$$C = \left\{ M \in \mathbb{R}^{k\times\ell} \mid (\forall j \in \{1,\ldots,k\}) \quad \|M_j\|_2 \le \delta \right\}, \tag{13}$$

and ι_C denotes the indicator function of C.[1]

3.3. An equivalent formulation

The objective function in (12) is convex with respect to W, but it is non-convex with respect to M. This last issue can be overcome by rewriting the ℓ_1-norm through its dual norm, i.e.

$$(\forall M \in \mathbb{R}^{k\times\ell}) \quad h(M) = \max_{U\in\mathcal{B}_\infty} \langle AM, U\rangle, \tag{14}$$

where $\mathcal{B}_\infty$ is the ℓ_∞-unit ball of $\mathbb{R}^{q\times k}$, and $\langle\cdot,\cdot\rangle$ denotes the Frobenius inner product. Therefore, Problem (12) is equivalent to

$$\begin{aligned} \underset{(M,W,U)}{\text{minimize}} \quad & f(YM - XW) + g(W) \\ & - \gamma\langle AM, U\rangle + \iota_C(M) + \iota_{\mathcal{B}^\infty}(U). \end{aligned} \tag{15}$$

The above problem is multi-convex, i.e. convex with respect to each variable M, W, or U, when the others are set. This property suggests to resort to an alternating proximal algorithm for solving Problem (15), as described in the following section.

4. PROPOSED ALGORITHM

4.1. Proximal alternating algorithm

To address Problem (15), one could think of using a standard alternating minimization approach. However, it is well-known that such an alternating minimization procedure requires quite restrictive conditions to guarantee its convergence toward a local minimizer (see for example [19]), even in a convex setting. Instead, we use an alternating proximal algorithm which was initially proposed in [20], for which sound convergence guarantees were established. We first recall the definition of the proximity operator.

Definition 1. Let $\mathcal{H}$ be a real Hilbert space. Let $\psi\colon \mathcal{H} \longrightarrow]-\infty, +\infty]$ be a proper, convex, lower semi-continuous function. Then for every $x \in \mathcal{H}$, the proximity operator of ψ at x is the unique vector defined as

$$\text{prox}_\psi(x) = \underset{y\in\mathcal{H}}{\text{argmin}}\ \psi(y) + \frac{1}{2}\|x - y\|^2. \tag{16}$$

Based on the expression of the objective function in Problem (15), we will also define the following functions:

$$\begin{aligned} \phi_U(M, W) = {} & f(YM - XW) + g(W) \\ & - \gamma\langle M, A^\top U\rangle + \iota_C(M) \end{aligned} \tag{17}$$

$$\phi_{(M,W)}(U) = -\gamma\langle AM, U\rangle + \iota_{\mathcal{B}^\infty}(U). \tag{18}$$

The chosen algorithm is given below, where λ and ν are parameters of the algorithm:

Algorithm 1: Alternating proximal algorithm

```
input : $M_0, W_0, U_0, \lambda > 0, \nu > 0$
for $n = 0, 1, \ldots$ do
    $(M_{n+1}, W_{n+1}) = \text{prox}_{\lambda\phi_{U_n}}(M_n, W_n)$
    $U_{n+1} = \text{prox}_{\nu\phi_{(M_{n+1},W_{n+1})}}(U_n)$
end
```

We can derive the proximity operator of $\nu\phi_{(M,W)}$ from the projection onto the ball $\mathcal{B}^\infty$:

$$\text{prox}_{\nu\phi_{(M,W)}}(U) = \text{proj}_{\mathcal{B}^\infty}(U + \nu\gamma AM) \tag{19}$$

However, there is no closed form expression for $\text{prox}_{\lambda\phi_U}$.

[1] $\iota_C(x) = 0$ if $x \in C$, $\iota_C(x) = +\infty$ otherwise.

		texture	sonar	pima	wdbc	banana	magic	satimage	titanic	bupa	AVG
APDNC	Train	87.0	85.4	74.7	94.1	57.3	77.1	78.3	77.3	59.7	76.8
	Test	86.4	72.0	73.0	94.0	54.9	77.2	77.3	77.3	58.2	74.5
Barlaud et al.	Train	72.8	83.1	76.5	88.0	56.0	66.0	74.7	77.6	69.3	73.8
	Test	72.3	68.1	75.5	87.2	54.4	65.7	72.1	77.6	67.8	71.2
NCM	Train	74.5	72.7	73.4	93.9	57.7	77.1	78.7	75.4	60.0	73.7
	Test	73.7	70.2	72.8	93.7	57.4	76.9	78.4	74.6	60.0	73.1

Table 1. Classification rate of APDNC compared to the state-of-the-art.

4.2. Computation of the proximal operator of $\lambda\phi_U$

For every $(\bar{M}, \bar{W}) \in \mathbb{R}^{k\times d} \times \mathbb{R}^{d\times\ell}$, we must compute

$$\begin{aligned}\operatorname{prox}_{\lambda\phi_U}(\bar{M},\bar{W}) = \underset{M,W}{\operatorname{argmin}}\ &\phi_U(M,W)\\ &+\frac{1}{2\lambda}(\|M-\bar{M}\|_{\mathrm{F}}^2 + \|W-\bar{W}\|_{\mathrm{F}}^2).\end{aligned}$$

Function ϕ_U can be split as $\phi_U(M,W) = f(L(M,W)) + r(M,W)$, where $L\colon (M,W) \mapsto YM - XW$ and r is a separable function of its two arguments M and W. L is a linear operator whose adjoint operator is

$$(\forall Z \in \mathbb{R}^{m\times\ell}) \quad L^*\colon Z \mapsto (Y^\top Z, -X^\top Z). \tag{20}$$

and whose spectral norm is $\|L\|_{\mathrm{S}} = \|XX^\top + YY^\top\|_{\mathrm{S}}^{1/2}$. Since the objective function involved in the definition of the proximity operator is strongly convex, we can employ an accelerated Primal-Dual algorithm [21, Alg. 2] to compute $\operatorname{prox}_{\lambda\phi_U}(\bar{M},\bar{W})$. This algorithm makes use of the conjugate function f^* of f, and two parameters τ_p and σ_p which are dynamically updated along the iterations.

Algorithm 2: Accelerated primal-dual algorithm

```
input :$\bar{M}, \bar{W}, U, M_0, W_0, V_0, \tau_0, \lambda, \gamma$
initialization : $\sigma_0 = \frac{1}{\tau_0\|L\|_{\mathrm{S}}^2}$
for $p = 0, 1, \ldots$ do
    $\theta_p = \left(\sqrt{1 + \frac{2\tau_p}{\lambda}}\right)^{-1}$
    $M_{p+1} =$
      $\operatorname{proj}_C\Big(\frac{\lambda}{1+\lambda}(M_p + \gamma A^\top U - \tau_p Y^\top V_p - \lambda^{-1}\bar{M})\Big)$
    $W_{p+1} =$
      $\operatorname{prox}_{\frac{\lambda\tau_p}{1+\lambda}g}\left(\frac{\lambda}{1+\lambda}(W_p + \tau_p X^\top V_p - \lambda^{-1}\bar{W})\right)$
    $Z_p = L\big((1+\theta_p)(M_{p+1}, W_{p+1}) - (M_p, W_p)\big)$
    $\tau_{p+1} = \theta_p\tau_p,\ \sigma_{p+1} = \frac{\sigma_p}{\theta_p}$
    $V_{p+1} = \operatorname{prox}_{\sigma_{p+1}f^*}(V_p + \sigma_{p+1}Z_p)$
end
```

4.3. Overall algorithm

Algorithm 2 is embedded as a sub-iteration in Algorithm 1. The resulting algorithm will be subsequently referred to as APDNC (alternating primal-dual non-convex minimization). Note that the initial variable M_0 and W_0 in Algorithm 2 are warm-restarted after each iteration of the main loop of Algorithm 1. This has been observed to be beneficial to the convergence speed.

5. NUMERICAL EXPERIMENTS

In this section, we evaluate the performance of our method on a set of supervised classification problems and we compare it to state-of-the-art methods.

Settings. The experimental results we present hereafter were obtained using $f = \|\cdot\|_1$ and $g = \alpha\|\cdot\|_1 + \frac{\beta}{2}\|\cdot\|_{\mathrm{F}}^2$. For all our experiments, we set the parameters of Algorithm 2 to $\nu = 1$ and the regularization parameters to $\alpha = 10^{-3}$, $\beta = 10^{-6}$. Parameter λ in the proximity operator, and penalty parameter γ on the centroid-separation term were chosen to reach the best classification performance for each dataset. Regarding the initialization of the primal variables, we set W_0 by performing a PCA on X, and then we computed M_0 by using the centroids of the transformed data. The dual variables were initialized as $V_0 = \operatorname{sign}(YM_0 - XW_0)$, $U_0 = \operatorname{sign}(AM_0)$.

Results. We used several datasets from Knowledge Extraction Evolutionary Learning (KEEL) [22], which include both simulated and real-word datasets. We used the 10-fold version, which allows us to perform cross-validation on our model. We compared our method with two state-of-the-art centroid-based methods, namely the Nearest Matrix Classification method (NCM) [23] and the method of Barlaud et al. [13, Alg. 7]. For more comparisons, refer to [24]. Table 1 reports the obtained results. One can observe that our method gives comparable results to its competitors and even outperforms them on some datasets.

6. CONCLUSION

We have proposed a sound variational formulation of centroid-based classification in a transformed domain. The resulting non-convex optimization problem has been tackled with recent proximal techniques. The performance obtained on standard datasets show the good performance of our approach. An advantage of this method is that it allows a flexible choice for the data fit term and the regularisation ones, which could be useful when dealing with difficult (e.g., corrupted) datasets.

7. REFERENCES

[1] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *VLDB*, Trondheim, Norway, Sep. 2005, vol. 5, pp. 901–909.

[2] M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Can shared-neighbor distances defeat the curse of dimensionality?," in *SSDBM*, Heidelberg, Germany, Jul. 2010, Springer, pp. 482–500.

[3] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, Wiley Series in Probability and Statistics. Wiley, 2011.

[4] W.-C. Chang, "On using principal components before separating a mixture of two multivariate normal distributions," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 32, no. 3, pp. 267–275, 1983.

[5] F. De la Torre and T. Kanade, "Discriminative cluster analysis," in *ICML*, Pittsburgh, USA, Jun. 2006, pp. 241–248.

[6] Q. Feng, Y. Zhou, and R. Lan, "Pairwise linear regression classification for image set retrieval," in *CVPR*, Hawai, USA, Jun. 2016, pp. 4865–4872.

[7] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2106–2112, 2010.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[9] A. Majumdar and Rabab K. Ward, "Fast group sparse classification," *Canadian Journal of Electrical and Computer Engineering*, vol. 34, no. 4, pp. 136–144, 2009.

[10] Y. Shao, N. Sang, C. Gao, and L. Ma, "Spatial and class structure regularized sparse representation graph for semi-supervised hyperspectral image classification," *Pattern Recognition*, vol. 81, pp. 81–94, 2018.

[11] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *NeurIPS*, Vancouver, Canada, Dec. 2010, vol. 23, Citeseer.

[12] Y. Zhou, J.-B. Caillau, M. Antonini, and M. Barlaud, "Robust classification with feature selection using alternating minimization and Douglas-Rachford splitting method," *hal preprint hal-01993753*, 2019.

[13] M. Barlaud, A. Chambolle, and J.-B. Caillau, "Robust supervised classification and feature selection using a primal-dual method," *arXiv preprint arXiv:1902.01600*, 2019.

[14] J. Maggu, E. Chouzenoux, G. Chierchia, and A. Majumdar, "Convolutional transform learning," in *ICONIP*, Siem Reap, Cambodia, Dec. 2018, Springer, pp. 162–174.

[15] F. Bach and Z. Harchaoui, "Diffrac: a discriminative and flexible framework for clustering," in *NeurIPS*, Vancouver, Canada, Dec. 2007, vol. 20, pp. 49–56.

[16] Z. Xu, G. Huang, K. Q. Weinberger, and A. X. Zheng, "Gradient boosted feature selection," in *KDD*, New York, USA, Aug. 2014, pp. 522–531.

[17] X. Wei, B. Cao, and P. S. Yu, "Nonlinear joint unsupervised feature selection," in *SDM*, Miami, USA, May 2016, SIAM, pp. 414–422.

[18] S. Hara and T. Maehara, "Enumerate lasso solutions for feature selection," in *AAAI*, San Francisco, USA, Feb. 2017, vol. 31.

[19] T. F. Chan and C.-K. Wong, "Convergence of the alternating minimization algorithm for blind deconvolution," *Linear Algebra and its Applications*, vol. 316, no. 1-3, pp. 259–285, 2000.

[20] J. Bolte, P. L. Combettes, and J.-C. Pesquet, "Alternating proximal algorithm for blind image recovery," in *IEEE ICIP*, Hong Kong, Hong Kong, Jan. 2010, pp. 1673–1676.

[21] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[22] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "Keel datamining software tool: data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic & Soft Computing*, vol. 17, 2011.

[23] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, "Distance-based image classification: generalizing to new classes at near-zero cost," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2624–2637, 2013.

[24] J. L. Suárez-Díaz, S. García, and F. Herrera, "A tutorial on distance metric learning: mathematical foundations, algorithms, experimental analysis, prospects and challenges," *Neurocomputing*, vol. 425, pp. 300–322, 2021.