Adaptive Quantile Low-Rank Matrix Factorization

Shuang Xu\textsuperscript{a}, Chunxia Zhang\textsuperscript{a,∗}, Jiangshe Zhang\textsuperscript{a}

\textsuperscript{a}School of Mathematics and Statistics, Xi’an Jiaotong University, Xian 710049, China

Abstract

Low-rank matrix factorization (LRMF) has received much popularity owing to its successful applications in both computer vision and data mining. By assuming the noise term to come from a Gaussian, Laplace or mixture of Gaussian distributions, significant efforts have been made on optimizing the (weighted) $L_1$ or $L_2$-norm loss between an observed matrix and its bilinear factorization. However, the type of noise distribution is generally unknown in real applications and inappropriate assumptions will inevitably deteriorate the behavior of LRMF. On the other hand, real data are often corrupted by skew rather than symmetric noise.

To tackle this problem, this paper presents a novel LRMF model called AQ-LRMF by modeling noise with a mixture of asymmetric Laplace distributions. An efficient algorithm based on the expectation-maximization (EM) algorithm is also offered to estimate the parameters involved in AQ-LRMF. The AQ-LRMF model possesses the advantage that it can approximate noise well no matter whether the real noise is symmetric or skew. The core idea of AQ-LRMF lies in solving a weighted $L_1$ problem with weights being learned from data. The experiments conducted with synthetic and real datasets show that AQ-LRMF outperforms several state-of-the-art techniques. Furthermore, AQ-LRMF also has the superiority over the other algorithms that it can capture local structural information contained in real images.

Keywords: Low-rank matrix factorization, Mixture of asymmetric Laplace distributions, Expectation maximization algorithm, Skew noise

∗Corresponding author

Email address: cxzhang@mail.xjtu.edu.cn (Chunxia Zhang)
1. Introduction

Researchers from machine learning [1], computer vision [2] and statistics [3] have paid increasing attention to low-rank matrix factorization (LRMF) [4]. Generally speaking, many real-world modeling tasks can be attributed as the problems of LRMF. The tasks include but are not limited to recommender systems [5], subspace clustering [6, 7], link prediction [8], face recognition [9] and image denoising [10].

The key idea of LRMF is to approximate a given matrix by the product of two low-rank matrices. Specifically, given an observed matrix $X \in \mathbb{R}^{m \times n}$, LRMF aims at solving the optimization problem

$$
\min_{U, V} ||\Omega \odot (X - UV^T)||,
$$

where $U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}$ (usually, $r \ll \min(m, n)$) and $\odot$ denotes the Hadamard product, that is, the element-wise product. The indicator matrix $\Omega = (\omega_{ij})_{m \times n}$ implies whether some data are missing, where $\omega_{ij} = 1$ if $x_{ij}$ is non-missing and 0 otherwise. The symbol $|| \cdot ||$ indicates a certain norm of a matrix, in which the most prevalent one is $L_2$ norm. It is well-known that singular value decomposition provides a closed-form solution for $L_2$-norm LRMF without missing entries. In addition, researchers have presented many fast algorithms to solve Eq. (1) when $X$ contains missing entries, as well. The $L_2$-norm LRMF greatly facilitates theoretical analysis, but it provides the best solution in sense of maximum likelihood principle only when noise is indeed sampled from a Gaussian distribution. If noise is from a heavy-tailed distribution or data are corrupted by outliers, $L_2$-norm LRMF is likely to perform badly. Thereafter, $L_1$-norm LRMF begins to gain increasing interest of both theoretical researchers and practitioners due to its robustness [11]. In fact, $L_1$-norm LRMF hypothesizes that noise is from a Laplace distribution. As is often the case with $L_2$-norm LRMF, $L_1$-norm LRMF may provide unexpected results as well if its assumptions are violated.

Because the noise in real data generally deviates far away from a Gaussian or Laplace distribution, analysts are no longer satisfied with $L_1$- or $L_2$-norm LRMF. To improve the robustness of LRMF, researchers attempt to directly model unknown noise via a mixture of Gaussians (MoG) due to its good property to universally approximate any continuous distribution [12, 13]. Nevertheless, the technique cannot fit real noise precisely in some complex cases. For example, in theory, infinite Gaussian components are required...
to approximate a Laplace distribution. In practice, we only utilize finite Gaussian components due to the characteristics of MoG. On the other hand, Gaussian, Laplace and MoG distributions are all symmetric. In the conditions with real noise being skew, they may provide unsatisfactory results.

As a matter of fact, there are no strictly symmetric noise in real images. For instance, Figure 1 illustrates several examples in which the real noise is either skewed to the left (e.g., (a-4) and (c-4)) or the right (e.g., (b-4)). In these situations, the symmetric distributions like Gaussian or Laplace are inadequate to approximate the noise. In statistics, to deal with an asymmetric noise distribution, a preliminary exploration called quantile regression has been made. Consider a simple case that there is only one covariate $X$, the quantile regression coefficient $\beta$ can be obtained by

$$\hat{\beta}_\kappa = \arg \min_{\beta} \sum_{i=1}^{n} \rho_\kappa(y_i - x_i \beta),$$

(2)

where $\{(y_x, x_i)\}_{i=1}^{n}$ are $n$ observations and $\kappa$ is a pre-defined asymmetry parameter. Moreover, the quantile loss $\rho_\kappa(\cdot)$ is defined as

$$\rho_\kappa(\epsilon) = \epsilon [\kappa - \mathbb{I}(\epsilon < 0)]$$

$$= |\epsilon| \mathbb{I}(\epsilon \geq 0) + (1 - \kappa)\mathbb{I}(\epsilon < 0)$$

(3)

with $\mathbb{I}(\cdot)$ being the indicator function. Evidently, the quantile loss with $\kappa = 1/2$ corresponds to the $L_1$-norm loss. From the Bayesian viewpoint, the estimate obtained by minimizing the quantile loss in (2) coincides with the result by assuming noise coming from an asymmetric Laplace distribution (ALD) [14, 15].

To overcome the shortcomings of existing LRMF methods that they assume a specific type of noise distribution, we present in this paper an adaptive quantile LRMF (AQ-LRMF) algorithm. The key idea of AQ-LRMF is to model noise via a mixture of asymmetric Laplace distributions (MoAL). The expectation maximization (EM) algorithm is employed to estimate parameters, under the maximum likelihood framework. The novelty of AQ-LRMF and our main contributions can be summarized as follows.

(1). The M-step of the EM algorithm corresponds to a weighted $L_1$-norm LRMF, where the weights encode the information about skewness and outliers.

(2). The weights are automatically learned from data under the framework of EM algorithm.
Figure 1: (a) and (b) illustrate two face images corresponding to underexposure and overexposure cases, respectively. In particular, (a-1) and (b-1) are face images captured with improper light sources while (a-2) and (b-2) are face images obtained with proper light sources. (a-3) and (b-3) are residual images in which the yellow (blue) locations indicate positive (negative) values. (a-4) and (b-4) illustrate the histograms of the residual images as well as the PDF curves fitted by ALD with $\alpha_a = 115, \kappa_a = 0.71, \lambda_a = 0.05$ and $\alpha_b = -9, \kappa_b = 0.44, \lambda_b = 0.11$, respectively. The skewness of the residual face in (a-3) is $-0.72$ whilst that for (b-3) is $0.69$. (c) shows a hyperspectral image. Similar to cases (a) and (b), the images from (c-1) to (c-4) are original, de-noised, noise images and the histogram of residuals, respectively. The skewness of the noise image (c-3) is $-0.55$. In (c-4), the fitted ALD is with $\alpha = 33, \kappa = 0.75$ and $\lambda = 0.05$. Obviously, the distributions of noise shown here are all asymmetric.
Different from quantile regression, our method does not need to pre-define the asymmetry parameter of quantile loss, because it is adaptively determined by data.

Our model can capture local structural information contained in some real images, although we do not encode it into our model.

The experiments show that our method can effectively approximate many different kinds of noise. If the noise has a strong tendency to take a particular sign, AQ-LRMF will produce better estimates than a method which assumes a symmetric noise distribution. In comparison with several state-of-the-art methods, the superiority of our method is demonstrated in both synthetic and real-data experiments such as image inpainting, face modeling, hyperspectral image (HSI) construction and so on.

The rest of the paper is organized as follows. Section 2 presents related work of LRMF. In section 3, we propose the AQ-LRMF model and also provide an efficient learning algorithm for it. Section 4 includes experimental studies. At last, some conclusions are drawn in section 5.

2. Related work

The study of robust LRMF has a long history. Srebro and Jaakkola [16] suggested to use a weighted $L_2$ loss to improve LRMF’s robustness. The problem can be solved by a simple but efficient EM algorithm. However, the choice of weights significantly affects its capability. Thereafter, Ke and Kanade [11] attempted to replace $L_2$ loss with $L_1$ loss and to solve the optimization by alternated linear or quadratic programming (ALP/AQP). In order to catalyze convergence, Eriksson and Hengel [17] developed the $L_1$-Wiberg algorithm. Kim et al. [18] used alternating rectified gradient method to solve a large-scale $L_1$-norm LRMF. The simulated experiments showed that this method performs well in terms of both matrix reconstruction performance and computational complexity. Okutomi et al. [19] modified the objective function of $L_1$-Wiberg by adding the nuclear norm of $V$ and the orthogonality constraints on $U$. This method has been shown to be effective in addressing structure from motion issue. Despite the non-convexity and non-smoothness of $L_1$-norm LRMF, Meng et al. [20] proposed a computationally efficient algorithm, cyclic weighted median (CWM) method, by solving a sequence of scalar minimization sub-problems to obtain the optimal.
solution. Inspired by majorization-minimization technique, Lin et al. [21] proposed LRMF-MM to solve an LRMF optimization task with $L_1$ loss plus the $L_2$-norm penalty placing on $U$ and $V$. In each step, they upper bound the original objective function by a strongly convex surrogate and then minimize the surrogate. Experiments on both simulated and real data sets testify the effectiveness of LRMF-MM. Li et al. [22] considered a similar problem, but they replace the $L_2$-norm penalty imposed on $U$ with $U^TU = I$. This model is solved by augmented Lagrange multiplier method. Furthermore, the authors of [22] designed a heuristic rank estimator for their model. As argued in introduction, $L_1$ loss actually corresponds to the Laplace-distributed noise. When the real distribution of noise deviates too far from Laplace, the robustness of $L_1$ LRMF will be suspectable.

Recently, the research community began to focus on probabilistic extensions of robust matrix factorizations. Generally speaking, it is assumed that $X = UV^T + E$, where $E$ is a noise matrix. Lakshminarayanan et al. [23] replaced Gaussian noise with Gaussian scale mixture noise. Nevertheless, it may be ineffective when processing heavy-tailed (such as Laplace-type) noise. Wang et al. [24] proposed a probabilistic $L_1$-norm LRMF, but they did not employ a fully Bayesian inference process. Beyond Laplace noise, Meng and Torre [12] presented a robust LRMF with unknown noise modeled by an MoG. In essence, the method iteratively optimizes $\min_{U,V,\theta} ||W(\theta) \odot (X - UV^T)||_{L_2}$, where $\theta$ are the MoG parameters which are automatically updated during optimization, and $W(\theta)$ is the weight function of $\theta$. Due to the benefit to adaptively assign small weights to corrupted entries, MoG-LRMF has been reported to be fairly effective. More recently, Cao et al. [25] presented a novel LRMF model by assuming noise as a mixture of exponential power (MoEP) distributions and also offered the corresponding learning algorithm.

On the other hand, robust principle component analysis (robust PCA) [26] considers an issue that is similar to LRMF, that is,

$$\min_{A,E} \text{rank}(A) + \lambda ||E||_{L_0} \quad \text{s.t.} \quad X = A + E.$$  

(4)

The underlying assumption of robust PCA is that the original data can be decomposed into the sum of a low-rank matrix and a sparse outlier matrix (i.e., the number of non-zero elements in $E$ is small). Clearly, $A$ plays the same role as the product of $U$ and $V^T$. Since Eq. (4) involves a non-convex objective function,
consider a tractable convex alternative, called principal component pursuit, to handle the corresponding problem, namely,
\[
\min_{A, E} \|A\|_* + \lambda \|E\|_{L_1} \quad \text{s.t. } X = A + E,
\]
where \(\|\cdot\|_*\) denotes the nuclear norm. It is worthwhile that principal component pursuit sometimes may fail to recover \(E\) when the real observation is also corrupted by a dense inlier matrix. To overcome this shortcoming, Zhou et al. [27] proposed the stable principal component pursuit (SPCP) by solving
\[
\min_{A, E} \|A\|_* + \lambda \|E\|_{L_1} \quad \text{s.t. } \|X - A - E\|_{L_2} \leq \varepsilon.
\]
Actually, the underlying assumption of SPCP is \(X = A + N + E\), where \(A\) is low-rank component, \(E\) is the sparse outliers and \(N\) is the small-magnitude noise that can be modeled by Gaussian. Both theory and experiments have shown that SPCP guarantees the stable recovery of \(E\) [27][26].

3. Adaptive Quantile LRMF (AQ-LRMF)

3.1. Motivation

Generally speaking, researchers employ the \(L_2\) or \(L_1\) loss function when solving a low-rank matrix factorization problem. As argued in introduction, \(L_2\) or \(L_1\) loss implicitly hypothesizes that the noise distribution is symmetric. Nevertheless, the noise in real data is often asymmetric and Fig. 1 illustrates several examples.

In Fig. 1 there are two face images and a hyperspectral image. Fig. 1 (a) displays a face image captured with a poor light source. There are cast shadows in a large area, while there exists overexposure phenomenon in a small area. As a result, the noise is negative skew. By contrast, Fig. 1 (b) illustrates a face image captured under a strong light source. Because of the camera range settings, there are saturated pixels, especially on the forehead. Under this circumstance, the noise is positive skew. Fig. 1 (c) shows a hyperspectral image that is mainly corrupted by stripe and Gaussian noise. Its residual image indicates that the signs of the noise are unbalanced, i.e., more pixels are corrupted by noise with negative values. Actually,
the skewness values of three residual (noise) images are −0.72, 0.69 and −0.55, respectively. Note that a symmetric distribution has skewness 0, the noise contained in these real data sets is thus asymmetric.

As a matter of fact, the noise in real data can hardly be governed by a strictly symmetric probability distribution. Therefore, it is natural to utilize an asymmetric distribution to model realistic noise. In statistics, researchers usually make use of a quantile loss function defined in [3] to address this issue. It has been shown that quantile loss function corresponds to the situation that noise is from an asymmetric Laplace distribution [14, 15]. In order to improve the performance of low-rank matrix factorization, we attempt to use a mixture of asymmetric Laplacian distributions (MoAL) to approximate noise.

3.2. Asymmetric Laplace distribution

In what follows, we use $AL(\epsilon|\alpha, \lambda, \kappa)$ to denote an ALD with location, scale and asymmetric parameters $\alpha, \lambda > 0$ and $0 < \kappa < 1$, respectively. Its probability distribution function (PDF) is

$$p(x; \alpha, \lambda, \kappa) = \lambda \kappa (1 - \kappa) \begin{cases} \exp(\lambda(1 - \kappa)(x - \alpha)), & \text{if } x < \alpha; \\ \exp(-\lambda \kappa(x - \alpha)), & \text{if } x \geq \alpha; \end{cases}$$

$$= \lambda \kappa (1 - \kappa) \exp(-|x - \alpha| \lambda \kappa (x - \alpha \geq 0)) + (1 - \kappa) I(x - \alpha < 0)).$$

(7)

Obviously, the location parameter $\alpha$ is exactly the mode of an ALD. In Fig. 2, we demonstrate the PDF curves for several ALDs with different parameters. The asymmetry parameter $\kappa$ controls the skewness of an ALD and $sk_{ALD} \in (-2, 2)$. In general, an ALD is positive skew if $0 < \kappa < 0.5$, and is negative skew if $0.5 < \kappa < 1$. If $\kappa = 0.5$, the ALD becomes a Laplace distribution. The smaller the scale parameter $\lambda$ is, the more heavy-tailed ALD is.

It is worthwhile that skew Gaussian distributions [28] are also prevailing in both theory and applications. However, it is not ideal for the analysis of LRMF. On one hand, the PDF of a skew Gaussian distribution is complex. On the other hand, its skewness lies in $(-1, 1)$ which is only a subset of the range of $sk_{ALD}$. Due to this fact, the fitting capability of an ALD is greater than that of a skew Gaussian distribution.
3.3. AQ-LRMF model

To enhance the robustness of LRMF in situations with skew and heavy-tailed noise, we propose an adaptive quantile LRMF (AQ-LRMF) model by modeling unknown noise as an MoAL. In particular, we consider a generative model of the observed matrix $X \in \mathbb{R}^{m \times n}$. For its each entry $x_{ij}$, suppose that there is

$$x_{ij} = u_i v_j^T + \epsilon_{ij}, \quad (8)$$

where $u_i$ is the $i$th row of $U$, $v_j$ is the $j$th row of $V$, and $\epsilon_{ij}$ is the noise term. In AQ-LRMF, we assume that $\epsilon_{ij}$ is distributed as an MoAL, namely,

$$p(\epsilon_{ij}) = \sum_{s=1}^{S} \pi_s AL_s(\epsilon_{ij} | 0, \lambda_s, \kappa_s), \quad (9)$$

in which $AL_s(\epsilon_{ij} | 0, \lambda_s, \kappa_s)$ stands for an asymmetric distribution with parameters $\alpha = 0, \lambda = \lambda_s$ and $\kappa = \kappa_s$. Meanwhile, $\pi_s$ indicates the mixing proportion with $\pi_s \geq 0$ and $\sum_{s=1}^{S} \pi_s = 1$, and $S$ means the number of mixture components.

To facilitate the estimation of unknown parameters, we further equip each noise $\epsilon_{ij}$ with an indicator vector $z_{ij} = (z_{ij1}, z_{ij2}, \cdots, z_{ijS})^T$ where $z_{ij} \in \{0, 1\}$ and $\sum_{s=1}^{S} z_{ij} = 1$. Here, $z_{ij} = 1$ indicates that the noise $\epsilon_{ij}$ is drawn from the $s$th AL distribution. Evidently, $z_{ij}$ follows a multinomial distribution, i.e., $z_{ij} \sim \mathcal{M}(\pi_1, \cdots, \pi_S)$. Under these assumptions, we can have

$$p(\epsilon_{ij}) = \prod_{s=1}^{S} [\pi_s AL_s(\epsilon_{ij} | 0, \lambda_s, \kappa_s)]^{z_{ij}}, \quad (10)$$
Now, it is easy to obtain the probability of $x_{ij}$ as
\[ p(x_{ij}|u_i, v_j, \lambda, K, \pi) = \prod_{s=1}^{S} \left[ \pi_s AL_s(x_{ij}|u_i, v_j^T, \lambda_s, \kappa_s) \right]^{z_{ij}s}, \]
(11)
where $\lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_S\}$, $K = \{\kappa_1, \kappa_2, \cdots, \kappa_S\}$ and $\pi = \{\pi_1, \pi_2, \cdots, \pi_S\}$ are unknown parameters. To estimate $U, V$ as well as $\lambda, K, \pi$, we employ the maximum likelihood principle. Consequently, the goal is to maximize the log-likelihood function of complete data shown below, namely,
\[ \ell(U, V, \lambda, K, \pi) = \sum_{(i,j) \in \Omega} \sum_{s=1}^{S} z_{ij}s \left[ \log AL_s(x_{ij}|u_i, v_j^T, \lambda_s, \kappa_s) + \log \pi_s \right], \]
(12)
where $\Omega$ denotes the index set of the non-missing entries of data. Subsequently, we will discuss how to maximize the log-likelihood function $\ell(U, V, \lambda, K, \pi)$ to get our interested items.

3.4. Learning of AQ-LRMF

Since each $x_{ij}$ associates with an indicator vector $z_{ij}$, the EM algorithm [29] is utilized to train the AQ-LRMF model. Particularly, the algorithm needs to iteratively implement the following two steps (i.e., E-step and M-step) until it converges. For ease of exposition, we let $e_{ij} = x_{ij} - u_i v_j^T$ and abbreviate $AL_s(e_{ij}|0, \lambda_s, \kappa_s)$ as $AL_s(e_{ij})$ in the following discussions.

**E-step:** Compute the conditional expectation of the latent variable $z_{ij}s$ as
\[ \gamma_{ij}s = E(z_{ij}s|x_{ij}) = \frac{\pi_s AL_s(e_{ij})}{\sum_{a=1}^{S} \pi_a AL_a(e_{ij})}. \]
(13)

In order to attain the updating rules of other parameters, we need to compute the $Q$-function. According to the working mechanism of EM algorithm, the $Q$-function can be obtained by taking expectation of the log-likelihood function shown in [12] with regard to the conditional distribution of the latent variables.
$z_{i_1 j_1}, z_{i_2 j_2}, \cdots, z_{i_S j_S}$. Specifically, it can be derived as

\[
Q = E_{Z|X}[\ell(U, V, \lambda, K, \pi)]
\]

\[
= E_{Z|X}\left\{ \sum_{(i,j) \in \Omega} \sum_{s=1}^{S} z_{ij s} \left[ \log AL_{s}(e_{ij}|0, \lambda_{s}, \kappa_{s}) + \log \pi_{s} \right] \right\}
\]

\[
= \sum_{(i,j) \in \Omega} \sum_{s=1}^{S} \gamma_{ij s} \left[ \log \pi_{s} + \log \lambda_{s} \kappa_{s}(1 - \kappa_{s}) - |e_{ij}| \lambda_{s} [(1 - \kappa_{s}) \mathbb{I}(e_{ij} < 0) + \kappa_{s} \mathbb{I}(e_{ij} \geq 0)] \right]
\]

\[
= \sum_{(i,j) \in \Omega} \sum_{s=1}^{S} \gamma_{ij s} \left[ \log \kappa_{s}(1 - \kappa_{s}) \lambda_{s} \pi_{s} - \lambda_{s} \rho_{ij s} |e_{ij}| \right],
\]

where

\[
\rho_{ij s} = [(1 - \kappa_{s}) \mathbb{I}(e_{ij} < 0) + \kappa_{s} \mathbb{I}(e_{ij} \geq 0)].
\]

**M-step:** Maximize the $Q$-function by iteratively updating its parameters as follows.

(1). **Update $\pi_s$:** To attain the update for $\pi_s$, we need to solve the following constrained optimization problem

\[
\max_{\pi_s} \sum_{(i,j) \in \Omega} \sum_{s=1}^{S} \gamma_{ij s} \log \pi_{s}, \quad \text{s.t.} \quad \sum_{s=1}^{S} \pi_{s} = 1,
\]

via the Lagrangian multiplier method. By some derivations, we have

\[
\pi_{s} = \frac{N_{s}}{N}, \quad \text{where} \quad N_{s} = \sum_{(i,j) \in \Omega} \gamma_{ij s},
\]

in which $N$ stands for the cardinality of $\Omega$.

(2). **Update $\lambda_s$:** Compute the gradient $\frac{\partial Q}{\partial \lambda_s}$ and let it be zero. Consequently, the update of $\lambda_s$ can be obtained as

\[
\lambda_{s} = \frac{N_{s}}{\sum_{(i,j) \in \Omega} \rho_{ij s} \gamma_{ij s} |e_{ij}|}.
\]

(3). **Update $\kappa_s$:** Compute the gradient $\frac{\partial Q}{\partial \kappa_s}$ and let it be zero, we can have

\[
\eta_{s} \kappa_{s}^{2} - (2N_{s} + \eta_{s}) \kappa_{s} + N_{s} = 0
\]
where the coefficients \( \eta_s = \lambda_s \sum_{(i,j) \in \Omega} \gamma_{ij} e_{ij} \). It is a two-order equation with regard to \( \kappa_s \). Obviously, Eq. (19) has a unique root satisfying \( 0 < \kappa_s < 1 \), that is,

\[
\kappa_s = \frac{2N_s + \eta_s - \sqrt{4N_s^2 + \eta_s^2}}{2\eta_s}.
\]

(20)

(4). **Update U, V:** By omitting some constants, the objective function to optimize \( U, V \) can be rewritten as

\[
\max \sum_{(i,j) \in \Omega} \sum_{s=1}^{S} \lambda_s \gamma_{ij} s \rho_{ij} |x_{ij} - u_i v_j^T| \\
\Leftrightarrow \min \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} |x_{ij} - u_i v_j^T| \\
\Leftrightarrow \min \| W \odot (X - UV^T) \|_{L_1},
\]

(21)

where the \((i,j)\)th entry of \( W \) is

\[
w_{ij} = \begin{cases} 
\sum_{s=1}^{S} \lambda_s \gamma_{ij} s \rho_{ij}, & \text{if } (i,j) \in \Omega, \\
0, & \text{if } (i,j) \notin \Omega.
\end{cases}
\]

(22)

Hence, the optimization problem in Eq. (21) is equivalent to the weighted \( L_1 \)-LRMF, which can be solved by a fast off-the-shelf algorithm such as the cyclic weighted median filter (CWM) [20].

On one hand, it is interesting that the M-step in AQ-LRMF is the same as that of MoG-LRMF [12], except that the latter one minimizes a weighted \( L_2 \) loss. Due to this feature, AQ-LRMF is more robust than MoG-LRMF. On the other hand, each weight of MoG-LRMF embodies the information about whether the corresponding entry is an outlier. For each weight of AQ-LRMF, it actually contains additional information about the sign of bias. In detail, \( \lambda_s \) is the scale parameter and the entries with smaller \( \lambda_s \) correspond to outliers. According to the definition of \( \rho_{ij} \) in Eq. (15), we know that \( \rho_{ij} \) is a function of the skewness parameter \( \kappa_s \). If the residual \( e_{ij} \geq 0 \), \( \rho_{ij} = \kappa_s \) and \( \rho_{ij} = 1 - \kappa_s \) otherwise. Hence, the weights assigned to two points still differ if two residuals with the same absolute value have different signs. In conclusion, AQ-LRMF has more capacity to process heavy-tailed skew data.

Based on the above analysis, we summarize the main steps to learn the parameters involved in AQ-LRMF in the following Algorithm [1]. We now discuss the computational complexity of Algorithm [1]. The complexity...
of updating $\gamma$ is $O(mnS)$. The same to updating $\pi, \lambda$. Then, complexity of updating $\kappa$ is $O(S)$. At last, updating $U, V$ needs $O(mnS)$ time costs if Eq. (21) is solved by CWM. Thus, the entire time complexity of Algorithm 1 is $O(T(mnS + S))$, where $T$ is the number of iterations for convergence. Note Algorithm 1 is derived by EM algorithm, and thus it can converge to a local optimum, since the likelihood does not decrease for each step. For ease of illustration, Fig. 3 reports the likelihood curve versus step in a synthetic experiment (see the detailed settings in section 4.1). It is shown that likelihood value quickly increases in first few iterations, and then it slowly rises. Finally algorithm converges at step 35.

### Algorithm 1 Learning algorithm of AQ-LRMF

**Input:**

The observed matrix $X$ of order $m \times n$; the index set $\Omega$ of non-missing entries of $X$; number of components $S$ in MoAL.

**Output:**

$U, V$.

1: Initialize $U, V, \lambda, K, \pi$;

2: (Initial E-step): Evaluate $\gamma_{ijs}$ by Eq. (13), $i = 1, ..., m$; $j = 1, ..., n$; $s = 1, \cdots, S$.

3: while the convergence criterion does not satisfy do

4: (M-step 1): Update $\pi_s, \lambda_s, \kappa_s$ ($s = 1, \cdots, S$) with Eqs. (17), (18) and (20), respectively.

5: (E-step 1): Evaluate $\gamma_{ijs}$ by Eq. (13), $i = 1, ..., m$; $j = 1, ..., n$; $s = 1, \cdots, S$.

6: (M-step 2): Update $U, V$ by Eq. (21).

7: (E-step 2): Evaluate $\gamma_{ijs}$ by Eq. (13), $i = 1, ..., m$; $j = 1, ..., n$; $s = 1, \cdots, S$.

8: (Tune $S$): For each pair $(i, j) \in \Omega$, compute its noise component index $C(i, j) = \arg \max_s \gamma_{ijs}$. Remove any ALD components which are not in $C$. Let $S$ be the current number of ALD components.

9: end while

#### 3.5. Solution of weighted $L_1$-LRMF

As stated in the last subsection, the learning of AQ-LRMF can be cast into a weighted $L_1$-LRMF problem. Now we will provide more details about how to solve it (i.e., how to update $U, V$ by Eq. (21))
with the CWM method [20].

Essentially, CWM minimizes the objective via solving a series of scalar minimization subproblems. Let $\tilde{u}_i$ and $\tilde{v}_i$ be the $i$th column of $U$ and $V$, respectively. To update $v_{ji}$, we assume that the other parameters have been estimated. As a result, the original problem can be rewritten as the optimization problem regarding $v_{ji}$, i.e.,

$$||W \odot (X - UV^T)||_{L_1} = ||W \odot (X - \sum_{j=1}^{r} \tilde{u}_j \tilde{v}_j^T)||_{L_1}$$

(23)

where $E_i = X - \sum_{j \neq i} u_j v_j^T$, and $\tilde{w}_j$ and $\tilde{e}_{ij}$ are $j$th column of $W$ and $E_i$, respectively. In Eq. (23), $c$ denotes a constant term not depending on $v_{ji}$. In this way, the optimal $v_{ji}$, say $v_{ji}^*$, can be easily attained by the weighted median filter when minimizing Eq. (23) can be provided by weighted median filter. Specifically, if let $e = \tilde{w}_j \odot \tilde{e}_{ij}$ and $u = \tilde{w}_j \odot \tilde{u}_i$, we can reformulate Eq. (23) as

$$||\tilde{w}_j \odot (\tilde{e}_{ij} - \tilde{u}_i v_{ji})||_{L_1} = ||e - u v_{ji}||_{L_1}$$

(24)

From the format of Eq. (24), it can be seen that the optimal $v_{ji}^*$ coincides with the weighted median of the sequence $\{e_l/u_l\}_{l=1}^{m}$ under weights $\{|u_l|\}_{l=1}^{m}$. As for the update of $u_{ij}$, it can be handled in a similar procedure. In short, the optimal $U, V$ can be obtained by employing CWM to repeatedly update $v_{ij}(i = 1, \cdots, n; j = 1, \cdots, r)$ and $u_{ij}(i = 1, \cdots, m; j = 1, \cdots, r)$ until the algorithm converges.

![Figure 3: Likelihood curve versus step.](image)
3.6. Some details of learning AQ-LRMF

**Tuning the number of components** $S$ in MoAL: Too large $S$ violates Occam Razor's principle, while small $S$ leads to poor performance. In consequence, as described in step 8 of Algorithm 1 we employ an effective method to tune $S$. To begin with, we initialize $S$ to be a small number such as 4, 5, \ldots, 8. After each iteration, we compute the cluster that $x_{ij}$ belongs to, by $C(i, j) = \arg \max_s \gamma_{ij,s}$. If there is no entry belonging to cluster $s$, we remove the corresponding ALD component.

**Initialization:** In Algorithm 1 we initialized the $(i, j)$th entry of $U$ as $2\xi_{ij}c - c$, where $\xi_{ij}$ denotes a random number sampled from the standard Gaussian distribution $\mathcal{N}(0, 1)$. In addition, $c = \sqrt{\bar{x}/r}$ where $\bar{x}$ is the median of all entries in $X$. Due to the characteristics of $U$ and $V$, we initialize $V$ similarly. Moreover, the parameters $\lambda_s$ and $\kappa_s$ is randomly sampled from $[0, 1]$.

**Convergence condition:** By following the common practice of EM algorithm, we stop the iteration if the change of $||U||$ is smaller than a pre-defined value or the maximum iteration number is reached.

4. Experimental Studies

We carried out experiments in this section to examine the performance of AQ-LRMF model. Several state-of-the-art methods were considered, including four robust LRMF methods (namely, MoG [12], CWM [20], Damped Wiberg (DW) [30], RegL1ALM [19]) and a robust PCA method (SPCP solved by quasi Newton method) [31]. We wrote the programming code for CWM. For the other compared algorithms, the codes provided by the corresponding authors were availed. Since SPCP is not available in presence of missing entries, it was thus excluded from some experiments which involve missing data. Notice that DW is only considered in section 4.1 because it meets the “out of memory” problem for large-scale datasets. In the meantime, we assigned the same rank to all the considered algorithms but SPCP since it can automatically determine the rank. To make the comparison more fair, all algorithms were initialized with the same values.

---

1. http://www.gr.xjtu.edu.cn/c/document_library/get_file?folderId=1816179&name=DLFE-32163.rar
2. http://www.vision.is.tohoku.ac.jp/ue/download/
3. https://sites.google.com/site/yinqiangzheng/
4. https://github.com/stephenbeckr/fastRPCA

---

15
Each algorithm was terminated when either 100 iterative steps are reached or the change of $\|U\|$ is less than $1 \times 10^{-50}$. In order to simplify notations, our proposed method AQ-LRMF was denoted as AQ in later discussions. All the experiments were conducted with Matlab R2015b and run on a computer with Intel Core CPU 2.30 GHz, 4.00 GB RAM and Windows 7(64-bit) system.

The remainder of this section has the following structure. Section 4.1 studies the performance of each algorithm on synthetic data in presence of various kinds of noise as well as missing values. Sections 4.2 and 4.3 employ some inpainted and multispectral images to investigate how the compared algorithms behave on real images which contain missing values and various kinds of noise, respectively. Finally, sections 4.4 and 4.5 examine the performance of all algorithms on face modeling and hyperspectral image processing tasks.

Table 1 summarizes the basic information of real-world data sets.

| Data set   | Size            | Section |
|------------|-----------------|---------|
| CAVE       | 262144 × 31     | 4.3     |
| Extended Yale B | 32256 × 64    | 4.4     |
| Urban      | 94249 × 210     | 4.5     |
| Terrain    | 153500 × 210    | 4.5     |

### 4.1. Synthetic experiments

First, we compared the behavior of each method with synthetic data containing different kinds of noise. For each case, we randomly generated 30 low rank matrices $X = UV^T$ of size $40 \times 20$, where $U \in \mathbb{R}^{40 \times r}$ and $V \in \mathbb{R}^{20 \times r}$ were sampled from the standard Gaussian distribution $\mathcal{N}(0, 1)$. In particular, we set $r$ to 4 and 8. Subsequently, we stochastically set 20% entries of $X$ as missing data and corrupted the non-missing entries with the following three groups of noise. (i) The first group include 4 kinds of heavy-tailed noise, i.e., $Lap(0, 1.5)$ (Laplace noise with scale parameter $b = 1.5$ and location parameter $\mu = 0$), Gaussian noise with $\sigma = 5, \mu = 0$ and Student’s $t$ noise with degrees of freedom 1 and 2, respectively. (ii) Two kinds of skew noise are included in the second group, i.e., asymmetric Laplace noise with $\lambda = 1, \kappa = 0.7$ and skew
normal with $\sigma = 3, \kappa = 0.7$. (iii) Two kinds of mixture noise are included in the last group. The one is $0.5\mathcal{N}(0, 1) + 0.3\text{Lap}(0, 1) + 0.2\text{Lap}(0, 2)$ and the other one is $0.5\mathcal{N}(0, 1) + 0.3\text{Lap}(0, 1) + 0.2\text{AL}(0, 1, 0.8)$. It is worthwhile to mention that the two mixture noises simulate the noise contained in real data, where most entries are corrupted by standard Gaussian noise and the rest entries are corrupted by heavy-tailed or skew noise. To evaluate the performance of each method, we employed the average $L_1$ error that is, 

$$\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} |x_{ij} - u_i v_j^T|.$$ 

In our experiments, all algorithms were implemented with true rank $r$. Tables 2 and 3 summarize the metrics averaged over 30 randomly generated matrices. When $r = 4$, it is quite obvious that our method reaches the minimum $L_1$ error in all situations, while MoG and CWM almost take the second place. And the approaches RegL1ALM and DW can hardly deal with the heavy-tailed and skew noise well. From the results of $r = 8$, similar conclusions can be drawn. However, CWM evidently outperforms MoG when $r = 8$, which indicates that MoG may be instable when the real rank in observed data is high. In addition, the running speed of AQ is fairly competitive, as shown in Fig. 4. In summary, our model performs very well to cope with different kinds of noise.

Table 2: The performance evaluation on synthetic data with rank 4. The best and second best results are highlighted in bold and italic typeface, respectively.

| Noise Type                  | AQ   | MoG  | CWM  | RegL1ALM | DW   |
|-----------------------------|------|------|------|-----------|------|
| Laplace Noise (b=1.5)       | 1.23 | 1.27 | 1.34 | 1.93      | 10.44|
| Gaussian Noise ($\sigma = 5$) | 2.95 | 3.46 | 3.16 | 4.97      | 4.56 |
| Student’s t Noise ($df = 1$) | 1.47 | 2.36 | 2.88 | 27.45     | 925.17|
| Student’s t Noise ($df = 2$) | 0.97 | 1.37 | 1.21 | 2.85      | 7.16 |
| AL Noise ($\lambda = 1, \kappa = 0.7$) | 1.99 | 2.58 | 2.33 | 3.99      | 3.65 |
| SN Noise ($\sigma = 3, \kappa = 0.7$) | 1.90 | 2.00 | 2.13 | 3.03      | 2.11 |
| Mixture Noise 1              | 0.84 | 0.93 | 1.02 | 1.28      | 1.06 |
| Mixture Noise 2              | 0.99 | 1.61 | 1.27 | 3.88      | 10.50|

mean: 1.54 1.95 1.92 6.17 120.58
median: 1.35 1.81 1.73 3.45 5.86
Figure 4: The running time (in seconds) on synthetic data.
Table 3: The performance evaluation on synthetic data with rank 8. The best and second best results are highlighted in bold and italic typeface, respectively.

|                           | AQ  | MoG | CWM | RegL1ALM | DW            |
|---------------------------|-----|-----|-----|----------|---------------|
| Laplace Noise (b=1.5)     | **1.82** | 2.13 | 1.90 | 3.81     | 12.71         |
| Gaussian Noise (σ = 5)    | 4.14 | 5.07 | **4.11** | 8.38 | 9.04         |
| Student’s t Noise (df = 1) | **2.54** | 3.91 | 3.60 | 22.38    | 2063.74       |
| Student’s t Noise (df = 2) | **1.63** | 2.32 | 1.81 | 4.20     | 40.82         |
| AL Noise (λ = 1, κ = 0.7) | **2.91** | 3.69 | 3.07 | 6.76     | 17.84         |
| SN Noise (σ = 3, κ = 0.7) | **2.61** | 3.14 | 2.68 | 4.97     | 4.26          |
| Mixture Noise 1           | **1.34** | 1.55 | 1.55 | 2.59     | 3.80          |
| Mixture Noise 2           | **1.68** | 2.32 | 1.91 | 5.29     | 22.19         |
| mean                      | **2.33** | 3.02 | 2.58 | 7.30     | 271.80        |
| median                    | **2.18** | 2.73 | 2.30 | 5.13     | 15.27         |

To delve into the difference between AQ and MoG, we further compared the distributions of residuals with real noises. Specifically, two symmetric and two asymmetric cases are illustrated in Figure 5. Here, the shown distributions fitted by AQ and MoG correspond to those reach the maximum likelihood over 30 random experiments. It is obvious that AQ does a much better job to approximate the real noise than MoG. Particularly, AQ almost provides a duplicate of real noise. In contrast, MoG is able to fit the tails, while, at the same time, it results in bad approximation to peaks. Hence, AQ has more power in fitting complex noise than MoG.

4.2. Image inpainting experiments

Image inpainting is a typical image processing task. In real applications, some parts of an image may be deteriorated so that the corresponding information is lost. To facilitate the understanding of the image, some sophisticated technique need to be adopted to recover the corrupted parts of the image. This is exactly the objective of image inpainting. There is evidence that many images are low-rank matrices so that the single image inpainting can be done by matrix completion [32]. In image inpainting, the corrupted pixels are viewed as missing values and then the image can be recovered by an LRMF algorithm. In this paper, three
Figure 5: The visual comparison of the PDFs for real noise and the ones fitted by AQ and MoG in the synthetic experiments.

Figure 6 visualizes the original, masked and reconstructed images, and Table 4 reports the average $L_1$ errors of each algorithm. It is obvious that removing a random mask is the easiest task. In this situation, there is no significantly visible difference among the reconstructed images. In terms of average $L_1$ error, MoG performs best and AQ can be ranked in second place. In contrast, the results shown in Figure 4 and Table 3 indicate that text mask removal is more difficult, especially when the images are corrupted with big fonts. The main reason lies in that the text mask is spatially correlated while it is difficult for any LRMF
algorithm to effectively utilize this type of information. Under these circumstances, it can be observed in Figure 4 and Table 3 that AQ outperforms the other methods to remove the text masks with regard to both average $L_1$ error and visualization. RegL1ALM and MoG perform badly and the clear text can often be seen in their reconstructed images. Although CWM produces slightly better results, its average $L_1$ error is still higher than that of AQ. In a word, AQ possesses the superiority over the other algorithms in our investigated image inpainting tasks. In particular, AQ achieves the smallest average $L_1$ error in 6 cases and the second smallest one in 2 cases.

Table 4: The average $L_1$ errors of each method on image inpainting experiments. The best and second best results are highlighted in bold and italic typeface, respectively.

|                  | Image A          | Image B          | Image C          |
|------------------|------------------|------------------|------------------|
|                  | AQ    | MoG    | CWM    | RegL1 | AQ    | MoG    | CWM    | RegL1 | AQ    | MoG    | CWM    | RegL1 |
| Small            | 2.3192 | 2.5905 | 2.3274 | 2.9108 | 5.6031 | 7.4266 | 6.8971 | 8.0762 | 5.1294 | 5.8328 | 6.2992 | 6.9657 |
| Large            | 5.5852 | 9.2519 | 7.8357 | 8.7736 | 6.8846 | 20.1643 | 8.7815 | 19.1101 | 6.8369 | 7.1456 | 7.6715 | 17.6973 |
| Random           | 2.8141 | 2.1774 | 3.0580 | 2.4519 | 5.9130 | 5.7458 | 6.6481 | 6.7950 | 5.6079 | 6.7389 | 6.7724 | 5.0700 |

4.3. Multispectral image experiments

In this subsection, we study the behavior of all algorithms in image denoising tasks. The Columbia Multispectral Image database, CAVE, was employed, where every scene contains 31 bands with size 512 x 512. To achieve our purpose, eight scenes out of them (i.e., Balloon, Paints, Flowers, Cloth, Feathers, Hairs, Pompoms and Clay) were utilized to test the effectiveness of our methods. The used images were resized by half and the pixels were rescaled to [0,1]. Analogous to the strategy used in image inpainting experiments, some noise was artificially added to the original images. Then, each LRMF algorithm was applied to remove the noise so that the corrupted images can be restored as accurate as possible. In the experiments, four different kinds of noise were considered, that is, Laplace noise with scale parameter $b = 10$, asymmetric Laplace noise with $\lambda = 10$, $\kappa = 0.7$ and mixture noise, i.e., $0.5 \mathcal{N}(0, 0.5) + 0.3 \mathcal{AL}(0, 8, 0.9) + 0.2 \mathcal{AL}(0, 8, 0.7)$. The rank was set to 4 for all algorithms.
Figure 6: The original, masked and inpainting images.

| Original | Masked | AQ | MoG | CWM | RegL1ALM |
|----------|--------|----|-----|-----|----------|
| ![Image A](image_a.png) | ![Masked Image A](masked_image_a.png) | ![AQ Image A](aq_image_a.png) | ![MoG Image A](mog_image_a.png) | ![CWM Image A](cwm_image_a.png) | ![RegL1ALM Image A](regl1alm_image_a.png) |
| ![Image B](image_b.png) | ![Masked Image B](masked_image_b.png) | ![AQ Image B](aq_image_b.png) | ![MoG Image B](mog_image_b.png) | ![CWM Image B](cwm_image_b.png) | ![RegL1ALM Image B](regl1alm_image_b.png) |
| ![Image C](image_c.png) | ![Masked Image C](masked_image_c.png) | ![AQ Image C](aq_image_c.png) | ![MoG Image C](mog_image_c.png) | ![CWM Image C](cwm_image_c.png) | ![RegL1ALM Image C](regl1alm_image_c.png) |
Table 5 reports the average $L_1$ errors of each method. For Laplace noise, it is found that RegL1ALM performs best, while AQ, MoG and CWM are ranked in the second, third and fourth places, respectively. The success of RegL1ALM can be attributed to the special format of its objective function, namely, $L_1$ norm loss plus two penalties on $U$ and $V$. On one hand, the $L_1$ norm loss is exactly compatible with Laplace noise. On the other hand, there is empirical evidence showing that its used penalties can lead to better performance on image datasets. For asymmetric Laplace and mixture noise, it is not surprising that AQ outperforms all the other methods. SPCP reaches the second lowest average $L_1$ error, while the other ones perform badly. The reason may be that SPCP does not rely on noise distribution, while the other approaches implicitly assume that the noise distribution is not skew.

4.4. Face modeling experiments

Here, we applied the LRMF techniques to address the face modeling task. The Extended Yale B database consisting of 64 images with size $192 \times 168$ of each subject was considered. Therefore, it leads to a $32256 \times 64$ matrix for each subject. Particularly, we used the face images of the third and fifth subjects. The first column of Figure 7 demonstrates some typical faces for illustration. We set the rank to 4 for all methods except for SPCP which determines the rank automatically. The second to sixth columns of Figure 7 display the faces reconstructed by the compared LRMF algorithms.

From Figure 7, we can observe that that all methods are able to remove the cast shadows, saturations and camera noise. However, the performance of SPCP seems to be worse in comparison with other algorithms. Evidently, AQ always outperforms the other methods due to its pretty reconstruction. As shown in Figure 1, there is an asymmetric distribution in the face with a large dark region. Because of this, the techniques MoG, CWM, RegL1ALM and SPCP which utilize the symmetric loss function lead to bad results, while AQ with the quantile loss function produces the best reconstructed images.

\footnote{http://cvc.yale.edu/projects/yalefaces/yalefaces.html}
Table 5: The average $L_1$ errors of each method on multispectral image experiments. The best and second best results are highlighted in bold and italic typeface, respectively. AL refers to asymmetric Laplace.

| Scene  | Type of Noise | AQ   | MoG   | CWM   | RegL1 | AL     | SPCP  |
|--------|---------------|------|-------|-------|-------|--------|-------|
| Balloon| Laplace       | 0.0342 | 0.0348 | 0.0398 | 0.0343 | 0.0487 |
|        | AL            | 0.0804 | 0.1964 | 0.1501 | 0.1379 | 0.1280 |
|        | Mixture       | 0.1974 | 0.2214 | 0.2453 | 0.2422 | 0.2026 |
|        | Laplace       | 0.0362 | 0.0395 | 0.0371 | 0.0343 | 0.0437 |
|        | AL            | 0.0761 | 0.1708 | 0.1450 | 0.1374 | 0.1289 |
|        | Mixture       | 0.1709 | 0.2339 | 0.2393 | 0.2437 | 0.2025 |
| Flowers| Laplace       | 0.0412 | 0.0399 | 0.0365 | 0.0323 | 0.0667 |
|        | AL            | 0.0975 | 0.1640 | 0.1424 | 0.1374 | 0.1314 |
|        | Mixture       | 0.1927 | 0.1898 | 0.2334 | 0.2512 | 0.2037 |
| Cloth  | Laplace       | 0.0392 | 0.0380 | 0.0358 | 0.0379 | 0.0824 |
|        | AL            | 0.0681 | 0.1969 | 0.1373 | 0.1346 | 0.1201 |
|        | Mixture       | 0.1412 | 0.2172 | 0.2305 | 0.2387 | 0.1979 |
| Feathers| Laplace      | 0.0493 | 0.0420 | 0.0414 | 0.0379 | 0.0824 |
|        | AL            | 0.1002 | 0.1937 | 0.1508 | 0.1422 | 0.1472 |
|        | Mixture       | 0.1630 | 0.2221 | 0.2489 | 0.2323 | 0.2153 |
| Hairs  | Laplace       | 0.0424 | 0.0370 | 0.0431 | 0.0354 | 0.0396 |
|        | AL            | 0.1009 | 0.1910 | 0.1432 | 0.1402 | 0.1259 |
|        | Mixture       | 0.2119 | 0.2311 | 0.2460 | 0.2399 | 0.2018 |
| Pompoms| Laplace       | 0.0335 | 0.0402 | 0.0362 | 0.0344 | 0.0596 |
|        | AL            | 0.0778 | 0.1787 | 0.1385 | 0.1342 | 0.1380 |
|        | Mixture       | 0.1569 | 0.2169 | 0.2403 | 0.2453 | 0.2097 |
| Paints | Laplace       | 0.0385 | 0.0388 | 0.0390 | 0.0344 | 0.0516 |
|        | AL            | 0.0865 | 0.1805 | 0.1445 | 0.1380 | 0.1312 |
|        | Mixture       | 0.1786 | 0.2212 | 0.2411 | 0.2420 | 0.2048 |
4.5. Hyperspectral image experiments

In this subsection, we employed two HSI datasets, Urban and Terrain to investigate the behavior of all algorithms. There are 210 bands, each of which is of size $307 \times 307$ for Urban and $500 \times 307$ for Terrain. Thus, the data matrix is of size $94249 \times 210$ for Urban and $153500 \times 210$ for Terrain. Here, we utilized the same experimental settings as those used in subsection 4.4. DW was still unavailable in this experiment due to the computational problem. As show in the first column of Figure 8 some parts of bands are seriously polluted by the atmosphere and water absorption.

The reconstructed images of bands 106 and 207 in the Terrain data set and the band 104 in the Urban data set are shown in Figure 8 (a), (c) and (e), respectively. Their residual images (i.e., $X - \hat{UV}^T$) are also demonstrated below the reconstructed ones. Obviously, the band 106 in Terrain is seriously polluted. Nevertheless, our proposed AQ method still effectively reconstructs a clean and smooth one. Although MoG, CWM and RegL1ALM remove most parts of noise, they miss a part of local information, that is, the line from upper left corner to bottom right hand side (i.e., the parallelogram marked in the original image). As for SPCP, it only removes few parts of noise. The residual images also reveal that AQ behaves better to deal with the detailed information. Note that the band 207 in Terrain and the band 104 in Urban are mainly corrupted by the stripe and Guassian-like noise. Under these circumstances, AQ still outperforms the others because the latter fails to remove the stripe noise. In particular, for the interested areas that are marked by rectangles and amplified areas, the bands reconstructed by MoG, CWM, RegL1ALM and SPCP contain evident stripes. As far as the reconstructed images produced by AQ are concerned, however, this phenomenon does not exist.

We conjectured that the main reason for the different behavior of these algorithms lies in their used loss function. For CWM, RegL1ALM and SPCP, too simple loss function lead them to work not well when encountering complicated noise. In contrast, AQ and MoG perform better because they use multiple distribution components to model noise. It is very interesting to study the difference between AQ and MoG. For these two algorithms, we found that they both approximate the noise in our considered three bands.

*http://www.erdc.usace.army.mil/Media/Fact-Sheets/Fact-Sheet-Article-View/Article/610433/hypercube/*
with two components. For AQ (MoG), we denoted them as AQ1 and AQ2 (MoG1 and MoG2), respectively. In Figure 9, we presented de-noised images and residual images produced by each component. Take the de-noised image in the column AQ1 as an example, it corresponds to $\hat{U}\hat{V}^T + AQ2$ and the residual image shown below it corresponds to AQ1 (i.e., $X - \hat{U}\hat{V}^T - AQ2$). The other images can be understood similarly. In doing so, we can further figure out the role that each component in AQ or MoG plays. When dealing with the band 106 in Terrain, the first AQ component is seen to de-noise the center parts, while the second one targets at the left and right edges. For the band 207 in Terrain, two AQ components de-noise the bottom and the rest parts, respectively. Regarding the band 104 in Urban, they focus on the right upper and center parts, respectively. By inspecting the results generated by MoG, however, we cannot discover some regular patterns for the role that two components play. Therefore, it can be concluded that AQ can capture the local structural information of real images, although we do not encode it into our model. The reason may be that the pixels with the same skewness in real images tend to cluster. In this aspect, AQ also possesses superiority over MoG.

5. Conclusions

Aiming at enhancing the performance of existing LRMF methods to cope with complicated noise in real applications, we propose in this work a new low-rank matrix factorization method AQ-LRMF to recover subspaces. The core idea of AQ-LRMF is to directly model unknown noise by a mixture of asymmetric Laplace distributions. We also present an efficient procedure based on the EM algorithm to estimate the parameters in AQ-LRMF. Actually, the objective function of AQ-LRMF corresponds to the adaptive quantile loss like those used in quantile regression. Compared with several state-of-the-art counterparts, the novel AQ-LRMF model always outperforms them in synthetic and real data experiments. In addition, AQ-LRMF also has the superiority to capture local structural information in real images. Therefore, AQ-LRMF can be deemed as a competitive tool to cope with complex real problems.
Statement

All authors declare that there is no conflict of interest, financial or otherwise.

Acknowledgements

This research was supported by the National Natural Science Foundation of China grant number [11671317, 61572393] and the Fundamental Research Funds for the Central Universities grant number [xzy022019059].

References

References

[1] J. Ye, Generalized low rank approximations of matrices, Machine Learning 61 (2005) 167–191.
[2] M. Udell, C. Horn, R. Zadeh, S. Boyd, Generalized low rank models, Foundations and Trends in Machine Learning 9 (2016) 1–118.
[3] V. Koltchinskii, K. Lounici, A. B. Tsybakov, Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion, Annals of Statistics 39 (2011) 2302–2329.
[4] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.
[5] H. Wu, Z. Zhang, K. Yue, B. Zhang, J. He, L. Sun, Dual-regularized matrix factorization with deep neural networks for recommender systems, Knowledge-Based Systems 145 (2018) 46 – 58.
[6] D. Tolić, N. Antulov-Fantulin, I. Kopriva, A nonlinear orthogonal non-negative matrix factorization approach to subspace clustering, Pattern Recognition 82 (2018) 40 – 55.
[7] H. Xiong, D. Kong, Elastic nonnegative matrix factorization, Pattern Recognition 90 (2019) 464 – 475.
[8] W. Wang, Y. Feng, P. Jiao, W. Yu, Kernel framework based on non-negative matrix factorization for networks reconstruction and link prediction, Knowledge-Based Systems 137 (2017) 104 – 114.
[9] J. Wright, Y. Peng, Y. Ma, A. Ganesh, S. Rao, Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization, in: International Conference on Neural Information Processing Systems (NIPS), 2009, pp. 2080–2088.
[10] X. Fei, Y. Chen, P. Chong, Y. Wang, X. Liu, G. He, Denoising of hyperspectral image using low-rank matrix factorization, IEEE Geoscience & Remote Sensing Letters 14 (2017) 1141–1145.
[11] Q. Ke, T. Kanade, Robust $l_1$ norm factorization in the presence of outliers and missing data by alternative convex programming, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 739–746.
[12] D. Meng, F. D. L. Torre, Robust matrix factorization with unknown noise, in: IEEE International Conference on Computer Vision (ICCV), 2014, pp. 1337–1344.

[13] V. Maz’Ya, G. Schmidt, On approximate approximations using gaussian kernels, IMA Journal of Numerical Analysis 16 (1996) 13–29.

[14] H. Kozumi, G. Kobayashi, Gibbs sampling methods for bayesian quantile regression, Journal of Statistical Computation & Simulation 81 (2011) 1565–1578.

[15] K. Yu, J. Zhang, A three-parameter asymmetric laplace distribution and its extension, Communications in Statistics - Theory and Methods 34 (2005) 1867–1879.

[16] N. Srebro, T. Jaakkola, Weighted low-rank approximations, in: International Conference on Machine Learning (ICML), 2003, pp. 720–727.

[17] A. Eriksson, A. van den Hengel, Efficient computation of robust low-rank matrix approximations in the presence of missing data using the $l_1$ norm, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 771–778.

[18] E. Kim, M. Lee, C. H. Choi, N. Kwak, S. Oh, Efficient $l_1$ -norm-based low-rank matrix approximations for large-scale problems using alternating rectified gradient method, IEEE Transactions on Neural Networks and Learning Systems 26 (2015) 237–251.

[19] M. Okutomi, S. Yan, S. Sugimoto, G. Liu, Y. Zheng, Practical low-rank matrix approximation under robust $l_1$-norm, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 2012, pp. 1410–1417.

[20] D. Meng, Z. Xu, L. Zhang, J. Zhao, A cyclic weighted median method for $l_1$ low-rank matrix factorization with missing entries, in: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013, pp. 704–710.

[21] Z. Lin, C. Xu, H. Zha, Robust matrix factorization by majorization minimization, IEEE Transactions on Pattern Analysis & Machine Intelligence 40 (2018) 208–220.

[22] S. Li, J. Zhang, X. Guo, Efficient low rank matrix approximation via orthogonality pursuit and $l_2$ regularization, in: IEEE International Conference on Multimedia and Expo, 2017, pp. 871–876.

[23] B. Lakshminarayanan, G. Bouchard, C. Archambeau, Robust bayesian matrix factorisation, Journal of Machine Learning Research 15 (2011) 425–433.

[24] N. Wang, T. Yao, J. Wang, D.-Y. Yeung, A probabilistic approach to robust matrix factorization, in: European Conference on Computer Vision (ECCV), Springer, 2012, pp. 126–139.

[25] X. Cao, Y. Chen, Q. Zhao, D. Meng, Low-rank matrix factorization under general mixture noise distributions, IEEE Transactions on Image Processing 25 (2016) 4677–4690.

[26] E. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis?, Journal of the ACM 58 (2011) 11.

[27] Z. Zhou, X. Li, J. Wright, E. Candès, Stable principal component pursuit, in: IEEE International Symposium on Information Theory Proceedings, Austin, TX, USA, 2010, pp. 1518–1522.

[28] A. Azzalini, A. D. Valle, The multivariate skew-normal distribution, Biometrika 83 (1996) 715–726.
[29] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society. Series B (Methodological) 39 (1977) 1–38.

[30] T. Okatani, T. Yoshida, K. Deguchi, Efficient algorithm for low-rank matrix factorization with missing components and performance comparison of latest algorithms, in: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 842–849.

[31] A. Aravkin, S. Becker, V. Cevher, P. Olsen, A variational approach to stable principal component pursuit, in: Conference on Uncertainty in Artificial Intelligence (UAI), Arlington, Virginia, United States, 2014, pp. 32–41.

[32] Y. Hu, D. Zhang, J. Ye, X. Li, X. He, Fast and accurate matrix completion via truncated nuclear norm regularization, IEEE Transactions on Pattern Analysis & Machine Intelligence 35 (2013) 2117–2130.
Figure 7: The original faces and the reconstructed ones.
Figure 8: The reconstructed and residual images.
Figure 9: The de-noised and residual images produced by the two components of AQ (i.e., columns marked with AQ1 and AQ2) and MoG (i.e., columns marked with MoG1 and MoG2). For example, the image lies in the first row and second column is the de-noised image which is obtained by removing the first AQ component from the original image.