L_1-Norm Low-Rank Matrix Factorization by Variational Bayesian Method

Qian Zhao, Deyu Meng, *Member, IEEE*, Zongben Xu, Wangmeng Zuo, *Member, IEEE*, and Yan Yan

***Abstract*— The L_1-norm low-rank matrix factorization (LRMF) has been attracting much attention due to its wide applications to computer vision and pattern recognition. In this paper, we construct a new hierarchical Bayesian generative model for the L_1-norm LRMF problem and design a mean-field variational method to automatically infer all the parameters involved in the model by closed-form equations. The variational Bayesian inference in the proposed method can be understood as solving a weighted LRMF problem with different weights on matrix elements based on their significance and with L_2-regularization penalties on parameters. Throughout the inference process of our method, the weights imposed on the matrix elements can be adaptively fitted so that the adverse influence of noises and outliers embedded in data can be largely suppressed, and the parameters can be appropriately regularized so that the generalization capability of the problem can be statistically guaranteed. The robustness and the efficiency of the proposed method are substantiated by a series of synthetic and real data experiments, as compared with the state-of-the-art L_1-norm LRMF methods. Especially, attributed to the intrinsic generalization capability of the Bayesian methodology, our method can always predict better on the unobserved ground truth data than existing methods.**

***Index Terms*— Background subtraction, face reconstruction, low-rank matrix factorization (LRMF), outlier detection, robustness, variational inference.**

I. Introduction

LOW-RANK matrix factorization (LRMF) is one of the fundamental problems in computer vision and pattern recognition and has wide applications including structure from motion (SFM) [1], shape from varying illumination [2], motion estimation [3], and object tracking [4]. Representing the observation data as an $m \times n$ matrix $\mathbf{X}$ with entries x_{ij}s, LRMF aims to factorize the matrix into two smaller ones $\mathbf{U} \in \mathbb{R}^{r\times m}$ and $\mathbf{V} \in \mathbb{R}^{r\times n}$, where $r \ll \min(m, n)$, such that

$$\mathbf{X} \approx \mathbf{U}^T\mathbf{V}. \tag{1}$$

This can be achieved by solving the optimization problem

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{U}^T\mathbf{V}\|_p \tag{2}$$

where $\|\mathbf{A}\|_p = (\sum_{i,j} |a_{ij}|^p)^{1/p}$ denotes the L_p-norm of a matrix. To deal with missing data problem in real applications, the above optimization is often reformulated as

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{W} \odot (\mathbf{X} - \mathbf{U}^T\mathbf{V})\|_p \tag{3}$$

where $\odot$ denotes the Hadamard product (the component-wise multiplication), and the element w_{ij} in the indicator matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$ equals 1 if the corresponding element x_{ij} in $\mathbf{X}$ is known, and 0 otherwise.

Under the assumption of Gaussian noises, it is natural to utilize the Frobenius norm (i.e., $p = 2$), which has been extensively studied in LRMF literatures. A unique approximation $\mathbf{U}^T\mathbf{V}$ can be easily attained by the well-known singular value decomposition (SVD) method [5], which is the global minimizer of the cost function (2). For the more difficult problem (3), although the global minimum cannot be guaranteed in general, there have already been a variety of methods to solve it effectively [6]–[10]. For example, Srebro and Jaakkola [6] proposed an expectation maximization (EM)-based method to solve the problem and applied it to collaborative filtering. Buchanan and Fitzgibbon [7] presented a damped Newton algorithm, using the information of second derivatives with a damping factor, and achieved satisfactory performance in computer vision applications. Mitra *et al.* [8] converted the original problem into a low-rank semidefinite programming, which can be solved efficiently, and gave good results on heavy missing data cases. Okatani and Deguchi [9] extended the Wiberg algorithm to this problem, and this approach has been further improved by Okatani *et al.* [10] via incorporating a damping factor to the conventional Wiberg algorithm.

As is well known, however, the L_2-norm minimization is sensitive to non-Gaussian noises and outliers, which is often the case in real problems due to the mechanism of data acquisition. To address this robustness issue, a common approach is to replace the L_2-norm with the L_1-norm [11]–[13], resulting in the following problem:

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{W} \odot (\mathbf{X} - \mathbf{U}^T\mathbf{V})\|_1. \tag{4}$$

Manuscript received October 24, 2013; revised August 4, 2014 and October 27, 2014; accepted December 18, 2014. Date of publication January 15, 2015; date of current version March 16, 2015. This work was supported in part by the National Basic Research Program (973 Program) of China under Grant 2013CB329404, in part by the National Natural Science Foundation of China under Contract 61373114, Contract 11131006, and Contract 91330204, and in part by the Civil Aviation Administration of China under Grant U1233110.

Q. Zhao, D. Meng, and Z. Xu are with the Institute for Information and System Sciences, School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China, and also with Beijing Center for Mathematics and Information Interdisciplinary Sciences, Beijing 100048, China (e-mail: timmy.zhaoqian@gmail.com; dymeng@mail.xjtu.edu.cn; zbxu@mail.xjtu.edu.cn).

W. Zuo is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: cswmzuo@gmail.com).

Y. Yan is with the Department of Information Engineering and Computer Science, University of Trento, Trento 38123, Italy (e-mail: yan@disi.unitn.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2014.2387376

Due to its nonconvex and nonsmooth properties, however, (4) is generally difficult to solve. Some researchers devoted to reformulate it into other robust formulations to simplify the problem. For example, de la Torre and Black [14] adopted a robust function instead of the L_1-norm and used iterative reweighted least squares algorithms. Ding *et al.* [15] employed the rotational invariant R_1-norm defined by $\|\mathbf{X}\|_{R_1} = \sum_{i=1}^{n}(\sum_{i=j}^{d} x_{ji}^2)^{1/2}$ to replace the L_1-norm. Kwak [16] proposed another approach to suppressing the influence of outliers by maximizing the L_1 dispersion of the data for this problem. Very recently, Meng *et al.* [17] modeled the noise as a mixture of Gaussians to make the model adaptable to a wide range of noise types. In contrast, some methods were designed to solve (4) directly. Ke and Kanade [11] presented an alternative linear/quadratic programming (ALP/AQP) method by decomposing the problem into a sequence of convex subproblems and then alternatively solving them by linear/quadratic programming. Eriksson and van den Hengel [12] designed the L_1-Wiberg method by extending the traditional Wiberg algorithm to L_1-norm setting. To further enhance the computational efficiency, Zheng *et al.* [13] proposed a RegL_1-ALM method by adding a convex trace-norm regularization term to the objective function of (4) and solving it by the augmented Lagrange multiplier (ALM) method, resulting in higher accuracy and faster convergence. Recently, Wang *et al.* [18] proposed a probabilistic model for L_1-norm LRMF and utilized the conditional EM (CEM) algorithm to solve the problem, obtaining high accuracy with less time consumption. Wang and Yeung [19] also extended it to the Bayesian framework and used Markov chain Monte Carlo sampling method to do inference. Besides, Meng *et al.* [20] proposed a novel cyclic weighted median method, which can solve the problem with high computational efficiency.

Most of the existing L_1-norm LRMF methods aim to well fit the partial observations of $\mathbf{X}$, while they do not consider the generalization of their results on the unobserved ground truth data. Although high approximation accuracy on the known observations of the input matrix might be attained by these methods, large deviations on the unobserved elements in $\mathbf{X}$ can also be simultaneously conducted, especially when outliers or heavy noise exist. This is essentially the well-known *overfitting* problem. To alleviate this issue, in this paper, we propose a new method for solving L_1-norm LRMF under the Bayesian framework. We first reformulate the problem as a hierarchical Bayesian generative model by virtue of the Gaussian scale mixture representation for the Laplace distribution. Then, we employ the mean-field variational inference method to infer the posteriors. The true matrix underlying the corrupted and incomplete input data can then be well fitted with high accuracy as verified by extensive experiments.

The rest of this paper is organized as follows. In Section II, we first formulate the L_1-norm LRMF as a hierarchical Bayesian model, and then compare it with the optimization-based approach and other probabilistic models. In Section III, we present the mean-field variational inference method for the model, together with the evaluation for its lower bound, and analyze its computational complexity. Section IV provides the empirical evaluation for the proposed method on various problems. We then conclude this paper in Section V. Throughout this paper, we denote matrices, vectors, and scalars by the upper-case bold-faced letters, lower-case bold-faced letters, and lower-case nonbold-faced letters, respectively.

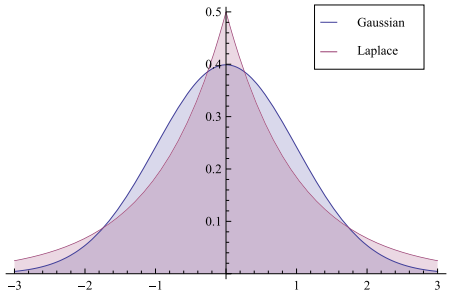


Fig. 1. Comparison of the PDFs of Gaussian and Laplace distributions with zero mean and unit variance.

II. Hierarchical Bayesian Model for L_1-Norm Low-Rank Matrix Factorization

In the following, we first discuss the L_1-norm LRMF problem in the maximum likelihood estimation (MLE) viewpoint, and then present our Bayesian formulation for the problem.

A. MLE Interpretation for L_1-Norm LRMF

Consider the generative model

$$x_{ij} = \mathbf{u}_i^T \mathbf{v}_j + \epsilon_{ij} \tag{5}$$

where x_{ij} is the element of $\mathbf{X}$ in its ith row and jth column, and $\mathbf{u}_i$ and $\mathbf{v}_j$ are the ith column of $\mathbf{U}$ and the jth column of $\mathbf{V}$, respectively. Assume that the noise ϵ_{ij} follows the Laplace distribution with zero mean:

$$\epsilon_{ij} \sim p(\epsilon_{ij}|0, b) \tag{6}$$

where

$$p(y|\mu, b) = \frac{1}{2b}\exp\left\{-\frac{|y-\mu|}{b}\right\} \tag{7}$$

is the probability density function (PDF) of the Laplace distribution, the log-likelihood function with respect to $\mathbf{U}$ and $\mathbf{V}$ can then be written as

$$\begin{aligned} L(\mathbf{U}, \mathbf{V}) &= \prod_{(i,j)\in\Omega} \ln p\left(x_{ij} - \mathbf{u}_i^T \mathbf{v}_j|0, b\right) \\ &= -\frac{1}{b}\sum_{i,j} w_{ij}\left|x_{ij} - \mathbf{u}_i^T \mathbf{v}_j\right| + C \\ &= -\frac{1}{b}\|\mathbf{W} \odot (\mathbf{X} - \mathbf{U}^T\mathbf{V})\|_1 + C \end{aligned} \tag{8}$$

where Ω denotes the set of the indices of the nonmissing data elements and C is a constant independent of $\mathbf{U}$ and $\mathbf{V}$. We can thus conclude that the L_1-norm LRMF problem is intrinsically equivalent to the MLE under the Laplace distributed noise.

As shown in Fig. 1, the Laplace distribution has a larger PDF value than the Gaussian distribution at the tail part, and thus it is known as a *heavy-tailed* distribution. It means that the Laplace distribution can better fit heavy noises and outliers as

compared with the Gaussian distribution in MLE. This is the intrinsic reason that the L_1-norm LRMF methods are always more robust than the L_2-norm ones on real data.

B. Hierarchical Bayesian Model Formulation

We aim to extend the aforementioned MLE framework for L_1-norm LRMF to a Bayesian formulation in this section. This is motivated by the fact that the Bayesian theory calls for the use of the posterior distribution to do predictive inference on unobserved data, and then the overfitting problem can thus always be alleviated [21]. To this aim, we first assume the Laplace distribution on the noise term ϵ_{ij} in (5) as

$$\epsilon_{ij} \sim \text{Laplace}(0, \sqrt{\lambda/2}). \tag{9}$$

However, due to the absolute value factor in its PDF as in (7), the Laplace distribution is inconvenient for posterior inference within the Bayesian framework. A commonly utilized strategy is to reformulate it as Gaussian scale mixtures with exponential distributed prior to the variance, as indicated in [22]. That is

$$\begin{aligned} &p(x|\mu, \sqrt{\lambda/2}) \\ &\quad = \frac{1}{2}\sqrt{\frac{2}{\lambda}}\exp\left(-\sqrt{\frac{2}{\lambda}}|x-\mu|\right) \\ &\quad = \int_0^\infty \frac{1}{\sqrt{2\pi z}}\exp\left(-\frac{(x-\mu)^2}{2z}\right)\frac{1}{\lambda}\exp\left(-\frac{z}{\lambda}\right)dz \\ &\quad = \int_0^\infty \mathcal{N}(x|\mu, z)p(z|\lambda)dz \end{aligned} \tag{10}$$

where $p(z|\lambda) = (1/\lambda)\exp(-(z/\lambda))$ is the PDF of the exponential distribution. Substituting (9) into the above equation, we have

$$p(\epsilon_{ij}|0, \sqrt{\lambda/2}) = \int_0^\infty \mathcal{N}(\epsilon_{ij}|0, z_{ij})p(z_{ij}|\lambda)dz_{ij}. \tag{11}$$

Therefore, we can impose a two-level hierarchical prior, instead of a single-level Laplace prior, on each ϵ_{ij}

$$\epsilon_{ij} \sim \mathcal{N}(0, z_{ij}), \quad z_{ij} \sim \text{Exponential}(\lambda). \tag{12}$$

To get a complete Bayesian model, we also need to introduce prior distributions for $\mathbf{U}$ and $\mathbf{V}$. As the conventional Bayesian methodology, we place two-level hierarchical priors to them: in the first level, the columns of $\mathbf{U}$ and $\mathbf{V}$ are set as Gaussian priors with zero means, and in the second level, Gamma priors are specified on the precision parameters of these Gaussian distributions. The generative model can then be constructed as

$$\begin{aligned} &\mathbf{u}_i \sim \mathcal{N}\big(\mathbf{0}, \tau_u^{-1}\mathbf{I}\big), \quad \mathbf{v}_j \sim \mathcal{N}\big(\mathbf{0}, \tau_v^{-1}\mathbf{I}\big) \\ &\tau_u \sim \Gamma(a_0, b_0), \quad \tau_v \sim \Gamma(c_0, d_0) \end{aligned} \tag{13}$$

where $a_0, b_0, c_0,$ and d_0 are the hyperparameters of the Gamma distributions, and can be easily specified by small values (e.g., 10^{-6}) in a noninformative fashion [21]. By combining (5), (12), and (13), a hierarchical Bayesian model (denoted as Model I) is constructed.

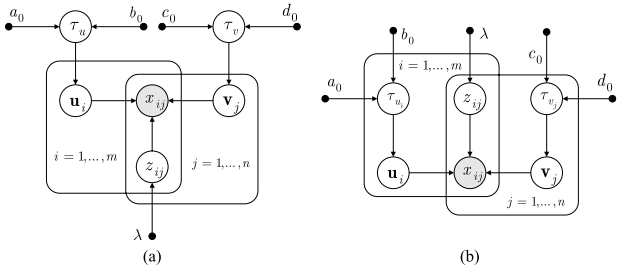


Fig. 2. Graphical models of (a) Model I and (b) Model II.

To make the model more flexible to real heterogeneous cases, we can further vary priors for different $\mathbf{u}_i$s and $\mathbf{v}_j$s as [referred to as Model II together with (5) and (12)]

$$\begin{aligned} &\mathbf{u}_i \sim \mathcal{N}\big(\mathbf{0}, \tau_{u_i}^{-1}\mathbf{I}\big), \quad \mathbf{v}_j \sim \mathcal{N}\big(\mathbf{0}, \tau_{v_j}^{-1}\mathbf{I}\big) \\ &\tau_{u_i} \sim \Gamma(a_0, b_0), \quad \tau_{v_j} \sim \Gamma(c_0, d_0). \end{aligned} \tag{14}$$

For easy visualization, Fig. 2 shows the graphical models for the two Bayesian formulations of L_1-norm LRMF as aforementioned.

The next aim is then to infer the posteriors of all parameters involved in Model I and Model II, given the observation $\mathbf{X}$. The posteriors for the proposed models can be expressed as

$$\begin{aligned} &p(\mathbf{U}, \mathbf{V}, \tau_u, \tau_v, \mathbf{Z}|\mathbf{X}) \\ &\quad \propto p(\mathbf{U}, \mathbf{V}, \tau_u, \tau_v, \mathbf{Z}, \mathbf{X}) \\ &\quad = p(\mathbf{X}|\mathbf{U}, \mathbf{V}, \mathbf{Z})p(\mathbf{U}|\tau_u)p(\mathbf{V}|\tau_v)p(\tau_u)p(\tau_v)p(\mathbf{Z}) \\ &\quad = \prod_{(i,j)\in\Omega} p\big(x_{ij}|\mathbf{u}_i^T\mathbf{v}_j, z_{ij}\big)\prod_{i=1}^m p(\mathbf{u}_i|\tau_u) \\ &\qquad \times \prod_{j=1}^n p(\mathbf{v}_j|\tau_v)p(\tau_u)p(\tau_v)\prod_{(i,j)\in\Omega} p(z_{ij}) \end{aligned} \tag{15}$$

and

$$\begin{aligned} &p(\mathbf{U}, \mathbf{V}, \boldsymbol{\tau}_u, \boldsymbol{\tau}_v, \mathbf{Z}|\mathbf{X}) \\ &\quad \propto p(\mathbf{U}, \mathbf{V}, \boldsymbol{\tau}_u, \boldsymbol{\tau}_v, \mathbf{Z}, \mathbf{X}) \\ &\quad = p(\mathbf{X}|\mathbf{U}, \mathbf{V}, \mathbf{Z})p(\mathbf{U}|\boldsymbol{\tau}_u)p(\mathbf{V}|\boldsymbol{\tau}_v)p(\boldsymbol{\tau}_u)p(\boldsymbol{\tau}_v)p(\mathbf{Z}) \\ &\quad = \prod_{(i,j)\in\Omega} p\big(x_{ij}|\mathbf{u}_i^T\mathbf{v}_j, z_{ij}\big)\prod_{i=1}^m p(\mathbf{u}_i|\tau_{u_i}) \\ &\qquad \times \prod_{j=1}^n p(\mathbf{v}_j|\tau_{v_j})\prod_{i=1}^m p(\tau_{u_i})\prod_{j=1}^n p(\tau_{v_j})\prod_{(i,j)\in\Omega} p(z_{ij}) \end{aligned} \tag{16}$$

respectively, where $\boldsymbol{\tau}_u = (\tau_{u_1}, \ldots, \tau_{u_m})$ and $\boldsymbol{\tau}_v = (\tau_{v_1}, \ldots, \tau_{v_n})$. Before introducing the details of how to infer both posteriors in Section III, we will first give more explanations on the insight of the presented models and discuss their relationship to the previous work.

C. Understanding the Proposed Models From Optimization Perspective

Here we only consider Model I, and a similar discussion can be easily extended to Model II. The negative logarithm of

the full posterior of Model I [i.e., (15)] is

$$\begin{aligned}-\ln p(\mathbf{U},\mathbf{V},\tau_u,\tau_v,\mathbf{Z}|\mathbf{X}) &= \frac{1}{2}\left\{\sum_{(i,j)\in\Omega} z_{ij}^{-1}\left(x_{ij}-\mathbf{u}_i^T\mathbf{v}_j\right)^2+\tau_u\sum_{i=1}^{m}\|\mathbf{u}_i\|_2^2\right. \\ &\quad \left.+\tau_v\sum_{j=1}^{n}\|\mathbf{v}_j\|_2^2+mr\ln\tau_u+nr\ln\tau_v\right\} \\ &\quad +\sum_{(i,j)\in\Omega}\left(\frac{1}{\lambda}z_{ij}-\frac{1}{2}\ln z_{ij}\right)-\ln p(\tau_u|a_0,b_0) \\ &\quad -\ln p(\tau_v|c_0,d_0)+C \end{aligned} \tag{17}$$

where C is a constant independent of the parameters to be estimated. Therefore, by applying the *maximum-a-posterior* (MAP) method to estimate the parameters in the model, we can get a weighted LRMF problem with L_2-regularization, regardless of the terms irrelevant to $\mathbf{u}_i$s and $\mathbf{v}_j$s.

By virtue of this understanding, the specificity of the proposed models is then evident.

1) The original nonsmooth L_1-norm optimization is converted to a smooth weighted L_2-norm problem, which is easier to solve. All weights z_{ij}s can be adaptively fitted to the elements x_{ij}s of $\mathbf{X}$, which makes the proposed models capable of suppressing the adverse effect caused by the noise or outlier elements embedded in $\mathbf{X}$.
2) L_2-regularization terms are imposed on all columns $\mathbf{u}_i$s and $\mathbf{v}_j$s of factorized matrices $\mathbf{U}$ and $\mathbf{V}$, respectively. Such terms are hopeful to bring statistical stability and good generalization capability to the model.
3) The penalty parameters τ_u and τ_v on $\mathbf{u}_i$s and $\mathbf{v}_i$s can be automatically inferred under the proposed Bayesian framework, without any manual operation. This largely avoids the difficulty of parameter tuning, as encountered by many other regularized models.
4) The proposed models seek full posterior distributions of $\mathbf{U}$ and $\mathbf{V}$, and thus the expectation for them can be given instead of point estimation by optimization-based approach. By doing this average, it is expected that the model can be less overfitted to data.

These properties explain why the proposed methods can always achieve robust performance against the noises and outliers and possess good generalization capability on unobserved data in our experiments.

D. Comparison With Previous Probabilistic LRMF-Related Methods

In this section, we discuss the relationship and difference between the proposed and the previous probabilistic models.

Mnih and Salakhutdinov [23] proposed a probabilistic formulation for LRMF and utilized MAP to estimate the factors. They also generalized the formulation to a fully Bayesian model and adopted MCMC to infer the posterior distributions of the factors [24]. The computational cost of this method, however, is very high. With the variational Bayesian approach, Lim and Teh [25] proposed an efficient algorithm for LRMF and also achieved good performance. To further speed up the algorithm, Nakajima *et al.* [26] proposed a variational method with global analytic solution to LRMF. However, this approach can be applied only to the matrix without missing entries. Designed for the collaborative filtering problem, nonparametric Bayesian techniques have also been incorporated into the probabilistic model, e.g., Dirichlet process in [27] and Gamma process in [28]. With appropriate choices of parameter and likelihood distributions, nonnegative matrix factorization has also been addressed within Bayesian framework [29] and applied to speech processing [30]. Most of these models are constructed under the Gaussian noise assumption, which is sensitive to outliers and heavy noises. To address this robustness issue, Lakshminarayanan *et al.* [31] proposed robust models with Gaussian scale mixture noise, alleviating the effect of non-Gaussian noise to a certain extent. However, this noise assumption is less effective for handling heavy outliers as compared with Laplace noise assumption.

Recently, Wang *et al.* [18] proposed a probabilistic model for L_1-norm LRMF, which looks somewhat similar to the proposed Model I. However, this model does not impose prior information on τ_u and τ_v, and let them fixed during the estimation process. Moreover, this model has not inferred the full posterior of the factors while employed the CEM algorithm to implement point estimation, which in fact solves the following optimization problem in each M-step:

$$\min_{\mathbf{U},\mathbf{V}}\sum_{(i,j)\in\Omega} z_{ij}^{-1}\left(x_{ij}-\mathbf{u}_i^T\mathbf{v}_j\right)^2+\tau_u\sum_{i=1}^{m}\|\mathbf{u}_i\|_2^2+\tau_v\sum_{j=1}^{n}\|\mathbf{v}_j\|_2^2. \tag{18}$$

As compared with our framework, this approach has not made use of the fully Bayesian inference. For example, the parameters τ_u and τ_v in [18] should be carefully preset since their function is to balance the approximation of the parameters to the observed data and the complexity of the parameters themselves. If they are set too small, $\mathbf{U}$ and $\mathbf{V}$ incline to overfit the corrupted data, and if they are set too large, the two factors will not approximate the true data well. In contrast, in our models, these parameters can be automatically learned from data under the Bayesian framework. This not only automates the parameter tuning in the proposed models but also benefits the adaptability of our models to data.

Very recently, Wang and Yeung [19] generalized this probabilistic model to the Bayesian framework and adopted MCMC for inference, which is similar to the one used in [24] except for the noise modeling to achieve robustness. Compared with our formulation, which is also Bayesian, this method, however, is computational inefficient since their utilized sampling method is generally time cumbersome. Besides, this method has not addressed the missing data issue, which often needs to be dealt with in practice, such as rigid and nonrigid SFM.

Bayesian approach has also been applied to principal component analysis (PCA) [32] and robust PCA (RPCA) [33] problems, which are closely related to LRMF since they also work on low-rank matrices. Bishop [34] applied

variational Bayesian method to PCA so that the number of the principal components can be automatically determined. Gao [35] proposed a robust version of PCA using a similar Gaussian scale mixture likelihood as our models. Luttinen *et al.* [36] extended this approach to deal with missing entries under student-t likelihood. These methods are designed to learn a linear transformation to map data from the original space to a low-dimensional space. This can be regarded as an indirect way to do matrix factorization, but always less competitive than conventional LRMF methods, which explicitly formulate the factorization as their goal, in the sense of reconstruction quality. Ding *et al.* [37] and Babacan *et al.* [38] formulated RPCA as Bayesian models and used MCMC and variational methods to infer the posteriors, respectively.[1] Nakajima *et al.* [39] further generalized this framework to model more complex structure of the additive noise matrices. Very recently, Zhao *et al.* [41] reformulated this problem by modeling noise as a mixture of Gaussian to adapt its availability under different kinds of noises. However, these methods are designed on nonmissing data, and thus they are inappropriate to be employed in missing data cases.

III. Approximate Bayesian Inference

As is widely known, the exact Bayesian inference for the posterior distributions such as (15) and (16) is intractable, since $p(\mathbf{X})$ cannot be analytically computed by marginalizing all of the other variables, and thus approximation methods are often used. Although sampling methods provide optimal approximation in theory, the computational complexity is always too high due to the large number of burn-in iterations, and the convergence generally cannot be easily monitored. Therefore, we adopt the well-known variational Bayesian method [21] to approximate the full posterior distributions involved in our models.

Before presenting the inference procedure for (15) and (16), we first briefly introduce the general framework of the mean-field variational technique.

A. General Framework of Variational Inference

Variational method is one of the most commonly utilized tools for approximating the intractable posterior. The method is constructed by minimizing the *Kullback-Leibler (KL) divergence* between an approximation distribution $q(\mathbf{x})$ and the true posterior $p(\mathbf{x}|\mathcal{D})$ through the following variational optimization model:

$$\min_{q\in\mathcal{C}} \mathrm{KL}(q\|p) = -\int q(\mathbf{x}) \ln\left\{\frac{p(\mathbf{x}|\mathcal{D})}{q(\mathbf{x})}\right\} d\mathbf{x} \tag{19}$$

where $\mathrm{KL}(q\|p)$ denotes the KL divergence between $q(\mathbf{x})$ and $p(\mathbf{x}|\mathcal{D})$, and $\mathcal{C}$ denotes the set of PDFs with certain restrictions to make the minimization tractable. Here q is generally set by partitioning the elements of $\mathbf{x}$ into disjoint groups $\{\mathbf{x}_i\}$, and then assuming that it can be factorized as $q(\mathbf{x}) = \prod_i q_i(\mathbf{x}_i)$. Under these assumptions, the closed-form solution for each group $\mathbf{x}_j$, with the others fixed, can be attained by

$$q_j^*(\mathbf{x}_j) = \frac{\exp(\mathbb{E}_{i\neq j}[\ln p(\mathbf{x},\mathcal{D})])}{\int \exp(\mathbb{E}_{i\neq j}[\ln p(\mathbf{x},\mathcal{D})])d\mathbf{x}_j} \tag{20}$$

where $p(\mathbf{x},\mathcal{D})$ is the joint distribution of parameters $\mathbf{x}$ and the observations $\mathcal{D}$, and $\mathbb{E}_{i\neq j}[\cdot]$ denotes the expectation with respect to $\mathbf{x}_i$s except $\mathbf{x}_j$. The solution to (19) can then be approached through alternatively optimizing each $q_j(\mathbf{x}_j)$ by (20).

Utilizing the general results above, the closed-form variational inference schemes for Model I and Model II of L_1-norm LRMF can then be derived.

B. Variational Inference for Model I

1) Estimation of $\boldsymbol{U}$ and τ_u: Based on the posterior distributions for $\mathbf{U}, \mathbf{V}, \tau_u, \tau_v$, and $\mathbf{Z}$, as shown in (15), its approximation $q(\mathbf{U}, \mathbf{V}, \tau_u, \tau_v, \mathbf{Z})$ can be assumed to have the following factorization form:

$$q(\mathbf{U},\mathbf{V},\tau_u,\tau_v,\mathbf{Z}) = \prod_{i=1}^{m} q(\mathbf{u}_i) \prod_{j=1}^{n} q(\mathbf{v}_j) q(\tau_u) q(\tau_v) \prod_{ij} q(z_{ij}). \tag{21}$$

Then by applying the general result (20) to (15) and (21) with respect to $\mathbf{u}_i$ and τ_u, respectively, we can get the following update equations:

$$q(\mathbf{u}_i) = \mathcal{N}\left(\mathbf{u}_i|\boldsymbol{\mu}_{u_i}, \boldsymbol{\Lambda}_{u_i}^{-1}\right) \tag{22}$$

$$q(\tau_u) = \Gamma(\tau_u|a,b) \tag{23}$$

with parameters

$$\boldsymbol{\Lambda}_{u_i} = \mathbb{E}[\tau_u]\mathbf{I} + \sum_{j=1}^{n} w_{ij}\mathbb{E}\left[z_{ij}^{-1}\right]\mathbb{E}\left[\mathbf{v}_j\mathbf{v}_j^T\right]$$

$$\boldsymbol{\mu}_{u_i} = \boldsymbol{\Lambda}_{u_i}^{-1} \sum_{j=1}^{n} w_{ij}x_{ij}\mathbb{E}\left[z_{ij}^{-1}\right]\mathbb{E}[\mathbf{v}_j]$$

$$a = a_0 + \frac{1}{2}rm, \quad b = b_0 + \frac{1}{2}\sum_{i=1}^{m}\mathbb{E}\left[\mathbf{u}_i^T\mathbf{u}_i\right].$$

The expectations included in the above equations (and the following sections) can be calculated with respect to the current parameter values of the variational distributions. The details are presented in Appendix.

2) Estimation of $\boldsymbol{V}$ and τ_v: Update equations for $\mathbf{v}_j$ and τ_v can be derived in a similar way as follows:

$$q(\mathbf{v}_j) = \mathcal{N}\left(\mathbf{v}_j|\boldsymbol{\mu}_{v_j}, \boldsymbol{\Lambda}_{v_j}^{-1}\right) \tag{24}$$

$$q(\tau_v) = \Gamma(\tau_v|c,d) \tag{25}$$

where

$$\boldsymbol{\Lambda}_{v_j} = \mathbb{E}[\tau_v]\mathbf{I} + \sum_{i=1}^{m} w_{ij}\mathbb{E}\left[z_{ij}^{-1}\right]\mathbb{E}\left[\mathbf{u}_i\mathbf{u}_i^T\right]$$

$$\boldsymbol{\mu}_{v_j} = \boldsymbol{\Lambda}_{v_j}^{-1} \sum_{j=1}^{m} w_{ij}x_{ij}\mathbb{E}\left[z_{ij}^{-1}\right]\mathbb{E}[\mathbf{u}_i]$$

$$c = c_0 + \frac{1}{2}rn, \quad d = d_0 + \frac{1}{2}\sum_{j=1}^{n}\mathbb{E}\left[\mathbf{v}_j^T\mathbf{v}_j\right].$$

[1]Note that Babacan *et al.* also proposed a method for low-rank matrix completion, which can handle missing data, in their paper. However, it does not deal with the robustness problem induced by outliers.

3) Estimation of z_{ij}s: To infer $q(z_{ij})$, we can substitute z_{ij} with its inverse $z_{ij} = (1/y_{ij})$ to get a transformed distribution $q(y_{ij})$, and then infer it with the general result (20) by the following inverse Gaussian distribution:

$$q(y_{ij}) = \mathcal{IG}(y_{ij}|\mu_{y_{ij}}, \lambda_y) \tag{26}$$

where

$$\mu_{y_{ij}} = \sqrt{\frac{2}{\lambda \mathbb{E}\left[\left(x_{ij} - \mathbf{u}_i^T \mathbf{v}_j\right)^2\right]}}, \qquad \lambda_y = \frac{2}{\lambda}.$$

4) Update of λ: As can be observed from (12), the parameter λ is directly related to the noise variance parameter z_{ij}. Therefore, it should be adjusted carefully to obtain reasonable results. Although it can be preset and fixed during the whole inference process, a better way is to make it adaptively tuned based on the noise information extracted from data. *Empirical Bayes* [21] provides an off-the-shelf tool to this aim, by updating it through

$$\lambda = \frac{\sum_{ij} w_{ij} \mathbb{E}(z_{ij})}{\sum_{ij} w_{ij}}. \tag{27}$$

Using this elaborate tool, λ can be properly adapted to real data variance.

C. Variational Inference for Model II

For Model II, the update equations are almost the same as those for Model I, while minor differences exist in the equations for $\mathbf{u}_i$, $\mathbf{v}_j$, τ_{u_i}, and τ_{v_j} due to the different priors imposed on different $\mathbf{u}_i$s and $\mathbf{v}_j$s. The updating equations are listed as follows:

$$q(\mathbf{u}_i) = \mathcal{N}\left(\mathbf{u}_i|\boldsymbol{\mu}_{u_i}, \boldsymbol{\Lambda}_{u_i}^{-1}\right) \tag{28}$$

$$q(\tau_{u_i}) = \Gamma(\tau_{u_i}|a, b_i) \tag{29}$$

$$q(\mathbf{v}_j) = \mathcal{N}\left(\mathbf{v}_j|\boldsymbol{\mu}_{v_j}, \boldsymbol{\Lambda}_{v_j}^{-1}\right) \tag{30}$$

$$q(\tau_{v_j}) = \Gamma(\tau_{v_j}|c, d_j) \tag{31}$$

where

$$\boldsymbol{\Lambda}_{u_i} = \mathbb{E}[\tau_{u_i}]\mathbf{I} + \sum_{j=1}^{n} w_{ij} \mathbb{E}\left[z_{ij}^{-1}\right] \mathbb{E}\left[\mathbf{v}_j \mathbf{v}_j^T\right]$$

$$\boldsymbol{\mu}_{u_i} = \boldsymbol{\Lambda}_{u_i}^{-1} \sum_{j=1}^{n} w_{ij} x_{ij} \mathbb{E}\left[z_{ij}^{-1}\right] \mathbb{E}[\mathbf{v}_j]$$

$$a = a_0 + \frac{1}{2}r, \quad b_i = b_0 + \frac{1}{2}\mathbb{E}\left[\mathbf{u}_i^T \mathbf{u}_i\right]$$

$$\boldsymbol{\Lambda}_{v_j} = \mathbb{E}[\tau_{v_j}]\mathbf{I} + \sum_{i=1}^{m} w_{ij} \mathbb{E}\left[z_{ij}^{-1}\right] \mathbb{E}\left[\mathbf{u}_i \mathbf{u}_i^T\right]$$

$$\boldsymbol{\mu}_{v_j} = \boldsymbol{\Lambda}_{v_j}^{-1} \sum_{j=1}^{m} w_{ij} x_{ij} \mathbb{E}\left[z_{ij}^{-1}\right] \mathbb{E}[\mathbf{u}_i]$$

$$c = c_0 + \frac{1}{2}r, \quad d = d_0 + \frac{1}{2}\mathbb{E}\left[\mathbf{v}_j^T \mathbf{v}_j\right].$$

D. Variational Lower Bound

In this section, we give a theoretical evaluation for variational lower bounds for both of the models, which are useful to assess the convergence behavior of the inference procedure. The general formulation for variational lower bound is

$$\mathcal{L}(q) = \int q(\mathbf{x}) \ln \left\{ \frac{p(\mathbf{x}, \mathcal{D})}{q(\mathbf{x})} \right\} d\mathbf{x}. \tag{32}$$

Applying the above equation to the complete data-likelihood (15) and variational distribution (21), we can get the variational lower bound for Model I as

$$\begin{aligned}
\mathcal{L}_{\mathrm{I}}(q) = & \sum_{(i,j)\in\Omega} \mathbb{E}\left[p\left(x_{ij}|\mathbf{u}_i^T \mathbf{v}_j, z_{ij}\right)\right] + \sum_{i=1}^{m} \mathbb{E}[\ln p(\mathbf{u}_i|\tau_u)] \\
& + \sum_{j=1}^{n} \mathbb{E}[\ln p(\mathbf{v}_j|\tau_v)] + \mathbb{E}[\ln p(\tau_u)] + \mathbb{E}[\ln p(\tau_v)] \\
& + \sum_{(i,j)\in\Omega} \mathbb{E}[\ln p(z_{ij})] - \sum_{i=1}^{m} \mathbb{E}[\ln q(\mathbf{u}_i)] \\
& - \sum_{j=1}^{n} \mathbb{E}[\ln q(\mathbf{v}_j)] - \mathbb{E}[\ln q(\tau_u)] - \mathbb{E}[\ln q(\tau_v)] \\
& - \sum_{(i,j)\in\Omega} \mathbb{E}[\ln q(z_{ij})]
\end{aligned} \tag{33}$$

where the expectations are taken with respect to current variational distribution q and have the following forms:

$$\begin{aligned}
\mathbb{E}\left[p\left(x_{ij}|\mathbf{u}_i^T \mathbf{v}_j, z_{ij}\right)\right] = & \frac{1}{2}\Big\{\mathbb{E}\left[\ln z_{ij}^{-1}\right] - \mathbb{E}\left[z_{ij}^{-1}\right] \\
& \times \mathbb{E}\left[\left(x_{ij} - \mathbf{u}_i^T \mathbf{v}_j\right)^2\right] - \ln 2\pi \Big\} \\
\mathbb{E}[\ln p(\mathbf{u}_i|\tau_u)] = & \frac{1}{2}\left\{r\mathbb{E}[\ln \tau_u] - \mathbb{E}[\tau_u]\mathbb{E}\left[\mathbf{u}_i^T \mathbf{u}_i\right] - r \ln 2\pi\right\} \\
\mathbb{E}[\ln p(\mathbf{v}_j|\tau_v)] = & \frac{1}{2}\left\{r\mathbb{E}[\ln \tau_v] - \mathbb{E}[\tau_v]\mathbb{E}\left[\mathbf{v}_j^T \mathbf{v}_j\right] - r \ln 2\pi\right\} \\
\mathbb{E}[\ln p(\tau_u)] = & a_0 \ln b_0 - \ln \Gamma(a_0) + (a_0 - 1)\mathbb{E}[\ln \tau_u] \\
& - b_0 \mathbb{E}[\tau_u] \\
\mathbb{E}[\ln p(\tau_v)] = & c_0 \ln d_0 - \ln \Gamma(c_0) + (c_0 - 1)\mathbb{E}[\ln \tau_v] \\
& - d_0 \mathbb{E}[\tau_v] \\
\mathbb{E}[\ln p(z_{ij})] = & -\ln \lambda - \frac{1}{\lambda}\mathbb{E}[z_{ij}] \\
\mathbb{E}[\ln q(\mathbf{u}_i)] = & \frac{1}{2}\Big\{\ln |\boldsymbol{\Lambda}_{u_i}| - r \ln 2\pi \\
& - \mathrm{Tr}\left(\boldsymbol{\Lambda}_{u_i} \mathbb{E}\left[\mathbf{u}_i \mathbf{u}_i^T\right]\right)\Big\} \\
\mathbb{E}[\ln q(\mathbf{v}_j)] = & \frac{1}{2}\Big\{\ln |\boldsymbol{\Lambda}_{v_j}| - r \ln 2\pi \\
& - \mathrm{Tr}\left(\boldsymbol{\Lambda}_{v_j} \mathbb{E}\left[\mathbf{v}_j \mathbf{v}_j^T\right]\right)\Big\} \\
\mathbb{E}[\ln q(\tau_u)] = & a \ln b - \ln \Gamma(a) + (a - 1)\mathbb{E}[\ln \tau_u] \\
& - b \mathbb{E}[\tau_u] \\
\mathbb{E}[\ln q(\tau_v)] = & c \ln d - \ln \Gamma(c) + (c - 1)\mathbb{E}[\ln \tau_v] \\
& - d \mathbb{E}[\tau_v] \\
\mathbb{E}[\ln q(z_{ij})] = & \frac{1}{2}\left\{\ln \lambda_y - \ln 2\pi - 3\mathbb{E}\left[\ln z_{ij}^{-1}\right]\right\} \\
& - \frac{\lambda_y}{2\mu_{y_{ij}}^2}\left\{\mathbb{E}\left[z_{ij}^{-1}\right] - 2\mu_{y_{ij}} + \mu_{y_{ij}}^2 \mathbb{E}[z_{ij}]\right\}.
\end{aligned}$$

All the expectations involved in the above equations, except $\mathbb{E}[\ln z_{ij}^{-1}]$, can be easily calculated, as aforementioned. Due to the more complex form of the variational distribution over z_{ij}, $\mathbb{E}[\ln z_{ij}^{-1}]$ cannot be calculated analytically. However, by doing some algebra, the expression for $\mathcal{L}_1(q)$ can be derived as

$$\begin{aligned}\mathcal{L}_1(q) &= 2\sum_{(i,j)\in\Omega}\mathbb{E}\big[\ln z_{ij}^{-1}\big]+\{\text{terms not including } \mathbb{E}\big[\ln z_{ij}^{-1}\big]\}\\ &\le 2\sum_{(i,j)\in\Omega}\ln\mathbb{E}\big[z_{ij}^{-1}\big]+\{\text{terms not including } \mathbb{E}\big[\ln z_{ij}^{-1}\big]\}\\ &\triangleq \mathcal{L}_{\text{App-I}}(q)\end{aligned}$$

where the inequality follows from the Jensen's inequality. Therefore, we can use $\mathcal{L}_{\text{App-I}}(q)$ as an approximation to the true lower bound $\mathcal{L}_1(q)$. A similar approximate lower bound $\mathcal{L}_{\text{App-II}}(q)$ for Model II can also be obtained, and we omit the details due to page limitation.

E. Computational Complexity

Now we give a brief discussion on the computational complexity of the proposed variational Bayesian methods for L_1-norm LRMF. It is easy to observe that only simple computations are involved in the variational inference of the parameters, except that inferring each of $\mathbf{u}_i$s or $\mathbf{v}_j$s needs to compute an inverse of a $r\times r$ matrix, leading to $\mathcal{O}((m+n)r^3)$ costs in total, where m and n are the number of matrix columns and rows, and r is the rank parameter. Altogether, each inference iteration needs $\mathcal{O}((m+n)r^3+mnr^2)$ costs, and the total complexity of the methods is thus $\mathcal{O}(T((m+n)r^3+mnr^2))$, where T is the upper bound of iterations. Since, in general, it holds that $r\ll\min(m,n)$, the proposed algorithm is always well suited for solving large-scale L_1-norm LRMF problems.

IV. Experiments

In this section, we evaluate the effectiveness of the proposed methods, denoted as VBMFL_1-I and VBMFL_1-II (for Model I and Model II, respectively) by experiments. The competing methods are the state-of-the-art methods for L_1-norm LRMF, including ALP, AQP [11], L_1-Wiberg [12], RegL_1-ALM [13], PRMF [18], CWM [20], and RBMF [31]. We use the publicly available toolboxes for L_1-Wiberg,[2] RegL_1-ALM,[3] PRMF,[4] and CWM,[5] author-provided implementation for RBMF, and write the codes for ALP and AQP using the LP and QP solvers in MATLAB Optimization Toolbox. All the programs were run under the MATLAB 8.0(R2012b) platform on the personal computer with Intel(R) Core(TM) i5-3230M at 2.60 GHz (CPU), 8 GB (memory), and Windows 8.1 64-bit (OS). For each experiment, we use random but same initializations for all the competing methods.

A. Correctness Verification for the Proposed Methods

Since our models are constructed based on the Laplace noise assumption, it is necessary to evaluate the behavior of the proposed methods under this kind of noise distribution.

[2]http://cs.adelaide.edu.au/~anders/code/cvpr2010.html.
[3]http://sites.google.com/site/yinqiangzheng/.
[4]http://winsty.net/prmf.html.
[5]http://gr.xjtu.edu.cn/web/dymeng/.

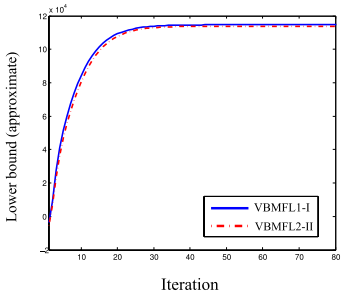


Fig. 3. Approximate lower bound curves for VBMFL_1-I and VBMFL_1-II on synthetic data introduced in Section IV-A.

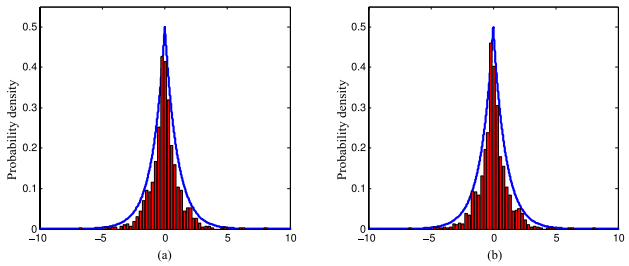


Fig. 4. True probability density (shown with blue curve) of the noise and its estimation (shown with red histogram) by the proposed methods on synthetic data introduced in Section IV-A. (a) VBMFL_1-I. (b) VBMFL_1-II.

To this aim, we designed a synthetic data set generated as follows: two matrices $\mathbf{U}\in\mathbb{R}^{4\times 40}$ and $\mathbf{V}\in\mathbb{R}^{4\times 40}$ were first randomly generated with each entry drawn from the Gaussian distribution $\mathcal{N}(0,5)$, resulting in the ground truth rank-4 matrix $\mathbf{M}_0=\mathbf{U}^T\mathbf{V}$. Then, 30% of the elements were randomly selected and specified as missing data, and the rest elements were mixed with noise based on (9) with $\lambda=2$. We implemented both of the proposed VBMFL_1-I and VBMFL_1-II methods on this data set for verification.

First, we plot the curves of the variational lower bounds for VBMFL_1-I and VBMFL_1-II, as derived in Section III-D, in Fig. 3, to see their convergence property. It is easy to observe that both the lower bound curves are monotonically increasing during the iterative process and quickly converge to a stable status within 40 iterations, which is consistent with the basic principle of variational inference.

Next, we show the ability of the proposed methods in discovering the true structure of the embedded noise. We subtracted the estimated matrices returned by the proposed methods from the observed matrix, and thus obtained the estimated noise. Then, we compared the empirical density of the estimated noise and the density of the true noise in Fig. 4. As can be observed from this figure, the shape of the empirical density is very close to the true Laplace density, implying that our methods are able to accurately recover the noise structure. Besides, the estimated values for λ are 2.0825 and 2.0851, by VBMFL_1-I and VBMFL_1-II, respectively, which are very close to its underlying true value.

We then show in Fig. 5 the pseudocolor images generated by the ground truth data and noise matrices and their estimations by VBMFL_1-I and VBMFL_1-II. As can be observed, both of the estimated data and noise matrices are very close to the

TABLE I

PERFORMANCE COMPARISON OF SEVEN EXISTING L_1-NORM LRMF METHODS AND THE PROPOSED METHODS ON SMALL-SCALE SYNTHETIC EXPERIMENTS. THE RESULTS ARE AVERAGED OVER 100 RUNS, AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

Method	Missing 20% Outlier 20%			Missing 30% Outlier 20%			Missing 40% Outlier 20%		
	MAE_{obv}	RMSE_{grd}	Time(s)	MAE_{obv}	RMSE_{grd}	Time(s)	MAE_{obv}	RMSE_{grd}	Time(s)
ALP	0.6606	0.8420	1.58	0.7576	1.6385	1.33	0.7992	2.3809	1.23
AQP	0.7883	3.8961	151.11	0.7948	33.4051	199.39	0.7557	64.4260	283.11
L_1-Wiberg	0.6308	3.1796	146.53	0.6747	17.8663	185.42	0.6757	21.5975	170.45
$RegL_1$-ALM	**0.6181**	0.7398	0.096	**0.6492**	1.7204	0.089	**0.6367**	2.8131	0.096
PRMF	0.7581	0.7886	0.102	0.8568	1.1174	0.096	0.9384	1.4208	0.104
CWM	0.7436	1.0417	**0.020**	0.8368	1.5441	**0.017**	0.9064	2.1341	**0.016**
RBMF	1.0252	1.0657	1.49	1.1914	1.2636	1.48	1.3616	1.4416	1.86
VBMFL_1-I	0.7114	0.5945	0.117	0.8466	0.8748	0.130	1.1317	1.2130	0.308
VBMFL_1-II	0.7079	**0.5902**	0.127	0.8394	**0.8695**	0.147	1.1163	**1.2012**	0.338
Method	Missing 20% Outlier 30%			Missing 30% Outlier 30%			Missing 40% Outlier 30%		
	MAE_{obv}	RMSE_{grd}	Time(s)	MAE_{obv}	RMSE_{grd}	Time(s)	MAE_{obv}	RMSE_{grd}	Time(s)
ALP	0.9090	1.4871	1.37	0.9662	2.0989	1.40	0.9951	2.8653	1.16
AQP	0.9957	5.5296	141.43	0.9795	36.3916	288.07	0.9147	54.5613	265.27
L_1-Wiberg	0.8637	13.6484	161.77	0.8735	16.8508	170.12	0.8549	78.3812	188.25
$RegL_1$-ALM	**0.8401**	1.5426	0.089	**0.8356**	2.4357	0.096	**0.8004**	3.5035	0.091
PRMF	1.0474	1.1895	0.093	1.1097	1.4250	0.106	1.1685	1.7355	0.100
CWM	0.9891	1.5054	**0.018**	1.0576	1.9382	**0.018**	1.1036	2.5626	**0.015**
RBMF	1.3634	1.2682	1.27	1.5608	1.4527	1.61	1.7946	1.6537	1.69
VBMFL_1-I	1.0566	0.9309	0.131	1.3257	1.2427	0.31	1.7648	1.6204	0.463
VBMFL_1-II	1.0523	**0.9284**	0.152	1.3175	**1.2370**	0.35	1.7477	**1.6087**	0.517

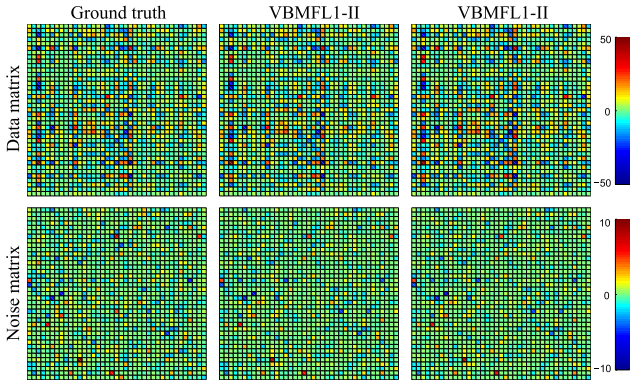


Fig. 5. Illustration of the ground truth data and noise matrices and those estimated by the proposed VBMFL_1-I and VBMFL_1-II methods on synthetic data introduced in Section IV-A.

ground truth, indicating that the proposed methods successfully deliver the true data information from the corrupted observations.

As can be observed from this simple simulation, the effectiveness of the proposed methods is substantiated, especially in terms of its ability in estimating the true noise structure, which naturally leads to its faithful reconstruction of the true data from the noisy observations.

B. Synthetic Experiments

We then evaluate the performance of the proposed methods on synthetic data with both small and large scales.

1) Small-Scale Data: The data were generated as follows: two matrices $\mathbf{U} \in \mathbb{R}^{4\times 20}$ and $\mathbf{V} \in \mathbb{R}^{4\times 30}$ were first randomly generated with each entry drawn from the Gaussian distribution $\mathcal{N}(0, 1)$, resulting in the ground truth rank-4 matrix $\mathbf{M}_0 = \mathbf{U}^T\mathbf{V}$, and a certain amount of the missing entries and outliers were then randomly specified on $\mathbf{M}_0$ to constitute the observation matrix $\mathbf{M}$. The outliers were then independently generated from the uniform distribution on $[-5, 5]$, which are extremely heavy compared with the clean data, following the settings in [12], [13], and [20]. To make a comprehensive comparison, we varied the missing data ratios and the outlier ratios to obtain a series of synthetic matrices.

Six series of synthetic matrices were generated by varying the outlier ratios and missing data ratios as $(20\%, 20\%)$, $(20\%, 30\%)$, $(20\%, 40\%)$, $(30\%, 20\%)$, $(30\%, 30\%)$, and $(30\%, 40\%)$, respectively. In each case, 100 matrices were generated, and the average performance of each competing method on these matrices, in terms of computational accuracy and time, was summarized in Table I. The accuracy of the competing method was measured by the root-mean-square error (RMSE), which is defined by

$$\text{RMSE}_{\text{grd}} = \sqrt{\frac{1}{mn}\|\mathbf{M}_0 - \hat{\mathbf{U}}^T\hat{\mathbf{V}}\|_2^2}$$

where $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are the factors estimated by a competing method, and $\mathbf{M}_0$ is the ground truth matrix. We also include the mean absolute error (MAE) between the corrupted matrix and the reconstructed one

$$\text{MAE}_{\text{obv}} = \frac{1}{\sum_{ij} w_{ij}}\|\mathbf{W} \odot (\mathbf{M} - \hat{\mathbf{U}}^T\hat{\mathbf{V}})\|_1$$

which actually corresponds to the objective function of the L_1-norm LRMF problem (4) in Table I. It should be noted that RMSE_{grd} is actually what we focus on, since it measures the difference between the estimated matrix and the true but unobserved data matrix, while MAE_{obv} only measures the closeness of the estimation to the observed matrix, which is embedded with noises/outliers.

It can be observed from Table I that although not achieving the lowest MAE_{obv}, the proposed VBMFL_1-I and VBMFL_1-II

TABLE II
PERFORMANCE COMPARISON OF FOUR EXISTING ROBUST LRMF METHODS AND THE PROPOSED METHODS ON LARGE-SCALE SYNTHETIC EXPERIMENTS. THE RESULTS ARE AVERAGED OVER 100 RUNS

Method	RMSE_{grd}	Time(s)
$RegL_1$-ALM	3.5808	**168.10**
PRMF	3.3320	172.56
CWM	3.4195	1865
RBMF	3.0061	3370
VBMFL_1-I	**2.9135**	802.96
VBMFL_1-II	2.9159	858.37

methods can always perform the best in terms of RMSE_{grd} while they consume reasonable time as compared with the other methods. We also observe that several methods, e.g., AQP and L_1-Wiberg, though they achieve a lower MAE_{obv}, yield very large RMSE_{grd}s, implying that they fail to recover the ground truth matrix in this scenario. This means that when specifically optimizing the objective function of the L_1-norm LRMF problem (4), these methods tend to overfit the corrupted observations, while they may not well handle the generalization capability of the results on the unobserved elements of the original uncorrupted matrix. As a comparison, the proposed methods are constructed within the Bayesian framework, under which the overfitting problem tends to be alleviated [21]. In addition, based on the relation to the optimization-based model, as discussed in Section II-C, the proposed methods tend to be more flexible to the data corruptions due to the automatically tuned weights z_{ij}s on matrix elements and generalize better attributed to its L_2-regularization penalties on $\mathbf{u}_i$s and $\mathbf{v}_i$s. It should be noted that RBMF, which is also based on the Baysian framework, can achieve better RMSE_{grd}s compared with most of the optimization-based methods, but worse compared with our methods. This means that, on the one hand, RBMF can alleviate the overfitting issue to a certain extent; on the other hand, the Laplace noise assumption adopted by our methods is more effective than that of RBMF in the cases of existing heavy outliers.

2) Large-Scale Data: The data were generated as follows: a ground truth matrix $\mathbf{M}_0 \in \mathbb{R}^{1000\times 1000}$ was first randomly generated with each entry drawn from the uniform distribution on $[0, 10]$. Then, 30% of the entries were specified as missing data, and another 30% entries were added to outliers i.i.d. generated from the uniform distribution on $[-50, 50]$. Note that the ground truth matrix here is of full rank, intending to test a method's ability of low-rank approximation to a full rank matrix. This setting is similar to performing SVD or PCA to reduce the dimensionality of data.

All competing methods except ALP, AQP, and L_1-Wiberg, suffering from the out of memory error due to their comparatively low scalability, were implemented. Table II summarizes the average performance of each method with rank-30 approximation, in terms of RMSE_{grd} and computational time, over 100 runs. It is easy to observe that the proposed methods attain better reconstruction accuracy than other utilized methods. It can also be observed that the proposed methods have a comparable computational speed with existing methods. Considering their better accuracy, it is rational to say that they are efficient.

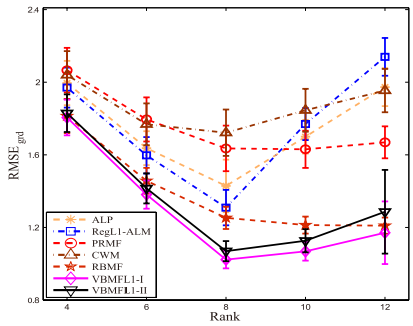


Fig. 6. Performance comparison of five existing robust LRMF methods and the proposed methods with rank parameter varying from 4 to 12.

3) Sensitivity to the Rank Parameter: Now we briefly analyze the sensitivity of the proposed methods to the rank parameter. We generated data in a similar way as before: two matrices $\mathbf{U} \in \mathbb{R}^{8\times 60}$ and $\mathbf{V} \in \mathbb{R}^{8\times 60}$ were first randomly generated with each entry drawn from the Gaussian distribution $\mathcal{N}(0, 1)$, resulting in the ground truth rank-8 matrix $\mathbf{M}_0 = \mathbf{U}^T\mathbf{V}$, and then 30% of the entries were randomly specified as missing data and another 30% were added outliers generated from the uniform distribution on $[-5, 5]$. The proposed methods, together with five existing methods, were run on this data with rank parameter varying from 4 to 12. Since the ground truth rank is eight, we can observe the behavior of each method when the rank was incorrectly set. The results in terms of RMSE_{grd} averaged over 100 runs were summarized in Fig. 6. We can observe that, most competing methods' performances degenerate when the rank is incorrectly specified. Our methods, however, can, on the one hand, achieve the best performance under the ground truth rank, and the one other hand, perform at least as robust as other competing methods under the incorrect rank.

C. Structure From Motion

SFM aims estimating 3-D structure from a sequence of 2-D images, which may be coupled with local motion information [1]. There are two types of SFM problems, namely, rigid and nonrigid SFM, both of which can be formulated as LRMF problems.

1) Rigid Structure From Motion: Rigid SFM is shown to be an intrinsic rank-3 matrix factorization problem after registering the image origin to the centroid of feature points in every frame [1]. However, the method cannot be directly used in real situations due to the missing components and noises embedded in each frame data. The L_1-norm LRMF with rank 4 is thus considered as a tool for this task [11]. We employ the *Dinosaur* sequence,[6] which contains 319 feature points tracked over 36 views, of the sequence, corresponding to a data matrix $\mathbf{M}_0$ of size 72×319 and with 76.92% missing data. To verify the robustness of the competing methods, 30% of the observed entries were randomly chosen and added to outliers, generated from the uniform distribution on $[-50, 50]$. We have run all of the competing methods on the problem,

[6]http://www.robots.ox.ac.uk/~vgg/data1.html.

TABLE III
PERFORMANCE COMPARISON OF FIVE EXISTING L_1-NORM LRMF METHODS AND THE PROPOSED METHODS ON RIGID SFM EXPERIMENTS. THE RESULTS ARE AVERAGED OVER 100 RUNS

Method	RMSE_{obv}	Time(s)
ALP	10.9707	378.29
$RegL_1$-ALM	6.1502	42.55
PRMF	17.8078	0.89
CWM	17.1055	**0.50**
RBMF	8.5091	53.12
VBMFL_1-I	**5.9778**	2.56
VBMFL_1-II	6.0922	2.85

TABLE IV
PERFORMANCE COMPARISON OF FIVE EXISTING L_1-NORM LRMF METHODS AND THE PROPOSED METHODS ON SFM EXPERIMENTS WITH TRAINING/TESTING SPLIT. THE RESULTS ARE AVERAGED OVER 100 RUNS

Method	RMSE_{test}	
	Rigid SFM	Nonrigid SFM
ALP	34.4444	17.2233
$RegL_1$-ALM	24.0307	5.1094
PRMF	29.8987	2.4855
CWM	31.5023	1.7105
RBMF	14.4753	4.2768
VBMFL_1-I	13.7926	0.8639
VBMFL_1-II	**12.2314**	**0.8520**

while the AQP and L_1-Wiberg failed to be implemented due to the out of memory problem. The computational accuracy and computation times, averaged over 100 runs, for all competing methods are reported in Table III, where the accuracy measure RMSE_{obv} is defined by

$$\text{RMSE}_{\text{obv}} = \sqrt{\frac{1}{\sum_{ij} w_{ij}} \|\mathbf{W} \odot (\mathbf{M}_0 - \hat{\mathbf{U}}^T \hat{\mathbf{V}})\|_2^2}.$$

It is easy to observe that the proposed methods attain more accurate results than the other utilized methods. Considering the computational time, the proposed methods run slower than PRMF and CWM while much faster than the other three.

Note that RMSE_{obv} can only measure the accuracy on the observed data, while we also care about the accuracy on the unobserved data. Therefore, we further split the observed data into two parts, one for training and one for testing, to see how well a method can approximate the ground truth matrix in this problem. Specifically, we randomly selected 80% of the observed data for training and used the rest 20% data for testing; 30% of the training data were randomly chosen and added to outliers from the uniform distribution on $[-50, 50]$. The performances, averaged over 100 runs, for all competing methods were reported in Table IV, where the accuracy measure $\text{RMSE}_{\text{test}}$ is defined similar to RMSE_{obv} but only calculated on the testing data. It can be observed that the proposed methods achieve the lowest $\text{RMSE}_{\text{test}}$ compared with other methods, which means that they are more accurate for recovering the unobserved data.

We also show the reconstructed tracks of different methods in Fig. 7. It is easy to observe that all of the results returned

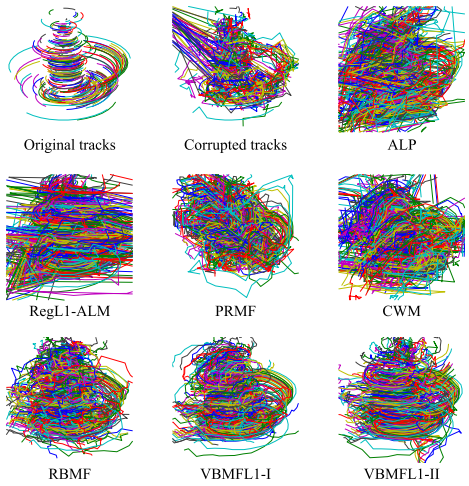


Fig. 7. Original incomplete, corrupted incomplete, and recovered tracks obtained by competing methods on the *Dinosaur* sequence.

by competing methods are affected by the outliers embedded in data, while our methods give a comparably better estimation, which complies with their good performance in terms of RMSE_{obv} and $\text{RMSE}_{\text{test}}$. It should be mentioned that the results of our methods are still far from perfect, since solving SFM problem requires additional constraints on the estimated matrices. We will further explore it in our future research.

2) Nonrigid Structure From Motion: Unlike the rigid SFM problem, the nonrigid SFM problem corresponds to an LRMF problem with rank $3k$, where k is the number of shape basis accounting for nonrigid deformation [40]. Here, we use the *Giraffe* sequence,[7] which includes 166 feature points tracked over 120 frames. The data matrix $\mathbf{M}_0$ is of size 240×166 with 30.24% entries missing; 20% of its elements were further randomly chosen and added to outliers, generated from the uniform distribution on $[-50, 50]$. We set k to 2 as in [7], leading to a rank-6 LRMF problem. The AQP and L_1-Wiberg encountered the out of memory problem on this data set again. Similar to rigid SFM, we also conducted two sets of experiments: one is to run a method on all the observed data and calculate the RMSE_{obv} value; the other is to split the observed data into 80% training/20% testing sets and run a method on the training set while calculating $\text{RMSE}_{\text{test}}$ on the test set. Each of the competing methods was run on the *Giraffe* sequence 100 times, and the averaged performance and computation time are reported in Tables IV and V. It is clear that the proposed methods get better accuracy among all the competing methods, while they are faster than ALP, RegL_1-ALM, and RBMF, and only unsubstantially slower than PRMF and CWM.

To further compare the performance of different methods, we depict the recovered points in three frames of the *Giraffe* sequence in Fig. 8. It is observed that the other competing methods produced disordered reconstruction in some frames more or less, while the proposed methods can more stably

[7] http://www.robots.ox.ac.uk/~abm/.

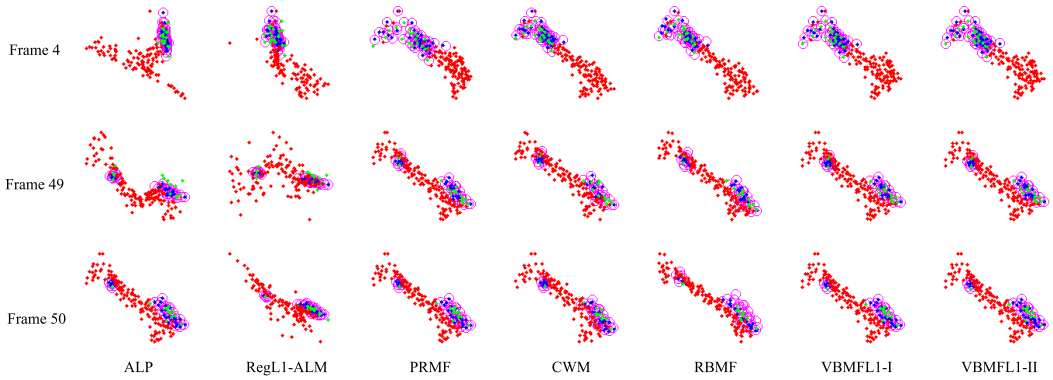


Fig. 8. Recovered points from the *Giraffe* sequence in the 4th, 49th, and 50th frames of the competing methods. From left to right: the results of ALP, RegL_1-ALM, PRMF, CWM, RBMF, VBMFL_1-I, and VBMFL_1-II, respectively. A point in green corresponds to an outlier, blue an observed entry, and red a missing entry. The magenta circles are the ground truth of observed points.

TABLE V
PERFORMANCE COMPARISON OF FIVE EXISTING L_1-NORM LRMF METHODS AND THE PROPOSED METHODS ON NONRIGID SFM EXPERIMENTS. THE RESULTS ARE AVERAGED OVER 100 RUNS

Method	RMSE_{obv}	Time(s)
ALP	1.5406	571.00
$RegL_1$-ALM	2.4979	72.12
PRMF	0.6424	**1.20**
CWM	0.9956	2.33
RBMF	2.8792	65.13
VBMFL_1-I	0.5604	3.33
VBMFL_1-II	**0.5415**	3.68

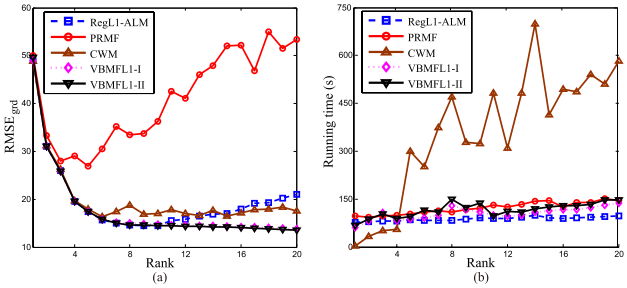


Fig. 9. (a) RMSE_{grd} curve and (b) running time curve for each competing method on faces of a single subject with rank varying from 1 to 20.

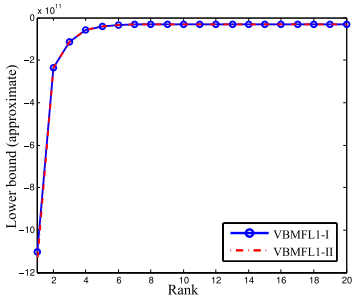


Fig. 10. Approximate lower bound curves for VBMFL_1-I and VBMFL_2-II on faces of a single subject with rank varying from 1 to 20.

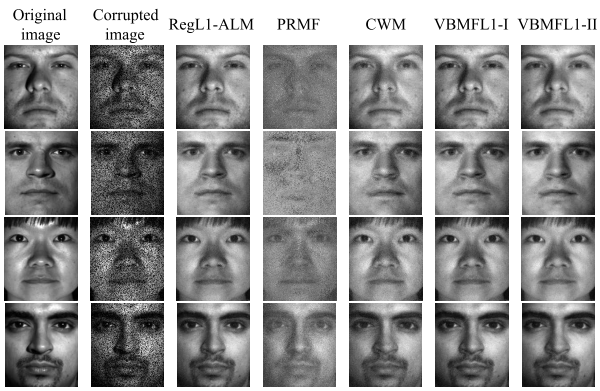


Fig. 11. From left to right: original face images, faces corrupted with 30% missing data and 30% outliers, and faces reconstructed by RegL_1-ALM, PRMF, CWM, VBMFL_1-I, and VBMFL_1-II.

recover all of the desired frames. This further verifies the effectiveness and the robustness of the proposed methods.

D. Face Reconstruction

As shown in [42], face images taken from one subject approximately lie on a low-dimensional subspace. Therefore, we can test the effectiveness of LRMF methods on face reconstruction problem, i.e., recovering face images from the corrupted ones with missing data and outliers. The faces were generated from the well-known Yale Face Database B [43]. We generated single-subject data as follows: 64 face images from the first subject were extracted, and all images were cropped to 192×168 pixels [44], resulting in the ground truth matrix $\mathbf{M}_0$ of size 32256×64. Then, 30% randomly chosen entries of $\mathbf{M}_0$ were designed as missing values, and 30% of the rest entries were added to uniform noise on $[-50, 50]$. All of the nine competing methods utilized on synthetic experiments have been tried, while four of them, including ALP, AQP, L_1-Wiberg, and RBMF, suffered from the out of memory error or cannot converge in a reasonable time. We varied the rank from 1 to 20 and recorded the RMSE_{grd} values and running time for each method, as summarized in Fig. 9.

As shown in Fig. 9, by taking a comparable computational time as existing methods, the proposed VBMFL_1-II

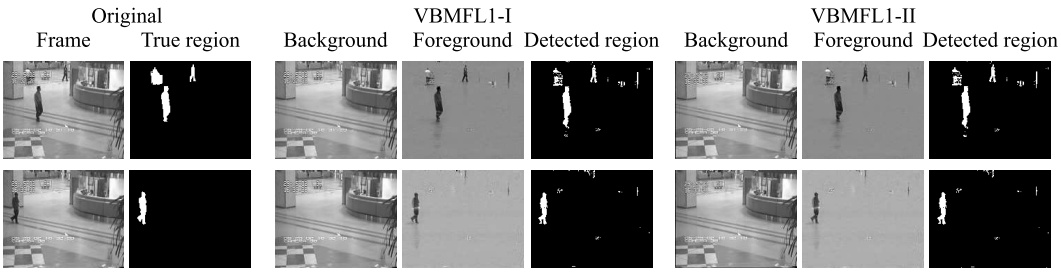


Fig. 12. Background subtraction results of the *Hall* sequence. From left to right: original video frames and ground truth foreground region, background and foreground separated by VBMFL_1-I and VBMFL_1-II.

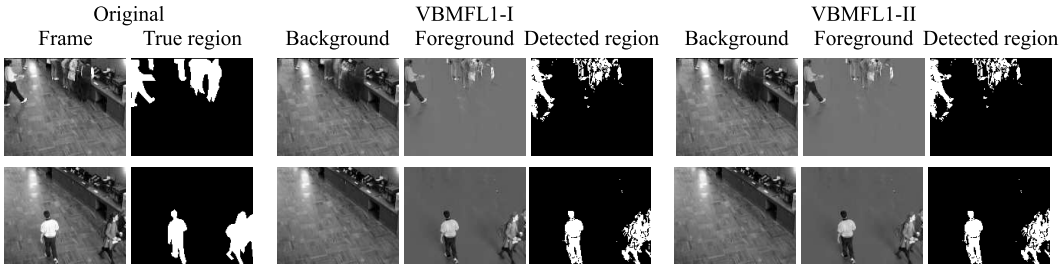


Fig. 13. Background subtraction results of the *Bootstrap* sequence. From left to right: original video frames and ground truth foreground region, background and foreground separated by VBMFL_1-I and VBMFL_1-II.

achieves the lowest RMSE_{grd} value among the four methods implemented for all rank configurations. It is interesting that the RMSE_{grd} values tend to become large for other competing methods as rank increasing, which implies that they are overfitting noise and are sensible to rank settings. In contrast, the proposed methods can stably obtain satisfactory reconstruction.

To further analyze the behavior of the proposed methods, we also plot the curves of variational lower bounds of the proposed methods versus rank, as shown in Fig. 10. It can be observed that the lower bounds keep increasing as the rank becomes higher and tend to be stable when the rank is larger than 8, which is consistent with tendency of the RMSE_{grd} curves shown in Fig. 9. It can also be observed that the lower bounds increase sharply before rank 4, while they change slowly after rank 4, indicating that the intrinsic rank should be 4 for this data set with faces of one person. This observation coincides with the analysis in [42], and thus suggests the potential application of our methods to model selection of LRMF problems.

Then, we implement the competing LRMF methods to reconstruct faces with multiperson. The data were generated as follows: 20 face images for each of the first 10 subjects in Yale Face Database B were randomly chosen, resulting in a total of 200 images. Then, the images were cropped and manually corrupted as the former experiments. Based on our aforementioned variational lower bound estimation, rank-40 approximation was used for this 10-person data set. The results in terms of RMSE_{grd} and computational time are summarized in Table VI. Some reconstructed faces are shown in Fig. 11 for easy comparison. It can be observed from Table VI and Fig. 11 that the proposed methods give better reconstruction for faces both quantitatively and visually.

TABLE VI
PERFORMANCE COMPARISON OF THREE EXISTING L_1-NORM LRMF METHODS AND THE PROPOSED METHODS ON FACE RECONSTRUCTION EXPERIMENTS

Method	RMSE_{grd}	Time(s)
$RegL_1$-ALM	17.3775	**416**
PRMF	43.0007	752
CWM	20.2280	4097
VBMFL_1-I	16.6978	841
VBMFL_1-II	**16.6926**	881

E. Background Subtraction

The background subtraction from a video sequence captured by a static camera can be modeled as a low-rank matrix analysis problem [33], and we thus verify the effectiveness of the proposed methods on this application. Two video sequences were adopted in our evaluation, including *Hall* and *Bootstrap* provided in [45].[8] Each sequence includes static background and intermittent movement in the foreground objects. RegL_1-ALM, PRMF, CWM, and the proposed VBMFL_1-I and VBMFL_1-II were implemented. Since the ground truth region for the foreground of some frames is provided [45], we can quantitatively compare the subtraction results given by different LRMF methods. To do this, we first extracted two subsequences for the two sequences, respectively, each including three frames with ground truth foreground. This resulted in a 700-frame subsequence for *Hall* and a 300-frame subsequence for *Bootstrap*. Then, rank-10 factorization was implemented by each method to obtain the background. Final subtraction was done by thresholding the absolute value of the difference between the original frame and the estimated background. Following [45], we use the following measure to

[8] http://perception.i2r.a-star.edu.sg/bk_model/bk_index.

TABLE VII
COMPARISON OF THREE EXISTING LRMF METHODS AND THE PROPOSED METHODS ON THE BACKGROUND SUBTRACTION EXPERIMENTS

Method	Hall		Bootstrap	
	Averaged S	Time(s)	Averaged S	Time(s)
$RegL_1$-ALM	0.6310	721.66	0.6038	218.42
PRMF	0.6216	**523.45**	0.5809	**153.74**
CWM	0.6335	4862	0.6122	1535
VBMFL_1-I	**0.6403**	771.35	**0.6143**	177.30
VBMFL_1-II	0.6384	830.96	0.6142	203.59

compare the performance of the competing methods:

$$S(A, B) = \frac{A \cap B}{A \cup B}$$

where A denotes the detected region and B is the corresponding ground truth region. Averaged S measure and running time for each method are summarized in Table VII.

As can be observed from Table VII, the proposed methods can more accurately detect the foreground region as compared with other competing methods. For visualization, we show the original video frames with ground truth foreground region and corresponding subtraction results of the proposed methods in Figs. 12 and 13. It can be observed that our methods can properly detect most of the true region. It is also expected to further improve the background subtraction performance by combining our methods with some more sophisticated techniques, e.g., Markov random field [46].

V. CONCLUSION

In this paper, we have proposed a new variational Beyesian approach to the L_1-norm LRMF problem. We have reformulated the original problem as a hierarchical Bayesian generative model and utilized the mean-field variational inference strategy to infer the parameters involved in the model. By virtue of the Bayesian framework, all the parameters can be automatically tuned to adapt to the data so that the generalization capability can be statistically guaranteed. A series of experimental results implemented on synthetic and real computer vision problems substantiates the efficiency and robustness of the proposed methods.

In our future work, we will try to further speed up the computation for the proposed models with the newly developed fast variational inference techniques, see [47], enhancing its applicability in more real large-scale problems. Besides, the proposed Bayesian framework can be further extended to other L_1-norm factorization tasks, such as nonnegative matrix factorization and tensor factorization [48], [49]. Moreover, extending current models to scenarios where the weight matrix $\mathbf{W}$ is unknown is a very interesting direction. These problems will be further investigated in our future research.

APPENDIX
CALCULATION OF EXPECTATIONS

In the following, we show how to calculate the expectations involved in the inference process.

The expectations utilized in Model I include: $\mathbb{E}[\mathbf{u}_i]$, $\mathbb{E}[\mathbf{u}_i\mathbf{u}_i^T]$, $\mathbb{E}[\mathbf{u}_i^T\mathbf{u}_i]$, $\mathbb{E}[\tau_u]$, $\mathbb{E}[\mathbf{v}_j]$, $\mathbb{E}[\mathbf{v}_j\mathbf{v}_j^T]$, $\mathbb{E}[\mathbf{v}_j^T\mathbf{v}_j]$, $\mathbb{E}[\tau_v]$, $\mathbb{E}[z_{ij}^{-1}]$, and $\mathbb{E}[(x_{ij} - \mathbf{u}_i^T\mathbf{v}_j)^2]$. Among them, $\mathbb{E}[\mathbf{u}_i]$, $\mathbb{E}[\tau_u]$, $\mathbb{E}[\mathbf{v}_j]$, and $\mathbb{E}[\tau_v]$ can be easily attained with respect to the current parameter values of the variational distributions by

$$\mathbb{E}[\mathbf{u}_i] = \boldsymbol{\mu}_{u_i}, \qquad \mathbb{E}[\tau_u] = \frac{a}{b}$$
$$\mathbb{E}[\mathbf{v}_j] = \boldsymbol{\mu}_{v_j}, \qquad \mathbb{E}[\tau_v] = \frac{c}{d}.$$

$\mathbb{E}[z_{ij}^{-1}]$ can also be easily calculated by

$$\mathbb{E}\big[z_{ij}^{-1}\big] = \mathbb{E}[y_{ij}] = \mu_{y_{ij}}.$$

Now we calculate $\mathbb{E}[\mathbf{u}_i\mathbf{u}_i^T]$ and $\mathbb{E}[\mathbf{u}_i^T\mathbf{u}_i]$. Note that

$$\begin{aligned}\boldsymbol{\Lambda}_{u_i}^{-1} &= \mathbb{E}[(\mathbf{u}_i - \mathbb{E}[\mathbf{u}_i])(\mathbf{u}_i - \mathbb{E}[\mathbf{u}_i])^T] \\ &= \mathbb{E}\big[\mathbf{u}_i\mathbf{u}_i^T\big] - \mathbb{E}[\mathbf{u}_i]\mathbb{E}[\mathbf{u}_i]^T\end{aligned}$$

and we thus have

$$\mathbb{E}\big[\mathbf{u}_i\mathbf{u}_i^T\big] = \boldsymbol{\Lambda}_{u_i}^{-1} + \mathbb{E}[\mathbf{u}_i]\mathbb{E}[\mathbf{u}_i]^T = \boldsymbol{\Lambda}_{u_i}^{-1} + \boldsymbol{\mu}_{u_i}\boldsymbol{\mu}_{u_i}^T.$$

Let $\mathrm{Tr}(\cdot)$ denote the trace of a matrix, and then we have

$$\mathbf{u}_i^T\mathbf{u}_i = \mathrm{Tr}\big(\mathbf{u}_i^T\mathbf{u}_i\big) = \mathrm{Tr}\big(\mathbf{u}_i\mathbf{u}_i^T\big).$$

Therefore

$$\begin{aligned}\mathbb{E}\big[\mathbf{u}_i^T\mathbf{u}_i\big] &= \mathbb{E}\big[\mathrm{Tr}\big(\mathbf{u}_i\mathbf{u}_i^T\big)\big] = \mathrm{Tr}\big(\mathbb{E}\big[\mathbf{u}_i\mathbf{u}_i^T\big]\big) \\ &= \mathrm{Tr}\big(\boldsymbol{\Lambda}_{u_i}^{-1} + \boldsymbol{\mu}_{u_i}\boldsymbol{\mu}_{u_i}^T\big) = \mathrm{Tr}\big(\boldsymbol{\Lambda}_{u_i}^{-1}\big) + \boldsymbol{\mu}_{u_i}^T\boldsymbol{\mu}_{u_i}.\end{aligned}$$

Similarly, $\mathbb{E}[\mathbf{v}_j\mathbf{v}_j^T]$ and $\mathbb{E}[\mathbf{v}_j^T\mathbf{v}_j]$ can be calculated by

$$\begin{aligned}\mathbb{E}\big[\mathbf{v}_j\mathbf{v}_j^T\big] &= \boldsymbol{\Lambda}_{v_j}^{-1} + \boldsymbol{\mu}_{v_j}\boldsymbol{\mu}_{v_j}^T \\ \mathbb{E}\big[\mathbf{v}_j^T\mathbf{v}_j\big] &= \mathrm{Tr}\big(\boldsymbol{\Lambda}_{v_j}^{-1}\big) + \boldsymbol{\mu}_{v_j}^T\boldsymbol{\mu}_{v_j}.\end{aligned}$$

Based on the above calculations, $\mathbb{E}[(x_{ij} - \mathbf{u}_i^T\mathbf{v}_j)^2]$ can be easily attained as follows:

$$\begin{aligned}&\mathbb{E}\big[\big(x_{ij} - \mathbf{u}_i^T\mathbf{v}_j\big)^2\big] \\ &= \mathbb{E}\big[x_{ij}^2 - 2x_{ij}\mathbf{u}_i^T\mathbf{v}_j + \mathbf{u}_i^T\mathbf{v}_j\mathbf{v}_j^T\mathbf{u}_i\big] \\ &= x_{ij}^2 - 2x_{ij}\mathbb{E}[\mathbf{u}_i]^T\mathbb{E}[\mathbf{v}_j] + \mathbb{E}\big[\mathrm{Tr}\big(\mathbf{u}_i\mathbf{u}_i^T\mathbf{v}_j\mathbf{v}_j^T\big)\big] \\ &= x_{ij}^2 - 2x_{ij}\boldsymbol{\mu}_{u_i}^T\boldsymbol{\mu}_{v_j} + \mathrm{Tr}\big(\mathbb{E}\big[\mathbf{u}_i\mathbf{u}_i^T\big]\mathbb{E}\big[\mathbf{v}_j\mathbf{v}_j^T\big]\big) \\ &= x_{ij}^2 - 2x_{ij}\boldsymbol{\mu}_{u_i}^T\boldsymbol{\mu}_{v_j} + \mathrm{Tr}\big(\big(\boldsymbol{\Lambda}_{u_i}^{-1} + \boldsymbol{\mu}_{u_i}\boldsymbol{\mu}_{u_i}^T\big)\big(\boldsymbol{\Lambda}_{v_j}^{-1} + \boldsymbol{\mu}_{v_j}\boldsymbol{\mu}_{v_j}^T\big)\big).\end{aligned}$$

The expectations for Model II can be calculated similar to those in Model I, except $\mathbb{E}[\tau_{u_i}]$ and $\mathbb{E}[\tau_{v_j}]$, which can be, respectively, calculated by

$$\mathbb{E}[\tau_{u_i}] = \frac{a}{b_i}, \qquad \mathbb{E}[\tau_{v_j}] = \frac{c}{d_i}.$$

Finally, we discuss the calculation of $\mathbb{E}[z_{ij}]$ involved in the update equation of λ. Note that

$$\mathbb{E}[z_{ij}] = \mathbb{E}\big[y_{ij}^{-1}\big]$$

and y_{ij} follows the inverse Gaussian distribution, and then using the result shown in [50], we can get:

$$\mathbb{E}\big[y_{ij}^{-1}\big] = \mu_{y_{ij}}^{-1} + \lambda_y^{-1}.$$

REFERENCES

[1] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *Int. J. Comput. Vis.*, vol. 9, no. 2, pp. 137–154, 1992.

[2] H. Hayakawa, "Photometric stereo under a light source with arbitrary motion," *J. Opt. Soc. Amer. A*, vol. 11, no. 11, pp. 3079–3089, 1994.

[3] M. Irani, "Multi-frame optical flow estimation using subspace constraints," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 1. Kerkyra, Greece, Sep. 1999, pp. 626–633.

[4] M. J. Black and A. D. Jepson, "EigenTracking: Robust matching and tracking of articulated objects using a view-based representation," *Int. J. Comput. Vis.*, vol. 26, no. 1, pp. 63–84, 1998.

[5] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural Netw.*, vol. 2, no. 1, pp. 53–58, 1989.

[6] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in *Proc. 20th Int. Conf. Mach. Learn.*, Washington, DC, USA, Aug. 2003, pp. 720–727.

[7] A. M. Buchanan and A. W. Fitzgibbon, "Damped Newton algorithms for matrix factorization with missing data," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. San Diego, CA, USA, Jun. 2005, pp. 316–322.

[8] K. Mitra, S. Sheorey, and R. Chellappa, "Large-scale matrix factorization with missing data under additional constraints," in *Advances in Neural Information Processing Systems 23*. Red Hook, NY, USA: Curran & Associates Inc., 2010, pp. 1651–1659.

[9] T. Okatani and K. Deguchi, "On the Wiberg algorithm for matrix factorization in the presence of missing components," *Int. J. Comput. Vis.*, vol. 72, no. 3, pp. 329–337, 2007.

[10] T. Okatani, T. Yoshida, and K. Deguchi, "Efficient algorithm for low-rank matrix factorization with missing components and performance comparison of latest algorithms," in *Proc. IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 842–849.

[11] Q. Ke and T. Kanade, "Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1. San Diego, CA, USA, Jun. 2005, pp. 739–746.

[12] A. Eriksson and A. van den Hengel, "Efficient computation of robust low-rank matrix approximations in the presence of missing data using the L_1 norm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 771–778.

[13] Y. Zheng, G. Liu, S. Sugimoto, S. Yan, and M. Okutomi, "Practical low-rank matrix approximation under robust L_1-norm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 1410–1417.

[14] F. De La Torre and M. J. Black, "A framework for robust subspace learning," *Int. J. Comput. Vis.*, vol. 54, no. 1, pp. 117–142, 2003.

[15] C. Ding, D. Zhou, X. He, and H. Zha, "R_1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, Jun. 2006, pp. 281–288.

[16] N. Kwak, "Principal component analysis based on L_1-norm maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1672–1680, Sep. 2008.

[17] D. Meng and F. De La Torre, "Robust matrix factorization with unknown noise," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 1337–1344.

[18] N. Wang, T. Yao, J. Wang, and D.-Y. Yeung, "A probabilistic approach to robust matrix factorization," in *Proc. 12th Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 126–139.

[19] N. Wang and D.-Y. Yeung, "Bayesian robust matrix factorization for image and video processing," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, VIC, Australia, Dec. 2013, pp. 1785–1792.

[20] D. Meng, Z. Xu, L. Zhang, and J. Zhao, "A cyclic weighted median method for L_1 low-rank matrix factorization with missing entries," in *Proc. 27th AAAI Conf. Artif. Intell.*, Bellevue, WA, USA, Jul. 2013, pp. 704–710.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.

[22] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *J. Roy. Statist. Soc., Ser. B (Methodological)*, vol. 36, no. 1, pp. 99–102, 1974.

[23] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *Advances in Neural Information Processing Systems 20*. Red Hook, NY, USA: Curran & Associates Inc., 2007, pp. 1257–1264.

[24] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo," in *Proc. 25th Int. Conf. Mach. Learn.*, Helsinki, Finland, Jul. 2008, pp. 880–887.

[25] Y. J. Lim and Y. W. Teh, "Variational Bayesian approach to movie rating prediction," in *Proc. KDD Cup Workshop*, San Jose, CA, USA, Aug. 2007, pp. 15–21.

[26] S. Nakajima, M. Sugiyama, and R. Tomioka, "Global analytic solution for variational Bayesian matrix factorization," in *Advances in Neural Information Processing Systems 23*. Red Hook, NY, USA: Curran & Associates Inc., 2010, pp. 1768–1776.

[27] I. Porteous, A. U. Asuncion, and M. Welling, "Bayesian matrix factorization with side information and Dirichlet process mixtures," in *Proc. 24th AAAI Conf. Artif. Intell.*, Atlanta, GA, USA, Jul. 2010, pp. 563–568.

[28] M. Hoffman, P. R. Cook, and D. M. Blei, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, Jun. 2010, pp. 439–446.

[29] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Comput. Intell. Neurosci.*, vol. 2009, Feb. 2009, Art. ID 785152.

[30] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[31] B. Lakshminarayanan, G. Bouchard, and C. Archambeau, "Robust Bayesian matrix factorisation," in *Proc. 4th Int. Conf. Artif. Intell. Statist.*, Fort Lauderdale, FL, USA, Apr. 2011, pp. 425–433.

[32] I. T. Jolliffe, *Principal Component Analysis* (Springer Series in Statistics), 2nd ed. New York, NY, USA: Springer-Verlag, 2002.

[33] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, 2011, Art. ID 11.

[34] C. M. Bishop, "Variational principal components," in *Proc. 9th Int. Conf. Artif. Neural Netw.*, vol. 1. Edinburgh, U.K., Sep. 1999, pp. 509–514.

[35] J. Gao, "Robust L_1 principal component analysis and its Bayesian variational inference," *Neural Comput.*, vol. 20, no. 2, pp. 555–572, 2008.

[36] J. Luttinen, A. Ilin, and J. Karhunen, "Bayesian robust PCA of incomplete data," *Neural Process. Lett.*, vol. 36, no. 2, pp. 189–202, 2012.

[37] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3419–3430, Dec. 2011.

[38] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3964–3977, Aug. 2012.

[39] S. Nakajima, M. Sugiyama, and S. Derin Babacan, "Variational Bayesian sparse additive matrix factorization," *Mach. Learn.*, vol. 92, nos. 2–3, pp. 319–347, 2013.

[40] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Hilton Head Island, SC, USA, Jun. 2000, pp. 690–696.

[41] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and L. Zhang, "Robust principal component analysis with complex noise," in *Proc. 31th Int. Conf. Mach. Learn.*, Beijing, China, Jun. 2014, pp. 55–63.

[42] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, Feb. 2003.

[43] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[44] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[45] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Trans. Image Process.*, vol. 13, no. 11, pp. 1459–1472, Nov. 2004.

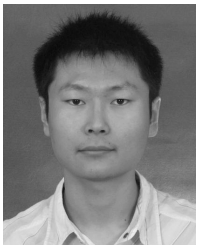
[46] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[47] J. Hensman, M. Rattray, and N. D. Lawrence, "Fast variational inference in the conjugate exponential family," in *Advances in Neural Information Processing Systems 25*. Red Hook, NY, USA: Curran & Associates Inc., 2012, pp. 2888–2896.

[48] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. West Sussex, U.K.: Wiley, 2009.

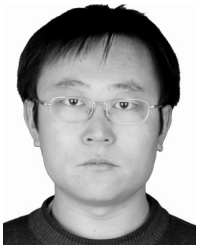
[49] Y. Peng, D. Meng, Z. Xu, C. Gao, Y. Yang, and B. Zhang, "Decomposable nonlocal tensor dictionary learning for multispectral image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2949–2956.

[50] M. C. K. Tweedie, "Statistical properties of inverse Gaussian distributions. I," *Ann. Math. Statist.*, vol. 28, no. 2, pp. 362–377, 1957.



Qian Zhao received the B.Sc. degree from Xi'an Jiaotong University, Xi'an, China, in 2009, where he is currently pursuing the Ph.D. degree.

His current research interests include low-rank matrix factorization, dimensionality reduction, and Bayesian method for machine learning.



Deyu Meng (M'13) received the B.Sc., M.Sc., and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2001, 2004, and 2008, respectively.

He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, from 2012 to 2014. He is currently an Associate Professor with the Institute for Information and System Sciences, School of Mathematics and Statistics, Xi'an Jiaotong University. His current research interests include principal component analysis, nonlinear dimensionality reduction, feature extraction and selection, compressed sensing, and sparse machine learning methods.

Zongben Xu received the Ph.D. degree in mathematics from Xi'an Jiaotong University, Xi'an, China, in 1987.

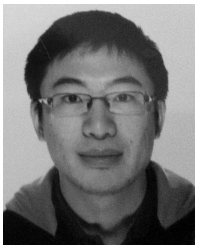
He currently serves as the Academician of the Chinese Academy of Sciences, the Chief Scientist of the National Basic Research Program of China (973 Project), and the Director of the Institute for Information and System Sciences with Xi'an Jiaotong University. His current research interests include nonlinear functional analysis and intelligent information processing.

Prof. Xu was a recipient of the National Natural Science Award of China in 2007 and the winner of the CSIAM Su Buchin Applied Mathematics Prize in 2008. He delivered a talk at the International Congress of Mathematicians in 2010.

Wangmeng Zuo (M'09) received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007.

He was a Research Assistant with the Department of Computing, Hong Kong Polytechnic University, Hong Kong, from 2004 to 2008, and a Visiting Professor with Microsoft Research Asia, Beijing, China, from 2009 to 2010. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology. He has authored about 60 papers in his research areas. His current research interests include image modeling and low-level vision, discriminative learning, biometrics, and computer vision.

Dr. Zuo is an Associate Editor of the *IET Biometrics* and *The Scientific World*.



Yan Yan received the Ph.D. degree from the University of Trento, Trento, Italy, in 2014.

He is currently a Post-Doctoral Researcher with the MHUG Group, University of Trento. His current research interests include machine learning and its application to computer vision and multimedia analysis.