Robust Low-Rank Correlation Fitting

Thu Ha Phi, Alexandre Hippert-Ferrer, Florent Bouchard, Arnaud Breloy

ROBUST LOW-RANK CORRELATION FITTING

Thu Ha Phi[1]*, Alexandre Hippert-Ferrer*[2]*, Florent Bouchard*[3]*, Arnaud Breloy*[1]

[1]: Univ Paris-Nanterre, LEME, IUT Ville d'Avray, Ville d'Avray, France
[2]: Univ Gustave Eiffel, IGN, ENSG, LASTIG, Marne-la-Vallée, France
[3]: Univ Paris-Saclay, CNRS, CentraleSupélec, L2S, Gif-sur-Yvette, France

ABSTRACT

This paper considers the problem of obtaining a low-rank factorization of a given correlation matrix. In order to handle possible spurious correlation coefficients within the input, a robust formulation is proposed with a criterion based on the Huber loss function. Minimizing this fitting criterion under the low-rank correlation structure constraint is then addressed using the block majorization-minimization framework. Several algorithm options are explored and compared in terms of computational complexity. The merits of the proposed correlation fitting method are then validated on simulations, and for the process of dimension reduction of microarray data.

***Index Terms*—** Correlation matrix, low-rank factorization, majorization-minimization, robust estimation.

1. INTRODUCTION

The correlation matrix is a fundamental quantity in multivariate statistical analysis, signal processing, and machine learning. Among many applications, we can mention its use as a discriminating feature in EEG classification processes [1], as an input variable to compute interferograms in remote sensing [2], and as weighting coefficients in financial engineering [3, 4] and chemometrics [5]. The most common estimate of this quantity is the sample correlation matrix, i.e., the sample covariance matrix of the standardized data. Obtaining a low-rank approximation of this matrix is often beneficial in order to perform correlation cleaning (improving the estimate). A low-rank factorization is also useful in terms of memory storage, as handling a $p \times p$ matrix can be problematic for large dimension p. In terms of dimension reduction, it also offers a structure-compliant alternative to principal component analysis (PCA) for standardized data (as the eigenvalue truncation of a correlation matrix does not yield an actual correlation matrix). In this scope [4, 6, 7] proposed algorithms to perform least squares (LS) low-rank correlation fitting. However, LS fitting is sensitive to outliers and thus susceptible to fit spurious correlations rather than the true underlying low-rank structure.

This work was supported by the MASSILIA project (ANR- 21-CE23-0038-01) of the French National Research Agency (ANR).

To overcome the aforementioned issue, we propose an alternate fitting criterion that leverages the Huber criterion, which offers a robust alternative to LS [8]. The corresponding low-rank covariance fitting problem is then tackled using the block majorization-minimization (MM) framework [9, 10, 11], resulting in a computationally efficient algorithm. The interest of the proposed method is then validated on synthetic data, and for the dimension reduction of microarray datasets [12].

2. LOW-RANK CORRELATION FITTING

2.1. Background

Given a n-sample of p-dimensional observations $\mathbf{X} \in \mathbb{R}^{p \times n}$, the empirical correlation matrix is obtained as $\hat{\mathbf{C}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top/n$, where $\tilde{\mathbf{X}}$ is the standardized data, i.e., each variable (line in $\mathbf{X}$) is centered and normalized by its standard deviation. The space of correlation matrices $\mathcal{C}_p$ is the space of $p \times p$ positive semi-definite symmetric matrices with unit diagonal, and off-diagonal elements in $[-1, 1]$. Finding the nearest low-rank correlation matrix to a plug-in estimate $\hat{\mathbf{C}} \in \mathcal{C}_p$ consists in tackling the optimization problem

$$\begin{aligned} \underset{\mathbf{C} \in \mathcal{C}_p}{\text{minimize}} \quad & \delta^2(\hat{\mathbf{C}}, \mathbf{C}) \\ \text{subject to} \quad & \text{rank}(\mathbf{C}) \leq k, \end{aligned} \tag{1}$$

where $\delta(\cdot, \cdot)$ is a matrix distance, and k denotes the desired maximal rank. In this scope, the least-squares (LS) fitting relies on the squared Frobenius norm as distance, i.e., $\delta^2_{\text{LS}} = ||\hat{\mathbf{C}} - \mathbf{C}||^2_F$. For this cost function, various algorithms have been proposed in order to evaluate the solution of (1) [4, 6]. More recently, [7] considered additionally masking entries in an LS fitting formulation in order to obtain a sparse solution.

2.2. Proposed robust formulation

As LS fitting is sensitive to outliers' abnormal values, we propose an alternative robust cost to alleviate this issue. We consider leveraging the Huber loss function [8], defined by:

$$\rho_c(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| < c \\ c\,|x| - \frac{1}{2}c^2 & \text{otherwise} \end{cases} \tag{2}$$

where $c \in \mathbb{R}_*^+$ is a chosen threshold that pilots whether a quadratic or linear loss is applied to the input x. This yields the corresponding matrix distance

$$\delta_H^2(\hat{\mathbf{C}}, \mathbf{C}) = \sum_{ij} \rho_c(\hat{\mathbf{C}}_{ij} - \mathbf{C}_{ij}), \tag{3}$$

which acts as a smooth trade-off between the two limit cases $c \to 0$ (ℓ_1-norm fitting, also referred to as median estimator), and $c \to \infty$ (LS fitting). In practice c is either set to a tolerance threshold to obtain a smooth mimic of the ℓ_1-norm, or adaptively tuned so that a desired proportion of the entries are treated with the LS criterion [8].

In order to derive computationally efficient optimization algorithms, we also consider a tall matrix as root parameterization for a rank k correlation matrix, i.e., $\mathbf{C} = \mathbf{W}\mathbf{W}^\top$, with $\mathbf{W} \in \mathbb{R}^{p \times k}$. The i^{th} line of $\mathbf{W}$ is denoted $\mathbf{w}_i^\top \in \mathbb{R}^{1 \times k}$, $\forall i \in [\![1, p]\!]$, i.e., we have $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_p]^\top$. Thus parameterized, the matrix $\mathbf{C} = \mathbf{W}\mathbf{W}^\top$ is a correlation matrix in $\mathcal{C}_p$ if and only if all the vectors $\mathbf{w}_i$, $\forall i \in [\![1, p]\!]$ are of unit norm.

Combining the considered cost function and parameterization, we formulate the Huber low-rank correlation fitting through the optimization problem

$$\begin{array}{ll} \underset{\mathbf{W} \in \mathbb{R}^{p \times k}}{\text{minimize}} & f(\mathbf{W}) \triangleq \sum_{ij} \rho_c(\hat{\mathbf{C}}_{ij} - \mathbf{w}_i^\top \mathbf{w}_j) \\ \text{subject to} & \mathbf{w}_i^\top \mathbf{w}_i = 1, \ \forall i \in [\![1, p]\!], \end{array} \tag{4}$$

which is tackled in the next section.

3. BLOCK MM ALGORITHMS

The problem (4) has no direct solution and requires to develop iterative algorithms for evaluating a local minimum. To do so, we leverage block MM framework, which is briefly reviewed below. More details can be found in the overviews [9, 11], and papers [13, 14] also provide some general guidelines for deriving algorithms under normalization constraints. In the context of our partitioned variable $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_p]^\top$, the block MM algorithm will operate a cyclic update of the blocks $\mathbf{w}_i$ by solving sub-problems of the form:

$$\begin{array}{ll} \underset{\mathbf{w}_i \in \mathbb{R}^k}{\text{minimize}} & g_i(\mathbf{w}_i | \mathbf{W}_t) \\ \text{subject to} & \mathbf{w}_i^\top \mathbf{w}_i = 1 \end{array} \tag{5}$$

where g_i is a *surrogate function*, i.e., a tight global upper-bound of the function f at the current point denoted $\mathbf{W}_t = [\mathbf{w}_1^t, \cdots, \mathbf{w}_p^t]^\top$. The goal is then to obtain a surrogate g_i such that it is easily minimized over the feasible set, and that it respects the conditions to ensure the convergence of the procedure: g_i is a global upper-bound of f, with equality and matching directional derivatives at the considered point.

3.1. Building blocks

This section provides the key propositions to obtain a surrogate g_i and solving its minimization step. The first proposition gives a global upper-bound of f that appears as a weighted least-squares correlation fitting criterion.

Proposition 1. *pThe objective f in* (4) *is majorized at point $\mathbf{W}_t$ by the surrogate function*

$$g_{\text{WLS}}(\mathbf{W} | \mathbf{W}_t) = \sum_{ij} \omega_{ij} (\hat{\mathbf{C}}_{ij} - \mathbf{w}_i^\top \mathbf{w}_j)^2 \tag{6}$$

with $w_{ij} = \rho_c'(\varepsilon_{ij}^t)/2(\varepsilon_{ij}^t)$ and $\varepsilon_{ij}^t = \hat{\mathbf{C}}_{ij} - (\mathbf{w}_i^t)^\top \mathbf{w}_j^t$. Equality is achieved at $\mathbf{W} = \mathbf{W}_t$.

Proof. The upper-bound is obtained on each term of the sum by using the majorization of the Huber function at point y [11]

$$\rho_c(x) \leq \frac{\rho_c'(y)}{2y} x^2 + \text{const.} \tag{7}$$

with equality at point y. □

The second proposition gives a block-wise linear upper-bound for a weighted least-squares correlation fitting criterion when restricting blocks $\mathbf{w}_i$ to lien on the unit k-sphere.

Proposition 2. *At point $\mathbf{W}_t$ and given an active block $\mathbf{w_i}$, the surrogate g_{WLS} in* (6) *is majorized on the unit k-sphere by the surrogate function*

$$g_i(\mathbf{w}_i | \mathbf{W}_t) = \mathbf{w}_i^\top \mathbf{v}_i^t + \text{const.} \tag{8}$$

with

$$\mathbf{v}_i^t = 2(\mathbf{B}_i^t - \lambda_i^t \mathbf{I}_k)\mathbf{w}_i^t - 2 \sum_{j \neq i} \omega_{ij} \hat{\mathbf{C}}_{ij} \mathbf{w}_j \tag{9}$$

with $\mathbf{B}_i^t = \sum_{j \neq i} \omega_{ij} \mathbf{w}_j \mathbf{w}_j^\top$ and where λ_i^t is any constant larger than $\lambda_{\max}^{\mathbf{B}_i^t}$, the largest eigenvalue of $\mathbf{B}_i^t$. Equality is achieved at $\mathbf{w}_i = \mathbf{w}_i^t$.

Proof. The function g_{WLS} in (6) with only $\mathbf{w}_i$ as active block reduces to the expression

$$g_{\text{WLS}}(\mathbf{w}_i | \mathbf{W}^t) = \mathbf{w}_i^\top \mathbf{B}_i^t \mathbf{w}_i - 2\mathbf{w}_i^\top \sum_{j \neq i} \omega_{ij} \hat{\mathbf{C}}_{ij} \mathbf{w}_j + \text{const.} \tag{10}$$

When restricting to the unit k-sphere, we have that the first term $\mathbf{w}_i^\top \mathbf{B}_i^t \mathbf{w}_i = \mathbf{w}_i^\top (\mathbf{B}_i^t - \lambda_i^t \mathbf{I}_k)\mathbf{w}_i + \text{const.}$, for any $\lambda_i^t \in \mathbb{R}$. Hence, the minimization of $g_{\text{WLS}}(\mathbf{w}_i | \mathbf{W}^t)$ is not affected by this constant shift. When setting $\lambda_i^t \geq \lambda_{\max}^{\mathbf{B}_i^t}$, the quadratic form $\mathbf{w}_i^\top (\mathbf{B}_i^t - \lambda_i^t \mathbf{I}_k)\mathbf{w}_i$ is concave so it is majorized by its first-order Taylor approximation [11]. This yields the desired result

$$g_{\text{WLS}}(\mathbf{w}_i | \mathbf{W}^t) \leq g_i(\mathbf{w}_i | \mathbf{W}^t) + \text{const.} \tag{11}$$

with g_i defined in (8) and (9). □

The last proposition concerns the minimization of a linear function on the unit k-sphere, which will allow us to compute simple updates when applying the previouly obtained surrogates.

Proposition 3. *The constrained optimization problem*

$$\begin{array}{ll} \underset{\mathbf{w}_i \in \mathbb{R}^k}{\text{minimize}} & \mathbf{w}^\top \mathbf{v} \\ \text{subject to} & \mathbf{w}^\top \mathbf{w} = 1 \end{array} \tag{12}$$

has for solution $\mathbf{w}^\star = -\mathbf{v}/|\mathbf{v}|$.

Proof. When constrained to have a unit-norm, the vector $\mathbf{w}^\star$ that minimizes the value of the inner product $\langle \mathbf{w}, \mathbf{v} \rangle$ is the one aligned with the direction of $-\mathbf{v}$ □

3.2. Algorithmic options

The chaining of Propositions 1 and 2 offers various possible MM strategies for the construction of the block updates, which are summed up in the box Algorithm 1. In details:

• Proposition 1 (updating ω_{ij}^t) can be applied at each global loop or at each block $\mathbf{w}_i$: the first option mimics the minimization of a sequence of weighted least-squares, while the second minimizes a tight upper-bound at each step, yielding a faster convergence. The computational trade-off is that the update of ω_{ij}^t at each step can be costly for large p.

• The eigenvalue upper-bound of $\mathbf{B}_i^t$ in Proposition 2 can be set to $\lambda_{\max}^{\mathbf{B}_i^t}$, which offers a tight upper-bound. By construction of this matrix (weighted sum of normalized vectors with weights $\omega_{ij} \leq 1$), an alternate loose upper-bound is $\lambda_i^t = p$. As previously, a tighter upper-bound offers a faster convergence, but the other option yields a lower computational cost per update (as it requires no SVD computation).

• In a block-coordinate descent fashion, the block index i can be hold for multiple applications of Proposition 2 (then minimization with Proposition 3). The resulting loop appears as a modified power method for minimizing a quadratic plus linear function on the k-sphere [14], which offers a slight acceleration of each block step for a low computational cost.

3.3. Algorithm analysis

Convergence: By construction of the MM procedure, each variable update yields a monotonic decrement of the objective [9, 11]. As the constraint set is not convex, the general results to a local minimum [10] do not directly apply. However, the guarantee of convergence to a local minimum still holds as the variables belong to a compact set satisfying the linear independence constraint qualification [15, 14].

Complexity and implementation: In the common case $n \leq p$, we notice that all matrix multiplications can be efficiently factorized using $\mathbf{X}^\top \mathbf{W}$ instead of directly computing $\hat{\mathbf{C}}\mathbf{W}$, hence the complexity of basic matrix operations is $\mathcal{O}(npk)$. The main and only computational bottleneck is the SVD of $\mathbf{B}_i^t$, which is $\mathcal{O}(k^3)$. This is not expected to be too demanding since we aim to address $k \ll p$ scenarios. This bottleneck can be avoided with the universal upper-bound $\lambda_i^t = p$, at the cost of requiring more total iterations.

Algorithm 1 MM for Huber low-rank correlation fitting

Input: plug-in $\hat{\mathbf{C}}$, Huber threshold c
MM options: inner-or-outer ω_{ij} updates $\text{in}_{\text{Hub}} \in \{\text{in}, \text{out}\}$, $\lambda_{\max}^{\mathbf{B}_i^t}$ upper-bound $\text{up}_\lambda \in \{\text{SVD}, \text{const.} = p\}$, inner-loops $n_{\text{loop}} \in \mathbb{N}^*$, convergence criterion.
while (convergence criterion) **do**
  if $\text{in}_{\text{Hub}} = \text{out}$ update ω_{ij}^t with Proposition 1
  for $i = 1, \ldots, p$ **do**
    if $\text{in}_{\text{Hub}} = \text{in}$ update ω_{ij}^t with Proposition 1
    Compute $\mathbf{B}_i^t$ with λ_i^t chosen according to up_λ
    for $\ell \in 1, \ldots, n_{\text{loop}}$ **do**
      Compute $\mathbf{v}_i^t$ with Proposition 2
      Update $\mathbf{w}_i^t$ with Proposition 3
      $t \leftarrow t + 1$
    end for
  end for
end while

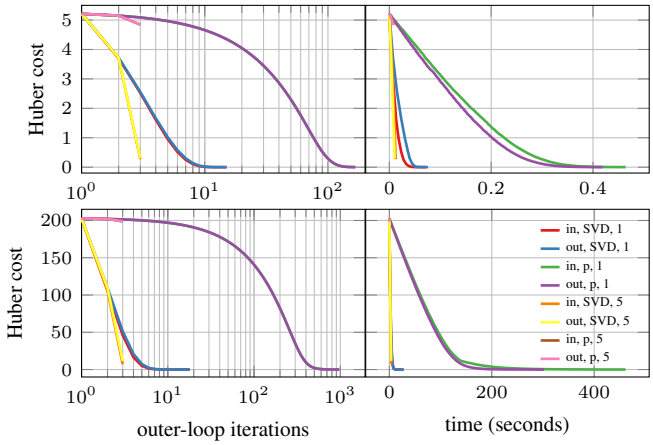

Fig. 1. Convergence of MM algorithms (options in Algorithm 1) versus number of outer-loop iterations (one point represents an update of all blocks $\mathbf{w}_i$) and time for $p = 100$, $k = 10$ (top), $p = 1000$, $k = 100$ (bottom). In the legend, the first parameter is in_{Hub}, the second up_λ and the third n_{loop}.

4. SIMULATIONS

Algorithms comparison: For a randomly generated plug-in $\hat{\mathbf{C}}$, Figure 1 displays the convergence of Algorithm 1 for various options (cf. Section 3.2) and dimensions p and k. It shows that the algorithms scale well for high dimensions. We observe that when using the loose upper-bound $\lambda_i^t = p$, the loss in convergence speed is not compensated by the gain in computational complexity of the universal when assessing total computation time (so $\lambda_i^t = \lambda_{\max}^{\mathbf{B}_i^t}$ should be favored). It also shows that the outer acceleration loops can be beneficial, but should be set depending on the dimensions.

Performance validation: We investigate the robustness of the proposed method to corruption of the sample correlation matrix. Firstly, a sample covariance matrix $\hat{\mathbf{\Sigma}} = \mathbf{X}\mathbf{X}^\top/n$

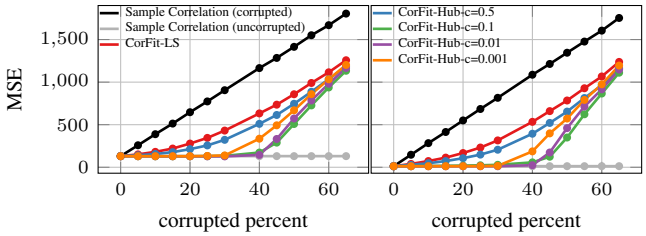

Fig. 2. MSE on the recovery of the low-rank correlation matrix with respect to the percentage of corrupted entries, SNR = 35 dB, number of trials = 50. $p = 100$, $k = 10$, $n = 50$ (left) or $n = 500$ (right).

	Datasets			
	(1)	(2)	(3)	(4)
PCA	13.66	16.94	28.73	27.4
CorFit-LS	13.66	17.42	28.9	**26.26**
Cor Fit-Hub	**13.29**	**15.87**	**28.34**	26.38

Table 1. Average clustering **error** in percent (%) on various embedding dimensions $k \in [3, 23]$. CuMiDa datasets are Brain_GSE50161[(1)] (p=54676, n=130, 5 cancer types), Colorectal_GSE21510[(2)] (p=54676, n=147, 3 cancer types), Colorectal_GSE41657[(3)] (p=33468, n=86, 4 cancer types), Leukemia_GSE28497[(4)] (p=22284, n=281, 7 cancer types).

is generated from n observations distributed according to the Gaussian probabilistic PCA model [16]

$$\mathbf{x} \stackrel{d}{=} \mathbf{W}_0\mathbf{s} + \mathbf{n} \tag{13}$$

with $\mathbf{W}_0 \in \mathbb{R}^{p\times k}$ and normalization $\mathrm{diag}(\mathbf{W}_0\mathbf{W}_0^\top) = \mathbf{I}_p$, $\mathbf{s} \sim \mathcal{N}(0, \sigma_s^2\mathbf{I}_k)$, and $\mathbf{n} \sim \mathcal{N}(0, \sigma_n^2\mathbf{I}_p)$. Thus $\mathbf{x} \sim \mathcal{N}(0, \mathbf{\Sigma})$, with $\mathbf{\Sigma} = \sigma_s^2\mathbf{W}_0\mathbf{W}_0^\top + \sigma_n^2\mathbf{I}_p$ and we have the signal to noise ratio $\mathrm{SNR} = \sigma_s^2/\sigma_n^2$. Secondly, a percentage $b \in [0, 100]$ of random entries of $\hat{\mathbf{\Sigma}}$ are corrupted by generating a uniform noise as follows

$$[\tilde{\mathbf{\Sigma}}]_{ij} = \begin{cases} a \in \pm\left[\tilde{\epsilon}, \sqrt{[\hat{\mathbf{\Sigma}}]_{ii}[\hat{\mathbf{\Sigma}}]_{jj}}\right] & \text{if } \left|[\hat{\mathbf{\Sigma}}]_{ij}\right| \leq \epsilon \\ a \in \pm[0, \tilde{\epsilon}] & \text{otherwise} \end{cases} \tag{14}$$

where $[\tilde{\mathbf{\Sigma}}]_{ji} = [\tilde{\mathbf{\Sigma}}]_{ij}$, and where $\tilde{\epsilon} = min(\epsilon, \sqrt{[\hat{\mathbf{\Sigma}}]_{ii}[\hat{\mathbf{\Sigma}}]_{jj}})$ with a low thres ϵ, set a bound so that the corrupted matrix remains symmetric positive semi-definite. The corrupted matrix $\tilde{\mathbf{\Sigma}}$ is then normalized to obtain a corrupted correlation matrix $\tilde{\mathbf{C}} = (\tilde{\mathbf{\Sigma}} \odot \mathbf{I})^{-1/2}\tilde{\mathbf{\Sigma}}(\tilde{\mathbf{\Sigma}} \odot \mathbf{I})^{-1/2}$. Finally, the $\tilde{\mathbf{C}}$ serves as a plug-in to various low-rank correlation factorization algorithms (LS fitting and Huber fitting with various threshold coefficient c). We then assess the merits of the algorithms in terms of mean squared error (MSE) for the recovery of the true underlying low-rank correlation $\mathbf{C}_0 = \mathbf{W}_0\mathbf{W}_0^\top$. Figure 2 displays the MSE with respect to the percentage of corrupted entries in $\tilde{\mathbf{C}}$. We observe that the low-rank factorization helps in lowering the error compared to the raw plug-in estimate. This observed behavior remains the same at low and large sample support n, though the floor error is naturally lower for larger n. Interestingly, Huber fitting with low c can support a higher percentage of corruption before breakdown (i.e. deviation from the MSE of the oracle uncorrupted plug-in).

5. APPLICATION TO MICROARRAY DATA

Context and protocol: Linear dimension reduction is a standard process when dealing with large p. The most common method is PCA, which produces linear embeddings from the k leading eigenvectors of the sample covariance matrix. For standardized data, the covariance matrix is also a correlation matrix, so it appears interesting to consider this additional structure, and rather project the data on $\mathrm{span}(\hat{\mathbf{W}})$, where $\hat{\mathbf{W}}$ is obtained from a low-rank correlation matrix factorization algorithm (such as the proposed MM for Huber fitting). In this scope, we compare the effectiveness of these approaches on the CuMiDa microarray dataset [12]. We assess the linear embedding method's ability to preserve *a priori* unknown class clusters in an unsupervised setting (actual labels are still used to validate the process afterward). This property is here quantified by the clustering accuracy of the K-means++ [17] applied on the projected data. For each compared method, we average the clustering error over 20 values of $k \in \{3, 4, ..., 23\}$ used as lower dimension.

Data Pre-processing: From the raw data, we first select variables with the highest variances (baseline approach to feature selection [18, 19]) up to a certain total variance percentage (30% gave the best overall PCA's results). The data is then standardized before applying dimension reduction methods.

Adaptive tuning of Huber threshold c: For Huber low-rank fitting, we compute a sequence of solutions for various c, and select the solution for which 85% of entries in the plug-in $\hat{\mathbf{C}}$ are treated with $|x| \geq c$ as in (2).

Results: Table 1 summarizes the mean clustering error for several CuMiDa datasets. We observe that the embedding based on correlation fitting can favor to PCA in this unsupervised clustering setting. The Huber fitting also slightly outperforms LS fitting in most cases, highlighting the interest in the proposed approach.

6. CONCLUSION

We proposed a robust low-rank correlation fitting method based on the Huber function, and derived MM algorithms to solve the corresponding problem. Simulations evidenced the method's interest in computational efficiency and robustness to data corruption. An application to microarray data clustering illustrated that the proposed method also offers an interesting alternative to the PCA.

7. REFERENCES

[1] Sylvain Chevallier, Emmanuel K Kalunga, Quentin Barthélemy, and Eric Monacelli, "Review of riemannian distances and divergences, applied to ssvep-based bci," *Neuroinformatics*, vol. 19, no. 1, pp. 93–106, 2021.

[2] Ning Cao, Hyongki Lee, and Hahn Chul Jung, "Mathematical framework for phase-triangulation algorithms in distributed-scatterer interferometry," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 9, pp. 1838–1842, 2015.

[3] Nicholas J Higham, "Computing the nearest correlation matrix—a problem from finance," *IMA journal of Numerical Analysis*, vol. 22, no. 3, pp. 329–343, 2002.

[4] Raoul Pietersz 4 and Patrick JF Groenen, "Rank reduction of correlation matrices by majorization," *Quantitative Finance*, vol. 4, no. 6, pp. 649–662, 2004.

[5] Jasper Engel, Lutgarde Buydens, and Lionel Blanchet, "An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics," *Journal of chemometrics*, vol. 31, no. 4, pp. e2880, 2017.

[6] Igor Grubišić and Raoul Pietersz, "Efficient rank reduction of correlation matrices," *Linear algebra and its applications*, vol. 422, no. 2-3, pp. 629–653, 2007.

[7] Kensuke Tanioka, Yuki Furotani, and Satoru Hiwa, "Thresholding approach for low-rank correlation matrix based on mm algorithm," *Entropy*, vol. 24, no. 5, pp. 579, 2022.

[8] Peter J Huber, *Robust statistics*, vol. 523, John Wiley & Sons, 2004.

[9] David R Hunter and Kenneth Lange, "A tutorial on mm algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[10] Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.

[11] Ying Sun, Prabhu Babu, and Daniel P Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2016.

[12] Bruno Cesar Feltes, Eduardo Bassani Chandelier, Bruno Iochins Grisci, and Márcio Dorn, "Cumida: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research," *Journal of Computational Biology*, vol. 26, no. 4, pp. 376–386, 2019.

[13] Henk AL Kiers, "Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems," *Computational statistics & data analysis*, vol. 41, no. 1, pp. 157–170, 2002.

[14] Arnaud Breloy, Sandeep Kumar, Ying Sun, and Daniel P Palomar, "Majorization-minimization on the stiefel manifold with application to robust sparse pca," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1507–1520, 2021.

[15] Konstantinos Benidis, Ying Sun, Prabhu Babu, and Daniel P Palomar, "Orthogonal sparse pca and covariance estimation via procrustes reformulation," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6211–6226, 2016.

[16] Michael E Tipping and Christopher M Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 61, no. 3, pp. 611–622, 1999.

[17] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, pp. 1295, 2020.

[18] Isabelle Guyon and André Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.