Information Sciences 728 (2026) 122823

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Distributionally robust nonnegative matrix factorization with self-paced adaptive multi-loss fusion

Wafa Barkhoda[a,b], Amjad Seyedi[c], Nicolas Gillis[c], Fardin Akhlaghian Tab[a,*]

[a] Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran
[b] Faculty of Information Technology, Kermanshah University of Technology, Kermanshah, Iran
[c] Department of Mathematics and Operational Research, University of Mons, Mons, Belgium

HIGHLIGHTS

- Robust NMF model combining distributional robustness and self-paced learning.
- Self-paced learning enhances resistance to outliers and heavy-tailed noise.
- Adaptive multi-loss fusion ensures balanced and stable model optimization.
- Outperforms leading robust NMF methods under diverse noise conditions.

ARTICLE INFO

Keywords:
Nonnegative matrix factorization
Distributionally robust optimization
Self-paced learning
Instance-wise representation

ABSTRACT

Nonnegative Matrix Factorization (NMF) is a widely used technique for parts-based data representation, but its sensitivity to non-Gaussian noise and outliers limits robustness. Existing robust models typically address this issue by modifying the loss function to mitigate such outliers; however, they often lack generalization across diverse noise distributions. This paper proposes a novel framework, Distributionally Robust NMF with Self-Paced Adaptive Multi-Loss Fusion (DRNMF-SP), to enhance robustness against both moderate and extreme outliers across various noise types. DRNMF-SP adopts a multi-objective optimization strategy that integrates multiple loss functions through a weighted sum, reflecting the uncertainty in selecting a single objective. It employs a distributionally robust optimization, minimizing the worst-case expected loss over a probabilistic ambiguity set. The integration of self-paced learning enables the model to progressively learn from clean instances while deferring to noisy samples, thereby enhancing its robustness to heavy-tailed distributions. Additionally, the instance-wise loss function shifts focus from individual features to the holistic structure of samples, improving performance in real-world datasets. An efficient iterative reweighted algorithm ensures computational feasibility, with costs comparable to basic NMF. Experimental evaluations on benchmark datasets confirm that DRNMF-SP consistently outperforms existing robust methods across noisy, complex scenarios. The implementation can be found at https://github.com/barkhoda/DRNMF-SP.

* Corresponding author.
Email addresses: barkhoda@kut.ac.ir (W. Barkhoda), seyedamjad.seyedi@umons.ac.be (A. Seyedi), nicolas.gillis@umons.ac.be (N. Gillis), f.akhlaghian@uok.ac.ir (F. Akhlaghian Tab).

https://doi.org/10.1016/j.ins.2025.122823
Received 11 August 2025; Received in revised form 22 October 2025; Accepted 22 October 2025
Available online 24 October 2025

1. Introduction

Nonnegative matrix factorization (NMF) has emerged as a prominent technique due to its interpretability and effectiveness in uncovering latent structures within nonnegative data [1]. In contrast to methods such as principal component analysis (PCA) and singular value decomposition (SVD), which rely on orthogonal projections, NMF approximates a given nonnegative matrix as the product of two smaller nonnegative matrices. This nonnegativity constraint promotes an additive composition of features, naturally leading to a parts-based data representation rather than a holistic one [2]. Due to these properties, NMF has been widely applied in dimensionality reduction tasks across domains such as clustering [3], network analysis [4], tensor embeddings [5], feature selection [6], matrix recovery [7], and hyperspectral unmixing [8]. The motivation for employing NMF arises from its unique capability to produce nonnegative, parts-based representations that are both interpretable and effective in uncovering latent structures. Unlike alternative factorizations such as PCA or SVD, which may yield negative or mixed-sign components, NMF ensures physically meaningful decompositions, particularly suitable for applications involving image, biological, and signal data. Therefore, improving the robustness of NMF under uncertain and noisy environments is essential for maintaining its interpretability and reliability in real-world scenarios. Despite its success, standard NMF models can perform suboptimally when faced with datasets containing outliers or non-Gaussian noise. Although least squares-based NMF is effective under Gaussian assumptions, its performance can deteriorate with heavy-tailed noise distributions such as Laplacian or Cauchy. This is largely due to the L_2 loss function's sensitivity to outlier values, which can distort the learned representation.

To overcome the sensitivity of traditional NMF to outliers and non-Gaussian noise, numerous studies have incorporated robust loss functions inspired by M-estimation theory. These approaches replace the standard squared Euclidean distance with alternatives that offer improved resistance to noise contamination. Common replacements for conventional L_2 loss include the L_1-norm [9], $L_{2,1}$-norm [10], Huber loss [11], and Correntropy [12]. For instance, Lam [9] proposed an NMF formulation based on L_1-norm minimization, which is more suitable for data affected by Laplacian noise or heavy-tailed distributions where the assumptions of the Central Limit Theorem may not hold. Kong et al. [10] extended this robustness by introducing $L_{2,1}$-NMF, which substitutes the Frobenius norm with the $L_{2,1}$ norm. This modification avoids squaring reconstruction errors, thereby reducing the undue influence of large deviations from individual data samples.

Liutkus et al. [13] proposed a Cauchy-based NMF, using the Cauchy distribution to model reconstruction errors within a maximum likelihood estimation framework. This approach naturally suppresses the influence of extreme values due to the heavy-tailed nature of the Cauchy distribution. In addition to this, several other robust NMF variants have been developed, including CIM-NMF (Correntropy-Induced Metric NMF) and its row-wise extension, rCIM-NMF [11]. Correntropy, based on the Welsch M-estimator, serves as a local and nonlinear similarity measure in information-theoretic learning (ITL) [14]. It reflects the statistical similarity between random variables near their joint support. Another family of robust methods focuses on explicitly controlling large reconstruction errors by either limiting their contribution or discarding them entirely. For example, Gao et al. [15] introduced a capped norm strategy, where reconstruction errors exceeding a fixed threshold are cut or ignored. However, a practical challenge is the difficulty in selecting a suitable threshold. To address this, Guan et al. [16] adopted the three-sigma principle for outlier detection and proposed a truncated Cauchy-based loss, which effectively handles both moderate and severe outliers but requires tuning of two distribution-specific parameters.

In summary, the robust methods discussed above improve the resistance to noise by replacing the standard squared loss with alternative formulations that reduce the influence of outliers on the reconstruction error. These techniques fall under the umbrella of robust optimization (RO). On the other hand, Distributionally Robust Optimization (DRO), originally introduced by Scarf [17], provides a broader probabilistic framework to manage data uncertainty. Rather than optimizing performance under a single known distribution, DRO seeks to ensure reliable outcomes across a family of distributions contained within an uncertainty set Ω. In essence, DRO aims to produce solutions that are stable in multiple likely noise scenarios. This is achieved by minimizing the worst-case expected loss over a probabilistic ambiguity set that is inferred from observed data and reflects partial knowledge of the true data-generating process. The appeal of DRO has grown rapidly in recent years, thanks to its solid theoretical foundation, adaptability to different ways of measuring distributional uncertainty, and strong performance in various applications [14].

DRO has gained popularity due to its ability to strike a balance between the overcautious nature of robust optimization and the data-specific demands of stochastic programming. In [18], the authors propose an approximation technique for DRO problems involving moment-based ambiguity sets by incorporating PCA to extract informative low-dimensional structures from data variability. The value of distributional robustness becomes even more pronounced in scenarios with limited training data, as it enhances the generalization capacity of predictive models. For example, Zhu et al. [19] investigate a minimax DRO formulation for weighted k-nearest neighbors, where optimal weighting schemes are derived to account for feature-level uncertainty. Classical classifiers often assume perfectly known training inputs, yet practical datasets frequently suffer from noise or perturbations. To address this, Faccini et al. [20] model data uncertainty using geometric sets, such as hyperrectangles or ellipsoids, and introduce a moment-based DRO framework that constrains deviations along principal axes. In the context of matrix factorization, Gillis et al. [21] developed a DRO-based NMF model utilizing the β-divergence family to increase robustness against types of noise represented in the ambiguity set. Extending this line of work, [22] proposed the instance-wise distributionally robust NMF (iDRNMF), a multi-objective formulation designed to accommodate a wide variety of noise distributions through adaptive loss function integration.

Conventional NMF methods typically incorporate noise modeling at the element level [23], assuming that each feature or entry in the data matrix can be treated independently. In contrast, instance-wise robust NMF approaches [11,24] focus on capturing noise patterns at the sample level, where each column (or instance) is considered as an integrated unit. This paradigm is especially effective for datasets with numerous samples or where the data exhibit continuity across features. By considering the full structure

of each instance, such models are better equipped to detect and preserve global trends, making them highly applicable in complex or noisy environments. Instance-level robustness improves the model's resilience and ensures consistent performance, even when entire samples are affected by noise or outliers. Moreover, it promotes a better understanding of the underlying relationships among samples, which is critical for interpretability in real-world applications. These advantages have led to the successful application of instance-wise robust NMF in diverse areas, including image restoration [25], data quality monitoring [26], randomized smoothing for robustness [27], biological data analysis [28], and hyperspectral signal separation [29].

One of the intrinsic challenges in NMF lies in the non-convexity of its objective function, rendering the task of finding a global optimum NP-hard. As a consequence, NMF algorithms often converge to suboptimal local solutions, especially in the presence of high noise levels or severe outliers. A widely adopted workaround is to execute the algorithm multiple times with different random initializations and select the best-performing outcome. However, this strategy is inefficient and impractical in unsupervised contexts, where there is no straightforward metric to identify the most suitable solution. To address this limitation, self-paced learning (SPL) [30] has emerged as a compelling strategy. SPL has been shown to reduce the risk of getting stuck in poor local minima and to enhance generalization performance [31]. The core idea of SPL is to prioritize training on simpler, cleaner data points before gradually including more complex or noisier ones—mirroring the incremental nature of human learning. This mechanism helps the model become more resilient to anomalies and improves overall training stability. SPL has demonstrated success across a range of computer vision and pattern recognition tasks [32,33]. For example, Zhao et al. [34] applied this principle to matrix factorization, proposing the SPMF method, which showed improved performance over classical techniques. Similarly, SPL has been incorporated into NMF-based models. Zhu and Zhang [35] proposed MSPNMF, which integrates SPL with Frobenius norm-based NMF. However, the reliance on the Frobenius norm still leaves the model susceptible to noisy inputs. To overcome this, Huang et al. [36] embedded SPL into the $L_{2,1}$-NMF framework, achieving enhanced robustness by reducing the influence of outliers during training.

In addition to the above-mentioned robust NMF approaches, robustness has also been studied in related representation learning frameworks such as robust principal component analysis (RPCA) [37] and robust matrix factorization [38], which explicitly model outliers or corrupted components. Within NMF, the current state-of-the-art can be divided into three main lines: (i) loss-based variants using L_1, $L_{2,1}$, Huber, or correntropy-based measures; (ii) distributionally robust formulations that optimize performance across ambiguity sets of noise distributions; and (iii) adaptive extensions such as self-paced NMF. Our work combines (ii) and (iii), providing a principled framework that enhances both robustness and optimization stability.

This paper introduces a novel model termed *Distributionally Robust Nonnegative Matrix Factorization with Self-Paced Adaptive Multi-Loss Fusion* (DRNMF-SP), aimed at delivering resilience to a diverse range of noise characteristics, including severe and heavy-tailed outliers. Formulated within a multi-objective optimization framework, DRNMF-SP is particularly suited for robust data representation under complex and uncertain conditions. The model jointly leverages two key principles: distributional robustness to account for varying noise distributions, and self-paced learning to suppress the impact of highly contaminated or anomalous data samples. A major difficulty in robust learning arises from outliers caused by heavy-tailed noise distributions, which can severely distort factorization quality. Existing robust NMF variants often fail to effectively mitigate such extreme deviations. By incorporating SPL, our approach gradually integrates samples into the learning process based on their reconstruction difficulty, thus naturally down-weighting samples with large residuals. This learning scheme not only improves robustness to outliers but also avoids dominance of any single objective function during the optimization process, which is particularly important when fusing multiple loss functions. Moreover, DRNMF-SP operates at the sample level, distinguishing it from conventional robust NMF methods that assume noise affects individual matrix entries. Our formulation evaluates the reconstruction fidelity using an instance-wise loss structure, allowing it to capture global sample-level corruption rather than isolated entry-wise distortions. In addition, the proposed model introduces a general and extensible strategy for constructing distributionally robust NMF formulations by integrating a user-defined collection of loss functions. Optimization is carried out using an efficient iterative reweighted scheme, which maintains computational complexity on par with classical NMF algorithms. This flexibility enables DRNMF-SP to seamlessly encompass a broad spectrum of robust objectives commonly encountered in literature, while offering scalability and simplicity in implementation. The main contributions of this work are summarized as follows:

- We propose a distributionally robust NMF model (DRNMF-SP) using a multi-objective framework that minimizes a worst-case expected loss under probabilistic ambiguity. This formulation accommodates the uncertainty in noise characteristics by combining multiple objective functions.
- Self-paced learning (SPL) is integrated to enhance robustness against both moderate and extreme outliers. SPL allows the model to learn from easier, cleaner samples first, gradually incorporating harder ones, improving resilience to heavy-tailed noise.
- In multi-objective settings, SPL ensures balanced optimization by preventing any single objective from dominating, maintaining diversity in learned solutions, and robustness in performance.
- Extensive experiments on benchmark datasets validate that DRNMF-SP consistently outperforms existing robust NMF models under various noise conditions, including mixed and single-distribution noise scenarios.

The proposed factorization model is general and can handle any noise type, but we focus on common real-world distributions, Gaussian, Laplacian, and Cauchy. Due to their heavy tails, Laplace and Cauchy distributions often contain more outliers. Our DRNMF-SP model is designed to robustly manage both moderate and extreme outliers across these noise types. To better position our work against recent advances, we emphasize the distinction of DRNMF-SP from the most relevant robust NMF models. The element-wise distributionally robust NMF (DRNMF) [21] achieves robustness by optimizing over an ambiguity set of noise distributions, yet it remains limited to entry-level corruption and does not address sample-wise noise patterns or extreme outliers. The instance-wise

DRNMF (iDRNMF) [22] extends this idea by introducing adaptive multi-loss fusion, enabling resilience across heterogeneous noise types, but it still lacks a mechanism to mitigate convergence to poor local minima in highly noisy settings. On the other hand, self-paced learning–based NMF variants such as SPLNMF [36] improve convergence stability and suppress outlier effects by gradually introducing difficult samples, but they are restricted to a single loss function and therefore cannot guarantee robustness under distributional uncertainty. In contrast, our proposed DRNMF-SP unifies these complementary directions: it combines distributionally robust optimization with self-paced adaptive multi-loss fusion at the instance level, thereby capturing sample-wise corruption, balancing multiple objectives during training, and progressively filtering heavy-tailed outliers. This integration yields a principled and generalizable framework that advances beyond existing robust NMF approaches in both theoretical formulation and empirical performance.

The rest of the paper is organized as follows: Section 2 covers background and preliminaries. Section 3 introduces the instance-wise distributionally robust NMF (iDRNMF), its extension with self-paced learning (DRNMF-SP), and a reweighted optimization strategy. Section 4 presents experimental results, and Section 5 concludes with a summary and future directions.

2. Preliminaries

In this section, we provide a review of key preliminaries, including the iterative reweighted algorithm, commonly employed for solving the general reconstruction problem. We also introduce the standard NMF and describe self-paced learning. Before introducing them, we provide the basic notation useful for understanding the paper.

2.1. Notation

In this paper, we follow a standard notation for mathematical symbols. Scalars are represented using lowercase letters, while vectors are denoted by lowercase letters in boldface. Matrices are indicated by uppercase letters. For a matrix M, we denote its i-th column as M_i, its j-th row as $M_{(j)}$, and the entry in the i-th row and j-th column as M_{ij}. The transpose of M is written as $M^\top$, and its trace is given by $\mathrm{Tr}(M)$. The Frobenius norm of a matrix $M \in \mathbb{R}^{m\times n}$ is defined as

$$\|M\|_F = \sqrt{\sum_{j=1}^{m}\sum_{i=1}^{n} M_{ji}^2} = \sqrt{\mathrm{Tr}(M^\top M)} = \sqrt{\mathrm{Tr}(M M^\top)}.$$

2.2. Iterative reweighted algorithm

The iterative reweighted least squares (IRLS) method is widely used for optimizing robust models. It avoids direct minimization of non-quadratic objectives by reformulating them as a series of weighted least squares problems. Let $e_i(v)$ denote the reconstruction error of the i-th instance, where v represents the model parameters. A general reconstruction problem can thus be written as:

$$\min_{v} \sum_{i=1}^{n} \Xi(e_i(v)), \tag{1}$$

where $\Xi(\cdot)$ denotes a monotonically increasing function applied to the nonnegative reconstruction error $e_i(v) \geq 0$, and the parameter vector $v = [v_1, v_2, \ldots, v_p]^\top$ includes the p variables to be estimated in solving problem (1). An optimal solution is obtained by setting the derivative of the objective in (1) to zero,

$$\sum_{i=1}^{n} \omega\left(e_i(v)\right) \frac{\partial e_i(v)}{\partial v_j} = 0, \quad j = 1, 2, \ldots, p, \tag{2}$$

where $\omega(e_i(v)) = \Xi'(e_i(v)) = \frac{d\Xi(e_i(v))}{d e_i(v)}$ is called the influence function. Furthermore, Eq. (2) can be rewritten as:

$$\sum_{i=1}^{n} \psi(e_i(v)) e_i(v) \frac{\partial e_i(v)}{\partial v_j} = 0, \quad j = 1, 2, \ldots, p, \tag{3}$$

where $\psi(e_i(v))$ is called the weight function. For effective optimization, the influence and weight functions in Eq. (3) are designed to ensure stability and convergence. When appropriately chosen, they allow Eq. (3) to be solved via the following iterative reweighted formulation [39]:

$$\min_{v} \sum_{i=1}^{n} \psi(e_i(v)^{[t-1]}) e_i(v)^2, \tag{4}$$

Here, $e_i(v)^{[t-1]}$ denotes the reconstruction error at the $(t-1)$-th iteration. Solving problem (4) proceeds iteratively in two steps. First, treat the weight $\psi(e_i(v)^{[t-1]})$ as fixed and solve for the optimal parameters based on the specific structure of (4). Then, update the weight using the current reconstruction error $e_i(v)^{[t]}$.

2.3. Iterative reweighted NMF

Given a nonnegative data matrix $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{m\times n}$, NMF seeks nonnegative matrices $W \in \mathbb{R}^{m\times r}$ and $H \in \mathbb{R}^{r\times n}$ such that $X \approx WH$. The general objective is formulated as

$$\min_{W,H} \sum_{i=1}^{n} \Xi(e(x_i, Wh_i)) \quad \text{s.t.} \quad W, H \geq 0, \tag{5}$$

where Ξ is a loss function applied to the reconstruction error $e(x_i, Wh_i)$. The basic NMF model uses the squared error, leading to:

$$\min_{W,H} \|X - WH\|_F^2 = \sum_{i=1}^{n} \|x_i - Wh_i\|^2 \quad \text{s.t. } W, H \geq 0. \tag{6}$$

Since the objective is bi-convex, alternating minimization methods are typically employed. Specifically, throughout each iteration, one of the two factors is held constant while the other is updated in such a way that decreases the objective function. A connection was established between the iterative reweighted algorithm and NMF, regardless of the used loss function [40]. More precisely, the NMF loss function and $\|x_i - Wh_i\|$ can be considered as $\Xi(\cdot)$ function and e_i in Eq. (1), respectively. Thus, in an NMF framework, Eq. (1) can be rewritten as:

$$\min_{W,H} \sum_{i=1}^{n} d_i \|x_i - Wh_i\|^2 = \mathrm{Tr}\left[(X - WH)D(X - WH)^\top\right], \quad \text{s.t. } W, H \geq 0, \tag{7}$$

where $d_i = \psi(\|x_i - Wh_i\|^{[t-1]})$ is a coefficient computed in each iteration according to the loss used by Eq. (3) and using the residue value in the previous iteration and D is a diagonal matrix with $D_{ii} = d_i$.

2.4. Self-paced learning

Self-Paced Learning is inspired by the way humans acquire knowledge—starting from simpler concepts and progressively moving toward more difficult ones. Unlike standard training approaches that treat all data equally, SPL prioritizes samples based on their reliability and difficulty. In classical learning frameworks, model parameters Θ are obtained by minimizing the empirical risk over all samples:

$$\min_{\Theta} \sum_{i=1}^{n} \ell_i(x_i, \Theta), \tag{8}$$

where $\ell_i(x_i, \Theta)$ denotes the loss associated with sample x_i, and Θ represents the model's parameters. However, not all data points are equally informative—especially in the presence of noise or outliers—so treating every instance uniformly can be suboptimal.

To address this, SPL introduces a mechanism that favors easier samples in the early stages of training. It jointly optimizes the model parameters Θ and a sample weighting vector $p = [p_1, \dots, p_n]^\top$ using the following objective:

$$\min_{(\Theta,p)} \sum_{i=1}^{n} p_i \ell_i(x_i, \Theta) + f(\alpha, p), \tag{9}$$

where each p_i quantifies the importance (or simplicity) of sample x_i, and $f(\alpha, p)$ is a self-paced regularizer controlled by the age parameter α. At early stages (small α), the algorithm focuses on samples with lower loss values. As training progresses and α increases, more complex instances are gradually included, allowing the model to mature over time [31].

The optimization of (9) typically proceeds by alternating updates of Θ and p. A common form of SPL, often referred to as hard self-paced learning, restricts the weights to binary values $p \in \{0, 1\}^n$ and defines the regularization term as:

$$f(\alpha, p) = -\alpha \sum_{i=1}^{n} p_i, \tag{10}$$

with the optimal value of p_i given by:

$$p_i^* = \begin{cases} 1, & \text{if } \ell_i < \alpha, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

In this scheme, a sample is selected for training only if its loss is below the current threshold α, which is increased gradually to incorporate more difficult examples in successive iterations.

3. Proposed model: DRNMF-SP

In this section, we introduce instance-wise Distributionally Robust Nonnegative Matrix Factorization (iDRNMF) with the incorporation of the Self-Paced Learning (SPL) framework, which fundamentally improves the learning process by adapting the model's sample selection strategy. The proposed method robustly represents data at the instance level, effectively managing noise from various distributions by dynamically balancing multiple objectives. The section begins with the formulation of the iDRNMF model, followed by the integration of SPL for iterative sample selection, and concludes with the optimization techniques employed, ensuring stable and generalizable data representations.

3.1. Instance-wise distributionally robust NMF

The iDRNMF model, introduced in [22], was developed to address the challenge of modeling diverse noise distributions in nonnegative matrix factorization. Its unified framework allows for the integration of multiple objective functions, each associated with a distinct noise distribution, as commonly studied in the distributionally robust optimization literature. Unlike element-wise models, iDRNMF adopts an instance-wise perspective, treating each column of the data matrix X as an individual data sample. Based on this view, a generalized loss function $\|\cdot\|_{(2,\tau)}$ was proposed, which applies the L_2 norm to each column of the residual matrix $E = X - WH$, followed by a loss function τ on the resulting values. Here, each τ is assumed to be related to a specific probability distribution, allowing $\|\cdot\|_{(2,\tau)}$ to reflect different types of noise.

Since the true underlying noise distribution is unknown but assumed to belong to a predefined ambiguity set Ω, a dynamic weighted sum of objective functions is employed, where the weights are optimized jointly. This formulation enhances robustness against a variety of noise types by solving the following min-max problem:

$$\min_{(W,H)} \max_{\lambda} \sum_{\tau\in\Omega} \lambda_\tau \|X - WH\|_{(2,\tau)}, \quad \text{s.t.} \quad W, H, \lambda \geq 0,\ \|\lambda\|_1 = 1, \tag{12}$$

where $\lambda \in \mathbb{R}^{|\Omega|}$ is a nonnegative weight vector summing to one, and each λ_τ reflects the contribution of the corresponding loss. In [22], the ambiguity set was chosen as $\Omega = \{1, 2, \text{cau}\}$ to ensure robustness under Laplacian, Gaussian, and Cauchy noise, including their mixtures. In this case, the formulation in Eq. (12) becomes:

$$\min_{(W,H)} \max_{\lambda} \lambda_1 \|X - WH\|_{2,1} + \lambda_2 \|X - WH\|_{2,2}^2 + \lambda_{\text{cau}} \|X - WH\|_{(2,\text{cau})}, \quad \text{s.t.} \quad W, H, \lambda \geq 0,\ \|\lambda\|_1 = 1. \tag{13}$$

This objective can also be represented in an instance-wise form, where x_i and h_i denote the i-th column of X and H, respectively:

$$\min_{(W,H)} \max_{\lambda} \left(\lambda_1 \sum_{i=1}^{n} \|x_i - Wh_i\| + \lambda_2 \sum_{i=1}^{n} \|x_i - Wh_i\|^2 + \lambda_{\text{cau}} \sum_{i=1}^{n} \ln\left(\|x_i - Wh_i\|^2 + \gamma^2\right) \right), \quad \text{s.t.} \quad W, H, \lambda \geq 0, \quad \|\lambda\|_1 = 1. \tag{14}$$

To solve this multi-objective formulation, the optimization process was designed to gradually reduce the total loss while prioritizing the term with the largest contribution in each iteration. Specifically, the coefficient of the most dominant error term is increased to focus more on reducing that loss, while others are adjusted downward accordingly. Over iterations, this strategy ensures balanced optimization across all loss components and convergence toward a stable solution. However, it was observed in [22] that due to the squared term in the Frobenius norm, the $\|\cdot\|_{2,2}$ loss tends to produce significantly higher values, while the Cauchy-based term $\|\cdot\|_{(2,\text{cau})}$ remains consistently smaller. This discrepancy causes an imbalance, making it difficult to transition smoothly between loss terms during optimization. To resolve this, the objective functions were normalized so that each term contributes comparably. This normalization is crucial for the iDRNMF model to avoid biased optimization and leverage the strengths of all constituent objectives. Following the procedure in [21], each single-objective problem $\zeta_\tau = \min_{(W,H\geq 0)} \|X - WH\|_{2,\tau}$ for $\tau \in \{1, 2, \text{cau}\}$ was solved individually. These values were then used to rescale the corresponding loss terms in the combined objective. Accordingly, the normalized instance-wise formulation becomes:

$$\min_{(W,H)} \max_{\lambda} \left(\frac{\lambda_1}{\zeta_1} \sum_{i=1}^{n} \|x_i - Wh_i\| + \frac{\lambda_2}{\zeta_2} \sum_{i=1}^{n} \|x_i - Wh_i\|^2 + \frac{\lambda_{\text{cau}}}{\zeta_{\text{cau}}} \sum_{i=1}^{n} \ln\left(\|x_i - Wh_i\|^2 + \gamma^2\right) \right), \quad \text{s.t.} \quad W, H, \lambda \geq 0, \quad \|\lambda\|_1 = 1. \tag{15}$$

This normalization ensures that the optimization does not become biased toward the objective with the largest absolute value. As a result, the model in Eq. (15) achieves a balanced trade-off among different loss functions. By accurately modeling the ambiguity set and maintaining robustness across heterogeneous noise conditions, the iDRNMF formulation provides more generalizable and reliable representations, as demonstrated in [22]. In this work, we select the ambiguity set $\Omega = \{1, 2, \text{cau}\}$, corresponding to Gaussian, Laplacian, and Cauchy noise. This choice is motivated by their complementary tail behaviors: the Gaussian distribution is light-tailed (mesokurtic) and represents the classical baseline; the Laplacian distribution is moderately heavy-tailed (leptokurtic), producing moderate outliers; and the Cauchy distribution is extremely heavy-tailed, generating extreme outliers. Together, they provide a representative spectrum of light, moderate, and severe contamination scenarios, enabling a meaningful case study for evaluating the robustness of the proposed DRNMF-SP framework. While we focus on this triplet for interpretability, the formulation is general and can be extended to other distributions in future work.

3.2. Distributionally robust NMF with self-paced adaptive multi-loss fusion

The iDRNMF framework offers a principled approach to NMF, aiming to decompose a given matrix X into nonnegative factors W and H, while exhibiting robustness against a wide range of noise distributions and moderate outliers. By minimizing the reconstruction error $\|X - WH\|$ under multiple loss functions, iDRNMF provides a powerful tool for data analysis across diverse scenarios. Its inherent robustness against noise and moderate outliers makes it particularly valuable in real-world applications, but it can still be trapped in local minima. Integrating SPL with iDRNMF enhances the model's generalization by helping it avoid suboptimal solutions, such as poor local minima.

Moreover, SPL significantly boosts the model's robustness against irregular data, including noise and outliers. While the basic iDRNMF framework is already robust against diverse noise and moderate outliers, SPL's structured approach further enhances performance. By gradually exposing the model to increasingly complex samples, starting with simpler, more regular data points, SPL

helps the model build a strong representation before tackling more challenging examples. This approach is particularly effective with data polluted by heavy-tailed noise distributions, as it improves the model's ability to distinguish between regular and irregular data, including extreme outliers and anomalies. Additionally, SPL allows for the optional exclusion of the most challenging samples, providing a comprehensive solution for managing noise and outliers in data analysis tasks. The combination of SPL and iDRNMF, therefore, results in the DRNMF-SP framework, offering advanced robustness and performance. We rewrite Eq. (12), combining SPL using Eqs. (9) and (10) as follows:

$$\min_{(W,H,p)} \max_{\lambda} \sum_{\tau\in\Omega} \lambda_\tau \sum_{i=1}^{n} p_i^{(\tau)} \mathcal{L}_{2,\tau}\left(\|x_i - Wh_i\|\right) - \alpha^{(\tau)} \sum_{i=1}^{n} p_i^{(\tau)}, \quad \text{s.t.} \quad W, H, \lambda \geq 0,\ \|\lambda\|_1 = 1. \tag{16}$$

Since the number of instances included in each objective $\mathcal{L}_{2,\tau}$ can differ in each iteration, directly summing them with a constant normalization coefficient ζ_τ would unfairly weight objectives that take into account more instances. Normalization by the total number of instances ensures that all objectives contribute equally to the final objective function, regardless of the specific instance allocation in a particular iteration. This facilitates a fairer comparison between the different objectives and avoids biases introduced by imbalanced instance usage.

Fig. 1 illustrates the overall structure of the proposed model, highlighting its main components, data flow, and functional blocks. In addition, Fig. 2 illustrates the behavior of the proposed DRNMF-SP model with self-paced filtering under three different loss functions (L_1, L_2, and Cauchy) across three stages (t_1, t_2, t_3). At each stage, the filtering mechanism gradually selects "easy" (clean) samples, shown as solid points, while discarding "hard" (noisy or outlier) samples, shown as transparent points. As training progresses from t_1 to t_3, more samples are incorporated, reflecting the curriculum learning effect of self-paced filtering. The comparison across loss functions highlights their different robustness properties, allowing the model to effectively adapt to various noise distributions. Overall, these plots demonstrate that the integration of distributionally robust losses with self-paced filtering not only enables the model to handle different types of noise but also significantly reduces the influence of outliers on the reconstruction quality.

For $\Omega = \{1, 2, \text{cau}\}$, the objective function is as follows:

$$\begin{aligned}\min_{(W,H,p)} \max_{\lambda} \Bigg(& \lambda_1 \sum_{i=1}^{n} \frac{p_i^{(1)}}{\epsilon_1} \|x_i - Wh_i\| + \lambda_2 \sum_{i=1}^{n} \frac{p_i^{(2)}}{\epsilon_2} \|x_i - Wh_i\|^2 + \lambda_{\text{cau}} \sum_{i=1}^{n} \frac{p_i^{(\text{cau})}}{\epsilon_{\text{cau}}} \ln\left(\|x_i - Wh_i\|^2 + \gamma^2\right) \\ & - \alpha^{(1)} \sum_{i=1}^{n} p_i^{(1)} - \alpha^{(2)} \sum_{i=1}^{n} p_i^{(2)} - \alpha^{(\text{cau})} \sum_{i=1}^{n} p_i^{(\text{cau})} \Bigg), \quad \text{s.t.} \quad W, H, \lambda \geq 0,\ \|\lambda\|_1 = 1,\end{aligned} \tag{17}$$

where $\epsilon_\tau = \zeta_\tau \cdot \sum_{i=1}^{n} p_i^{(\tau)}$ is an instance-wise normalization coefficient. At each iteration, we calculate $p_i^{(\tau)}$ as follows:

$$p_i^{(\tau)} = \begin{cases} 1 & \text{if } e_i^{(\tau)} \leq \alpha^{(\tau)} \times \frac{\epsilon^{(\tau)}}{\lambda_\tau}, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

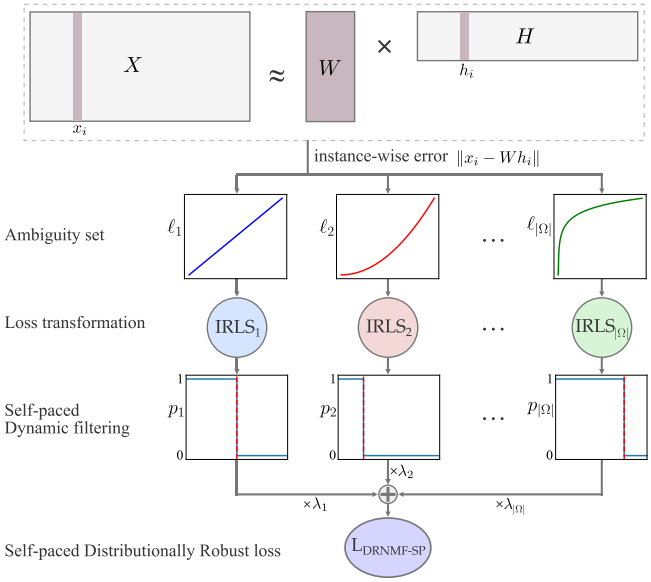


Fig. 1. Overview of the Distributionally Robust NMF with Self-Paced learning (DRNMF-SP), showing the main components and data flow.

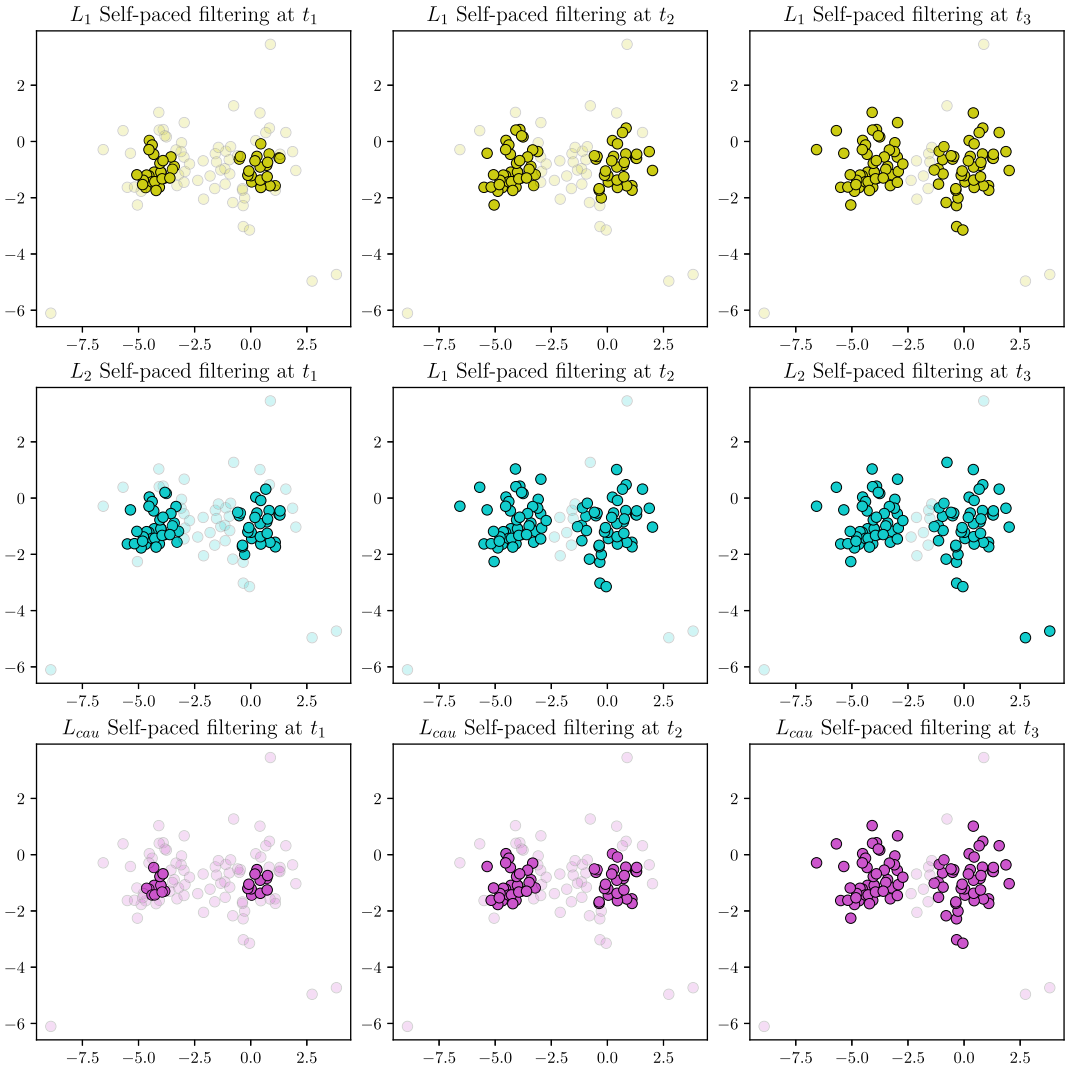


Fig. 2. Self-paced filtering with DRNMF under L_1, L_2, and Cauchy losses, showing progressive inclusion of clean samples (solid) and suppression of noisy ones (transparent) from t_1 to t_3.

Notice that in each iteration of the learning process, an increase in the value of λ_τ indicates a larger residual error for the corresponding loss function. Consequently, the threshold level becomes stricter, resulting in fewer samples being incorporated into the training process for that particular function. In contrast, a decrease in the value of λ_τ leads to a less stringent threshold, allowing more samples to contribute to the training process. Due to the inherent difficulty of optimizing the nonconvex and nonlinear objective function (17), we reformulate the model as an iteratively reweighted NMF problem.

It is important to note that DRNMF-SP directly extends the iDRNMF framework [22]. While iDRNMF already achieves robustness through adaptive multi-loss fusion under distributional uncertainty, it remains vulnerable to convergence toward poor local optima and struggles with extreme outliers. By embedding self-paced learning, DRNMF-SP introduces a principled mechanism to gradually incorporate more complex and noisy samples into training, which mitigates these weaknesses. A detailed empirical comparison between DRNMF-SP and iDRNMF, including their relative performance under extreme outlier conditions, is presented in Section 4.6.

3.3. Iterative reweighted DRNMF-SP

In subsections 2.2 and 2.3, we presented an iterative reweighted algorithm and showed that, based on it, certain objective functions satisfying specific conditions can be converted into a weighted basic NMF. Therefore, by utilizing the reweighted framework (7), our multi-objective formulation can also be converted as follows:

$$\min_{(W,H,p)} \max_{\lambda} \left(\sum_{i=1}^{n} d_i^{(\Omega)} \|x_i - W h_i\|^2 - \alpha^\tau \sum_{i=1}^{n} p_i^{(\tau)} \right), \quad \text{s.t.} \quad W, H, \lambda \geq 0, \quad \|\lambda\|_1 = 1, \tag{19}$$

where $d_i^{(\Omega)}$ is an instance weight calculated according to Eq. (4) as follows:

$$d_i^{(\Omega)} = \sum_{\tau \in \Omega} \frac{\lambda_\tau p_i^{(\tau)}}{\epsilon_\tau} \psi_\tau \left(\|x_i - W h_i\|^{[t-1]} \right). \tag{20}$$

Indeed, in Eq. (19) each part of the objective function is assigned a weight $d^{(\tau)}$. Consequently, we can rewrite it as:

$$\min_{(W,H,p)} \max_{\lambda} \sum_{\tau \in \Omega} \frac{\lambda_\tau p_i^{(\tau)}}{\epsilon_\tau} \sum_{i=1}^{n} d_i^{(\tau)} \|x_i - W h_i\|^2 - \alpha^\tau \sum_{i=1}^{n} p_i^{(\tau)}, \quad \text{s.t.} \quad W, H, \lambda \geq 0, \quad \|\lambda\|_1 = 1, \tag{21}$$

where the sample weight is $d_i^{(\tau)} = \psi_\tau \left(\|x_i - W h_i\|^{[t-1]} \right)$. If $\Omega = \{1, 2, \text{cau}\}$, we have:

$$\begin{aligned} \min_{(W,H,p)} \max_{\lambda} \Bigg(& \frac{\lambda_1 p_i^{(1)}}{\epsilon_1} \sum_{i=1}^{n} d_i^{(1)} \|x_i - W h_i\|^2 + \frac{\lambda_2 p_i^{(2)}}{\epsilon_2} \sum_{i=1}^{n} \|x_i - W h_i\|^2 + \frac{\lambda_{\text{cau}} p_i^{(\text{cau})}}{\epsilon_{\text{cau}}} \sum_{i=1}^{n} d_i^{(\text{cau})} \|x_i - W h_i\|^2 \\ & - \alpha^{(1)} \sum_{i=1}^{n} p_i^{(1)} - \alpha^{(2)} \sum_{i=1}^{n} p_i^{(2)} - \alpha^{(\text{cau})} \sum_{i=1}^{n} p_i^{(\text{cau})} \Bigg) \quad \text{s.t.} \quad W, H, \lambda \geq 0, \quad \|\lambda\|_1 = 1, \end{aligned} \tag{22}$$

where $d_i^{(1)} = \frac{1}{\|x_i - W h_i\|}$ and $d_i^{(\text{cau})} = \frac{1}{\|x_i - W h_i\|^2 + \gamma^2}$. Finally, we can rearrange the formula as follows:

$$\begin{aligned} & \min_{(W,H,p)} \max_{\lambda} \sum_{i=1}^{n} \left(\frac{\lambda_1 p_i^{(1)} d_i^{(1)}}{\epsilon_1} + \frac{\lambda_2 p_i^{(2)}}{\epsilon_2} + \frac{\lambda_{\text{cau}} p_i^{(\text{cau})} d_i^{(\text{cau})}}{\epsilon_{\text{cau}}} \right) \|x_i - W h_i\|^2 - \alpha^{(1)} \sum_{i=1}^{n} p_i^{(1)} - \alpha^{(2)} \sum_{i=1}^{n} p_i^{(2)} - \alpha^{(\text{cau})} \sum_{i=1}^{n} p_i^{(\text{cau})} \\ & = \min_{(W,H,p)} \max_{\lambda} \sum_{i=1}^{n} d_i^{(\Omega)} \|x_i - W h_i\|^2 - \alpha^{(\tau)} \sum_{i=1}^{n} p_i^{(\tau)} \quad \text{s.t.} \quad W, H, \lambda \geq 0, \quad \|\lambda\|_1 = 1. \end{aligned} \tag{23}$$

It is important to mention that we can expand ω to encompass any preferred distribution and utilize this unified weighted formulation to manage it.

3.4. Optimization

The cost function of our DRNMF-SP model is not jointly convex in W and H, which makes optimization challenging. To facilitate successful factorizations, we can break this problem into two smaller convex subproblems. We can tackle problem (23) using alternating minimization (Algorithm 1), which allows us to iteratively refine the variables until we reach a satisfactory solution. In each iteration, we use the Multiplicative Update Rule (MUR) to update one factor while keeping the other fixed.

3.4.1. Updating factors

To derive the update rules for the W and H factors, we rewrite the objective function in trace form. This allows us to solve it using the MUR method within a weighted NMF framework:

$$\min_{W,H,p} \sum_{i=1}^{n} d_i^{(\Omega)} \|x_i - W h_i\|^2 - \alpha^\tau \sum_{i=1}^{n} p_i^{(\tau)} = \text{Tr}\left[(X - WH) D (X - WH)^\top \right] - \alpha^\tau \sum_{i=1}^{n} p_i^{(\tau)}, \quad \text{s.t.} \quad W, H \geq 0. \tag{24}$$

where $D_{ii} = d_i^{(\Omega)}$ can be computed as:

$$D_{ii} = \frac{\lambda_1 p_i^{(1)}}{\epsilon_1 \|x_i - W h_i\|} + \frac{\lambda_2 p_i^{(2)}}{\epsilon_2} + \frac{\lambda_{\text{cau}} p_i^{(\text{cau})}}{\epsilon_{\text{cau}} \|x_i - W h_i\|^2 + \gamma^2}. \tag{25}$$

Algorithm 1 Distributionally Robust NMF with Self-Paced adaptive multi-loss fusion (DRNMF-SP).

Input: data matrix X, rank r, self-paced parameters α^τ, a finite ambiguity set Ω;
Output: basis matrix W and representation matrix H;
Initialize W and H randomly, $\lambda_\tau^{[0]} = \frac{1}{|\Omega|}$ $\forall \tau \in \Omega$;
Calculate ϵ_τ for all objectives;
while convergence not reached **do**
  Calculate $p_i^{(\tau)}$ according to (18);
  Update instance weight matrix D according to (25);
  Update basis matrix W according to (26);
  Update representation matrix H according to (26);
  Update weights $\lambda^{[t+1]}$ according to (27);
end while

To solve (24) with the non-negativity constraint, we use the following updating rules for W and H:

$$W \leftarrow W \odot \frac{XDH^\top}{WHDH^\top}, \quad H \leftarrow H \odot \frac{W^\top XD}{W^\top WHD}, \tag{26}$$

where the division is element-wise, and each entry of the numerator is divided by the corresponding entry of the denominator.

3.4.2. Updating weights

To solve the optimization problem in (16), we adopt an iterative strategy that prioritizes the worst-case objective at each step. Specifically, we identify the maximum loss value and focus on minimizing it in the next iteration. This encourages all objectives to be reduced, while giving extra emphasis to the largest one. Based on [21], we initialize the weights as $\lambda_\tau^{[0]} = \frac{1}{|\Omega|}$ for all $\tau \in \Omega$. At each iteration t, we find:

$$p^* = \arg\max_{\tau \in \Omega} \|X - WH\|_{(2,\tau)},$$

and define $\lambda_*^{[t]}$ as a one-hot vector with value 1 at position p^*. The weights are then updated via:

$$\lambda^{[t+1]} = (1-\eta)\lambda^{[t]} + \eta\lambda_*^{[t]}. \tag{27}$$

Since $\|\lambda\| = 1$, this update increases the weight corresponding to the highest loss while slightly decreasing the others. The parameter $\eta \in (0,1)$ controls the adjustment rate in the adaptive weight update scheme: larger values give more priority to the maximum loss, whereas smaller values result in slower updates. In our implementation, η is not treated as a fixed hyperparameter but is instead updated dynamically according to the iteration number t. Specifically, we set

$$\eta^{[t]} = \frac{1}{t+1},$$

which means that η starts from $\frac{1}{2}$ in the first iteration and gradually decreases toward zero as t increases. This schedule allows the algorithm to initially focus more on the dominant loss while ensuring that no single objective completely dominates the optimization process in later iterations.

3.5. Computational complexity

Assume $X \in \mathbb{R}_+^{m\times n}$ with m features, n samples, and target rank r. Each iteration of DRNMF-SP involves the following major steps: (i) computing residuals and sample-wise norms, (ii) constructing the diagonal weight matrix D, (iii) updating self-paced weights $p^{(\tau)}$ for each loss component, and (iv) updating the factors W and H via weighted multiplicative rules. As in standard NMF, the dominant operations are matrix–matrix multiplications of size $m \times n$ with rank r. The additional instance-weighting and self-paced modules introduce only linear-time overhead in n, which does not affect the asymptotic order. Table 1 summarizes the arithmetic costs for standard NMF and DRNMF-SP. For DRNMF-SP, the extra computations of D and p contribute $O(mnr)$ and $O(n|\Omega|)$ respectively, where $|\Omega|$ is the number of candidate loss functions in the ambiguity set (usually small). Overall, these terms are dominated by the $O(mnr)$ cost of the update rules. Table 1 shows that both methods share the same asymptotic order $O(mnr)$. The running-time results in Section 4.9 confirm the theoretical complexity, showing that DRNMF-SP remains as efficient as standard NMF.

4. Experimental results

In this section, we comprehensively evaluate the robustness and effectiveness of the proposed method by conducting extensive experiments on 12 benchmark datasets and comparing it with 10 baseline and state-of-the-art NMF-based methods. To ensure a fair and reliable comparison, each method is executed 10 times with different random initializations to mitigate the impact of initialization sensitivity, and we report the average results. The MUR for factor matrices is performed for 300 iterations. For each compared method, we carefully set the hyperparameters according to the original papers in which they were first introduced, ensuring consistency with prior studies. The number of latent components is fixed to match the number of clusters in each dataset. To evaluate clustering performance, we apply the standard k-means algorithm to the learned representation matrix H. Performance is assessed using three widely adopted clustering evaluation metrics [41]: Normalized Mutual Information (NMI), Accuracy (ACC), and Adjusted Rand Index (ARI). These metrics provide a comprehensive assessment of clustering quality by measuring the consistency between predicted labels

Table 1
Per-iteration cost (arithmetic order) for standard NMF and DRNMF-SP.

Operation	Standard NMF	DRNMF-SP
Update W (MUR)	$O(mnr)$	$O(mnr)$
Update H (MUR)	$O(mnr)$	$O(mnr)$
Form residuals / norms	$O(mn)$	$O(mn)$
Compute diagonal weights D	–	$O(mnr)$
Compute SPL sample weights $p^{(\tau)}$	–	$O(n\|\Omega\|)$
Update λ	–	$O(\|\Omega\|)$
Per-iteration total	$O(mnr)$	$O(mnr)$

and ground truth annotations. By adhering to these rigorous experimental protocols, we ensure a robust and meaningful comparison of the proposed method against existing approaches. All datasets were preprocessed following standard practices: the image datasets were converted to grayscale (if applicable) and resized to the dimensions specified in Section 4.1. The non-image datasets (Seeds, Ecoli) were normalized feature-wise to the [0,1] range for consistent scaling across attributes. For all datasets, the number of latent components r was set equal to the number of ground-truth classes. The factor matrices were initialized randomly and updated using the multiplicative update rules for 300 iterations. For the Cauchy loss, we set $\gamma = 1$, following prior works. Unless otherwise stated, all other hyperparameters were fixed across datasets. For baseline methods, we used the recommended hyperparameters reported in their original papers.

4.1. Datasets

We evaluated our methods on 12 benchmark datasets from various domains, including face recognition (Yale, ORL, UMIST), object recognition (COIL20), handwriting digit and fashion classification (MNIST, USPS, Fashion-MNIST), biological analysis (Seeds, Ecoli), and medical imaging (OrganA, Blood, Pneumonia). Face datasets comprise grayscale images with varying illumination and perspectives, resized for computational efficiency (Yale, ORL: 32×32; UMIST: 28×23). COIL20 features object images resized to 32×32 pixels. MNIST, USPS, and Fashion-MNIST were subsetted to 1000–1100 samples each for streamlined testing. Biological datasets (Seeds: 210 samples, 7 attributes; Ecoli: 336 samples, 7 features) cover distinct classification tasks. The medical datasets from MedMNIST [42] (OrganA: 58,830 CT images; Blood: 17,092 RGB microscope images; Pneumonia: 5856 chest X-rays) highlight scalability and adaptability to high-dimensional, complex data. Fig. 3 presents representative dataset samples.

By leveraging this diverse set of datasets, including those with high-dimensional feature spaces and large sample sizes, we ensure a rigorous and comprehensive evaluation of the robustness and effectiveness of the proposed method in various real-world applications. The details of the datasets are described and summarized in Table 2.

4.2. Comparison of methods

To verify the superior performance of the proposed DRNMF-SP method for data representation, we compare it against 10 NMF models, including conventional, element-wise, sample-wise, and deep learning-based approaches. These methods represent a diverse set of techniques designed to enhance the robustness of NMF in different ways:

- **Frobenius-NMF** [43]: The standard NMF formulation that minimizes the squared error loss. It serves as a benchmark for evaluating improvements brought by robust techniques.
- $L_{2,1}$**-NMF** [10]: A robust formulation of NMF that replaces the conventional least squares loss with an $L_{2,1}$ norm-based loss.

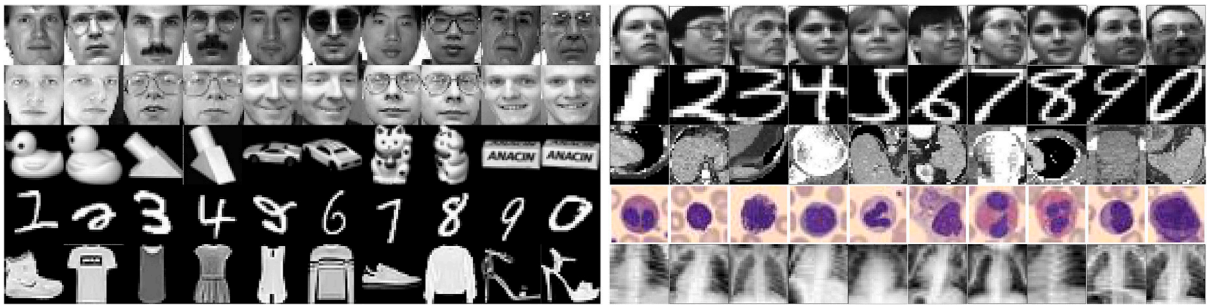


Fig. 3. Example images from ten benchmark datasets: Yale, ORL, COIL20, MNIST, Fashion-MNIST, UMIST, USPS, OrganA, Blood, and Pneumonia, shown in order from top to bottom and left to right.

Table 2
Descriptions of 12 datasets under test.

Dataset	#Sample	#Feature	#Class	Application
Yale	165	1024	15	Face
ORL	400	1024	40	Face
COIL20	1440	1024	20	Object
MNIST	1000	784	10	Handwriting digit
Fashion	1000	784	10	Cloth
Seeds	210	7	3	Biology
Ecoli	336	7	8	Biology
USPS	1100	256	10	Handwriting digit
UMIST	575	644	20	Face
OrganA	58,830	784	11	Abdominal CT
Blood	17,092	2352	8	Blood Cell Microscope
Pneumonia	5,856	784	2	Chest X-Ray

- **Cauchy-NMF** [13]: An element-wise approach that assumes an isotropic Cauchy distribution to model the reconstruction error, making it highly effective against heavy-tailed noise.
- **EWNMF** [44]: Entropy-Weighted NMF introduces attribute-wise robustness by assigning optimizable weights to each feature of each instance.
- **rCIM-NMF** [11]: A sample-wise robust NMF method that leverages the Correntropy Induced Metric (CIM) function to minimize the impact of outlier samples on the factorization process.
- **Huber-NMF** [11]: Another element-wise method that incorporates the Huber function to balance squared and absolute errors, offering robustness against moderate outliers.
- **Elastic-NMF** [45]: Proposes an elastic loss function that smoothly transitions between the Frobenius and $L_{2,1}$ norms, offering a flexible trade-off between standard and robust factorization.
- **DANMF** [46]: A deep-learning-based extension of NMF that utilizes multiple layers to extract hierarchical representations while preserving the interpretability of nonnegative factorization.
- **DRNMF** [21]: An element-wise distributionally robust NMF model that covers the set of Gaussian, Poisson, and Gamma noise distributions.
- **SPLNMF** [36]: A robust NMF method that integrates self-paced learning into $L_{2,1}$-NMF to avoid bad local minima and improve convergence stability.

To enhance clarity, we summarize all baseline methods in Table 3, highlighting their publication year, central idea, and primary differences from our proposed DRNMF-SP. This overview helps position our contribution within the evolution of robust and distributionally robust NMF models.

4.3. Noisy datasets

We verify the robustness of DRNMF-SP by running it on datasets contaminated with noise having uncertain probability distributions, as well as with noise composed of a combination of different distributions within Ω.

Scenario a: occluded image datasets. One way to add noise is by occluding 40 % of images using an occlusion square, where the value of each occluded pixel is set to zero. Note that the indices of the masked images are randomly selected following a uniform distribution, and the positions of the occluded pixels within each image are also uniformly selected at random. It is evident that the noise $\epsilon_i = x_i^\star - x_i$ introduced by the masked image $x_i^\star$ is heterogeneous and follows an uncertain probability distribution. Fig. 4 shows sample occluded images from the ORL, OrganA, and Pneumonia datasets.

Table 3
Summary of baseline methods compared with DRNMF-SP.

Method	Year	Core idea	Key Difference from DRNMF-SP
Frobenius-NMF [43]	2000	Using squared reconstruction error	Sensitive to noise and outliers
$L_{2,1}$-NMF [10]	2011	Employs $L_{2,1}$-norm loss	Robust to moderate outliers only
Cauchy-NMF [13]	2015	Using a Cauchy distribution	fixed to a single distribution
EWNMF [44]	2023	Assign adaptive weights to samples	not distributional uncertainty
rCIM-NMF [11]	2012	Using Correntropy criterion	lacks multi-loss formulation
Huber-NMF [11]	2012	Utilizes Huber norms	Not adaptive to multiple noise sources
Elastic-NMF [45]	2019	Introduces adversarial regularization	Ignores distributional robustness
DANMF [46]	2023	Using deep architecture	Lacks robust optimization
DRNMF [21]	2022	Distributionally robust NMF	Operates at element-wise level; no self-paced learning.
SPLNMF [36]	2019	Using self-paced learning	Fixed to a single noise model

Fig. 4. Noise introduced by occlusion.

Scenario b: the combination of noise with different distributions. To evaluate the robustness and efficacy of a distributionally robust model that accounts for noise distributions in Ω, it is common to contaminate clean data with noise from a single distribution $\tau \in \Omega$ and assess performance. However, in real-world scenarios, noise rarely adheres to a single distribution; instead, it often results from a combination of multiple noise sources. Therefore, it is more realistic to examine the robustness of a method under such mixed-distribution noise conditions. We define the combined noise $\hat{N}$ as:

$$\hat{N} = \sum_{\tau \in \Omega} \frac{N_\tau}{\|N_\tau\|_F} \tag{28}$$

where N_τ is the noise generated from the distribution associated with $\tau \in \Omega$. The final noise matrix N is defined as:

$$N = \varrho \cdot \frac{\|X\|_F}{\|\hat{N}\|_F} \cdot \hat{N}, \quad 0 < \varrho < 1 \tag{29}$$

where ϱ denotes the desired noise intensity. The noisy data matrix $\hat{X}$ is then constructed as $\hat{X} = \max(0, X + N)$ where X is a clean low-rank instance and $\hat{X}$ is the corresponding noisy version.

4.4. Visualization

To provide qualitative evidence of the effectiveness of the proposed model, we present visual comparisons on the Yale and ORL datasets. For each dataset, mixed noise composed of Gaussian, Laplacian, and Cauchy was added to the images to simulate realistic degradation conditions. The qualitative results are shown in Figs. 5 and 6. Each figure displays ten representative samples arranged horizontally. From top to bottom, the rows illustrate the clean input images, the corresponding noisy versions, and the reconstructed outputs generated by the proposed DRNMF-SP model. It can be observed that the reconstructed faces effectively preserve global structure and key facial features while substantially suppressing the mixed noise components.

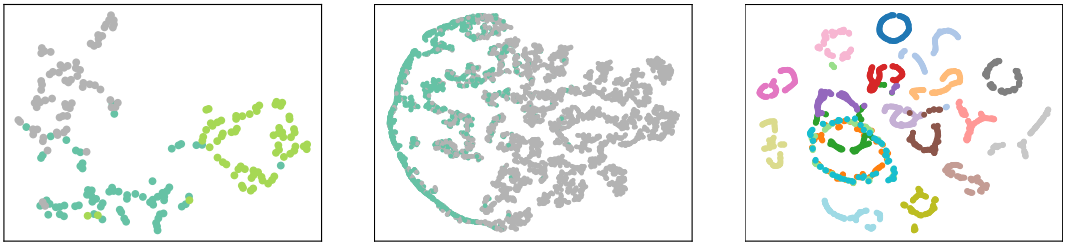
In addition, Fig. 7 shows the t-SNE visualization of the latent features H learned by the proposed model for the Seeds, COIL20, and Pneumonia datasets. The clusters are reasonably well-formed, though not perfectly separated, which can be attributed to the linearity of the model and its limited capacity to capture nonlinear data structures. Since the primary focus of this work is on the distributionally robust loss function, we did not extend the representation model. Future work could consider incorporating manifold- or graph-regularized NMF to potentially improve cluster separation.

4.5. Clustering results

In this section, we evaluate the clustering performance of our DRNMF-SP method compared to 10 other algorithms under different noise conditions. To assess robustness, we introduce noise into the datasets using two defined scenarios from the previous subsection.

Fig. 5. Visual comparison of reconstructed images on the Yale dataset under mixed noise.

Fig. 6. Visual comparison of reconstructed images on the ORL dataset under mixed noise.



(a) Seeds (b) Pneumonia (c) COIL20

Fig. 7. t-SNE visualizations of the learned representations on three datasets using the proposed DRNMF-SPL method.

Table 4
NMI under Scenario A. Best result in **bold**, second-best is underlined.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.4248	0.4340	0.4419	0.4455	0.4182	0.4196	0.3923	0.4115	0.3897	0.4573	**0.4686**
ORL	0.6173	0.6215	0.6159	0.6163	0.6233	0.6154	0.6150	0.6190	0.6114	0.6211	**0.6320**
COIL20	0.7642	0.7567	0.7115	0.7441	0.7561	0.7561	0.7566	0.7079	0.7101	0.7699	**0.7873**
MNIST	0.4309	0.4385	0.4549	0.4545	0.4404	0.4359	0.4373	0.4161	0.4012	0.4506	**0.4623**
Fashion	0.5155	0.5199	0.5535	0.5226	0.5238	0.5165	0.5209	0.4746	0.5087	0.5399	**0.5695**
USPS	0.4118	0.3902	0.3583	0.3768	0.3994	0.3857	0.3835	0.4105	0.3562	0.3853	**0.4291**
UMIST	0.5823	0.5908	0.5953	0.5910	0.5979	0.5605	0.6130	0.5922	0.5718	0.6313	**0.6368**
OrganA	0.6205	0.6589	0.6122	0.6481	0.6048	0.6261	0.6313	0.6247	0.6021	0.6026	**0.6788**
Blood	0.3439	0.3449	0.3449	0.3560	0.3451	0.3343	0.3461	0.3368	0.3357	0.3461	**0.3608**
Pneumonia	0.2835	0.2835	0.2365	0.2365	0.2465	0.2788	0.2565	0.2040	0.2115	0.2908	**0.3130**

Table 5
ARI under Scenario A. Best result in **bold**, second-best is underlined.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.1493	0.1608	0.1754	0.1801	0.1535	0.1506	0.1276	0.1491	0.1465	0.1860	**0.2013**
ORL	0.2131	0.2091	0.2094	0.2005	0.2080	0.2000	0.2042	0.2055	0.2012	0.2119	**0.2358**
COIL20	0.5846	0.5590	0.4948	0.5446	0.5564	0.5705	0.5713	0.5014	0.5011	0.6039	**0.6352**
MNIST	0.2838	0.2788	0.3057	0.3058	0.2871	0.2847	0.2887	0.2671	0.2701	0.3036	**0.3232**
Fashion	0.3530	0.3519	0.3843	0.3845	0.3584	0.3679	0.3460	0.3305	0.3456	0.3737	**0.3995**
USPS	0.2576	0.2256	0.2026	0.2273	0.2427	0.2326	0.2217	0.2523	0.2135	0.2336	**0.2841**
UMIST	0.2775	0.2945	0.2959	0.2985	0.3060	0.2576	0.3040	0.2726	0.2045	0.3282	**0.3468**
OrganA	0.4632	0.4185	0.4471	0.5217	0.3871	0.4496	0.4918	0.4483	0.4078	0.4525	**0.5621**
Blood	0.2486	0.2468	0.2422	0.2523	0.2533	0.2459	0.2451	0.2275	0.1965	0.1876	**0.2559**
Pneumonia	0.3028	0.3028	0.2189	0.2189	0.2209	0.2947	0.2227	0.3352	0.2541	0.2952	**0.3623**

Each scenario simulates different types of noise contamination to analyze the effectiveness of the methods in handling real-world data imperfections. We use three widely adopted clustering metrics, NMI, ARI, and ACC, to quantitatively compare the clustering results. Higher values indicate better alignment between the predicted clusters and the ground-truth labels. The results are presented in two subsections, where Scenario A examines occluded images with an uncertain noise distribution, and Scenario B focuses on data contamination by a mixture of known noise distributions from the ambiguity set. The performance of all methods is reported across multiple datasets to provide a comprehensive evaluation.

4.5.1. Scenario a: occlusion-based noise contamination

In this scenario, we assess the robustness of clustering methods when the image data is partially occluded. To simulate real-world cases where missing or corrupted regions occur, we randomly select 40 % of the images in the examined dataset and occlude 10 % of their pixels by applying a square mask of size $m \times m$. The position of this mask is randomly chosen for each affected image, ensuring variability in the occlusion pattern. We evaluate the clustering performance using NMI (Table 4), ARI (Table 5), and ACC (Table 6). These results provide insights into how different methods handle partial occlusion and whether they can effectively recover meaningful cluster structures despite missing information.

The results in Tables 4–6 show that our DRNMF-SP consistently outperforms other methods, achieving the highest scores in all cases. This highlights its ability to handle structured occlusions and uncertain noise distributions more effectively than conventional robust NMF methods. Across various datasets, DRNMF-SP consistently outperforms other methods, particularly on the Yale, ORL, COIL20, and Pneumonia datasets, where other approaches struggle with missing pixel information. Even in datasets where the performance gap is smaller, such as USPS and Blood, DRNMF-SP remains competitive, often leading in clustering accuracy. Overall, these findings confirm that DRNMF-SP is highly resilient to occlusion-based distortions, preserving the cluster structure even when a substantial portion of the data is missing.

Table 6
ACC under Scenario A. Best result in **bold**, second-best is underlined.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.3757	0.3575	0.3939	0.4091	0.3939	0.3818	0.3575	0.3696	0.3656	0.4121	**0.4242**
ORL	0.4400	0.4387	0.4350	0.4363	0.4350	0.4187	0.4325	0.4300	0.4215	0.4375	**0.4600**
COIL20	0.6774	0.6733	0.6111	0.6698	0.6861	0.6799	0.6799	0.6201	0.6357	0.7014	**0.7257**
MNIST	0.5265	0.5230	0.5385	0.5410	0.5345	0.5180	0.5345	0.5020	0.5125	0.5390	**0.5670**
Fashion	0.5580	0.5530	0.5765	0.5885	0.5610	0.5540	0.5425	0.5190	0.5314	0.5730	**0.5900**
USPS	0.4768	0.4450	0.4145	0.4600	0.4655	0.4514	0.4441	0.4759	0.4215	0.4645	**0.5100**
UMIST	0.4417	0.4522	0.4652	0.4748	0.4609	0.4252	0.4791	0.4504	0.4058	0.4939	**0.5200**
OrganA	0.6598	0.6069	0.6632	0.7176	0.6512	0.6800	0.6989	0.6790	0.6669	0.6368	**0.7323**
Blood	0.4632	0.4684	0.4789	0.4824	0.4731	0.4556	0.4889	0.4445	0.4625	0.4696	**0.4982**
Pneumonia	0.7767	0.7767	0.7423	0.7423	0.7424	0.7729	0.7423	0.8015	0.7598	0.7729	**0.8034**

Table 7
The comparison results for the Gaussian noise, evaluated based on NMI.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.4427	0.4532	0.4738	0.4700	0.4597	0.4698	0.4556	0.4508	0.4465	0.4971	**0.5134**
ORL	0.7970	0.7947	0.7892	0.7764	0.7663	0.7815	0.7906	0.6926	0.7015	0.7813	**0.7993**
COIL20	0.7425	0.7255	0.7378	0.7388	0.7028	0.7270	0.7231	0.7324	0.7145	0.7261	**0.7753**
MNIST	0.4513	0.4231	0.4359	0.4480	0.3900	0.4070	0.4364	0.4208	0.3968	0.4480	**0.4523**
Fashion	0.5236	0.5533	0.5340	0.5207	0.5353	0.5297	0.5255	0.5334	0.5189	0.5434	**0.5724**
Seeds	0.6101	0.5535	0.5630	0.5123	0.5795	0.6220	0.4757	0.5801	0.4964	0.5377	**0.6797**
Ecoli	0.5388	0.5240	0.5258	0.5364	0.5222	0.5034	0.5238	0.4942	0.5146	0.5311	**0.5800**
USPS	0.3914	0.3662	0.3821	0.3657	0.3658	0.3696	0.3674	0.3788	0.3627	0.3963	**0.4391**
UMIST	**0.6170**	0.5975	0.5723	0.5987	0.5980	0.5847	0.6059	0.5932	0.5713	0.6032	0.6126
OrganA	0.6144	0.6220	0.6282	0.5868	0.6238	0.6131	0.5868	0.6085	0.5902	0.6187	**0.6404**
Blood	0.2673	0.2509	0.2622	0.2536	0.2640	0.2732	0.2789	0.2690	0.2510	0.3550	**0.3656**
Pneumonia	0.3098	0.3108	0.3013	0.2525	0.2141	0.3109	0.3099	0.3013	0.2616	0.2989	**0.3305**

Table 8
The comparison results for the Gaussian noise, evaluated based on ARI.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.1599	0.1725	0.2047	0.1978	0.1871	0.1925	0.1845	0.1849	0.1932	0.2322	**0.2556**
ORL	0.4866	0.4868	0.4749	0.4445	0.4115	0.4528	0.4623	0.3996	0.4100	0.4447	**0.4954**
COIL20	0.5436	0.5398	0.5447	0.5518	0.4819	0.5306	0.5442	0.5217	0.5001	0.5112	**0.6050**
MNIST	0.2878	0.2697	0.2721	0.2911	0.2222	0.2437	0.2805	0.2676	0.2399	0.2957	**0.3035**
Fashion	0.3451	0.4138	0.3725	0.3742	0.4057	0.3429	0.3832	0.3854	0.3625	0.3824	**0.4266**
Seeds	0.6454	0.5873	0.5962	0.5540	0.6057	0.6606	0.5013	0.6053	0.6201	0.5809	**0.7266**
Ecoli	0.4392	0.3910	0.4112	0.4463	0.4300	0.3815	0.4050	0.3742	0.4077	0.4335	**0.5765**
USPS	0.2272	0.2081	0.2253	0.2050	0.2080	0.2085	0.2137	0.2176	0.2264	0.2432	**0.2904**
UMIST	0.2979	0.3032	0.2589	0.2968	0.2827	0.2731	0.3063	0.2791	0.2679	0.3023	**0.3168**
OrganA	0.4495	0.4818	0.4992	0.3964	0.4789	0.4891	0.3964	0.4900	0.4031	0.4715	**0.5160**
Blood	0.1923	0.1841	0.1429	0.1848	0.1934	0.2071	0.2003	0.1961	0.1749	0.2602	**0.2654**
Pneumonia	0.3285	0.3080	0.3239	0.2666	0.2149	0.3080	0.3285	0.3240	0.2480	0.3198	**0.3578**

4.5.2. Scenario b: known probable noise distributions and their

In Scenario B, we define the ambiguity set as $\Omega = \{\text{Laplacian}, \text{Gaussian}, \text{Cauchy}\}$ and conduct four experiments to simulate diverse noise environments. The first three experiments individually inject Laplacian, Gaussian, and Cauchy noise, respectively, into the datasets. The fourth experiment considers a mixed-noise setting by combining all three distributions, as detailed in Section 4.3. The noise parameters are adapted to each dataset to ensure a consistent and meaningful level of corruption across experiments.

The performance of our algorithm under these different noise conditions is evaluated using clustering metrics—NMI, ARI, and ACC—and the results are presented in Tables 7–18. These results show the robustness of DRNMF-SP, especially when data is contaminated by a mixture of distributions, showcasing its ability to handle realistic and complex noise scenarios.

The results for Gaussian noise (Tables 7–9), where $\Omega = \{\text{Gaussian}\}$, reveal that DRNMF-SP stands out in handling diverse data distributions, securing the top rank in most datasets. Its exceptional performance is particularly evident in challenging datasets such as Fashion, Seeds, Ecoli, and OrganA. While some methods show competitive results in certain cases, DRNMF-SP consistently delivers superior clustering results, confirming its robustness in real-world scenarios where Gaussian noise is prevalent.

When confronted with Laplacian noise, DRNMF-SP demonstrates its resilience by outperforming the other methods across a wide range of datasets (Tables 10–12). It excels in datasets like Seeds, Ecoli, OrganA, and Pneumonia, illustrating its adaptability to Laplacian noise. Although some methods excel in specific datasets, DRNMF-SP consistently delivers the best overall performance, reinforcing its reliability in clustering tasks subject to Laplacian noise.

Table 9
The comparison results for the Gaussian noise, evaluated based on ACC.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.3988	0.4145	0.4267	0.4315	0.4194	0.4133	0.4073	0.4000	0.4235	0.4667	**0.5030**
ORL	**0.6900**	0.6758	0.6700	0.6450	0.6292	0.6600	0.6700	0.5325	0.5568	0.6550	0.6883
COIL20	0.6562	0.6576	0.6660	0.6639	0.6285	0.6382	0.6597	0.6569	0.5986	0.6340	**0.7000**
MNIST	0.5425	0.5040	0.5170	0.5345	0.4595	0.4850	0.5205	0.4950	0.4766	0.5460	**0.5540**
Fashion	0.5585	0.6340	0.5865	0.5845	0.5890	0.5570	0.5665	0.5850	0.5999	0.5880	**0.6400**
Seeds	0.8667	0.8429	0.8476	0.8238	0.8476	0.8662	0.8000	0.8524	0.8102	0.8381	**0.9000**
Ecoli	0.7506	0.7491	0.7315	0.7479	0.7384	0.7188	0.7467	0.7530	0.7259	0.7530	**0.8036**
USPS	0.4409	0.4168	0.4559	0.4268	0.4145	0.4214	0.4182	0.4409	0.4126	0.4536	**0.5000**
UMIST	0.4809	0.4783	0.4583	0.4696	0.4817	0.4678	0.4817	0.4713	0.4565	0.4878	**0.4974**
OrganA	0.6513	0.6778	0.6847	0.6700	0.6686	0.6887	0.6700	0.6977	0.6578	0.6938	**0.7294**
Blood	0.4102	0.3884	0.4036	0.3931	0.3936	0.4316	0.4129	0.4129	0.4200	0.4883	**0.5040**
Pneumonia	0.7881	0.7786	0.7862	0.7595	0.7424	0.7786	0.7882	0.7863	0.7391	0.7844	**0.8034**

Table 10
The comparison results for the Laplacian noise, evaluated based on NMI.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.4462	0.4611	0.4508	0.4339	0.4651	0.4635	0.4298	0.4653	0.4369	0.4382	**0.4914**
ORL	0.7920	0.7960	0.7896	0.7955	0.7889	0.7942	0.7933	0.7640	0.7845	0.7735	**0.8009**
COIL20	0.7273	0.7592	0.7459	0.7327	0.7426	0.7593	0.7505	0.7446	0.7000	0.7028	**0.7724**
MNIST	0.4340	0.4316	0.4387	0.4382	0.3920	0.4350	0.4303	0.4452	0.4188	0.4091	**0.4825**
Fashion	0.5251	0.5399	0.5165	0.5421	0.5274	0.5258	0.5099	0.5173	0.5224	0.5318	**0.5808**
Seeds	0.4858	0.5485	0.5068	0.4825	0.5360	0.5595	0.4839	0.4982	0.5142	0.5605	**0.5610**
Ecoli	0.5514	0.5225	0.5581	0.5070	0.5394	0.5537	0.5418	0.5403	0.5050	0.5299	**0.6104**
USPS	0.3912	0.3956	0.3511	0.3446	0.3918	0.3864	0.3757	0.3656	0.3670	0.3788	**0.4422**
UMIST	0.5968	0.5813	0.6069	0.6068	0.6028	0.5955	0.6067	0.5962	0.5864	0.6067	**0.6294**
OrganA	0.6124	0.5926	0.6209	0.6134	0.6154	0.6157	0.6032	0.5936	0.6032	0.5953	**0.6590**
Blood	0.3297	0.3189	0.3272	0.3294	0.3256	0.3161	0.3269	0.3277	0.3159	0.3306	**0.3629**
Pneumonia	0.2026	0.2710	0.1502	0.2820	0.2107	0.1525	0.2710	0.2711	0.1953	0.2297	**0.3050**

Table 11
The comparison results for the Laplacian noise, evaluated based on ARI.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.1682	0.1947	0.1736	0.1603	0.1819	0.1924	0.1381	0.2092	0.1746	0.1446	**0.2347**
ORL	0.4761	0.4888	0.4807	0.4957	0.4671	0.4790	0.4877	0.4345	0.4603	0.4407	**0.4973**
COIL20	0.5292	0.5864	0.5426	0.4985	0.5558	**0.5958**	0.5527	0.5436	0.5197	0.4819	0.5930
MNIST	0.2934	0.2970	0.2874	0.2944	0.2500	0.3051	0.2936	0.3003	0.2661	0.2653	**0.3509**
Fashion	0.3657	0.3775	0.3591	0.3850	0.3495	0.3650	0.3487	0.3722	0.3512	0.3656	**0.4196**
Seeds	0.4976	0.5653	0.5015	0.4825	0.5449	0.5896	0.5051	0.5121	0.4936	0.5783	**0.5937**
Ecoli	0.4429	0.3873	0.4292	0.3744	0.4168	0.4354	0.4330	0.4951	0.3921	0.3948	**0.5068**
USPS	0.2371	0.2344	0.1988	0.1929	0.2327	0.2296	0.2149	0.2109	0.2074	0.2132	**0.2930**
UMIST	0.2866	0.2774	0.3047	0.2992	0.3002	0.2876	0.2959	0.2931	0.2710	0.2954	**0.3127**
OrganA	0.4936	0.3717	0.4615	0.4968	0.5056	0.4267	0.4799	0.3823	0.3845	0.4502	**0.5623**
Blood	0.2052	0.1840	0.2295	0.1794	0.2233	0.2261	0.2295	0.2306	0.2078	0.1783	**0.2601**
Pneumonia	0.3286	0.2903	0.1218	0.3267	0.2061	0.1472	0.2903	0.2904	0.1954	0.2258	**0.3574**

Table 12
The comparison results for the Laplacian noise, evaluated based on ACC.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.4061	0.4212	0.4030	0.4000	0.4091	0.4212	0.3879	0.4363	0.4008	0.3939	**0.4667**
ORL	0.6833	0.6800	0.6650	0.6792	0.6600	0.6725	0.6767	0.6375	0.6547	0.6500	**0.6875**
COIL20	0.6375	0.7063	0.6701	0.6313	0.6854	0.7056	0.6667	0.6645	0.6519	0.6285	**0.7194**
MNIST	0.5275	0.5510	0.5275	0.5280	0.4720	0.5430	0.5330	0.5510	0.5173	0.4960	**0.5700**
Fashion	0.5600	0.5720	0.5640	0.5600	0.5625	0.5635	0.5425	0.5700	0.5607	0.5620	**0.6260**
Seeds	0.7905	0.8286	0.8048	0.7905	0.8238	0.8381	0.8000	0.8095	0.8162	0.8333	**0.8429**
Ecoli	0.7690	0.7423	0.7542	0.7384	0.7583	0.7634	0.7524	0.7714	0.7106	0.7262	**0.8214**
USPS	0.4442	0.4736	0.4327	0.4176	0.4412	0.4418	0.4352	0.4445	0.4316	0.4212	**0.4900**
UMIST	0.4470	0.4470	0.4754	0.4881	0.4736	0.4580	0.4719	0.4556	0.4418	0.4539	**0.4939**
OrganA	0.6718	0.6484	0.6934	0.6794	0.6815	0.6800	0.6558	0.6501	0.6523	0.6590	**0.7213**
Blood	0.4608	0.4468	0.4585	0.4632	0.4521	0.4491	0.4661	0.4655	0.4502	0.4480	**0.4830**
Pneumonia	0.7987	0.7709	0.7423	0.7881	0.7567	0.7424	0.7709	0.7710	0.7505	0.7423	**0.8015**

Table 13
The comparison results for the Cauchy noise, evaluated based on NMI.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.4337	0.4501	0.4799	0.4546	0.4453	0.4580	0.4498	0.4699	0.4495	0.4798	**0.5125**
ORL	0.7754	0.7687	0.7725	0.7801	0.7770	0.7851	0.7776	0.7366	0.7514	0.7866	**0.7997**
COIL20	0.7522	0.7780	0.7633	0.7369	0.7333	0.7307	0.7255	0.7575	0.7381	0.7524	**0.7871**
MNIST	0.4236	0.4204	0.4177	0.4133	0.4210	0.3999	0.4118	0.4167	0.4000	0.4203	**0.4654**
Fashion	0.5275	0.5394	0.5434	0.5276	0.5114	0.5130	0.5258	0.5393	0.5166	0.5482	**0.5946**
Seeds	0.4753	0.4354	0.6020	0.5801	0.4240	0.5312	0.4316	0.5100	0.4925	0.6074	**0.6243**
Ecoli	0.5070	0.5193	0.5342	0.5148	0.5209	0.5207	0.5270	0.5116	0.5142	0.5188	**0.5877**
USPS	0.3715	0.3915	0.3919	0.3614	0.3704	0.3709	0.3647	0.3744	0.3888	0.3740	**0.4721**
UMIST	0.5961	0.5917	0.6054	0.5987	0.5969	0.5923	0.5928	0.5886	0.5936	0.5944	**0.6236**
OrganA	0.5507	0.5630	0.5850	0.5630	0.5609	0.5541	0.5951	0.5690	0.5524	0.5581	**0.6290**
Blood	0.2960	0.2963	0.2898	0.2863	0.2821	0.3026	0.3051	0.3043	0.2763	0.3226	**0.3443**
Pneumonia	0.2465	0.1681	0.1732	0.2772	0.2719	0.1984	0.2691	0.2374	0.2156	0.2367	**0.2853**

Table 14
The comparison results for the Cauchy noise, evaluated based on ARI.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.1654	0.1869	0.2111	0.1862	0.1796	0.1901	0.1848	0.1869	0.1969	0.2083	**0.2452**
ORL	0.4499	0.4449	0.4371	0.4519	0.4264	0.4526	0.4498	0.3259	0.4625	0.4656	**0.5058**
COIL20	0.5600	0.5838	0.5827	0.5459	0.5215	0.5375	0.5398	0.5551	0.5748	0.5445	**0.6208**
MNIST	0.2640	0.2744	0.2654	0.2567	0.2628	0.2523	0.2664	0.2860	0.2598	0.2771	**0.3324**
Fashion	0.3531	0.3956	0.3824	0.3950	0.3456	0.3474	0.3650	0.3626	0.3704	0.4106	**0.4654**
Seeds	0.4908	0.4709	0.6583	0.6238	0.4571	0.5088	0.4690	0.4887	0.5230	0.6500	**0.6715**
Ecoli	0.3782	0.3803	0.4318	0.3843	0.3999	0.3951	0.3991	0.3542	0.3915	0.4480	**0.5741**
USPS	0.2141	0.2271	0.2351	0.2039	0.2133	0.2098	0.2003	0.2285	0.2301	0.2075	**0.3308**
UMIST	0.2937	0.3019	0.3116	0.3047	0.2950	0.2827	0.2879	0.2908	0.2988	0.3015	**0.3364**
OrganA	0.3671	0.3469	0.4431	0.3899	0.4144	0.3682	0.4403	0.3528	0.3745	0.4121	**0.5270**
Blood	0.1812	0.1974	0.2123	0.2087	0.2056	0.1718	0.1935	0.1742	0.1922	0.2297	**0.2522**
Pneumonia	0.2811	0.2995	0.1714	0.3545	0.3458	0.1592	0.2691	0.1880	0.1603	0.1982	**0.3678**

Table 15
The comparison results for the Cauchy noise, evaluated based on ACC.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.3980	0.4141	0.4424	0.4202	0.4182	0.4182	0.4242	0.4242	0.4485	0.4384	**0.4848**
ORL	0.6350	0.6463	0.6488	0.6575	0.6463	0.6687	0.6600	0.5375	0.6302	0.6463	**0.6800**
COIL20	0.6611	0.6687	0.6812	0.6438	0.6306	0.6556	0.6576	0.6756	0.6477	0.6556	**0.7174**
MNIST	0.5080	0.5207	0.5120	0.5060	0.5053	0.4897	0.5037	0.5550	0.4920	0.5193	**0.5920**
Fashion	0.5560	0.5990	0.5880	0.5850	0.5600	0.5500	0.5635	0.5720	0.5700	0.6000	**0.6800**
Seeds	0.8000	0.7762	0.8714	0.8571	0.7714	0.8095	0.7762	0.8000	0.7936	0.8667	**0.8762**
Ecoli	0.7369	0.7455	0.7550	0.7542	0.7446	0.7455	0.7494	0.7470	0.7198	0.7520	**0.7857**
USPS	0.4236	0.4330	0.4627	0.4076	0.4279	0.4297	0.4052	0.4590	0.4355	0.4218	**0.5409**
UMIST	0.4557	0.4704	0.4896	0.4765	0.4626	0.4557	0.4835	0.4608	0.4471	0.4991	**0.5096**
OrganA	0.5897	0.5710	0.6260	0.6190	0.6066	0.5891	0.6268	0.5755	0.5926	0.5938	**0.6775**
Blood	0.4351	0.4252	0.4363	0.4293	0.4211	0.4357	0.4433	0.4398	0.4241	0.4579	**0.4941**
Pneumonia	0.7671	0.7901	0.7423	0.8015	0.7977	0.7423	0.8015	0.7423	0.7756	0.7423	**0.8072**

Table 16
The comparison results for the mixture of Gaussian, Laplacian, and Cauchy noise, evaluated based on NMI.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.4619	0.4655	0.4488	0.4576	0.4511	0.4543	0.4505	0.4577	0.4426	0.4805	**0.5363**
ORL	0.7828	0.8008	0.7932	0.7932	0.7780	0.7956	0.7828	0.7374	0.7845	0.7887	**0.8284**
COIL20	0.7529	0.7273	0.7481	0.7028	0.7524	0.7261	0.7578	0.7389	0.7163	0.7327	**0.7624**
MNIST	0.4087	0.4286	0.4047	0.4154	0.4287	0.3949	0.4047	0.4315	0.3985	0.4154	**0.4525**
Fashion	0.5331	0.5169	0.5099	0.5152	0.5115	0.5087	0.5100	0.5178	0.5041	0.5165	**0.5411**
Seeds	0.5123	0.5650	0.5795	0.5187	0.5377	0.4021	0.5377	0.5377	0.4782	0.5187	**0.6105**
Ecoli	0.5096	0.5095	0.5231	0.5032	0.4811	0.5062	0.5182	0.5283	0.5171	0.5006	**0.6063**
USPS	0.3728	0.3765	0.3914	0.3823	0.3777	0.3729	0.3793	0.3805	0.3862	0.3977	**0.4371**
UMIST	0.5911	0.6096	0.6194	0.6058	0.5834	0.5941	0.5783	0.6119	0.5843	0.5824	**0.6255**
OrganA	0.5907	0.5818	0.5540	0.5806	0.5564	0.5826	0.5791	0.5931	0.5691	0.5951	**0.6201**
Blood	0.2327	0.2425	0.2240	0.2438	0.2558	0.2507	0.3195	0.2447	0.2129	0.2383	**0.3234**
Pneumonia	0.2506	0.2605	0.2665	0.2958	0.2785	0.2774	0.2666	0.2881	0.2748	0.2197	**0.3139**

In the presence of Cauchy noise, which is characterized by its heavy-tailed distribution, DRNMF-SP consistently outperforms other methods across all datasets (Tables 13–15). Heavy-tailed distributions, such as Cauchy, are known for their propensity to generate extreme outliers, particularly in the tail part of the distribution, where very large errors can occur in the data. These extreme outliers pose significant challenges for many robust methods, which struggle to handle such outliers effectively. However, by integrating self-paced learning into our distributionally robust NMF method, DRNMF-SP demonstrates a remarkable ability to manage these extreme outliers. This enhancement allows the method to handle noisy and challenging data more effectively, as shown in the results, where DRNMF-SP consistently delivers superior performance across NMI, ARI, and ACC metrics. The results highlight the effectiveness of this approach in clustering tasks where extreme outliers are prevalent due to Cauchy-distributed noise.

In real-world scenarios, the data is often contaminated by a combination of different types of noise, rather than by a single noise distribution. This combinational noise is common in many practical applications, making it crucial for robust clustering methods to handle such complex noise mixtures. The results of our experiment clearly show that DRNMF-SP excels under these conditions, outperforming all other methods on all three metrics (NMI, ARI, and ACC) in nearly every dataset (Tables 16–18). Specifically, DRNMF-SP consistently demonstrates superior clustering quality performance, achieving the highest NMI, ARI, and ACC in most cases. These results emphasize the effectiveness of DRNMF-SP in real-world applications, where data are often subject to multifaceted noise environments.

The experimental results confirm the robustness and adaptability of our DRNMF-SP algorithm under various noise conditions. As expected, methods that are explicitly designed for a particular noise distribution, such as the Frobenius norm for Gaussian noise, the $L_{2,1}$ norm for Laplacian noise, and the Cauchy loss for Cauchy noise, tend to perform best when the noise follows their respective assumptions. However, our proposed method, which employs a weighted sum of these three loss functions, consistently outperforms all other methods across all noise scenarios. This superior performance can be attributed to the fact that in each scenario, the most appropriate loss function dominates, while the remaining two act as regularization terms that enhance stability and improve overall robustness. For instance, in the presence of Laplacian noise, the $L_{2,1}$ norm serves as the primary loss, while the Frobenius and Cauchy losses provide complementary regularization, refining the learned representations and mitigating potential overfitting to extreme values.

Moreover, when the data are contaminated by a mixture of noise distributions, our method exhibits even greater advantages. Unlike traditional approaches that rely on a single predefined loss function, our weighted sum formulation acts as an ensemble model of loss functions, dynamically adjusting its weights to best match the underlying noise characteristics. The key strength of DRNMF-SP lies in its ability to learn these weights adaptively during the optimization process, ensuring optimal alignment with the given noise structure. This self-adjusting mechanism enables our method to effectively handle complex and uncertain noise conditions, demonstrating its suitability for real-world applications where noise distributions are often unknown and heterogeneous.

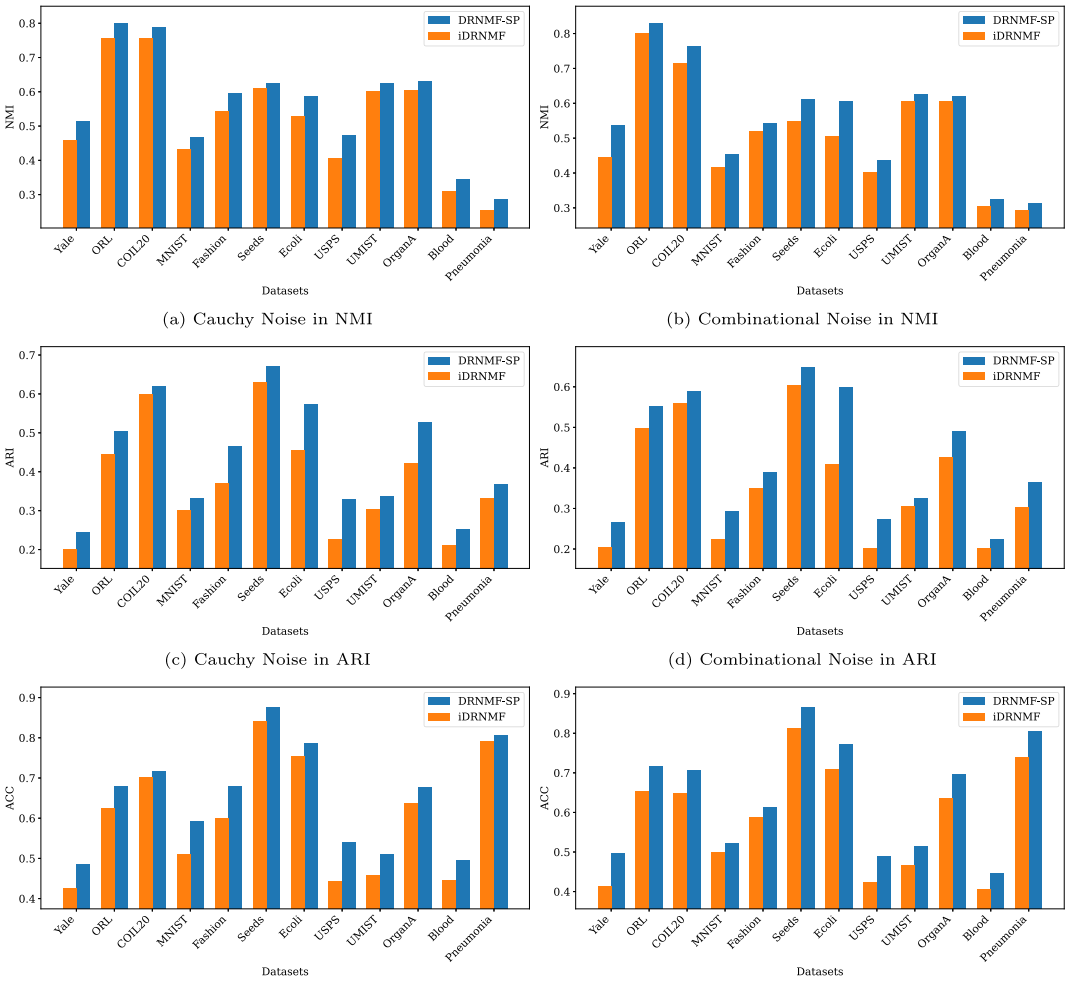
4.6. Improving robustness to extreme outliers: iDRNMF vs. DRNMF-SP

In this paper, we present an enhanced version of the instance-wise distributionally robust nonnegative matrix factorization (iDRNMF) method by integrating self-paced learning (SPL) to create the DRNMF-SP method. Although iDRNMF is robust against a variety of noise types and their combinations, it struggles when extreme outliers are present. These outliers, often appearing as extreme deviations in the tail of heavy-tailed noise distributions, can significantly disrupt the performance of many robust methods, including iDRNMF. This issue becomes especially critical when the data contains severe anomalies, where most traditional approaches fail to handle such extremes effectively. To address this limitation, we introduce SPL, a technique designed to progressively focus on easier samples and avoid the influence of extreme outliers early in the learning process. By leveraging SPL, our DRNMF-SP method gains the ability to better manage these outliers, while retaining the robustness against various noise distributions that iDRNMF offers. Additionally, SPL enhances the optimization process by preventing the model from getting stuck in bad local minima, leading to better overall generalization performance. In this subsection, we compare the original iDRNMF method with the DRNMF-SP. Given the large number of figures and the need for effective summarization without loss of generalization, we focus on presenting results for Cauchy noise (due to its heavy-tailed nature and propensity to produce extreme outliers) and the combined noise distributions in Fig. 8. Through this analysis, we highlight the improvements offered by DRNMF-SP, particularly in its ability to handle extreme outliers, achieve better clustering performance, and maintain robustness against diverse noise distributions.

These results clearly demonstrate that self-paced learning provides a decisive advantage when extending iDRNMF to DRNMF-SP. The progressive sample selection mechanism enables the model to first build a reliable structure from clean data and then incorporate noisier or more contaminated samples, rather than fitting all data at once. This strategy not only enhances resistance to heavy-tailed and extreme outliers but also stabilizes the optimization process by reducing the risk of poor local minima. Consequently, DRNMF-SP consistently outperforms iDRNMF in the most challenging scenarios, highlighting the specific benefits introduced by integrating self-paced learning into the distributionally robust framework.

4.7. Contribution of DRNMF-SP components

In order to further analyze the individual contributions of the proposed model's terms, we conducted an ablation study focusing on the adaptive multi-loss fusion and distributionally robust optimization. The previous subsection already investigated the role of self-paced learning by comparing DRNMF-SP with its non-self-paced variant (iDRNMF). Here, we isolate and quantify the effect of the multi-objective formulation and distributional robustness by examining the performance of DRNMF-SP under different configurations of the ambiguity set Ω on the Yale dataset with four types of noise contamination: Gaussian, Laplacian, Cauchy, and their mixture. The results are reported in Table 19.



(a) Cauchy Noise in NMI (b) Combinational Noise in NMI

(c) Cauchy Noise in ARI (d) Combinational Noise in ARI

(e) Cauchy Noise in ACC (f) Combinational Noise in ACC

Fig. 8. Comparison of DRNMF-SP and iDRNMF for different noise types and metrics.

Table 17
The comparison results for the mixture of Gaussian, Laplacian, and Cauchy noise, evaluated based on ARI.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.1880	0.1901	0.1765	0.1867	0.1760	0.1827	0.1799	0.1787	0.1846	0.2105	**0.2670**
ORL	0.4430	0.4889	0.4830	0.4844	0.4404	0.4776	0.4655	0.3877	0.4431	0.4898	**0.5516**
COIL20	0.5740	0.5292	0.5460	0.4819	0.5445	0.5112	0.5586	0.5416	0.5087	0.4985	**0.5901**
MNIST	0.2542	0.2579	0.2408	0.2535	0.2607	0.2422	0.2408	0.2688	0.2321	0.2535	**0.2927**
Fashion	0.3638	0.3790	0.3487	0.3558	0.3369	0.3326	0.3470	0.3529	0.3454	0.3591	**0.3903**
Seeds	0.5540	0.6041	0.6057	0.5487	0.5809	0.3955	0.5809	0.5717	0.4219	0.5487	**0.6491**
Ecoli	0.4141	0.4023	0.4360	0.3776	0.3575	0.3923	0.4262	0.3330	0.3837	0.4292	**0.5985**
USPS	0.2115	0.2183	0.2316	0.2374	0.2131	0.2085	0.2208	0.2356	0.2161	0.2406	**0.2726**
UMIST	0.2838	0.3171	0.3109	0.3063	0.2829	0.2892	0.2736	0.3031	0.2837	0.2815	**0.3267**
OrganA	0.4575	0.4239	0.4043	0.4276	0.3836	0.3898	0.3925	0.4185	0.3915	0.4002	**0.4900**
Blood	0.1495	0.1586	0.1363	0.1365	0.1730	0.1415	0.1678	0.1592	0.1364	0.1275	**0.2256**
Pneumonia	0.2121	0.2300	0.2824	0.3435	0.3628	0.3104	0.2824	0.3724	0.2658	0.2819	**0.3650**

The comparison highlights several important observations. First, models based on a single loss function ($\ell_{2,1}$, Frobenius, or Cauchy) yield the weakest performance across all noise settings, indicating that no single loss is sufficiently robust to handle the diverse noise distributions. Second, combining two losses consistently improves the results, demonstrating that multi-objective optimization captures complementary strengths of different noise models. Third, the use of all three losses with adaptive weighting achieves the best performance in every case, confirming the advantage of distributionally robust optimization where the model automatically adjusts the importance of each objective to match the underlying data distribution. Interestingly, when the three losses are combined

Table 18
The comparison results for the mixture of Gaussian, Laplacian, and Cauchy noise, evaluated based on ACC.

Dataset	Frobenius	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.4133	0.4267	0.4024	0.4194	0.3988	0.4085	0.3927	0.4121	0.4028	0.4182	**0.4970**
ORL	0.6508	0.6767	0.6733	0.6592	0.6500	0.6733	0.6625	0.5900	0.6639	0.6750	**0.7175**
COIL20	0.6861	0.6375	0.6625	0.6285	0.6556	0.6340	0.6764	0.6541	0.6247	0.6313	**0.7063**
MNIST	0.4890	0.5020	0.4600	0.4860	0.5050	0.4740	0.4600	0.5060	0.4708	0.4860	**0.5220**
Fashion	0.5790	0.5940	0.5425	0.5690	0.5420	0.5370	0.5340	0.5680	0.5348	0.5660	**0.6140**
Seeds	0.8238	0.8471	0.8476	0.8190	0.8381	0.7476	0.8381	0.8333	0.8255	0.8190	**0.8667**
Ecoli	0.7299	0.7449	0.7369	0.7396	0.7095	0.7330	0.7318	0.7649	0.7100	0.7262	**0.7708**
USPS	0.4361	0.4479	0.4570	0.4561	0.4339	0.4255	0.4321	0.4700	0.4416	0.4594	**0.4891**
UMIST	0.4707	0.4922	0.5032	0.4893	0.4684	0.4632	0.4638	0.4713	0.4635	0.4817	**0.5154**
OrganA	0.6250	0.6442	0.6028	0.6068	0.6080	0.6509	0.6347	0.6456	0.6005	0.6532	**0.6955**
Blood	0.3785	0.3925	0.3738	0.3802	0.4047	0.3802	0.4398	0.3966	0.3920	0.3732	**0.4450**
Pneumonia	0.7424	0.7424	0.7671	0.7958	0.8051	0.7805	0.7672	**0.8092**	0.7711	0.7690	0.8053

Table 19
Ablation study results of the proposed DRNMF-SLP (in terms of NMI) on the Yale dataset under various noise types.

Model	λ	Noise			
		G	L	C	G+L+C
$L_{2,1}$	–	0.4392	0.4205	0.4163	0.4302
Frobenius	–	0.4123	0.4048	0.3971	0.3952
Cauchy	–	0.4294	0.4312	0.4362	0.4213
$\Omega = \{1, 2\}$	adaptive	0.4527	0.4492	0.4306	0.4591
$\Omega = \{1, \text{Cauchy}\}$	adaptive	0.4476	0.4581	0.4557	0.4489
$\Omega = \{2, \text{Cauchy}\}$	adaptive	0.4579	0.4446	0.4540	0.4403
$\Omega = \{1, 2, \text{Cauchy}\}$	adaptive	**0.4722**	**0.4654**	**0.4628**	**0.4653**
$\Omega = \{1, 2, \text{Cauchy}\}$	fixed	0.4441	0.4395	0.4432	0.4360

but their weights are fixed ($\lambda = 1/3$), the performance improves over single-loss settings but falls short of the adaptive scheme, showing that adaptivity is a crucial factor in achieving robustness.

Another noteworthy trend is that the relative effectiveness of each loss aligns with its expected robustness property. Under Gaussian noise, configurations including the Frobenius term perform slightly better, while under Laplacian noise the $\ell_{2,1}$-based terms are stronger, and under Cauchy noise the Cauchy loss dominates. This observation validates the design principle of DRNMF-SP: by dynamically balancing multiple objectives, the model adapts to different noise scenarios without requiring prior knowledge of the noise distribution. Taken together, this subsection and the preceding analysis of self-paced learning form a comprehensive ablation study that examines the contribution of all major components of DRNMF-SP. The results clearly demonstrate that self-paced learning, adaptive multi-loss fusion, and distributional robustness each provide significant and complementary benefits, ultimately leading to the superior robustness and clustering accuracy of the full model.

4.8. Generalization to unseen noise distributions

Although our model is developed within a distributionally robust framework—where noise is assumed to come from a predefined ambiguity set (in this case, $\Omega = \{\text{Laplacian}, \text{Gaussian}, \text{Cauchy}\}$)—real-world data contamination does not always conform to these specific distributions. In traditional distributionally robust optimization, the goal is to ensure performance across all distributions $\tau \in \Omega$, yet robustness is not guaranteed when the actual noise follows a different distribution $\tau' \notin \Omega$. To address this limitation, our DRNMF-SP method adopts a multi-objective formulation that combines multiple loss functions through a weighted sum. Crucially, these weights are adaptively adjusted during training based on the observed data. This self-adjusting mechanism enables the model to align itself with the underlying characteristics of the noise, suggesting that DRNMF-SP can potentially maintain strong performance even when facing noise types outside the predefined ambiguity set.

To explore this hypothesis, we extend our evaluation to include four additional noise types not covered by the original ambiguity set: Poisson noise, Rayleigh noise, Gamma noise, and impulse noise. Each of these introduces distinct statistical properties and challenges, providing a rigorous test of the model's generalization capabilities beyond its intended robustness domain. As part of this extended evaluation, we first apply DRNMF-SP to the USPS dataset corrupted with Poisson noise. Unlike the distributions in our ambiguity set, Poisson noise introduces signal-dependent fluctuations that are common in low-light or photon-limited imaging. Nevertheless, as illustrated in Fig. 9a, DRNMF-SP demonstrates strong performance, underscoring its adaptability and potential for handling a broad range of real-world noise conditions.

To further evaluate the robustness of DRNMF-SP, we first apply it to the Yale dataset contaminated with speckle noise, simulated using a multiplicative Gamma distribution. This type of noise, often encountered in imaging applications, introduces intensity-dependent distortions that are not explicitly covered by our ambiguity set. As shown in Fig. 9b, DRNMF-SP continues to perform well despite this mismatch. We attribute this resilience to the interplay between the $L_{2,1}$ and Cauchy loss components, which help manage the

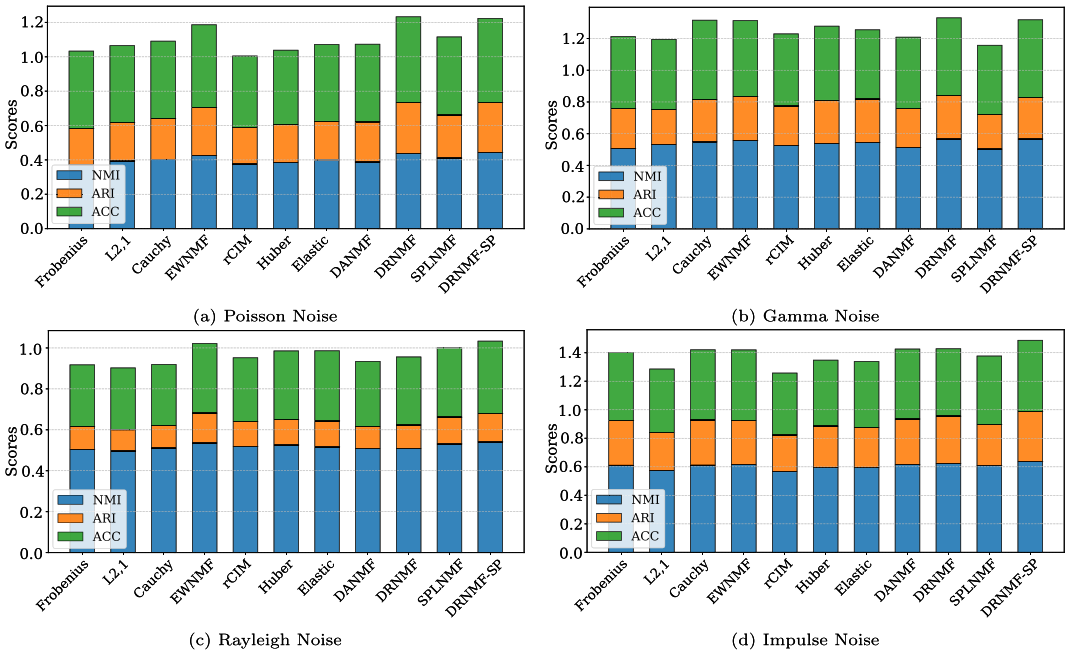


Fig. 9. Clustering result for different noise types on the USPS dataset.

heavy-tailed characteristics of Gamma noise, while the Frobenius norm contributes additional stability. These findings suggest that our method generalizes effectively even under unseen noise conditions.

We also test DRNMF-SP on the ORL dataset, this time contaminated with speckle noise modeled by a Rayleigh distribution. Rayleigh noise, commonly found in radar and medical imaging, introduces multiplicative distortions and exhibits an asymmetric distribution. Although this noise type also lies outside our predefined ambiguity set, DRNMF-SP again delivers strong results, as illustrated in Fig. 9c. The $L_{2,1}$ norm likely plays a key role in suppressing the asymmetric effects of Rayleigh noise, while the remaining loss terms contribute to the model's overall robustness. These outcomes further emphasize the adaptive nature of our approach in handling diverse and unanticipated noise scenarios. Finally, we examine the performance of DRNMF-SP on the Pneumonia dataset corrupted with impulse noise at a contamination rate of 0.15. Impulse noise, which introduces sharp, random spikes in pixel values, is common in image transmission errors and faulty sensor readings. Despite being absent from our ambiguity set, DRNMF-SP remains robust, as depicted in Fig. 9d. The self-paced learning mechanism is especially effective here, enabling the model to downweight the influence of extreme outliers during training. Overall, these results highlight the flexibility and generalization capacity of DRNMF-SP in the presence of challenging, real-world noise types not seen during training.

4.9. Running time

To assess computational efficiency, we compare the average runtimes (in seconds) of all methods across twelve datasets in Table 20. All experiments were performed on an Intel Core i7-3520 M CPU (2.9 GHz, 8 GB RAM). As expected, the standard Frobenius and $L_{2,1}$-NMF methods are the fastest, while deep (DANMF) and highly robust methods (Huber, DRNMF) incur higher costs. Our proposed DRNMF-SP achieves a favorable trade-off: it is consistently faster than Huber, DANMF, and DRNMF, while remaining close in efficiency to simpler robust baselines. This balance is especially evident on large-scale datasets (OrganA, Blood, Pneumonia), where DRNMF-SP substantially reduces runtime compared to DRNMF. Given the additional complexity introduced by its distributionally robust formulation, these results highlight DRNMF-SP's ability to achieve enhanced robustness while maintaining feasible computation times. It is also important to note that DRNMF and DRNMF-SP involve an initialization phase for single-objective optimization, which is not separately reported in the table. Nevertheless, the overall runtime results confirm that DRNMF-SP strikes a practical balance between computational efficiency and robustness, making it a viable choice for high-dimensional and noisy data scenarios.

4.10. Robustness on various noise rates

To further evaluate the robustness of DRNMF-SP under different levels of noise contamination, we conducted experiments across varying noise intensities. This analysis directly addresses the important question of how performance trends evolve as noise levels increase. For this, we evaluate the robustness of the proposed DRNMF-SP method by introducing impulse noise at varying levels on the Yale dataset. This experiment presents a significant challenge, as impulse noise can severely corrupt data by randomly replacing pixel values, making it difficult to extract a clean subspace. The noise contamination levels range from 0 % to 50 %, representing the proportion of affected pixels in the dataset. Fig. 10 illustrates the NMI, ARI, and ACC scores of DRNMF-SP compared to alternative

Table 20
Average running time (seconds) of different methods on twelve datasets.

Dataset	Fro	$L_{2,1}$	Cauchy	EWNMF	rCIM	Huber	Elastic	DANMF	DRNMF	SPLNMF	DRNMF-SP
Yale	0.129	0.162	0.161	0.168	0.178	0.589	0.163	1.81	0.601	0.338	0.195
ORL	0.745	0.931	0.824	0.779	0.934	1.82	1.01	4.46	3.20	1.09	1.13
COIL20	1.80	1.93	1.92	1.95	2.14	2.89	2.26	5.73	7.29	2.41	2.52
MNIST	0.507	0.597	0.563	0.648	0.743	1.08	0.702	3.78	2.31	0.98	0.79
Fashion	0.420	0.611	0.529	0.669	0.741	1.05	0.728	2.86	2.29	0.94	0.82
Seeds	0.052	0.070	0.068	0.074	0.081	0.245	0.076	0.65	0.49	0.19	0.12
Ecoli	0.085	0.112	0.111	0.118	0.127	0.356	0.124	0.92	0.68	0.29	0.21
USPS	0.332	0.421	0.405	0.438	0.481	0.985	0.512	2.46	1.86	0.74	0.58
UMIST	0.266	0.347	0.341	0.354	0.389	0.841	0.405	2.01	1.44	0.62	0.47
OrganA	15.50	20.55	21.54	20.65	25.19	35.21	25.15	44.9	82.2	32.5	31.2
Blood	12.42	17.25	18.56	17.53	21.61	31.08	20.21	38.7	69.7	27.1	24.9
Pneumonia	2.38	3.10	3.05	3.18	3.45	6.89	3.55	11.8	9.62	4.72	4.05

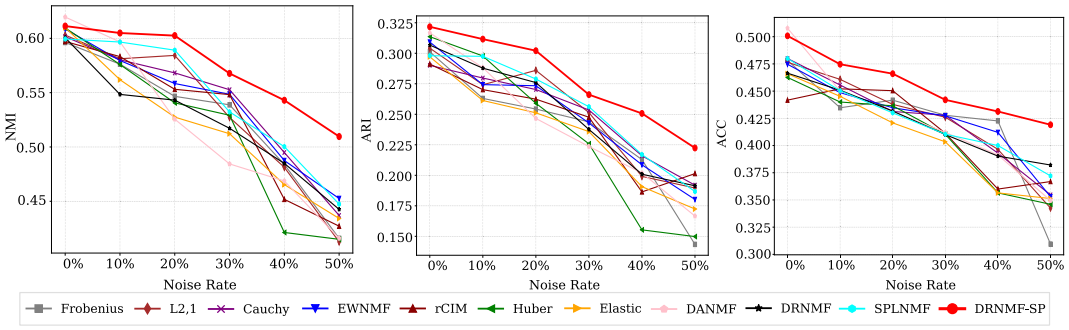


Fig. 10. NMI, ARI, and ACC Results on the Yale dataset with different impulse noise intensities.

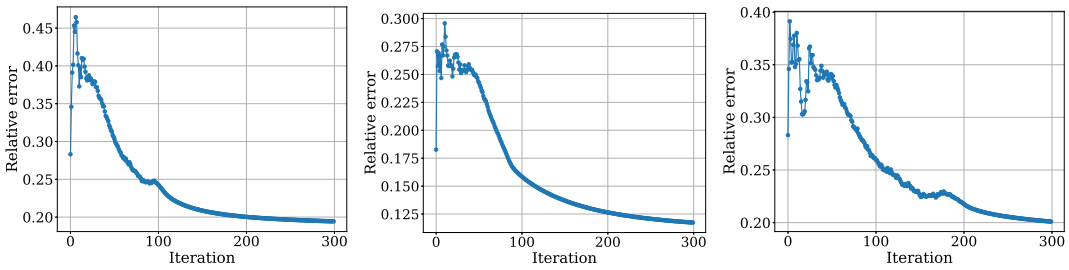


Fig. 11. Convergence analysis of DRNMF-SP over 300 iterations.

methods. As the noise level increases, the performance of most methods deteriorates significantly. However, DRNMF-SP consistently outperforms competing approaches, demonstrating greater resilience in maintaining clustering accuracy under high noise conditions. This highlights the method's ability to mitigate the adverse effects of impulse noise and preserve meaningful data structures. The results confirm the effectiveness and robustness of DRNMF-SP, which consistently outperforms other methods even as noise intensity increases. While all approaches experience some degradation under higher corruption levels, DRNMF-SP exhibits a slower performance decline, achieving state-of-the-art results under moderate noise and remaining stable and effective even in severely corrupted scenarios. These findings underscore its superiority and suitability for robust subspace learning in noisy environments.

4.11. Convergence analysis

To further assess the behavior of the proposed DRNMF-SP model, we conduct a convergence analysis by reporting the evolution of the objective function over 300 iterations on three representative datasets: Yale, ORL, and COIL20. The corresponding convergence curves are shown in Fig. 11. As illustrated, DRNMF-SP starts with a relatively low error value, followed by a slight increase during the first 30–40 iterations, after which the error decreases steadily until convergence. This temporary rise can be attributed to two intrinsic properties of the method. First, DRNMF-SP solves a multi-objective optimization problem whose relative weights are adaptively adjusted according to the data distribution and noise characteristics, which may cause short-term fluctuations before stabilization. Second, the model adopts a self-paced learning strategy, progressively incorporating samples from clean to noisy. Consequently, early iterations reflect low error values for cleaner data, while subsequent inclusion of noisier samples temporarily raises the error

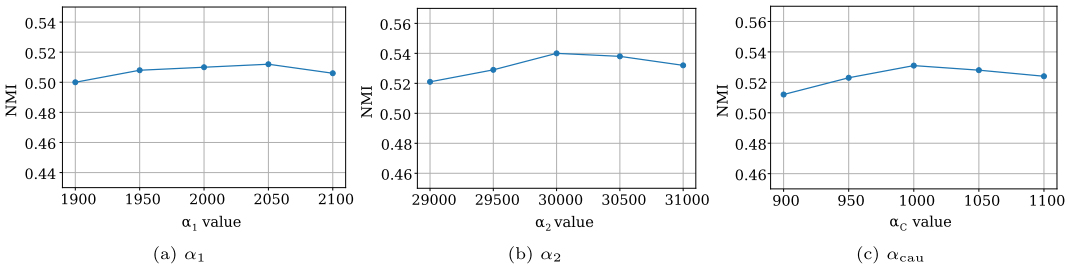


Fig. 12. Sensitivity of DRNMF-SP to the self-paced learning parameters α_τ on the Yale dataset.

before convergence. Overall, despite these initial fluctuations, DRNMF-SP consistently converges to a stable solution, demonstrating an effective balance between reconstruction accuracy and robustness to noise.

4.12. Hyperparameter sensitivity

An important practical consideration is the sensitivity of DRNMF-SP to its hyperparameters. The adaptive weight update η is dynamically set as $\eta^{[t]} = 1/(t+1)$, starting from $\frac{1}{2}$, which automatically decreases the learning rate as training progresses and thus requires no manual adjustment. This leaves the self-paced learning parameters α_τ as the main hyperparameters to be tuned. These parameters control the pace at which harder samples are incorporated into training, directly influencing both convergence stability and robustness to noise. We systematically evaluated the effect of each α_τ on the Yale dataset under three loss formulations: $\ell_{2,1}$, Frobenius, and Cauchy. In each case, one α_τ was varied while keeping the others fixed to assess its isolated impact. Fig. 12 summarizes these results. We observed that performance drops markedly when α_τ is too small—leading to overly conservative sample inclusion—or too large, which can prematurely include noisy or outlier data. However, the overall performance remains stable within a broad intermediate range, indicating that DRNMF-SP is not overly sensitive to precise hyperparameter tuning. This robustness simplifies practical deployment, as approximate values determined by cross-validation or empirical heuristics are generally sufficient to achieve strong results across different datasets and loss settings.

4.13. Discussion

The proposed DRNMF-SP model provides not only improved robustness in factorization tasks but also broader implications for the scientific community. Its relevance lies in reinforcing the usability of NMF in domains where data is inherently imperfect, such as medical imaging, bioinformatics, and computer vision. In such applications, the interpretability of nonnegative factors is a decisive advantage, but conventional NMF often fails under heterogeneous or heavy-tailed noise. By integrating distributionally robust optimization with self-paced learning, our framework addresses this gap, enabling interpretable decompositions to remain stable even in the presence of severe data contamination. A key distinction of DRNMF-SP compared with earlier robust NMF variants is its ability to adaptively balance multiple loss functions rather than committing to a single noise model. This flexibility allows the method to operate effectively across diverse data conditions, which is highly relevant for real-world practice where noise distributions are rarely known in advance. Furthermore, the self-paced scheme enhances optimization stability, providing a mechanism to gradually incorporate complex samples while avoiding convergence to poor local minima. Beyond its immediate results, the framework is extensible: the ambiguity set can be augmented with additional loss functions tailored to specific domains, and the self-paced mechanism can be embedded in other matrix factorization or representation learning models. These features make DRNMF-SP a promising foundation for further research in interpretable and robust learning, supporting the community's growing interest in models that are both reliable and adaptable.

A comparative view of the results across all datasets confirms that DRNMF-SP consistently achieves higher clustering performance than classical and recent robust NMF models. In particular, the method demonstrates superior resilience under both extreme outliers and various distribution noise, where competing approaches often suffer from degraded accuracy. These findings highlight the advantage of simultaneously leveraging multiple loss functions and self-paced learning, in contrast to prior methods that rely on a single loss or lack adaptive instance weighting. Beyond benchmark comparisons, the improvements have clear implications for applications where robust and interpretable representations are crucial. Examples include image recognition under partial occlusion, medical imaging where scans are frequently corrupted by noise or missing regions, and biological data analysis where measurements often contain heterogeneous errors. In such contexts, the ability of DRNMF-SP to maintain stable and meaningful decompositions makes it a promising tool for reliable downstream tasks such as clustering, classification, and anomaly detection.

5. Conclusion and future works

This paper introduces Distributionally Robust Nonnegative Matrix Factorization with Self-Paced Adaptive Multi-Loss Fusion (DRNMF-SP), a robust matrix factorization framework that combines distributionally robust optimization with self-paced learning. The proposed method integrates multiple loss functions through adaptive weighting to model uncertainty in noise types, while self-paced learning allows the algorithm to focus on clean and informative samples before gradually incorporating more challenging

ones. This dual strategy enables DRNMF-SP to effectively handle heterogeneous noise and extreme outliers. An efficient iterative reweighted algorithm ensures low computational cost, comparable to standard NMF. Extensive experiments across benchmark datasets demonstrate that DRNMF-SP consistently outperforms existing robust and distributionally robust NMF methods, highlighting the effectiveness of unifying robustness, adaptive loss design, and curriculum-style learning. A key strength of DRNMF-SP lies in its double adaptive mechanism: the distributionally robust component adjusts the importance of different loss functions based on the underlying noise characteristics, while the self-paced component dynamically controls the order in which data samples are learned. This coordinated adjustment at both the loss and the data levels improves the flexibility and robustness of the model, leading to more reliable low-dimensional representations under diverse noise conditions.

This work opens several avenues for future exploration. A key direction is to conduct a formal convergence analysis of DRNMF-SP. Although the current focus has been on algorithm design and empirical validation, studying the theoretical behavior of the optimization process can further strengthen the method's foundations.

Also, exploring such an Online NMF-inspired online extension, such as the ONMFO approach [47], along with its computational trade-offs and convergence properties, remains a promising avenue for future research. In an ONMF-based extension, the coefficient matrix would be updated locally for each new data sample, while the basis would be refined globally through incremental updates using accumulated statistics. This approach enables the model to adapt continuously to incoming data with minimal memory and computational cost, making it suitable for large-scale or real-time applications. In addition, although class-imbalanced data were not explicitly examined in this study, such scenarios are highly relevant in many real-world applications. The proposed framework is general and can be naturally extended to handle imbalanced data distributions, making the investigation of its performance under severe class imbalance an important direction for future work.

Finally, since DRNMF-SP is currently a shallow model, another future direction is to extend it into a deep version. For example, one could design a multi-layer architecture that stacks several DRNMF-SP layers or integrates it into deep learning frameworks such as autoencoders. This could improve its ability to capture more complex structures in data while preserving robustness. Exploring such deep extensions may improve performance in challenging tasks such as image analysis, speech recognition, or recommendation systems with noisy or weak supervision.

CRediT authorship contribution statement

Wafa Barkhoda: Writing – original draft, Software, Methodology, Investigation. **Amjad Seyedi:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Nicolas Gillis:** Writing – review & editing, Validation, Investigation, Conceptualization. **Fardin Akhlaghian Tab:** Writing – review & editing, Validation, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered potential competing interests:

Amjad Seyedi reports that financial support was provided by the European Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Amjad Seyedi acknowledges the support of the European Union (ERC consolidator, eLinoR, no 101085607).

Data availability

Data will be made available on request.

References

[1] N. Gillis, Nonnegative Matrix Factorization, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2020.
[2] Z. Li, J. Tang, X. He, Robust structured nonnegative matrix factorization for image representation, IEEE Trans. Neural Netw. Learn. Syst. 29 (5) (2018) 1947–1960.
[3] W. Luo, Z. Wu, N. Zhou, Hypergraph-based convex semi-supervised unconstraint symmetric matrix factorization for image clustering, Inf. Sci. 680 (2024) 121138.
[4] S. Yu, B. Mao, Y. Zhou, Y. Liu, C. Yi, F. Li, D. Yao, P. Xu, X. San Liang, T. Zhang, Large-scale cortical network analysis and classification of MI-BCI tasks based on Bayesian nonnegative matrix factorization, IEEE Trans. Neural Syst. Rehabil. Eng. 32 (2024) 2187–2197.
[5] B. Zhong, J.-S. Wu, W. Huang, W.-S. Zheng, Cluster structure augmented deep nonnegative matrix factorization with low-rank tensor learning, Inf. Sci. 670 (2024) 120585.
[6] Z. He, Y. Lin, Z. Lin, C. Wang, Multi-label feature selection via similarity constraints with non-negative matrix factorization, Knowl.-Based Syst. 297 (2024) 111948.
[7] B. Wang, J. Fan, Robust matrix completion with heavy-tailed noise, J. Am. Stat. Assoc. 120 (550) (2025) 922–934.
[8] J. Peng, W. Sun, F. Jiang, H. Chen, Y. Zhou, Q. Du, A general loss-based nonnegative matrix factorization for hyperspectral unmixing, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5.
[9] E.Y. Lam, Non-negative matrix factorization for images with Laplacian noise, in: 2008 IEEE Asia Pacific Conference on Circuits and Systems, 2008, pp. 798–801.
[10] D. Kong, C. Ding, H. Huang, Robust nonnegative matrix factorization using L21-norm, in: International Conference on Information and Knowledge Management, 2011, pp. 673–682.
[11] L. Du, X. Li, Y.-D. Shen, Robust nonnegative matrix factorization via half-quadratic minimization, in: 2012 IEEE 12th International Conference on Data Mining, 2012, pp. 201–210.

[12] S. Peng, W. Ser, Z. Lin, B. Chen, Robust sparse nonnegative matrix factorization based on maximum correntropy criterion, in: 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 2018, pp. 1–5.
[13] A. Liutkus, D. Fitzgerald, R. Badeau, Cauchy nonnegative matrix factorization, in: 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2015, pp. 1–5.
[14] D. Bertsimas, V. Gupta, N. Kallus, Data-driven robust optimization, Math. Program. 167 (2018) 235–292.
[15] H. Gao, F. Nie, W. Cai, H. Huang, Robust capped norm nonnegative matrix factorization: capped norm NMF, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015, pp. 871–880.
[16] N. Guan, T. Liu, Y. Zhang, D. Tao, L.S. Davis, Truncated Cauchy non-negative matrix factorization, IEEE Trans. Pattern Anal. Mach. Intell. 41 (1) (2019) 246–259.
[17] H. Scarf, A min-max solution of an inventory problem, in: Studies in the Mathematical Theory of Inventory and Production, Stanford University Press, Redwood City, CA, 1958, pp. 201–209.
[18] J. Cheng, R. Li-Yang Chen, H.N. Najm, A. Pinar, C. Safta, J.-P. Watson, Distributionally robust optimization with principal component analysis, SIAM J. Optim. 28 (2) (2018) 1817–1841.
[19] S. Zhu, L. Xie, M. Zhang, R. Gao, Y. Xie, Distributionally robust weighted k-nearest neighbors, in: Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 29088–29100.
[20] D. Faccini, F. Maggioni, F.A. Potra, Robust and distributionally robust optimization models for linear support vector machine, Comput. Oper. Res. 147 (2022) 105930.
[21] N. Gillis, L.T.K. Hien, V. Leplat, V.Y.F. Tan, Distributionally robust and multi-objective nonnegative matrix factorization, IEEE Trans. Pattern Anal. Mach. Intell. 44 (8) (2022) 4052–4064.
[22] W. Barkhoda, A. Seyedi, N. Gillis, F. Akhlaghian Tab, Instance-wise distributionally robust nonnegative matrix factorization, Pattern Recognit. 169 (2026) 111732.
[23] S. Soleymanbaigi, A. Seyedi, F. Akhlaghian Tab, F. Daneshfar, Encoder-decoder nonnegative matrix factorization with β-divergence for data clustering, Pattern Recognit. 171 (2026) 112211.
[24] H. Xiong, D. Kong, F. Nie, Cauchy balanced nonnegative matrix factorization, Artif. Intell. Rev. 56 (10) (2023) 11867–11903.
[25] J.S. Cavazos, J.A. Fessler, L. Balzano, ALPCAH: sample-wise heteroscedastic PCA with tail singular value regularization, in: 2023 International Conference on Sampling Theory and Applications (SAMPTA), 2023, pp. 1–6.
[26] Z. Yu, G. Dong, H. Liu, Sar image quality assessment: from sample-wise to class-wise, Remote Sens. 15 (8) (2023).
[27] F. Eiras, M. Alfarra, P. Torr, M.P. Kumar, P.K. Dokania, B. Ghanem, A. Bibi, Ancer: anisotropic certification via sample-wise volume maximization, Trans. Mach. Learn. Res. (2022).
[28] L. Chen, C.-T. Wu, C.-H. Lin, R. Dai, C. Liu, R. Clarke, G. Yu, J.E. Van Eyk, D.M. Herrington, Y. Wang, Swcam: estimation of subtype-specific expressions in individual samples with unsupervised sample-wise deconvolution, Bioinformatics 38 (5) (2021) 1403–1410.
[29] H. Wang, W. Yang, N. Guan, Cauchy sparse NMF with manifold regularization: a robust method for hyperspectral unmixing, Knowl.-Based Syst. 184 (2019) 104898.
[30] M. Kumar, B. Packer, D. Koller, Self-paced learning for latent variable models, in: Advances in Neural Information Processing Systems, vol. 23, 2010, pp. 1189–1197.
[31] Z. Liu, X. Feng, Y. Wang, W. Zuo, Self-paced learning enhanced neural matrix factorization for noise-aware recommendation, Knowl.-Based Syst. 213 (2021) 106660.
[32] Z. Huang, Y. Ren, X. Pu, L. Pan, D. Yao, G. Yu, Dual self-paced multi-view clustering, Neural Networks 140 (2021) 184–192.
[33] S.A. Seyedi, S.S. Ghodsi, F. Akhlaghian, M. Jalili, P. Moradi, Self-paced multi-label learning with diversity, in: Asian Conference on Machine Learning, vol. 101, PMLR, 2019, pp. 790–805.
[34] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, A. Hauptmann, Self-paced learning for matrix factorization, Proc. AAAI Conf. Artif. Intell. 29 (1) (Feb. 2015).
[35] X. Zhu, Z. Zhang, Improved self-paced learning framework for nonnegative matrix factorization, Pattern Recognit. Lett. 97 (2017) 1–7.
[36] S. Huang, P. Zhao, Y. Ren, T. Li, Z. Xu, Self-paced and soft-weighted nonnegative matrix factorization for data representation, Knowl.-Based Syst. 164 (2019) 29–37.
[37] E.J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, J. ACM 58 (3) (2011).
[38] L. Zhang, Z. Chen, M. Zheng, X. He, Robust non-negative matrix factorization, Front. Electr. Electron. Eng. China 6 (2) (2011) 192–200.
[39] I. Daubechies, R. DeVore, M. Fornasier, C.S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery, Commun. Pure Appl. Math. 63 (1) (2010) 1–38.
[40] Q. Wang, X. He, X. Jiang, X. Li, Robust bi-stochastic graph regularized matrix factorization for data clustering, IEEE Trans. Pattern Anal. Mach. Intell. 44 (1) (2022) 390–403.
[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[42] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, MedMNIST v2 - a large-scale lightweight benchmark for 2D and 3D biomedical image classification, Sci. Data 10 (1) (2023) 41.
[43] D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Advances in Neural Information Processing Systems, vol. 13, MIT Press, 2000, pp. 535–541.
[44] J. Wei, C. Tong, B. Wu, Q. He, S. Qi, Y. Yao, Y. Teng, An entropy weighted nonnegative matrix factorization algorithm for feature representation, IEEE Trans. Neural Netw. Learn. Syst. 34 (9) (2023) 5381–5391.
[45] H. Xiong, D. Kong, Elastic nonnegative matrix factorization, Pattern Recognit. 90 (2019) 464–475.
[46] N. Salahian, F.A. Tab, S.A. Seyedi, J. Chavoshinejad, Deep autoencoder-like NMF with contrastive regularization and feature relationship preservation, Expert Syst. Appl. 214 (2023) 119051.
[47] R. Zhao, V.Y.F. Tan, Online nonnegative matrix factorization with outliers, IEEE Trans. Signal Process. 65 (3) (2017) 555–570.