

Applications of Mathematics

Guoqiang Huan; Ying Li; Zhanjie Song

A novel robust principal component analysis method for image and video processing

Applications of Mathematics, Vol. 61 (2016), No. 2, 197–214

Persistent URL: <http://dml.cz/dmlcz/144844>

Terms of use:

© Institute of Mathematics AS CR, 2016

Institute of Mathematics of the Czech Academy of Sciences provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.



This document has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* <http://dml.cz>

A NOVEL ROBUST PRINCIPAL COMPONENT ANALYSIS
METHOD FOR IMAGE AND VIDEO PROCESSING

GUOQIANG HUAN, YING LI, ZHANJIE SONG, Tianjin

(Received May 10, 2015)

Abstract. The research on the robust principal component analysis has been attracting much attention recently. Generally, the model assumes sparse noise and characterizes the error term by the ℓ_1 -norm. However, the sparse noise has clustering effect in practice so using a certain ℓ_p -norm simply is not appropriate for modeling. In this paper, we propose a novel method based on sparse Bayesian learning principles and Markov random fields. The method is proved to be very effective for low-rank matrix recovery and contiguous outliers detection, by enforcing the low-rank constraint in a matrix factorization formulation and incorporating the contiguity prior as a sparsity constraint. The experiments on both synthetic data and some practical computer vision applications show that the novel method proposed in this paper is competitive when compared with other state-of-the-art methods.

Keywords: robust principal component analysis; sparse Bayesian learning; Markov random fields; matrix factorization; contiguity prior

MSC 2010: 62H25, 60J20, 68Q87

1. INTRODUCTION

As a classic method for data analysis, the principal component analysis (PCA) has a wide range of applications in science and engineering [11]. Motivated by recent advances in low-rank matrix analysis [5], [6], [20], the so-called robust principal component analysis (RPCA) [22] has been proposed to decompose a given data matrix into a low-rank matrix and a sparse matrix. It can be mathematically described as the following convex optimization problem:

$$(1.1) \quad \min_{L, E} (\|L\|_* + \lambda \|E\|_1) \quad \text{s.t. } Y = L + E,$$

This work was partially supported by the National Natural Science Foundation of China (Grant No. 61379014) and Tianjin Research Program of Application Foundation and Advanced Technology (15JCYBJC21700).

where $Y \in \mathbb{R}^{m \times n}$, $L \in \mathbb{R}^{m \times n}$, and $E \in \mathbb{R}^{m \times n}$ denote the original data matrix, the low-rank component, and the sparse component, respectively. The symbol $\|L\|_* = \sum_r \sigma_r(L)$ denotes the nuclear norm of L , $\sigma_r(L)$ ($r = 1, 2, \dots, \min(m, n)$) is the r^{th} singular value of L , $\|E\|_1 = \sum_{ij} |e_{ij}|$ denotes the ℓ_1 -norm of E and e_{ij} is the element in the i^{th} row and j^{th} column of E .

Under the assumptions of certain noise sparsity and rank upper-bound, it has been proved that one can exactly recover L and E from Y with overwhelming probability [4]. RPCA has been successfully applied to many machine learning and computer vision problems, such as video surveillance [22], face modeling [19], and subspace clustering [17].

In practice, however, measurement noise exists everywhere within the matrix and sometimes this dense (distributed) noise contribution cannot be ignored. Thus, Zhou et al. [25] modified (1.1) to include the additive noise of this type, that is,

$$(1.2) \quad \min_{L, E} \left(\frac{1}{2} \|Y - L - E\|_F^2 + \lambda_1 \|L\|_* + \lambda_2 \|E\|_1 \right) \quad \text{s.t. } Y = L + E + N,$$

where $N \in \mathbb{R}^{m \times n}$ denotes the dense noise, which has generally small magnitude but dense support. The norm $\|Y - L - E\|_F = \|N\|_F = \sqrt{\sum_{ij} n_{ij}^2}$ is the Frobenius norm of N and n_{ij} is the element in the i^{th} row and j^{th} column of N .

This paper presents a novel RPCA method based on sparse Bayesian learning (SBL) principles and Markov random fields (MRFs), in which the low-rank matrix and outlier support are estimated simultaneously. Starting from the low-rank factorization of the unknown matrix, we employ independent sparsity priors on the individual factors with a common sparsity profile which favors low-rank solutions. Subsequently, we use the variational inference method to infer the posterior. In other methods of RPCA, no prior knowledge on the spatial distribution of outliers has been considered. Since the outlier support is modeled explicitly in our formulation, we can naturally incorporate such contiguity prior using MRFs. The proposed method gives competitive experimental results when compared with other state-of-the-art methods.

2. RELATED WORK

Early attempts to solve the RPCA problem replace the ℓ_2 -norm error by some robust losses. De la Torre & Black [7] utilized the Geman-McClure function in robust statistics to improve the robustness of PCA; Ding et al. [9] used a smoothed R_1 -norm to this end; Kwak [13] introduced the ℓ_1 -norm variance and designed an efficient

algorithm to optimize it. These methods, however, are sensitive to initialization, and only perform well on Laplacian-like noise.

In recent years, low-rank matrix analysis methods have been rapidly developed. Wright et al. [22] initially formulated the RPCA model as shown in (1.1). Some variants have also been proposed, e.g., Xu et al. [23] used the ℓ_{12} -norm to handle data corrupted by column. The iterative thresholding method [4] was proposed to solve the RPCA model but it converges very slow. To speed up the computation, Lin et al. [15], [16] proposed the accelerated proximal gradient (APG) and the augmented Lagrangian multiplier (ALM) methods. ALM leads to state-of-the-art performance in terms of both speed and accuracy.

Besides, Bayesian approaches to RPCA have also been investigated. Ding et al. [8] modeled the singular values of L and the entries of E with beta-Bernoulli priors, using a Markov chain Monte Carlo (MCMC) sampling scheme to perform inference (BRPCA). This method needs many sampling iterations to converge to a stationary distribution, if at all, or may even get stuck at local optima. Babacan et al. [1] adopted the automatic relevance determination (ARD) approach to model both L and E , and utilized the variational Bayes method to do inference (VBLR). This method is more computationally efficient. Zhao et al. [24] proposed a generative RPCA model under the Bayesian framework by modeling data noise as a mixture of Gaussian (MoG). Although MoG is able to fit a wide range of noises such as Laplacian, Gaussian, sparse noise, and any combinations of them, it cannot effectively model the dynamic background.

One problem closely related to RPCA is the low-rank matrix factorization (LRMF). The principal component pursuit (PCP) [4] and the stable principal component pursuit (SPCP) [25] utilized the nuclear norm for normalization. By using the nuclear norm, many extensions have been proposed subsequently, e.g., Zhou, et al. [26] made the outliers contiguous¹ using a graph cut algorithm (DECOLOR). Wang et al. [21] proposed a full Bayesian approach to the robust matrix factorization (BRMF). Besides the basic model, they also proposed an extension assuming that the outliers exhibit spatial and temporal proximity (MBRMF). But all of them are not robust to the scene of dynamic background in practice.

The rest of this paper is organized as follows. The Bayesian mixture model is proposed in Section 3, followed by the algorithm based on SBL and MRFs in Section 4. We present an analysis of the proposed method and empirical results with synthetic and real data in Section 5, and finally draw a conclusion in Section 6.

¹ Contiguous outliers refer to foreground objects or the sparse matrix here.

3. BAYESIAN MIXTURE MODEL

3.1. Notations. Three norms of a matrix are used throughout this paper. The symbol $\|X\|_1 = \sum_{ij} |X_{ij}|$ denotes the ℓ_1 -norm, $\|X\|_F = \sqrt{\sum_{ij} X_{ij}^2}$ is the Frobenius norm, $\|X\|_* = \sum_{i=1}^r \sigma_i$ means the nuclear norm, where σ_i is the i^{th} singular value of the matrix X .

We assume that the observed data matrix $Y \in \mathbb{R}^{m \times n}$ is the superposition of three parts: low-rank component $L \in \mathbb{R}^{m \times n}$, sparse component $E \in \mathbb{R}^{m \times n}$, and noise term $N \in \mathbb{R}^{m \times n}$. $E = Y \circ S$, where $S \in \{0, 1\}^{m \times n}$ is a binary matrix denoting the foreground support, and \circ denotes the Hadamard product. We use $\mathcal{P}_S(Y)$ to represent the orthogonal projection of a matrix Y onto the linear space supported by S :

$$(3.1) \quad S_{ij} = \begin{cases} 0 & \text{if } Y_{ij} \text{ is background,} \\ 1 & \text{if } Y_{ij} \text{ is foreground,} \end{cases} \quad \mathcal{P}_S(Y)(i, j) = \begin{cases} 0 & \text{if } S_{ij} = 0, \\ Y_{ij} & \text{if } S_{ij} = 1, \end{cases}$$

where $\mathcal{P}_{S^\perp}(Y)$ is its complementary projection, i.e., $\mathcal{P}_S(Y) + \mathcal{P}_{S^\perp}(Y) = Y$.

3.2. Low-rank component. Our modeling is based on the low-rank parametrization of the unknown matrix L ($\text{rank}(L) = r$), given by

$$(3.2) \quad L = UV' = \sum_{j=1}^k u_{.j} v'_{.j},$$

where $r \leq k \leq \min(m, n)$. U and V are $m \times k$ and $n \times k$ matrices, respectively. Besides, we use $u_{.i}$ and $u_{.j}$ to denote the i^{th} row and the j^{th} column of U , respectively.

Since a low-rank estimate of L is sought, most columns in U and V are set equal to zero to achieve the column sparsity. To enforce this constraint, we associate the columns of U and V with Gaussian priors of precisions α_j , that is,

$$(3.3) \quad \begin{aligned} u_{.j} | \alpha &\sim \mathcal{N}(u_{.j} | 0, \alpha_j^{-1} I_m), \\ v_{.j} | \alpha &\sim \mathcal{N}(v_{.j} | 0, \alpha_j^{-1} I_n), \end{aligned}$$

where I_m denotes the $m \times m$ identity matrix.

In addition to (3.3), we incorporate the conjugate Gamma hyperprior on the precisions α_j

$$(3.4) \quad \alpha_j \sim \text{Gamma}\left(a, \frac{1}{b}\right).$$

The parameters a and b are regarded as deterministic and they are set to very small values (e.g., 10^{-6}) to obtain broad hyperpriors.

3.3. Noise component. We follow the standard assumption and incorporate white Gaussian noise in the observation, such that

$$(3.5) \quad n_{ij} | \beta \sim \mathcal{N}(n_{ij} | 0, \beta^{-1}),$$

with $\beta = 1/\varepsilon$ denoting the noise precision. The noise precision β is assigned the noninformative Jeffrey’s prior

$$(3.6) \quad p(\beta) = \beta^{-1},$$

which has been applied successfully in variable selection [1].

3.4. Formulation. To make the problem well-posed, we have the following models to describe it:

Bayesian model: Given observation $Y = \mathcal{P}_{S^\perp}(L + N)$, our objective is to estimate the underlying background² L as well as the foreground³ support S . The cardinality of the set S^\perp is pmn , with p the fraction of observed coefficients. According to the probabilistic interpretation above, the joint distribution is given by

$$(3.7) \quad p(Y, U, V, \alpha, \beta) = p(U | \alpha)p(V | \alpha)p(Y | U, V, \beta)p(\alpha)p(\beta).$$

Equality (3.7) is abbreviated to $p(Y, \Theta) = p(Y | \Theta)p(\Theta)$, where $\Theta = \{U, V, \alpha, \beta\}$ represents all model parameters.

Ising model: The foreground is defined as any object that moves differently from the background and that can be detected as outlier in the low-rank representation. In other words, the sparse noise has a clustering effect. Thus, we prefer to detect contiguous regions. The binary states of entries in foreground support S can be naturally modeled by MRFs. Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} is the set of vertices denoting all $m \times n$ pixels in the sequence and \mathcal{E} is the set of edges connecting spatially or temporally neighboring pixels. Then, the energy of S is given by the Ising model [14]

$$(3.8) \quad \sum_{ij \in \mathcal{V}} \lambda_{ij} S_{ij} + \sum_{(ij, xy) \in \mathcal{E}} \lambda_{ij, xy} |S_{ij} - S_{xy}|.$$

² Background refers to the background scene that needs to be reconstructed.

³ Foreground refers to the foreground objects that we intend to detect or track.

For simplicity, we set $\lambda_{ij} = \xi$ and $\lambda_{ij,xy} = \eta$, where $\xi > 0$ and $\eta > 0$ are positive constants penalizing $S_{ij} = 1$ and $|S_{ij} - S_{xy}|_{(ij,xy) \in \mathcal{E}}$, respectively.

Signal model: The signal model [26] describes the formation of Y , for given L and S . As we have assumed that $n_{ij} \sim \mathcal{N}(0, \beta^{-1})$, n_{ij} should be the best fitting to zero in the least-squares sense, when $S_{ij} = 0$. Besides, in the other regions where $S_{ij} = 1$, the background scene is occluded by the foreground. Thus, S_{ij} is not constrained.

Combining the three models above, we propose a mixture model Bayesian-Ising-Signal (BIS) to estimate L and S :

$$(3.9) \quad \min_{L,S} \left(\frac{1}{2} \|\mathcal{P}_{S^\perp}(Y - L)\|_F^2 + \xi \|S\|_1 + \eta \|A \text{vec}(S)\|_1 \right) \\ \text{s.t. } Y = \mathcal{P}_{S^\perp}(L + N), \quad L = UV', \quad p(Y, \Theta) = p(Y | \Theta)p(\Theta).$$

Here, A is the node-edge incidence matrix of \mathcal{G} and $\text{vec}(S)$ makes the matrix S vectorized.

3.5. Relation to the RPCA. As discussed in [25], SPCP tries to find the decomposition by minimizing the following energy:

$$(3.10) \quad \min_{L,E} \left(\frac{1}{2} \|Y - L - E\|_F^2 + \lambda_1 \text{rank}(L) + \lambda_2 \|E\|_0 \right).$$

We must have $E_{ij} = Y_{ij} - L_{ij}$ to minimize (3.10) when $E_{ij} \neq 0$. Noticing that $\|E\|_0 = \|S\|_1$ and replacing $\text{rank}(L)$ with $\|L\|_*$, (3.10) has the same minimizer with the following energy:

$$(3.11) \quad \min_{L,S} \left(\frac{1}{2} \|\mathcal{P}_{S^\perp}(Y - L)\|_F^2 + \lambda_1 \|L\|_* + \lambda_2 \|S\|_1 \right).$$

Lemma 3.1. *For any matrix $L \in \mathbb{R}^{m \times n}$, the following holds [18]:*

$$(3.12) \quad \|L\|_* = \arg \min_{U,V,L=UV'} \frac{1}{2} (\|U\|_F^2 + \|V\|_F^2).$$

If $\text{rank}(L) = k \leq \min(m, n)$, then the minimum above is attained at a factor decomposition $L = U_{m \times k} V'_{n \times k}$.

Theorem 3.2. *The solution space of (3.11) is contained in that of (3.9) given the outlier support \widehat{S} .*

Proof. By treating U, V as model parameters and α, β as hyperparameters with fixed values, we use maximum a posterior (MAP) estimation to compute U and V . From the Bayes' rule, (3.9) can be written as

$$(3.13) \quad \begin{aligned} \log p(U, V | Y, \alpha, \beta) &= \log p(Y, U, V, \alpha, \beta) - \log p(Y, \alpha, \beta) \\ &= -\frac{\beta}{2} \|\mathcal{P}_{\widehat{S}^\perp}(Y - UV')\|_F^2 - \frac{\alpha}{2} (\|U\|_F^2 + \|V\|_F^2) + C, \end{aligned}$$

where C is a constant term independent of U and V . Obviously, the problem of maximizing $\log p(U, V | Y, \alpha, \beta)$ w.r.t. U and V is equivalent to the following minimization problem:

$$(3.14) \quad \min_{U, V} \left(\frac{1}{2} \|\mathcal{P}_{\widehat{S}^\perp}(Y - UV')\|_F^2 + \frac{\lambda}{2} (\|U\|_F^2 + \|V\|_F^2) \right),$$

where $\lambda = \alpha/\beta$.

Suppose L is a solution to (3.11) with $\text{rank}(L) = k$ and $\lambda = \lambda_1$, we can find $U_{m \times k}$ and $V_{n \times k}$ satisfying $L = UV'$. Then we have

$$(3.15) \quad \begin{aligned} \arg \min_{U, V} \left(\frac{1}{2} \|\mathcal{P}_{\widehat{S}^\perp}(Y - UV')\|_F^2 + \frac{\lambda}{2} (\|U\|_F^2 + \|V\|_F^2) \right) \\ = \arg \min_{U, V} \left(\frac{1}{2} \|\mathcal{P}_{\widehat{S}^\perp}(Y - UV')\|_F^2 + \lambda_1 \|UV'\|_* \right) \\ = \arg \min_{L, \text{rank}(L)=k} \left(\frac{1}{2} \|\mathcal{P}_{\widehat{S}^\perp}(Y - L)\|_F^2 + \lambda_1 \|L\|_* \right). \end{aligned}$$

□

3.6. Relation to the truncated SVD. By considering the SVD, any matrix of rank k can be decomposed in the form

$$(3.16) \quad L = \widetilde{U} \widetilde{S} \widetilde{V}' = \sum_{i=1}^k \tilde{u}_i \sigma_i \tilde{v}_i',$$

where \widetilde{U} and \widetilde{V} are $m \times k$ and $n \times k$ matrices with orthogonal columns, and \widetilde{S} is a $k \times k$ diagonal matrix of the nonzero singular values.

According to the truncated SVD [10], the low-rank matrix L is approximated by a lower-rank matrix

$$(3.17) \quad L = \sum_{i=1}^s \tilde{u}_i \sigma_i \tilde{v}_i^T,$$

where $s < k = \text{rank } L$. The problem of (3.17) is equivalent to the following minimization problem:

$$(3.18) \quad \min_L \|Y - L\|_F \quad \text{s.t.} \quad \text{rank}(L) \leq s,$$

where $\|\cdot\|_F$ denotes the Frobenius norm. However, (3.18) is NP-hard and no known polynomial-time algorithms exist. A popular technique is to utilize convex relaxation based on the nuclear norm [4]. To make the optimization tractable, (3.18) is relaxed by replacing $\text{rank}(L)$ with $\|L\|_*$, given by

$$(3.19) \quad \min_L \left(\frac{1}{2} \|Y - L\|_F^2 + \lambda_1 \|L\|_* \right) \quad \text{s.t.} \quad Y = L + N,$$

where $\|\cdot\|_*$ means the nuclear norm.

Furthermore, the sparse matrix E is not included in this model. We can derive the form of model (1.2) by adding the term $\|E\|_1$ to (3.19).

4. ALGORITHM

The objective function defined in (3.9) is convex with continuous and discrete variables. Hence, we adopt an alternating algorithm that separates the energy minimization over L and S into two steps. L -step is a low-rank matrix estimation problem, and S -step is a combinatorial optimization problem. It turns out that the optimal solutions of L -step and S -step can be computed efficiently.

4.1. Estimation of the low-rank matrix L .

4.1.1. Variational Bayesian inference. In this work, we use the variational Bayes (VB) [2] method to infer the posterior approximation for each latent variable. Let Θ be the vector of all latent variables such that $\Theta = \{U, V, \alpha, \beta\}$. VB seeks an approximation distribution $q(\Theta)$ to the true posterior $p(\Theta | Y)$ by solving the following variational optimization:

$$(4.1) \quad \min_{q \in \mathcal{C}} KL(q \| p) = - \int q(\Theta) \ln \frac{p(\Theta | Y)}{q(\Theta)} d\Theta,$$

where $KL(q \| p)$ denotes KL divergence between $q(\Theta)$ and $p(\Theta | Y)$, and \mathcal{C} denotes the set of probability densities with certain restrictions to make the minimization tractable. Taking $q(\Theta) = \prod_k q(\Theta_k)$, the closed-form solution to $q(\Theta_k)$, with other factors fixed, can be attained by:

$$(4.2) \quad q^*(\Theta_k) = \frac{\exp\{\langle \ln p(\Theta, Y) \rangle_{\Theta \setminus \Theta_k}\}}{\int \exp\{\langle \ln p(\Theta, Y) \rangle_{\Theta \setminus \Theta_k}\} d\Theta_k},$$

where $\langle \cdot \rangle$ denotes the expectation, and $\Theta \setminus \Theta_k$ denotes the set of Θ with Θ_k removed. Problem (4.1) can be solved by alternatively calculating (4.2).

4.1.2. Inference for Bayesian model. Estimation of factors U and V . With some algebra, it follows from (4.2) that the approximation to the posterior distribution of U and V decomposes as independent distributions of their rows. By combining the priors in (3.3), the posterior density of u_i is given by

$$(4.3) \quad q(u_{i\cdot}) = \mathcal{N}(u_{i\cdot} \mid \langle u_{i\cdot} \rangle, \Sigma_i^u)$$

with mean and covariance

$$(4.4) \quad \begin{aligned} \langle u_{i\cdot} \rangle &= \langle \beta \rangle y_{i\cdot} \langle V_i \rangle \Sigma_i^u, \\ \Sigma_i^u &= (\langle \beta \rangle \langle V_i' V_i \rangle + \Lambda)^{-1}, \end{aligned}$$

where $\Lambda = \text{diag}(\langle \alpha \rangle)$ and the matrix V_i contains only the j^{th} rows of V for $(i, j) \in S^\perp$, such that

$$(4.5) \quad \langle V_i' V_i \rangle = \sum_{j: (i,j) \in S^\perp} \langle v_j' v_{j\cdot} \rangle = \sum_{j: (i,j) \in S^\perp} (\langle v_j' \rangle \langle v_{j\cdot} \rangle + \Sigma_j^v)$$

with Σ_j^v denoting the posterior covariance of the j^{th} row of V . Additionally, the row vector $y_{i\cdot}$ contains the observed entries in the i^{th} row of Y . Similarly, the posterior approximation of v_j is given by a normal distribution

$$(4.6) \quad q(v_{j\cdot}) = \mathcal{N}(v_{j\cdot} \mid \langle v_{j\cdot} \rangle, \Sigma_j^v)$$

with parameters

$$(4.7) \quad \begin{aligned} \langle v_{j\cdot} \rangle &= \langle \beta \rangle y_{\cdot j}' \langle U_j \rangle \Sigma_j^v, \\ \Sigma_j^v &= (\langle \beta \rangle \langle U_j' U_j \rangle + \Lambda)^{-1}, \end{aligned}$$

where $y_{\cdot j}$ contains the observed entries in the j^{th} column of Y , and U_j contains the i^{th} row of U for $(i, j) \in S^\perp$. It can be seen that the covariances Σ_i^v of the estimate of V are incorporated in the estimation of U (and vice versa).

Estimation of hyperparameters α . By combining $p(U \mid \alpha)$, $p(V \mid \alpha)$, and $p(\alpha_i)$, the posterior density of α_i becomes a Gamma distribution

$$(4.8) \quad q(\alpha_i) \propto \alpha_i^{a-1+(m+n)/2} \exp\left(-\alpha_i \frac{2b + \langle u_i' u_{i\cdot} \rangle + \langle v_i' v_{i\cdot} \rangle}{2}\right)$$

with mean

$$(4.9) \quad \langle \alpha_i \rangle = \frac{2a + m + n}{2b + \langle u_i' u_i \rangle + \langle v_i' v_i \rangle}.$$

The required expectations are given by

$$(4.10) \quad \langle u_i' u_i \rangle = \langle u_i \rangle' \langle u_i \rangle + \sum_j (\Sigma_j^u)_{ii},$$

$$(4.11) \quad \langle v_i' v_i \rangle = \langle v_i \rangle' \langle v_i \rangle + \sum_j (\Sigma_j^v)_{ii}.$$

Estimation of noise precision β . Finally, the posterior approximation of the noise precision assumes a Gamma distribution with mean

$$(4.12) \quad \langle \beta \rangle = \frac{pmn}{\langle \|Y - \mathcal{P}_{S^\perp}(UV')\|_F^2 \rangle},$$

where

$$(4.13) \quad \begin{aligned} \langle \|Y - \mathcal{P}_{S^\perp}(UV')\|_F^2 \rangle &= \|Y - \mathcal{P}_{S^\perp}(\langle U \rangle \langle V \rangle')\|_F^2 \\ &+ \sum_j \left(\text{Tr}(\langle u_i \rangle' \langle u_i \rangle \Sigma_j^v) + \text{Tr}(\langle v_j \rangle' \langle v_j \rangle \sum_i \Sigma_i^u) \right) \\ &+ \text{Tr} \left(\sum_i \Sigma_i^u \Sigma_j^v \right), \quad (i, j) \in S^\perp. \end{aligned}$$

In summary, the L -step proceeds by first estimating the rows of U and V using (4.4) and (4.7), respectively, followed by the estimation of the precision α_i using (4.9), and the noise precision β using (4.12).

4.2. Estimation of the outlier support S . Note that $S_{ij} \in \{0, 1\}$, the energy in (3.9) can be rewritten over S for the low-rank matrix \widehat{L} as follows:

$$(4.14) \quad \begin{aligned} &\frac{1}{2} \|\mathcal{P}_{S^\perp}(Y - \widehat{L})\|_F^2 + \xi \|S\|_1 + \eta \|A \text{vec}(S)\|_1 \\ &= \frac{1}{2} \sum_{ij} (Y_{ij} - \widehat{L}_{ij})^2 (1 - S_{ij}) + \xi \sum_{ij} S_{ij} + \eta \|A \text{vec}(S)\|_1 \\ &= \sum_{ij} \left(\xi - \frac{1}{2} (Y_{ij} - \widehat{L}_{ij})^2 \right) S_{ij} + \eta \|A \text{vec}(S)\|_1 + C, \end{aligned}$$

where $C = \frac{1}{2} \sum_{ij} (Y_{ij} - \widehat{L}_{ij})^2$ is a constant when \widehat{L} is fixed. The energy above is in the standard form of the first-order MRFs with binary labels, which can be solved exactly using graph cuts [3], [12].

All steps of BIS are summarized in Algorithm 1.

Algorithm 1: BIS algorithm for RPCA

1. Input: $Y \in \mathbb{R}^{m \times n}$
 2. Initialize: $\widehat{L}, \widehat{S}, \widehat{U}, \widehat{V}, \alpha, \beta$ randomly
 3. repeat
 4. repeat
 5. L -step: $\widehat{L} = \arg \min_L \frac{1}{2} \|\mathcal{P}_{\widehat{S}}^\perp(Y - L)\|_F^2$
 6. until convergence
 7. S -step:
 8. $\widehat{S} = \arg \min_S \frac{1}{2} \|\mathcal{P}_{S^\perp}(Y - \widehat{L})\|_F^2 + \xi \|S\|_1 + \eta \|A \text{vec}(S)\|_1$
 9. until convergence
 10. $\widehat{E} = Y \circ \widehat{S}$
 11. Output: $\widehat{L}, \widehat{E}, \widehat{S}$
-

4.3. Initialization and parameter tuning. Although randomly initializing the matrix U and V generally provided satisfactory results, faster convergence and better reconstruction performance can be achieved by carefully selecting the initial values. During the implementation, we calculate the SVD of the matrix $Y = USV'$ and set $U_0 = US^{\frac{1}{2}}$ and $V_0 = S^{\frac{1}{2}}V'$. With this choice, the algorithm is initialized with a full-rank matrix. The empirical results show negligible difference in performance if a reasonable initial rank (larger than the true rank) is chosen, whereas the computational complexity can be significantly reduced.

The parameter ξ in (3.9) controls the sparsity of the outlier support. From (4.14), it can be seen that \widehat{S}_{ij} is more likely to be 1 if $\frac{1}{2}(Y_{ij} - \widehat{L}_{ij})^2 > \xi$. Thus the choice of ξ should depend on the noise level in images. Typically we set $\xi = 4.5\widehat{\sigma}^2$, where $\widehat{\sigma}^2$ is estimated online by the variance of $Y_{ij} - \widehat{L}_{ij}$. Since the estimation of \widehat{L} and $\widehat{\sigma}$ is biased at the beginning, we propose to start our algorithm with a relatively large ξ , and then reduce ξ by a factor $\varepsilon = 0.5$ after each iteration until $\xi = 4.5\widehat{\sigma}^2$. In other words, the proposed algorithm here tolerates more error in model fitting at the beginning, since the model itself is not accurate enough. With the model estimation getting better and better, we decrease the threshold and declare more and more outliers. In Section 5.1.3 it can be found that our method performs stably when $\eta \in [0.5\xi, 2.5\xi]$. In the experiments, we set $\eta = \xi$ for simulation and choose optimal η for real sequences.

4.4. Convergence and computational complexity. With the properties of the variational Bayes methods, the algorithm is guaranteed to converge to a local minimum of the variational bound [2]. The complexity of the algorithm is $O(mk^3 + nk^3)$

per iteration, where k is equal to the estimated rank at each iteration. For fixed parameters, we also minimize a single lower-bound energy in each step. Therefore, the algorithm must converge to a local minimum. For adaptive parameter tuning, our strategy guarantees that the coefficients (ξ, η) keep decreasing for each change. Thus, the energy in (3.9) decreases monotonically with the algorithm running.

5. EXPERIMENTS

5.1. Simulation. In this section, we first demonstrate the performance of the proposed method with some existing algorithms based on simulated data. The objective of this task is to recover background and estimate outlier entries at the same time.

We generate the observed matrix Y by adding a foreground occlusion with support S_0 to a background matrix L_0 . The background matrix L_0 with rank r is generated as $L_0 = UV'$, where U and V are $m \times r$ and $n \times r$ matrices randomly sampled from a standard normal distribution $\mathcal{N}(0, 1)$. We choose $m = 100$, $n = 50$ and $r = 5$ for all experiments. Then, an object with width $W = 40$ is superposed on each column of L_0 and shifted downwards for 1 pixel per column. The intensity of this object is independently sampled from a uniform distribution $U(-c, c)$, where c is chosen to be the largest magnitude of entries in L_0 . Moreover, i.i.d. Gaussian noise is added to the original data with signal noise ratio (SNR)=10(dB).

We measure the accuracy of low-rank recovery by calculating the difference between \widehat{L} and L_0 for quantitative evaluation. The root mean square error (RMSE) is used to measure the difference:

$$(5.1) \quad \text{RMSE} = \frac{\|\widehat{L} - L_0\|_F}{\|L_0\|_F}.$$

Furthermore, we measure the accuracy of outlier detection by comparing \widehat{S} with S_0 . F -measure that combines precision and recall is used to evaluate the results:

$$(5.2) \quad F\text{-measure} = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

where

$$(5.3) \quad \text{precision} = \frac{TP}{TP + FP}, \quad \text{recall} = \frac{TP}{TP + FN},$$

where TP , FP , TN , and FN are the numbers of true positives, false positives, true negatives, and false negatives, respectively. The higher the F -measure is, the better the detection accuracy is.

5.1.1. Comparison with state-of-the-art. The results obtained from different methods are shown in Figure 1. BIS provides competitive performance in terms of low-rank matrix recovery, which is followed by DECOLOR, PCP and ALM. As far as outlier detection is concerned, BIS and DECOLOR obtain the best masks, not only visually but also quantitatively, and other methods except BRPCA and VBLR give reasonable results. BRPCA and VBLR totally fail in this task. Although PCP, MBRMF and ALM perform well in detecting the outliers, they fail to recover the low-rank matrix well. Quantitative results are shown in Table 1. It can be seen that BIS outperforms other methods for both low-rank matrix recovery and outlier detection.

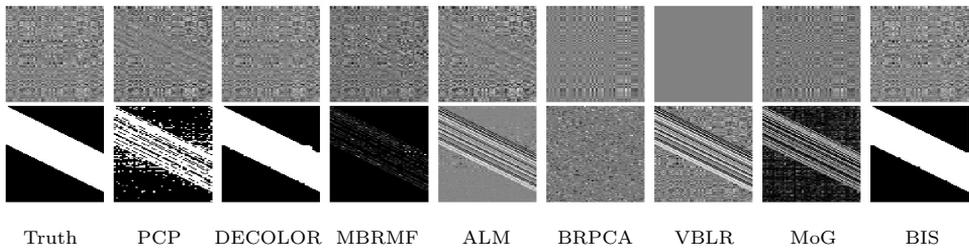


Figure 1. Numerical experiments results. The underlying background images L_0 and the foreground support S_0 are shown in Column 1. The corresponding results by 8 methods are presented from Column 2 to Column 9. The top panel is the estimated background and the bottom is the foreground mask.

Methods	PCP	DECOLOR	MBRMF	ALM	BRPCA	VBLR	MoG	BIS
RMSE	0.578	0.252	0.742	0.579	0.940	0.999	0.879	0.069
F -measure	0.797	0.996	0.783	0.794	0.495	0.682	0.688	0.999

Table 1. Comparison of different methods in the numerical experiment. The first row shows the recovery results based on RMSE and the second row shows the outlier detection results based on the F -measure.

5.1.2. Quantitative evaluation. We perform random experiments with different true rank, object width W and SNR. RMSE, F -measure and estimated rank as functions of the true rank are shown in Figure 2(a), (b), (c). BIS gives the best result on the aspect of low-rank matrix recovery, which is followed by DECOLOR and PCP. With respect to outlier detection, BIS and DECOLOR provide more accurate estimates than other methods, and all other methods except for BRPCA give reasonable results. Although ALM is a very attractive optimization-based method due to its recovery performance and fast convergence, it does not estimate the correct rank in all cases. BIS correctly estimates the unknown rank and the sparsity level where $r \leq 15$ (under the condition of low-rank).

From Figure 2(d), it can be seen that BIS keeps effective and stable as W increases with $r = 5$ and $\text{SNR} = 10$. Moreover, the accuracy of low-rank matrix recovery drops obviously when $W \geq 40$. The reason is that some background pixels are always occluded when the foreground is too large, so they cannot be recovered even when the foreground can be detected accurately.

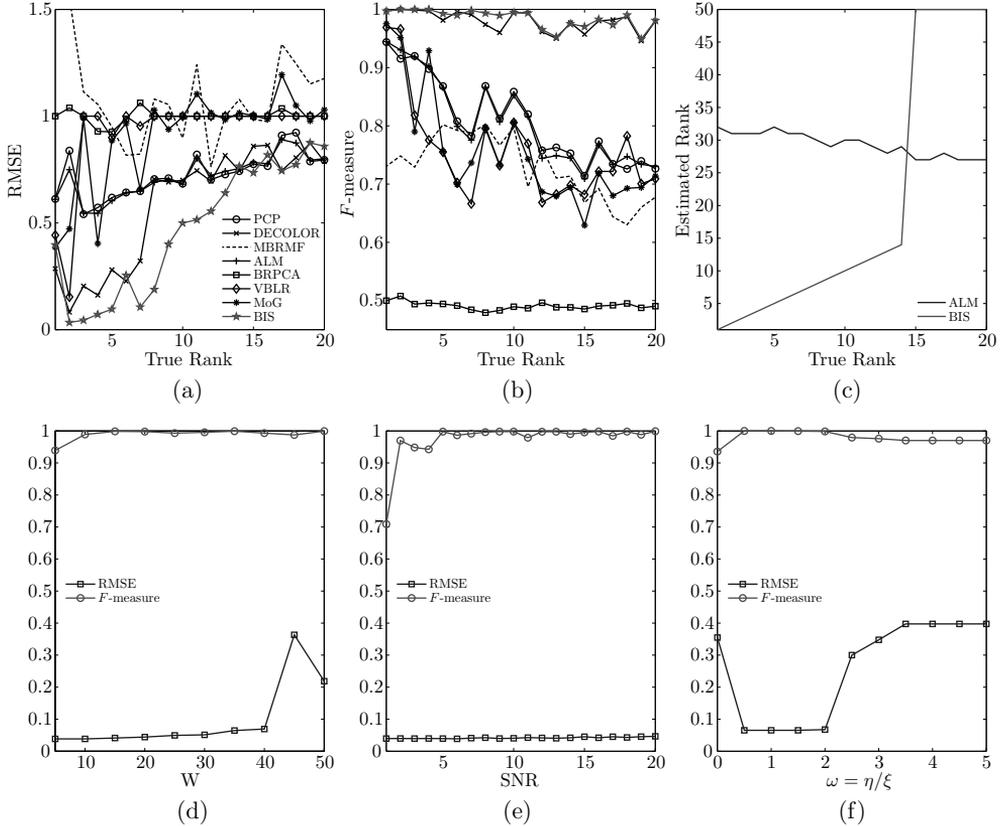


Figure 2. Quantitative evaluation. Top row: RMSE, F -measure and estimated rank as functions of true rank, bottom row: RMSE and F -measure as functions of W , SNR and ω .

RMSE and F -measure as functions of SNR show the results under different noise levels in Figure 2(e). As SNR increases with $r = 5$ and $W = 40$, the performance of BIS keeps less affected, which demonstrates the robustness of our method.

5.1.3. Effect of the parameter η . Figure 2(f) demonstrates the effect of the parameter η . It controls the strength of interaction between neighboring pixels. It can be seen that the performance of our method keeps very stable when $\eta \in [0.5\xi, 2.5\xi]$.

5.2. Background modeling. Background modeling, subtracting background from video sequences captured by a static camera, can be modeled as a low-rank matrix analysis problem. Four commonly utilized video sequences, namely an indoor scene (Hall), an outdoor scene (Pedestrian) and two dynamic background scenes (Airport and WaterSurface), were adopted in our experiment. We compare the BIS with other competing methods, namely PCP, DECOLOR, MBRMF, ALM, BRPCA, VBLR and MoG. The last frame of each sequence and results are shown in Figure 3.

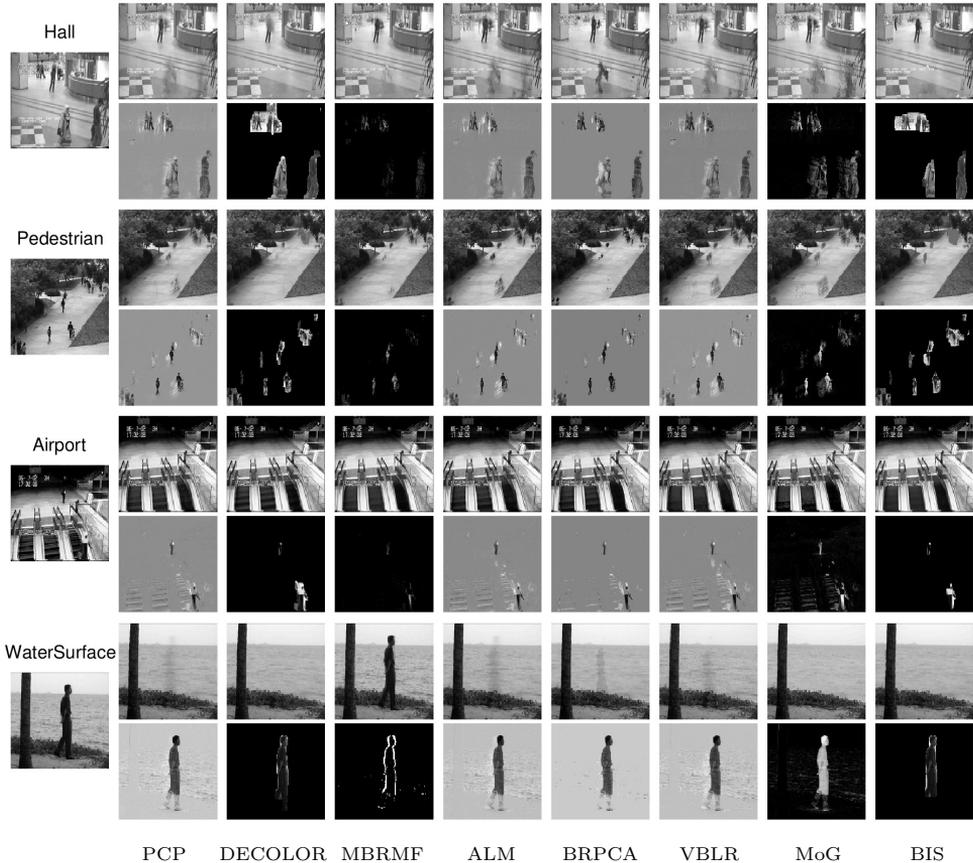


Figure 3. Four subsequences of surveillance videos. The corresponding results by 8 methods are presented from Column 1 to Column 8. The top panel is the estimated background and the bottom is the detected outliers.

It can be seen that all the competing methods can extract the background from videos with slight difference in visualization. The method proposed in this paper, however, can extract more elaborate background and foreground information. More specifically, our method can subtract the dynamic background (Airport and WaterSurface) and remove the foreground noise (all subsequences).

5.3. Face modeling. This experiment aims to test the effectiveness of BIS in face modeling applications. The second subset of the Extended Yale B database, consisting of 64 faces of one subject with size 192×168 , was used to generate the data matrix. The reconstructed faces by all methods are compared in Figure 4.



Figure 4. From left to right: original faces, reconstructed faces and extracted noise by 8 methods. The heat-map images (blue = dark pixels, red = bright pixels) are better seen in the electronic version by zooming on a computer screen.

The proposed method, as well as the other competing methods, is able to remove the cast shadows and saturations in faces. BIS and BRPCA perform better on faces with a large region. But BRPCA has not completely separated the reconstructed face and noise with a small region. Such face images contain both significant cast shadow and saturations noises, which correspond to the highly dark and bright areas in the face, and camera/red noise which is much amplified in the dark areas. It is very interesting that the proposed method is capable of accurately extracting these two kinds of noise. The better noise fitting capability of the proposed method thus leads to better face reconstruction performance.

6. CONCLUSION

We proposed a novel RPCA method based on sparse Bayesian learning principles and Markov random fields. Starting from the low-rank factorization of the unknown matrix, we enforce a common sparsity profile on its underlying components using a probabilistic formulation. The sparse component is modeled by incorporating the contiguity prior as a sparsity constraint and exactly solved using graph cuts. Moreover, we modeled the remaining unknown variables and observations within the hierarchical Bayesian framework and developed inference methods based on the variational inference method. The effectiveness of our method was demonstrated by synthetic and real data. The proposed method outperforms previous methods in terms of accurately recovering the low-rank structure and detecting contiguous outliers from the observed data.

References

- [1] *S. D. Babacan, M. Luessi, R. Molina, A. K. Katsaggelos*: Sparse Bayesian methods for low-rank matrix estimation. *IEEE Trans. Signal Process.* *60* (2012), 3964–3977.
- [2] *C. M. Bishop*: *Pattern Recognition and Machine Learning*. Information Science and Statistics, Springer, New York, 2006.
- [3] *Y. Boykov, O. Veksler, R. Zabih*: Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* *23* (2010), 1222–1239.
- [4] *E. J. Candès, X. Li, Y. Ma, J. Wright*: Robust principal component analysis? *J. ACM* *58* (2011), Article No. 11, 37 pages.
- [5] *E. J. Candès, B. Recht*: Exact matrix completion via convex optimization. *Found. Comput. Math.* *9* (2009), 717–772.
- [6] *E. J. Candès, T. Tao*: The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inf. Theory* *56* (2010), 2053–2080.
- [7] *F. De la Torre, M. J. Black*: A framework for robust subspace learning. *Int. J. Comput. Vis.* *54* (2003), 117–142.
- [8] *X. Ding, L. He, L. Carin*: Bayesian robust principal component analysis. *IEEE Trans. Image Process.* *20* (2011), 3419–3430.
- [9] *C. Ding, D. Zhou, X. He, H. Zha*: R_1 -PCA: rotational invariant L_1 -norm principal component analysis for robust subspace factorization. *Proc. 23rd Int. Conf. on Machine Learning*. Pittsburgh. ACM, New York, 2006, pp. 281–288.
- [10] *P. C. Hansen*: *Rank-Deficient and Discrete Ill-Posed Problems. Numerical Aspects of Linear Inversion*. SIAM Monographs on Mathematical Modeling and Computation 4, Society for Industrial and Applied Mathematics, Philadelphia, 1998.
- [11] *I. T. Jolliffe*: *Principal Component Analysis*. Springer Series in Statistics, Springer, New York, 2002.
- [12] *V. Kolmogorov, R. Zabih*: What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.* *26* (2004), 147–159.
- [13] *N. Kwak*: Principal component analysis based on L_1 -norm maximization. *IEEE Trans. Pattern Anal. Mach. Intell.* *30* (2008), 1672–1680.
- [14] *S. Z. Li*: *Markov Random Field Modeling in Image Analysis*. *Advances in Pattern Recognition*, Springer, London, 2009.

- [15] *Z. Lin, M. Chen, Y. Ma*: The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. ArXiv preprint arXiv:1009.5055, 2010.
- [16] *Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, Y. Ma*: Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. Computational Advances in Multi-Sensor Adaptive Processing. Aruba, Dutch Antilles, Proceedings. IEEE, 2009.
- [17] *G. Liu, Z. Lin, Y. Yu*: Robust subspace segmentation by low-rank representation. Proc. 27th Int. Conf. on Machine Learning, Haifa. Proceedings Omni Press., 2010, pp. 663–670.
- [18] *R. Mazumder, T. Hastie, R. Tibshirani*: Spectral regularization algorithms for learning large incomplete matrices. J. Mach. Learn. Res. 11 (2010), 2287–2322.
- [19] *Y. Peng, A. Ganesh, J. Wright, W. Xu, Y. Ma*: RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012), 2233–2246.
- [20] *B. Recht, M. Fazel, P. A. Parrilo*: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. 52 (2010), 471–501.
- [21] *N. Wang, D. Y. Yeung*: Bayesian robust matrix factorization for image and video processing. Computer Vision, 2013 IEEE International Conference, Sydney. Proceedings. IEEE, 2013, pp. 1785–1792.
- [22] *J. Wright, A. Ganesh, S. Rao, Y. Peng, Y. Ma*: Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. Advances in Neural Information Processing Systems 22, Vancouver. Proceedings. Curran Associates, 2009, pp. 2080–2088.
- [23] *H. Xu, C. Caramanis, S. Sanghavi*: Robust PCA via outlier pursuit. IEEE Trans. Inform. Theory 58 (2012), 3047–3064.
- [24] *Q. Zhao, D. Meng, Z. Xu, W. Zuo, L. Zhang*: Robust principal component analysis with complex noise. Proc. 31st Int. Conf. on Machine Learning, Beijing. Proceedings. J. Mach. Learn. Res., 2014, pp. 55–63.
- [25] *Z. Zhou, X. Li, J. Wright, E. J. Candès, Y. Ma*: Stable principal component pursuit. 2010 IEEE International Symposium on Information Theory Proceedings, Austin. Proceedings. IEEE, 2010, pp. 1518–1522.
- [26] *X. Zhou, C. Yang, W. Yu*: Moving object detection by detecting contiguous outliers in the low-rank representation. IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013), 597–610.

Authors' addresses: Guoqiang Huan, Zhanjie Song, School of Science, Tianjin University, Tianjin 300072, China; Center for Applied Mathematics, Tianjin University, Tianjin 300072, China, e-mail: huanguoqiang@tju.edu.cn, zhanjiesong@tju.edu.cn; Ying Li (corresponding author), State Key Laboratory of Hydraulic Engineering Simulation and Safety, Tianjin 300072, China, e-mail: yingli2011@tju.edu.cn.