A Variational Approach for Sparse Component Estimation and Low-Rank Matrix Recovery

Zhaofu Chen[1], Rafael Molina[2], and Aggelos K. Katsaggelos[1]
[1]Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, 60208, USA
[2]Deptartmento de Ciencias de la Computación e I. A., Universidad de Granada, 18071 Granada, Spain
Email: zhaofuchen2014@u.northwestern.edu; rms@decsai.ugr.es; aggk@eecs.northwestern.edu

***Abstract* — We propose a variational Bayesian based algorithm for the estimation of the sparse component of an outlier-corrupted low-rank matrix, when linearly transformed composite data are observed. The model constitutes a generalization of robust principal component analysis. The problem considered herein is applicable in various practical scenarios, such as foreground detection in blurred and noisy video sequences and detection of network anomalies among others. The proposed algorithm models the low-rank matrix and the sparse component using a hierarchical Bayesian framework, and employs a variational approach for inference of the unknowns. The effectiveness of the proposed algorithm is demonstrated using real life experiments, and its performance improvement over regularization based approaches is shown.**

***Index Terms* — Bayesian inference, variational approach, robust principal component analysis, foreground detection, network anomaly detection**

I. Introduction

The problem of estimating low-rank matrices in the presence of sparse outliers has drawn significant attention recently. A typical example is robust principal component analysis (RPCA), where the high-dimensional data lying in a low-dimensional subspace are subject to the perturbation of a few outliers. Recently theoretical performance guarantees for RPCA have been provided in [1], where it is shown that the RPCA problem can be solved using convex optimization. In addition, RPCA has been applied to solve a wide range of problems [2]–[5] and its advantage has been demonstrated [6]–[9].

In this paper we consider a generalization of the original RPCA problem, where a linear transformation through the use of a known measurement matrix, is applied to the outlier-corrupted data. The goal is to estimate the outlier amplitudes given the transformed observation. This problem stems from several practical scenarios, which we will discuss in detail shortly. A regularization based algorithm, which requires the manual tuning of its parameters, was proposed to solve this problem [10]. In this work, we propose a variational Bayesian based approach that provides approximate posterior distributions of all the model unknowns. Experiments using real life datasets as well as computer simulations show performance improvement of the proposed algorithm over its regularization based counterpart.

This paper is organized as follows. In Section II we present the general data model and several areas of applications. A brief overview of the related work in each of these areas is also provided. In Section III we introduce the proposed hierarchical Bayesian model. Details of the variational inference procedure are provided in Section IV. Numerical examples are presented in Section V. Finally we draw conclusion remarks in Section VI.

Notation: Matrices and vectors are denoted by uppercase and lowercase boldface letters, respectively. $\mathrm{vec}(\cdot)$, $\mathrm{diag}(\cdot)$ and $\mathrm{Tr}(\cdot)$ are vectorization, diagonalization and trace operators, respectively. Given a matrix $\mathbf{X}$, we denote as $\mathbf{x}_{i\cdot}$, $\mathbf{x}_{\cdot j}$ and X_{ij} its i^{th} row, j^{th} column and $(i,j)^{\text{th}}$ element, respectively.

II. Problem Statement and Data Model

In this section we formulate the sparse component detection and low-rank matrix estimation problem. We first present a general data model that covers a wide range of applications, and then consider specific scenarios as its special cases.

Let $\mathbf{X} \in \mathbb{R}^{L\times T}$ be a low-rank matrix with $\mathrm{rank}(\mathbf{X}) \ll \min(L,T)$, and $\mathbf{E} \in \mathbb{R}^{F\times T}$ be a sparse matrix with entries of arbitrarily large magnitudes. Matrix $\mathbf{R} \in \mathbb{R}^{L\times F}$ models the linear transformation performed on the data, that is, $\mathbf{R}$ is a known measurement matrix, where $L \leq F$. $\mathbf{R}$ bears specific physical meanings in the scenarios discussed below. Consider also dense measurement noise $\mathbf{N} \in \mathbb{R}^{L\times T}$ added to the observations. $\mathbf{N}$ is assumed to have small amplitudes compared with $\mathbf{X}$ and $\mathbf{E}$. With the quantities defined above, we have the general data model:

$$\mathbf{Y} = \mathbf{X} + \mathbf{R}\mathbf{E} + \mathbf{N}\,. \tag{1}$$

The goal of the problem is to obtain accurate estimates for the sparse term $\mathbf{E}$ and the low-rank term $\mathbf{X}$, given the noise corrupted observation $\mathbf{Y}$. Although $\mathbf{E}$ is sparse, the multiplication with a wide matrix $\mathbf{R}$ has an effect of compression, and hence the product $\mathbf{RE}$ is not necessarily sparse. Therefore conventional RPCA approaches cannot be applied directly to solve this problem.

Given (1), we present below examples of $\mathbf{R}$, which stem from different application scenarios.

Manuscript received July 6, 2013; revised September 20, 2013.
Please address correspondences to A. K. Katsaggelos (email: aggk@eecs.northwestern.edu)
This work has been supported in part by "Ministerio de Ciencia e Innovacion" under contract TIN2010-15137 and a grant from the Department of Energy (DE-NA0000457).
doi:10.12720/jcm.8.9.600-611

A. Robust PCA

With $L = F$ and $\mathbf{R} = \mathbf{I}_F$, (1) reduces to

$$\mathbf{Y} = \mathbf{X} + \mathbf{E} + \mathbf{N}, \tag{2}$$

where all matrices involved are of dimensions $F \times T$. This is the classical RPCA problem, where the goal is to recover the low-rank component $\mathbf{X}$ and the sparse component $\mathbf{E}$. The RPCA problem has recently drawn significant attention from the research community, and a wealth of literature has been devoted to related studies. Algorithms for solving the RPCA problem can be broadly classified into two categories: regularization based approaches or statistical inference based approaches.

For the former category, the problem is formulated as regularized fitting, where the regularizers are convex surrogates for rank and sparsity. Analysis on the exact recovery in the RPCA problem is given in [1]. Examples of algorithms in this category include the singular value thresholding (SVT) algorithm [11], the accelerated proximal gradient (APG) method [12], and the augmented Lagrange multiplier (ALM) method in [13]. In addition, extensive efforts have been made to analyze the RPCA problem from various perspectives [14]–[18].

For the latter category, hierarchical statistical models are introduced to formulate the data generation process and prior distributions are selected to capture the low-rank and sparse properties of the respective terms. The joint distribution involving the observations, unknown variables and hyperparameters can be determined from the priors and conditional distributions. Posterior distributions of the unknowns are approximated using Bayesian based approaches. Representative algorithms in this category include [19]–[21]. As an example, [21] employs hierarchical model to capture the properties of the data and applies variational approach for inference. It is therefore a special case of the algorithm proposed herein when $\mathbf{R} = \mathbf{I}_F$.

B. Foreground Detection in Blurred and Noisy Video Sequsenc

Foreground (FG) detection is an important computer vision problem [22], [23]. Denote a video sequence as an $F \times T$ matrix $\mathbf{D}$, where F is the number of pixels per frame and T is the number of frames. Each frame, represented as a column in $\mathbf{D}$, is the superposition of moving FG objects and relatively static background (BG) scene. Since the FG objects are usually small relative to the entire frame and do not persist across the entire sequence, they can be represented as an $F \times T$ sparse matrix $\mathbf{E}$ [20], [21], [24]. On the other hand, the background is more static, and its time invariance can be captured by a low-rank matrix $\mathbf{X}_0$. Let $\mathbf{R}$ model a linear transformation on a frame, such as blurring and resolution scaling. With the above specifications, the transformed and noisy video sequence can be represented as

$$\begin{aligned} \mathbf{Y} &= \mathbf{RD} + \mathbf{N} = \mathbf{R}(\mathbf{X}_0 + \mathbf{E}) + \mathbf{N} \\ &= \mathbf{RX}_0 + \mathbf{RE} + \mathbf{N} = \mathbf{X} + \mathbf{RE} + \mathbf{N}, \end{aligned} \tag{3}$$

where we have combined $\mathbf{RX}_0$ as $\mathbf{X}$ since the primary objective is to detect the foreground $\mathbf{E}$. Note $\mathbf{X}$ inherits the low-rank property from $\mathbf{X}_0$. It is clear (3) is identical to (1), with $\mathbf{R}$ being an $L \times F$ real-valued matrix modeling the linear transformation performed on video pixels.

C. Network Anomaly Detection From Link Observations

Another application emerging from (1) is network security and anomaly detection [25], [26]. Consider a network consisting of N nodes (e.g., routers) that can send and forward data packets. A link is the physical connection between a pair of nodes, and L denotes the number of links. An origin-destination (OD) flow is defined as a stream of packets sent from one node and received by another node, and F denotes the number of flows. Since the network in general is not strongly connected (i.e., there is not a link between every pair of nodes), we have $L \ll F$. In the network scenario, $\mathbf{R}$ is an $L \times F$ wide routing matrix consisting of binary entries r_{ij}, with $r_{ij} = 1$ denoting flow j passes through link i and $r_{ij} = 0$ otherwise [27].

Let the $F \times T$ matrix $\mathbf{X}_0$ denote the OD flow traffic during normal network operation, where its $(i, j)^{\text{th}}$ component is the sampled traffic of flow i at time j. $\mathbf{X}_0$ is low-rank because the normal network flows roughly follow a temporal pattern, rendering the columns of $\mathbf{X}_0$ approximately linearly dependent. On the other hand, network anomalies can occur sporadically in the OD flows and at different times, possibly due, for instance, to network failures, external attacks, etc. The anomaly induced traffic can be of large magnitude compared with the normal traffic, and is represented by a sparse $F \times T$ matrix $\mathbf{E}$. The multiplication of the total OD flow $\mathbf{X}_0 + \mathbf{E}$ by $\mathbf{R}$ maps the composite flow traffic into the link measurements. It can be seen that the data model is again the same as (1), with $\mathbf{R}$ being an $L \times F$ wide binary routing matrix.

Researchers have performed extensive and in-depth studies into OD flow anomaly detection. For example, [28] presents analysis of OD flow time series using PCA and the decomposition of flows into the normal and abnormal constituents. [29] applies the principal component pursuit (PCP) algorithm [1] to analyze the OD flows. Along the same line of research as [28], authors of [30] apply subspace methods to decompose the link measurements and discusses the subsequent identification of anomalies in the OD flows. From a slightly different perspective, [31] examines the possible types of attacks specifically targeted at PCA-based detectors, which provides valuable insights as to the design of more effective detection technologies. In addition, the authors of [10] extend the APG algorithm [12] to analyze the link measurements for flow anomaly detection. For an overview of anomaly detection techniques the reader is referred to [25].

In the next section we propose a variational Bayesian based approach to solve the recovery problem (1) for general $\mathbf{R}$. Within this framework, the variational Bayesian robust PCA algorithm introduced in [21] can be regarded as a special case of our proposed algorithm when $\mathbf{R} = \mathbf{I}_F$.

III. Bayesian Modeling

In this section we present the Bayesian modeling for the sparse component estimation and low-rank matrix recovery

problem. Both unknowns and observed quantities are treated as random variables. A hierarchical Bayesian framework is employed to model the data generation process, where the joint distribution of all quantities is factorized into the product of priors and conditional probabilities.

A. Low-Rank Term Modeling

Given the observation model in (1) our goal is to estimate a low-rank matrix $\mathbf{X}$ and a sparse matrix $\mathbf{E}$ from the noisy $\mathbf{Y}$. A natural estimator is to fit $\mathbf{Y}$ in the least-squares (LS) sense as well as to minimize the rank of $\mathbf{X}$ and the number of non-zero entries in $\mathbf{E}$ measured by its l_0-(pseudo) norm. Unfortunately, both l_0-norm and rank minimizations are in general NP-hard problems [32], [33]. Therefore, we start by replacing the l_0-norm with its convex surrogate $||\mathbf{E}||_1$, and the rank of $\mathbf{X}$ with $||\mathbf{X}||_*$, which is equal to the sum of the singular values of $\mathbf{X}$. The nuclear norm $||\mathbf{X}||_*$ can be neatly characterized as [34]

$$\begin{aligned} ||\mathbf{X}||_* = \min_{\{\mathbf{A},\mathbf{B}\}} \frac{1}{2}\{||\mathbf{A}||_{\mathrm{F}}^2 + ||\mathbf{B}||_{\mathrm{F}}^2\} \\ \text{subject to} \quad \mathbf{X} = \mathbf{A}\mathbf{B}^{\mathrm{T}} \end{aligned} \tag{4}$$

Adopting these relaxation and parametrization, one obtains the following optimization problem

$$\min_{\{\mathbf{A},\mathbf{B},\mathbf{E}\}} \frac{1}{2}||\mathbf{Y}-\mathbf{A}\mathbf{B}^{\mathrm{T}}-\mathbf{R}\mathbf{E}||_{\mathrm{F}}^2+\lambda_* \left(||\mathbf{A}||_{\mathrm{F}}^2+||\mathbf{B}||_{\mathrm{F}}^2\right)+\lambda_1||\mathbf{E}||_1, \tag{5}$$

where λ_* and λ_1 are regularization parameters.

We denote an estimated upper bound for the rank of $\mathbf{X}$ as k and let $\mathbf{A}$ and $\mathbf{B}$ be $L \times k$ and $T \times k$ matrices, respectively. In this way $\mathbf{X}$ can be expressed as the sum of k outer-products, i.e.,

$$\mathbf{X} = \mathbf{A}\mathbf{B}^{\mathrm{T}} = \sum_{i=1}^{k} \mathbf{a}_{\cdot i}\mathbf{b}_{\cdot i}^{\mathrm{T}} . \tag{6}$$

Note that with the factorization in (6) we are able to characterize the rank of $\mathbf{X}$ while accommodating $\mathbf{X}$ of arbitrary dimensions.

To embody the low-rank property of $\mathbf{X}$, we aim at promoting column sparsity in $\mathbf{A}$ and $\mathbf{B}$ such that the sum in (6) has only a small number of non-zero terms. To enforce this constraint, we associate the columns of $\mathbf{A}$ and $\mathbf{B}$ with Gaussian priors of precisions $\{\gamma_i\}_{i=1}^k$, i.e.,

$$p(\mathbf{A}|\boldsymbol{\gamma}) = \prod_{i=1}^{k} p(\mathbf{a}_{\cdot i}|\gamma_i) = \prod_{i=1}^{k} \mathcal{N}(\mathbf{a}_{\cdot i}|\mathbf{0}, \gamma_i^{-1}\mathbf{I}_L), \tag{7a}$$

$$p(\mathbf{B}|\boldsymbol{\gamma}) = \prod_{i=1}^{k} p(\mathbf{b}_{\cdot i}|\gamma_i) = \prod_{i=1}^{k} \mathcal{N}(\mathbf{b}_{\cdot i}|\mathbf{0}, \gamma_i^{-1}\mathbf{I}_T), \tag{7b}$$

where $\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_k]$. Since the columns of $\mathbf{A}$ and $\mathbf{B}$ have the same sparsity profile enforced by the common precisions $\{\gamma_i\}_{i=1}^k$, they become small simultaneously when the corresponding γ_i assumes large value. During the inference process, many of the precision parameters $\boldsymbol{\gamma}_i$ will take large values, effectively eliminating the respective outer products from $\mathbf{X}$. In this fashion, we achieve the goal of promoting low-rank property of $\mathbf{X}$.

In addition to (7), we incorporate the i.i.d. conjugate Gamma hyperprior on the precisions $\{\gamma_i\}_{i=1}^k$, i.e.,

$$p(\boldsymbol{\gamma}) = \prod_{i=1}^{k} p(\gamma_i) = \prod_{i=1}^{k} \gamma_i^{a-1} \exp(-b\gamma_i), \tag{8}$$

where a and b are treated as deterministic and they are assigned small values to yield broad hyperpriors.

B. Sparse Term Modeling

To model the sparse term $\mathbf{E}$, we let the entries be independent of each other, and their amplitudes be governed by zero-mean Gaussian distributions with independent precisions. Specifically, we have

$$p(\mathbf{E}|\boldsymbol{\alpha}) = \prod_{i=1}^{F}\prod_{j=1}^{T} p(e_{ij}|\alpha_{ij}) = \prod_{i=1}^{F}\prod_{j=1}^{T} \mathcal{N}(e_{ij}|0, \alpha_{ij}^{-1}), \tag{9}$$

where $\boldsymbol{\alpha}$ is the matrix containing all the FT values of the precisions α_{ij}. For large α_{ij}, the corresponding e_{ij} is close to zero with high probability. During inference, a large number of α_{ij} will be set to high values, and consequently the corresponding e_{ij} will be literally zeros. The precisions α_{ij} are assigned i.i.d. non-informative Jeffrey's prior

$$p(\boldsymbol{\alpha}) = \prod_{i=1}^{F}\prod_{j=1}^{T} p(\alpha_{ij}) = \prod_{i=1}^{F}\prod_{j=1}^{T} \alpha_{ij}^{-1} . \tag{10}$$

C. Noisy Observation Modeling

We employ the common i.i.d. Gaussian priors with zero mean and precision β to model the dense observation noise $\mathbf{N}$:

$$p(\mathbf{N}|\beta) = \prod_{i=1}^{L}\prod_{j=1}^{T} p(n_{ij}|\beta) = \prod_{i=1}^{L}\prod_{j=1}^{T} \mathcal{N}(n_{ij}|0, \beta^{-1}), \tag{11}$$

Assigning Jeffrey's prior on β, we have

$$p(\beta) = \beta^{-1} . \tag{12}$$

Given the components defined above, the observation model is given by

$$\begin{aligned} p(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{E},\beta) &= \mathcal{N}(\mathrm{vec}(\mathbf{Y})|\mathrm{vec}(\mathbf{A}\mathbf{B}^{\mathrm{T}}+\mathbf{R}\mathbf{E}), \beta^{-1}\mathbf{I}_{LT}) \\ &\propto \beta^{\frac{LT}{2}} \exp\left\{-\frac{\beta}{2}||\mathbf{Y}-\mathbf{A}\mathbf{B}^{\mathrm{T}}-\mathbf{R}\mathbf{E}||_{\mathrm{F}}^2\right\} \end{aligned} \tag{13}$$

D. Joint Distribution

By combining the stages of the hierarchical Bayesian model, the joint distribution of the observation and all the unknowns is expressed as

$$\begin{aligned} p(\mathbf{Y},\mathbf{A},\mathbf{B},\mathbf{E},\boldsymbol{\gamma},\boldsymbol{\alpha},\beta) &= p(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{E},\beta)p(\mathbf{A}|\boldsymbol{\gamma})p(\mathbf{B}|\boldsymbol{\gamma}) \\ &\quad \times p(\boldsymbol{\gamma})p(\mathbf{E}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\beta), \end{aligned} \tag{14}$$

where $p(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{E},\beta)$, $p(\mathbf{A}|\boldsymbol{\gamma})$, $p(\mathbf{B}|\boldsymbol{\gamma})$, $p(\boldsymbol{\gamma})$, $p(\mathbf{E}|\boldsymbol{\alpha})$, $p(\boldsymbol{\alpha})$ and $p(\beta)$ are given in (13), (7a), (7b), (8), (9), (10) and (12), respectively. The dependencies in this joint probability model are shown in graphical form in Figure 1, where the arrows are used to denote the generative model.

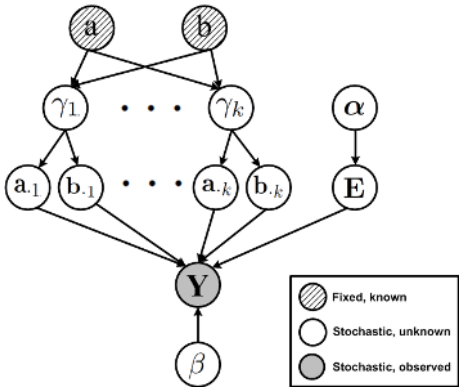


Figure. 1: Directed acyclic graph representation of the hierarchical Bayesian model.

IV. Approximate Bayesian Inference

As is widely known, Bayesian inference is based on the posterior distribution of unknowns given the observations. Let $\mathbf{z}$ be the vector of all the unknowns such that

$$\mathbf{z} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\gamma}, \mathbf{E}, \boldsymbol{\alpha}, \beta)\,. \tag{15}$$

The posterior is expressed using Bayes rule as

$$p(\mathbf{z}|\mathbf{Y}) = \frac{p(\mathbf{Y}, \mathbf{z})}{p(\mathbf{Y})}\,. \tag{16}$$

However, the posterior $p(\mathbf{z}|\mathbf{Y})$ is computationally intractable because the denominator

$$p(\mathbf{Y}) = \int p(\mathbf{Y}, \mathbf{z}) \mathrm{d}\mathbf{z} \tag{17}$$

cannot be calculated analytically by marginalizing all unknowns. Therefore, approximation methods must be utilized. Common approaches for approximation include maximum *a posteriori* (MAP) estimation, evidenced-based analysis, and variational Bayesian. Among these options, Bayesian inference is generally more effective in avoiding local minima than deterministic approaches such as MAP, due to the fact that Bayesian methods approximate the full posterior distributions instead of merely providing point estimates of the modes.

In this section we present an inference procedure based on mean field variational Bayes [35], [36] for approximating the posterior distributions. Let $q(\mathbf{z})$ denote the approximate posterior distribution. The goal is to minimize the Kullback-Leibler (KL) divergence between $q(\mathbf{z})$ and the true posterior distribution $p(\mathbf{z}|\mathbf{Y})$, given by

$$\begin{aligned} \mathrm{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{Y})) &= \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{Y})} \mathrm{d}\mathbf{z} \\ &= \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{Y}, \mathbf{z})} \mathrm{d}\mathbf{z} + \mathrm{C}, \end{aligned} \tag{18}$$

where $\mathrm{C} = \log p(\mathbf{Y})$ does not involve $\mathbf{z}$. To simplify notation, we use C to denote constants that do not depend on the variables currently under consideration, and C in different equations may have different meanings. The non-negative $\mathrm{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{Y}))$ measures the difference between $q(\mathbf{z})$ and $p(\mathbf{z}|\mathbf{Y})$ and equals 0 if and only if $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{Y})$. In this sense, minimizing the KL divergence with respect to $q(\mathbf{z})$ is equivalent to finding an approximation to the unknown $p(\mathbf{z}|\mathbf{Y})$.

Another motivation for minimizing the KL divergence is to derive a lower bound for the evidence of the observed data $p(\mathbf{Y})$, with marginalization performed over the unknown variables. This can be seen by decomposing $\log p(\mathbf{Y})$ as follows

$$\begin{aligned} \log p(\mathbf{Y}) &= \log \frac{p(\mathbf{z}|\mathbf{Y}) p(\mathbf{Y})}{p(\mathbf{z}|\mathbf{Y})} = \log \frac{p(\mathbf{Y}, \mathbf{z}) q(\mathbf{z})}{p(\mathbf{z}|\mathbf{Y}) q(\mathbf{z})} \\ &= \int q(\mathbf{z}) \log \frac{p(\mathbf{Y}, \mathbf{z}) q(\mathbf{z})}{p(\mathbf{z}|\mathbf{Y}) q(\mathbf{z})} \mathrm{d}\mathbf{z} \\ &= Q(q(\mathbf{z})) + \mathrm{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{Y})), \end{aligned} \tag{19}$$

where

$$Q(q(\mathbf{z})) = \int q(\mathbf{z}) \log \frac{p(\mathbf{Y}, \mathbf{z})}{q(\mathbf{z})} \mathrm{d}\mathbf{z} \tag{20}$$

is a lower bound of $\log p(\mathbf{Y})$ because $\mathrm{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{Y})) \geq 0$. Since $\log p(\mathbf{Y})$ does not depend on $q(\mathbf{z})$, we have

$$\begin{aligned} \underset{q(\mathbf{z})}{\operatorname{argmax}}\ Q(q(\mathbf{z})) &= \underset{q(\mathbf{z})}{\operatorname{argmin}}\ \mathrm{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{Y})) \\ &= \underset{q(\mathbf{z})}{\operatorname{argmin}} \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{Y}, \mathbf{z})} \mathrm{d}\mathbf{z}, \end{aligned} \tag{21}$$

where the second equality follows from (18).

In the mean field approach, we make the assumption on the factorization of the approximate posterior distributions

$$q(\mathbf{z}) = \prod_m q(\mathbf{z}_m), \tag{22}$$

where $\mathbf{z}_m$ denote the components of $\mathbf{z}$, as is given in (15).

With this factorization we have

$$\log q(\mathbf{z}_m) = \mathrm{E}_{\mathbf{z}\setminus\mathbf{z}_m}\left[\log p(\mathbf{Y}, \mathbf{z})\right] + \mathrm{C}, \tag{23}$$

from which we can determine the functional form of the approximate posterior of each unknown. The expectation $\mathrm{E}_{\mathbf{z}\setminus\mathbf{z}_m}[\cdot]$ in (23) is taken with respect to $q(\mathbf{z}\setminus\mathbf{z}_m)$. In the following we present the update rules resulting from this inference scheme (23) for each unknown.

A. Inference for $\mathbf{A}$ and $\mathbf{B}$

Letting $\mathbf{z}_m = \mathbf{A}$ and invoking (23), we have

$$\begin{aligned} \log q(\mathbf{A}) &= \mathrm{E}_{\mathbf{z}\setminus\mathbf{A}}[\log p(\mathbf{Y}, \mathbf{z})] + \mathrm{C} \\ &= \mathrm{E}_{\mathbf{z}\setminus\mathbf{A}}[\log p(\mathbf{Y}|\mathbf{A}, \mathbf{B}, \mathbf{E}, \beta) + \log p(\mathbf{A}|\boldsymbol{\gamma})] + \mathrm{C} \\ &= \sum_{i=1}^{L} \mathrm{E}_{\mathbf{z}\setminus\mathbf{A}}\left[-\frac{1}{2}\left(\beta||\mathbf{y}_{i\cdot} - \mathbf{a}_{i\cdot}\mathbf{B}^{\mathrm{T}} - \mathbf{r}_{i\cdot}\mathbf{E}||_2^2 + \mathbf{a}_{i\cdot}\boldsymbol{\Gamma}\mathbf{a}_{i\cdot}^{\mathrm{T}}\right)\right] + \mathrm{C} \\ &= \sum_{i=1}^{L} \log q(\mathbf{a}_{i\cdot}) + \mathrm{C}, \end{aligned} \tag{24}$$

where $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$. From (24) we see that the approximate posterior distribution of $\mathbf{A}$ decomposes as independent distributions of its rows. Moreover, by multiplying out the terms and taking the appropriate expectations, it follows that

$$\begin{aligned} \log q(\mathbf{a}_{i\cdot}) &= \mathrm{E}_{\mathbf{z}\setminus\mathbf{A}}\left[-\frac{1}{2}\left(\beta||\mathbf{y}_{i\cdot} - \mathbf{a}_{i\cdot}\mathbf{B}^{\mathrm{T}} - \mathbf{r}_{i\cdot}\mathbf{E}||_2^2 + \mathbf{a}_{i\cdot}\boldsymbol{\Gamma}\mathbf{a}_{i\cdot}^{\mathrm{T}}\right)\right] + \mathrm{C} \\ &= -\frac{1}{2}\left[(\mathbf{a}_{i\cdot} - \langle\mathbf{a}_{i}\rangle)\left(\boldsymbol{\Sigma}^{A}\right)^{-1}(\mathbf{a}_{i\cdot} - \langle\mathbf{a}_{i\cdot}\rangle)^{\mathrm{T}}\right] + \mathrm{C}\,. \end{aligned} \tag{25}$$

Recognizing the right hand side of (25) as the energy of a Gaussian distribution, we have

$$q(\mathbf{a}_{i\cdot}) = \mathcal{N}(\mathbf{a}_{i\cdot}|\langle\mathbf{a}_{i\cdot}\rangle, \mathbf{\Sigma}^A), \tag{26}$$

where

$$\langle\mathbf{a}_{i\cdot}\rangle^{\mathrm{T}} = \langle\beta\rangle\mathbf{\Sigma}^A\langle\mathbf{B}\rangle^{\mathrm{T}}(\mathbf{y}_{i\cdot} - \mathbf{r}_{i\cdot}\langle\mathbf{E}\rangle)^{\mathrm{T}} \tag{27}$$

and

$$\mathbf{\Sigma}^A = \left(\langle\beta\rangle\langle\mathbf{B}^{\mathrm{T}}\mathbf{B}\rangle + \langle\mathbf{\Gamma}\rangle\right)^{-1} \tag{28}$$

are the approximate posterior mean and covariance matrix, respectively. In (27) and (28) the expectations on the right hand side of the equalities are taken with respect to the most recent approximate posterior distributions of the respective terms.

Following steps similar to (24), we can see that $q(\mathbf{B})$ also decomposes into the product of Gaussian distributions $q(\mathbf{b}_{i\cdot})$, i.e.,

$$q(\mathbf{B}) = \prod_{i=1}^{T} q(\mathbf{b}_{i\cdot}) = \prod_{i=1}^{T} \mathcal{N}(\mathbf{b}_{i\cdot}|\langle\mathbf{b}_{i\cdot}\rangle, \mathbf{\Sigma}^B), \tag{29}$$

where

$$\langle\mathbf{b}_{i\cdot}\rangle^{\mathrm{T}} = \langle\beta\rangle\mathbf{\Sigma}^B\langle\mathbf{A}\rangle^{\mathrm{T}}(\mathbf{y}_{\cdot i} - \mathbf{R}\langle\mathbf{e}_{\cdot i}\rangle) \tag{30}$$

and

$$\mathbf{\Sigma}^B = \left(\langle\beta\rangle\langle\mathbf{A}^{\mathrm{T}}\mathbf{A}\rangle + \langle\mathbf{\Gamma}\rangle\right)^{-1} . \tag{31}$$

The required expectations in (28) and (31) can be found as

$$\begin{aligned}\langle\mathbf{A}^{\mathrm{T}}\mathbf{A}\rangle &= \langle\mathbf{A}\rangle^{\mathrm{T}}\langle\mathbf{A}\rangle + L\mathbf{\Sigma}^A \\ \langle\mathbf{B}^{\mathrm{T}}\mathbf{B}\rangle &= \langle\mathbf{B}\rangle^{\mathrm{T}}\langle\mathbf{B}\rangle + T\mathbf{\Sigma}^B .\end{aligned} \tag{32}$$

B. Inference for $\mathbf{E}$

Substituting $\mathbf{z}_m = \mathbf{E}$ into (23), we have

$$\begin{aligned}\log q(\mathbf{E}) &= \mathrm{E}_{\mathbf{z}\setminus\mathbf{E}}[\log p(\mathbf{Y}, \mathbf{z})] + \mathrm{C} \\ &= \mathrm{E}_{\mathbf{z}\setminus\mathbf{E}}[\log p(\mathbf{Y}|\mathbf{A}, \mathbf{B}, \mathbf{E}, \beta) + \log p(\mathbf{E}|\boldsymbol{\alpha})] + \mathrm{C} \\ &= \sum_{i=1}^{T} \mathrm{E}_{\mathbf{z}\setminus\mathbf{E}}\left[-\frac{1}{2}\left\{\beta||\mathbf{y}_{\cdot i} - \mathbf{A}\mathbf{b}_{i\cdot}^{\mathrm{T}} - \mathbf{R}\mathbf{e}_{\cdot i}||_2^2 + \mathbf{e}_{\cdot i}^{\mathrm{T}}\mathbf{\Omega}^i\mathbf{e}_{\cdot i}\right\}\right] + \mathrm{C},\end{aligned} \tag{33}$$

where $\mathbf{\Omega}^i = \mathrm{diag}([\alpha_{1i}, \alpha_{2i}, \cdots, \alpha_{Fi}])$ contains the prior precisions for the i^{th} column of $\mathbf{E}$.

From (33) we see that $q(\mathbf{E})$ factors into the product of the approximate posterior distributions $q(\mathbf{e}_{\cdot i})$, where

$$q(\mathbf{e}_{\cdot i}) = \mathcal{N}(\mathbf{e}_{\cdot i}|\langle\mathbf{e}_{\cdot i}\rangle, \mathbf{\Sigma}^{E_i}) \tag{34}$$

is a Gaussian distribution. With some algebra, we obtain that the approximate posterior mean and covariance matrix of (34) are given by

$$\langle\mathbf{e}_{\cdot i}\rangle = \langle\beta\rangle\mathbf{\Sigma}^{E_i}\mathbf{R}^{\mathrm{T}}\left(\mathbf{y}_{\cdot i} - \langle\mathbf{A}\rangle\langle\mathbf{b}_{i\cdot}\rangle^{\mathrm{T}}\right) \tag{35}$$

and

$$\mathbf{\Sigma}^{E_i} = \left(\langle\beta\rangle\mathbf{R}^{\mathrm{T}}\mathbf{R} + \langle\mathbf{\Omega}^i\rangle\right)^{-1} . \tag{36}$$

From (36) we see that although the elements of $\mathbf{e}_{\cdot i}$ are independent of each other in the prior distribution, they are correlated with each other in the posterior distribution. This is due to the coupling of $\mathbf{R}$, which in effect takes a weighted sum of a column in $\mathbf{E}$ and maps it into a single entry in the observation.

Note that when $\mathbf{R} = \mathbf{I}_F$, as is the case in robust principal component analysis, the elements in $\mathbf{E}$ are independent of each other and their distributions are single-variable Gaussians.

C. Estimation of hyperparameters $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$

By keeping $p(\mathbf{A}|\boldsymbol{\gamma})$, $p(\mathbf{B}|\boldsymbol{\gamma})$ and $p(\gamma)$ in (23), we have

$$\begin{aligned}\log q(\boldsymbol{\gamma}) &= \mathrm{E}_{\mathbf{z}\setminus\boldsymbol{\gamma}}[\log p(\mathbf{Y}, \mathbf{z})] + \mathrm{C} \\ &= \mathrm{E}_{\mathbf{z}\setminus\boldsymbol{\gamma}}[\log p(\mathbf{A}|\boldsymbol{\gamma}) + \log p(\mathbf{B}|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma})] + \mathrm{C} \\ &= \sum_{i=1}^{k} \mathrm{E}_{\mathbf{z}\setminus\boldsymbol{\gamma}}[\log p(\mathbf{a}_{\cdot i}|\gamma_i) + \log p(\mathbf{b}_{\cdot i}|\gamma_i) + \log p(\gamma_i)] + \mathrm{C} .\end{aligned} \tag{37}$$

Taking the logarithms of (7) and (8), we have

$$\begin{aligned}\log p(\mathbf{a}_{\cdot i}|\gamma_i) &= \frac{L}{2}\log\gamma_i - \frac{\gamma_i}{2}||\mathbf{a}_{\cdot i}||_2^2, \\ \log p(\mathbf{b}_{\cdot i}|\gamma_i) &= \frac{T}{2}\log\gamma_i - \frac{\gamma_i}{2}||\mathbf{b}_{\cdot i}||_2^2, \\ \log p(\gamma_i) &= (a-1)\log\gamma_i - b\gamma_i .\end{aligned} \tag{38}$$

Substituting (38) into (37), it follows that

$$\begin{aligned}\log q(\boldsymbol{\gamma}) &= \sum_{i=1}^{k} \mathrm{E}_{\mathbf{z}\setminus\boldsymbol{\gamma}}\left[\left(\frac{L+T+2a}{2} - 1\right)\log\gamma_i \right. \\ &\qquad \left. - \frac{||\mathbf{a}_{\cdot i}||_2^2 + ||\mathbf{b}_{\cdot i}||_2^2 + 2b}{2}\gamma_i\right] + \mathrm{C} \\ &= \sum_{i=1}^{k}\left[\left(\frac{L+T+2a}{2} - 1\right)\log\gamma_i \right. \\ &\qquad \left. - \frac{\langle||\mathbf{a}_{\cdot i}||_2^2\rangle + \langle||\mathbf{b}_{\cdot i}||_2^2\rangle + 2b}{2}\gamma_i\right] + \mathrm{C} \\ &= \sum_{i=1}^{k}\log q(\gamma_i) + \mathrm{C},\end{aligned} \tag{39}$$

where

$$q(\gamma_i) \propto \gamma_i^{\frac{L+T+2a}{2}-1}\exp\left(-\gamma_i\frac{\langle||\mathbf{a}_{\cdot i}||_2^2\rangle + \langle||\mathbf{b}_{\cdot i}||_2^2\rangle + 2b}{2}\right) . \tag{40}$$

is recognized as a Gamma distribution. After resolving terms $\langle||\mathbf{a}_{\cdot i}||_2^2\rangle$ and $\langle||\mathbf{b}_{\cdot i}||_2^2\rangle$, we have

$$\langle\gamma_i\rangle = \frac{L+T+2a}{||\langle\mathbf{a}_{\cdot i}\rangle||_2^2 + ||\langle\mathbf{b}_{\cdot i}\rangle||_2^2 + L\sigma_{ii}^A + T\sigma_{ii}^B + 2b}, \tag{41}$$

where σ_{ii}^A and σ_{ii}^B denote the $(i,i)^{\text{th}}$ element of $\mathbf{\Sigma}^A$ and $\mathbf{\Sigma}^B$, respectively.

Similarly, setting $\mathbf{z}_m = \boldsymbol{\alpha}$ in (23), it follows that

$$\begin{aligned}\log q(\boldsymbol{\alpha}) &= \mathrm{E}_{\mathbf{z}\setminus\boldsymbol{\alpha}}\left[\log p(\mathbf{Y}, \mathbf{z})\right] + \mathrm{C} \\ &= \mathrm{E}_{\mathbf{z}\setminus\boldsymbol{\alpha}}\left[\log p(\mathbf{E}|\boldsymbol{\alpha}) + \log p(\boldsymbol{\alpha})\right] + \mathrm{C} \\ &= \sum_{i=1}^{F}\sum_{j=1}^{T}\mathrm{E}_{\mathbf{z}\setminus\boldsymbol{\alpha}}\left[\log p(e_{ij}|\alpha_{ij}) + \log p(\alpha_{ij})\right] + \mathrm{C} \\ &= \sum_{i=1}^{F}\sum_{j=1}^{T}\mathrm{E}_{\mathbf{z}\setminus\boldsymbol{\alpha}}\left[\frac{1}{2}\log\alpha_{ij} - \frac{1}{2}e_{ij}^2\alpha_{ij} - \log\alpha_{ij}\right] + \mathrm{C} \\ &= \sum_{i=1}^{F}\sum_{j=1}^{T}\left[-\frac{1}{2}\log\alpha_{ij} - \frac{1}{2}\langle e_{ij}^2\rangle\alpha_{ij}\right] + \mathrm{C} \\ &= \sum_{i=1}^{F}\sum_{j=1}^{T}\log q(\alpha_{ij}) + \mathrm{C} .\end{aligned} \tag{42}$$

From (42) $q(\alpha_{ij})$ is recognized as a Gamma distribution with mean

$$\langle \alpha_{ij} \rangle = \frac{1}{\langle e_{ij}^2 \rangle} = \frac{1}{\langle e_{ij} \rangle^2 + (\sigma_{ii}^{E_j})^2}, \tag{43}$$

where $(\sigma_{ii}^{E_j})^2$ denotes the $(i,i)^{\text{th}}$ element in $\mathbf{\Sigma}^{E_j}$, the approximate posterior covariance matrix for the j^{th} column of $\mathbf{E}$.

D. Estimation of noise precision β

By including the components of $p(\mathbf{Y}, \mathbf{z})$ that depend on β in (23), it follows that

$$\begin{aligned} \log q(\beta) &= \mathrm{E}_{\mathbf{z}\setminus\beta} \left[\log p(\mathbf{Y}|\mathbf{A},\mathbf{B},\mathbf{E},\beta) + \log p(\beta)\right] + \mathrm{C} \\ &= \mathrm{E}_{\mathbf{z}\setminus\beta} \left[\frac{LT}{2}\log\beta - \frac{\beta}{2}||\mathbf{Y}-\mathbf{A}\mathbf{B}^{\mathrm{T}}-\mathbf{R}\mathbf{E}||_{\mathrm{F}}^2 - \log\beta\right] + \mathrm{C} \\ &= \left(\frac{LT}{2}-1\right)\log\beta - \frac{1}{2}\langle||\mathbf{Y}-\mathbf{A}\mathbf{B}^{\mathrm{T}}-\mathbf{R}\mathbf{E}||_{\mathrm{F}}^2\rangle\beta + \mathrm{C} \end{aligned} \tag{44}$$

Therefore, $q(\beta)$ is a Gamma distribution

$$q(\beta) \propto \beta^{\frac{LT}{2}-1} \exp\left\{-\frac{\langle||\mathbf{Y}-\mathbf{A}\mathbf{B}^{\mathrm{T}}-\mathbf{R}\mathbf{E}||_{\mathrm{F}}^2\rangle}{2}\beta\right\} \tag{45}$$

with mean given by

$$\langle\beta\rangle = \frac{LT}{\langle||\mathbf{Y}-\mathbf{A}\mathbf{B}^{\mathrm{T}}-\mathbf{R}\mathbf{E}||_{\mathrm{F}}^2\rangle}. \tag{46}$$

Eq. (46) intuitively makes sense because the denominator measures the energy of the dense noise and the entire expression is the reciprocal of the noise power (variance), which is exactly the definition of precision.

To evaluate the expectation in the denominator of (46), additional steps are needed. We present the result here while leaving the derivation to Appendix A, in order to make the text more readable. With some algebra, the denominator can be computed in closed form using the posterior means of other quantities as

$$\begin{aligned} \langle||\mathbf{Y}-\mathbf{A}\mathbf{B}^{\mathrm{T}}-\mathbf{R}\mathbf{E}||_{\mathrm{F}}^2\rangle &= ||\mathbf{Y}-\langle\mathbf{A}\rangle\langle\mathbf{B}\rangle^{\mathrm{T}}-\mathbf{R}\langle\mathbf{E}\rangle||_{\mathrm{F}}^2 \\ &+ T\mathrm{Tr}\left(\langle\mathbf{A}\rangle^{\mathrm{T}}\langle\mathbf{A}\rangle\mathbf{\Sigma}^B\right) + L\mathrm{Tr}\left(\langle\mathbf{B}\rangle^{\mathrm{T}}\langle\mathbf{B}\rangle\mathbf{\Sigma}^A\right) \\ &+ LT\mathrm{Tr}\left(\mathbf{\Sigma}^A\mathbf{\Sigma}^B\right) + \sum_{i=1}^{T}\mathrm{Tr}\left(\mathbf{R}\mathbf{\Sigma}^{E_i}\mathbf{R}^{\mathrm{T}}\right) \end{aligned} \tag{47}$$

E. Summary

The proposed variational Bayesian inference procedure estimates the posterior distribution of the unknowns iteratively, where in each iteration the algorithm first computes $q(\mathbf{A})$, $q(\mathbf{B})$ and $q(\mathbf{E})$ using (26), (29) and (34), respectively. Then the hyperparameters are estimated by taking their approximate posterior means according to (41), (43) and (46), respectively. A stopping criterion is that the relative change in the estimated $\langle\mathbf{E}\rangle$ falls below a pre-specified threshold.

V. Numerical Examples

In this section we test the proposed algorithm with numerical experiments. We consider both real life datasets to demonstrate the effectiveness of the proposed approach in solving practical problems, and simulated experiments to evaluate its performance under various conditions.

To make the discussion clear, we name the proposed algorithm that solves (1) as variational Bayesian based sparse estimator (VBSE). The algorithm presented in [21] is a special case of VBSE when $\mathbf{R} = \mathbf{I}_F$. In comparison, we consider two popular regularization based alternatives to validate the performance of VBSE and furthermore to demonstrate their respective merits and limitations. The algorithm presented in [10] is named as accelerated proximal gradient sparse estimator (APGSE). We have also extended the ALM algorithm, which is the state-of-the-art solver for the RPCA problem, to solve (1), and name it augmented Lagrange multiplier sparse estimator (ALMSE). The algorithmic details are left to Appendix B. Note that both APGSE and ALMSE have user parameters controlling the achieved sparsity in $\mathbf{E}$ and the estimated rank of $\mathbf{X}$, respectively. The selection of these parameters has direct impact on the performance of the regularization based approaches. In each of the experiments, we have determined the empirical optimal values of these parameters. The VBSE algorithm is free of user parameters.

F We consider two applications, namely the foreground detection in blurred and noisy video sequences and the detection of network anomalies, whose details have been presented in Sections II-B and II-C, respectively. The experimental description and performance evaluation of the proposed and alternative algorithms are given below.

A. Foreground Detection in Blurred and Noisy Video

In this experiments, the *CAVIAR* test video sequence [37] was used. A sample video frame is presented in Figure 3(a), showing the hallway in a shopping mall with people moving as the foreground. The original 180×140 frame was divided into 30 non-overlapping blocks of size 30×28, and each block was treated as the basic unit in the experiment. The pixels in each block were vectorized into a column of size $= 30 \times 28 = 840$, and $T = 100$ such columns were stacked into an $F \times T$ matrix $\mathbf{D}$. The $F \times F$ matrix $\mathbf{R}$ models pixel-wise blurring, where each output was computed as the average of the 13 inputs within its radius-2 neighborhood, as is illustrated in Figure 2. Dense Gaussian noise $\mathbf{N}$ was added to the blurred video $\mathbf{RD}$, resulting in the observed $\mathbf{Y}$ with an SNR value at 23.5 dB. A blurred and noisy sample frame is shown in Figure 3(b). Note that due to the presence of noise, the direct application of inverse filter will magnify the noise and render poor results. Therefore, approaches such as the proposed VBSE have to be employed to deal with the blurred and noisy data.

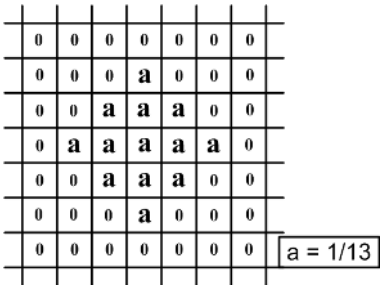
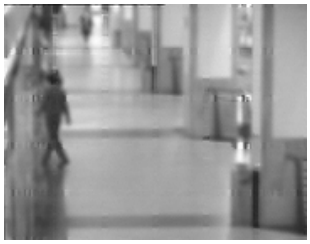


Figure. 2: Illustration of blurring kernel.

In order to obtain a reference for performance evaluation, we estimate the ground truth foreground $\mathbf{E}$ by applying VBSE to the original video $\mathbf{D}$ with $\mathbf{R} = \mathbf{I}_F$. This estimated ground truth, denoted as $\mathbf{E}_{\text{GT}}$ and shown in Figure 3(c), provides a reasonable representation of the moving FG objects. From the corresponding binarized FG map in Figure 3(d) we see that the FG contains two moving shoppers and their reflected images cast by the floor and the glass. Figures 3(c) and 3(d) serve as the reference for visually evaluating the performance of the algorithms.

(a) Original



(b) Blurred and noisy

(c) $\mathbf{E}_{\text{GT}}$

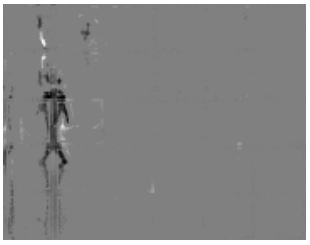
(d) FG map from $\mathbf{E}_{\text{GT}}$

Figure. 3: Sample frame of *CAVIAR* video sequence and estimated ground truth.

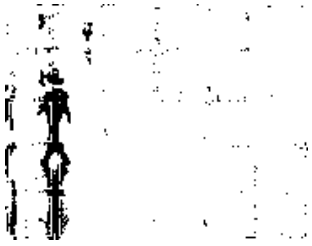
VBSE, APGSE and ALMSE were then applied on the blurred and noisy $\mathbf{Y}$, and the results are shown in Figures 4(a)-4(f). From the figures we see that both VBSE and its regularization based counterparts produce reasonable results that highlight the moving objects in the foreground. The VBSE algorithm gives cleaner FG maps, while the results of APGSE and ALMSE contain more isolated pixels falsely classified as foreground. The presence of such pixels may result in an increased false detection probability, if the FG maps are to be used subsequently for applications such as surveillance and intruder detection. Also note that the performance of APGSE and ALMSE has been optimized via the manual tuning of input parameters, while the VBSE approach is free of input parameters and is hence more amenable to automated deployment.

(a) $\mathbf{E}_{\text{VBSE}}$

(b) FG map from $\mathbf{E}_{\text{VBSE}}$

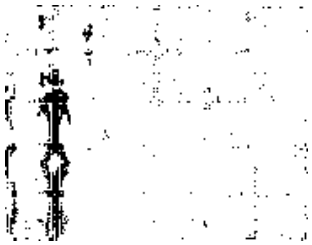


(c) $\mathbf{E}_{\text{APGSE}}$



(d) FG map from $\mathbf{E}_{\text{APGSE}}$

(e) $\mathbf{E}_{\text{ALMSE}}$



(f) FG map from $\mathbf{E}_{\text{ALMSE}}$

Figure. 4: Estimated foreground from *CAVIAR* video seqyebce

B. Detection of Network Anomalies

In this subsection we apply the VBSE algorithm to the network anomaly detection problem and examine its performance under various experimental conditions. First we consider the real life *Internet2* dataset [38], which records the composite OD flow traffic across the Abilene backbone network.

Figure 5 gives an illustration of the network, which consists of 11 nodes located at various cities across the United States, represented by red dots in the figure. The number of OD flows is given by $F = N^2 = 121$, including flows to and from the same nodes. A solid line denotes a pair of bi-directional links between two nodes, and there are a total of 15 such link pairs. In addition, there is a link from every node to itself. Therefore, the total number of links present in the network is $L = 2 \times 15 + 11 = 41$. A binary routing matrix $\mathbf{R}$ maps the $F = 121$ flows onto the $L = 41$ network links. In the experiment we took $T = 210$ temporal snapshots from the data. Given the $F \times T$ OD flow traffic $\mathbf{D}$, the link load $\mathbf{Y}$ is obtained through multiplication with $\mathbf{R}$. Note that besides obtaining $\mathbf{Y}$ from flow measurements, link loads can also be determined using the simple network management protocol (SNMP) traces [39].

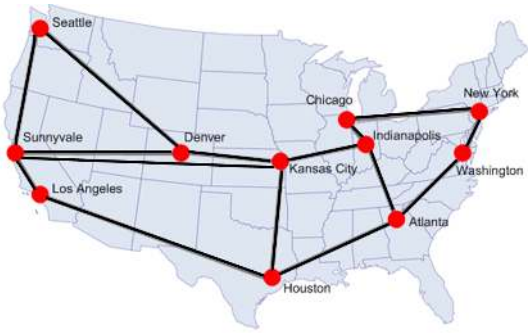


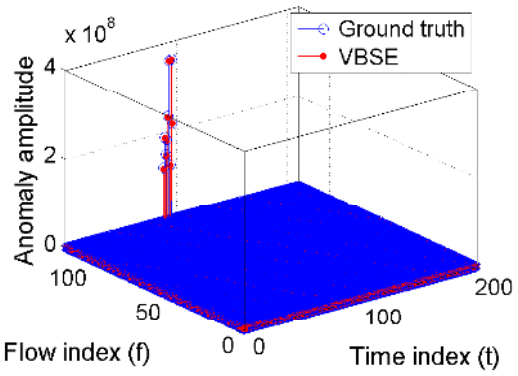
Figure. 5: *Internet2* backbone network map.

In order to quantitatively evaluate the performance of the algorithms, we need to estimate the ground truth from the data. This was done by applying the VBSE algorithm to $\mathbf{D}$ by setting $\mathbf{R} = \mathbf{I}_F$. As a validation, we also applied the APG [12] and ALM [13] algorithms to $\mathbf{D}$. These three algorithms produced comparable results, with the estimated anomalies comprising 9 numerically non-zero entries at the same spatio-temporal locations and the estimated clean traffic being of rank 4. We define as the estimated ground truth anomalies $\mathbf{E}_{\text{GT}}$ the average of the results produced by VBSE (with $\mathbf{R} = \mathbf{I}_F$), APG and ALM in what follows.

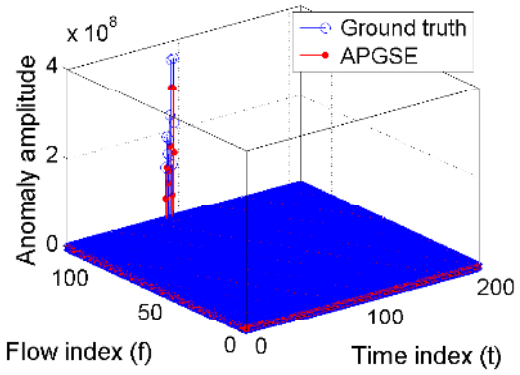
The link load data $\mathbf{Y}$ were fed into the algorithms for anomaly detection and amplitude estimation. Figure 6 shows the algorithmic results superimposed on the ground truth. As can be seen in the figures, VBSE detects all the anomalies present in the ground truth and accurately estimates their amplitudes. In contrast, APGSE tends to produce biased estimates and the amplitudes differ from the ground truth by a margin. ALMSE, although accurately estimating the amplitudes of anomalies present in the ground truth, has additionally detected many spurious anomalies. This will result in an increased false alarm probability, if ALMSE is used for network monitoring.

To further investigate the performance of the various algorithms, we considered additional numerical experiments. Specifically, we artificially added dense Gaussian noise to the link measurements. The performance of VBSE, APGSE and ALMSE at various SNR levels are recorded in Figures 7(a) and 7(b). As can be seen, VBSE is more robust to noise and gives uniformly lower estimation error than APGSE and ALMSE. VBSE is also able to precisely identify the number of anomalies, while APGSE yields a noticeably higher number of false detections.

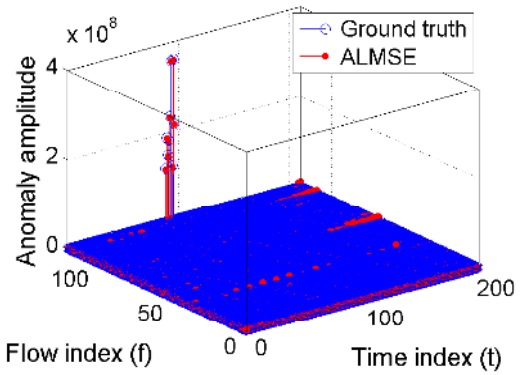
Finally, we carried out a separate simulated experiment to examine the performances at different anomaly densities (i.e., various degrees of sparsity of $\mathbf{E}$). The data generation process is described as follows. A low-rank $\mathbf{X}_0$ was simulated as the product of $F \times r$ and $r \times T$ matrices with i.i.d. entries drawn from $\mathcal{N}(0, 100/F)$ and $\mathcal{N}(0, 100/T)$, respectively. Synthesized $\mathbf{E}$ with amplitudes drawn from i.i.d. $\mathcal{U}(-10, 10)$ was added to $\mathbf{X}_0$. $\mathbf{R}$ is simulated as a random binary matrix with 95% of zeros. The anomaly density was varied across a wide range. As can be seen from Figures 8(a) and 8(b), VBSE gives uniformly lower estimation error and more accurate estimate of anomaly density than APGSE and ALMSE.



(a) $\mathbf{E}_{\text{VBSE}}$



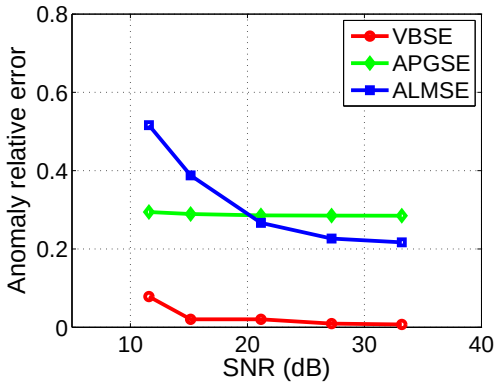
(b) $\mathbf{E}_{\text{APGSE}}$



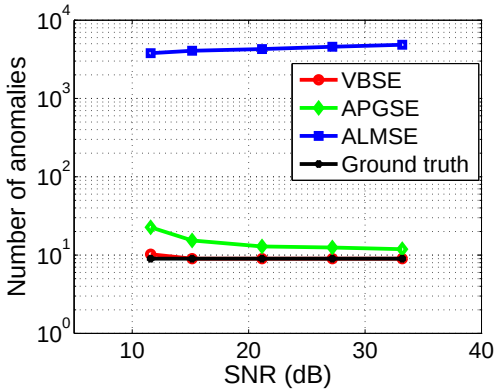
(c) $\mathbf{E}_{\text{ALMSE}}$

Figure. 6: Estimated flow anomalies from *Internet2* data.

The numerical examples presented above demonstrate the effectiveness of VBSE in solving real life problems. Compared with its regularization based counterparts, VBSE has satisfactory performance under a broad range of experimental conditions. Moreover, VBSE is free of user parameters and is hence especially amenable for automated deployment. Last but not least, VBSE provides the posterior distributions of the unknowns rather than only point estimates.

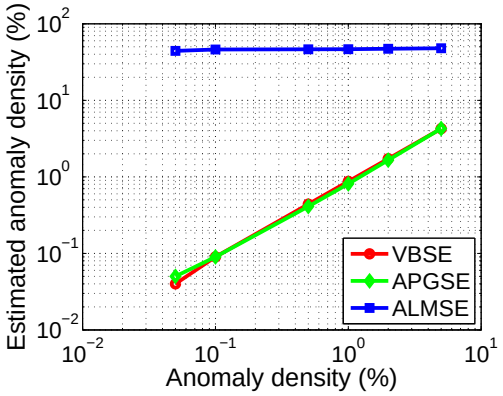


(a) Error v.s. SNR



(b) Sparsity v.s. SNR

Figure. 7: Performance of VBSE, APGSE and ALMSE at various SNR levels.



(b) Sparsity v.s. anomaly density

Figure. 8: Performance of VBSE, APGSE and ALMSE at various degrees of sparsity of $\mathbf{E}$.

VI. Conclusions

In this paper we proposed a variational Bayesian based algorithm for estimation of the sparse component from an outlier-corrupted low-rank matrix. A general data model was formulated and several specific application scenarios were discussed. The proposed algorithm is based on a hierarchical Bayesian model, and employs a variational approach for inference. The proposed algorithm is free of user parameters and its effectiveness was demonstrated using both real life and simulated experiments.

Appendix A
Derivation of Expected Noise Energy

The expected noise energy

$$W = \langle ||\mathbf{Y} - \mathbf{A}\mathbf{B}^{\mathrm{T}} - \mathbf{R}\mathbf{E}||_{\mathrm{F}}^2 \rangle \tag{A.1}$$

in the denominator of (46) can be derived as follows.

First, note that $\langle\cdot\rangle$ denotes the expectation taken with respect to $q(\mathbf{z}\backslash\beta)$, i.e., the approximate posterior distribution of all unknowns except β. Since the unknowns that W depends on are $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{E}$, the expectation reduces to

$$W = \mathrm{E}_{\mathbf{A},\mathbf{B},\mathbf{E}} \langle ||\mathbf{Y} - \mathbf{A}\mathbf{B}^{\mathrm{T}} - \mathbf{R}\mathbf{E}||_{\mathrm{F}}^2 \rangle \,. \tag{A.2}$$

From the definition of Frobenius norm, it follows from (A.1) that

$$\begin{aligned} W &= \langle ||\mathbf{Y} - \mathbf{A}\mathbf{B}^{\mathrm{T}} - \mathbf{R}\mathbf{E}||_{\mathrm{F}}^2 \rangle \\ &= \langle \mathrm{Tr}\left((\mathbf{Y} - \mathbf{A}\mathbf{B}^{\mathrm{T}} - \mathbf{R}\mathbf{E})^{\mathrm{T}}(\mathbf{Y} - \mathbf{A}\mathbf{B}^{\mathrm{T}} - \mathbf{R}\mathbf{E})\right) \rangle \\ &= \langle \mathrm{Tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}) \rangle - \langle \mathrm{Tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{A}\mathbf{B}^{\mathrm{T}}) \rangle - \langle \mathrm{Tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{R}\mathbf{E}) \rangle \\ &\quad - \langle \mathrm{Tr}(\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{Y}) \rangle + \langle \mathrm{Tr}(\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{B}^{\mathrm{T}}) \rangle + \langle \mathrm{Tr}(\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{R}\mathbf{E}) \rangle \\ &\quad - \langle \mathrm{Tr}(\mathbf{E}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{Y}) \rangle + \langle \mathrm{Tr}(\mathbf{E}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{A}\mathbf{B}^{\mathrm{T}}) \rangle + \langle \mathrm{Tr}(\mathbf{E}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{R}\mathbf{E}) \rangle \end{aligned} \tag{A.3}$$

Since both trace and expectation are linear operations, and the unknowns $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{E}$ are assumed to be independent of each other in the approximate posterior distribution (see (22)), (A.3) can be further written as

$$\begin{aligned} W &= \mathrm{Tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}) - \mathrm{Tr}(\mathbf{Y}^{\mathrm{T}}\langle\mathbf{A}\rangle\langle\mathbf{B}\rangle^{\mathrm{T}}) - \mathrm{Tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{R}\langle\mathbf{E}\rangle) \\ &\quad - \mathrm{Tr}(\langle\mathbf{B}\rangle\langle\mathbf{A}\rangle^{\mathrm{T}}\mathbf{Y}) + \mathrm{Tr}(\langle\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{B}^{\mathrm{T}}\rangle) + \mathrm{Tr}(\langle\mathbf{B}\rangle\langle\mathbf{A}\rangle^{\mathrm{T}}\mathbf{R}\langle\mathbf{E}\rangle) \\ &\quad - \mathrm{Tr}(\langle\mathbf{E}\rangle^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{Y}) + \mathrm{Tr}(\langle\mathbf{E}\rangle^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\langle\mathbf{A}\rangle\langle\mathbf{B}\rangle^{\mathrm{T}}) + \mathrm{Tr}(\langle\mathbf{E}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{R}\mathbf{E}\rangle) \\ &= ||\mathbf{Y} - \langle\mathbf{A}\rangle\langle\mathbf{B}\rangle^{\mathrm{T}} - \mathbf{R}\langle\mathbf{E}\rangle||_{\mathrm{F}}^2 - \mathrm{Tr}(\langle\mathbf{B}\rangle\langle\mathbf{A}\rangle^{\mathrm{T}}\langle\mathbf{A}\rangle\langle\mathbf{B}\rangle^{\mathrm{T}}) \\ &\quad - \mathrm{Tr}(\langle\mathbf{E}\rangle^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{R}\langle\mathbf{E}\rangle) + \mathrm{Tr}(\langle\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{B}^{\mathrm{T}}\rangle) + \mathrm{Tr}(\langle\mathbf{E}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{R}\mathbf{E}\rangle) \end{aligned} \tag{A.4}$$

For notational simplicity, denoting

$$\begin{aligned} S_1 &= \mathrm{Tr}(\langle\mathbf{B}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{B}^{\mathrm{T}}\rangle) - \mathrm{Tr}(\langle\mathbf{B}\rangle\langle\mathbf{A}\rangle^{\mathrm{T}}\langle\mathbf{A}\rangle\langle\mathbf{B}\rangle^{\mathrm{T}}) \\ S_2 &= \mathrm{Tr}(\langle\mathbf{E}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{R}\mathbf{E}\rangle) - \mathrm{Tr}(\langle\mathbf{E}\rangle^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{R}\langle\mathbf{E}\rangle), \end{aligned}$$

we have from (A.4) that

$$W = ||\mathbf{Y} - \langle\mathbf{A}\rangle\langle\mathbf{B}\rangle^{\mathrm{T}} - \mathbf{R}\langle\mathbf{E}\rangle||_{\mathrm{F}}^2 + S_1 + S_2 \,. \tag{A.5}$$

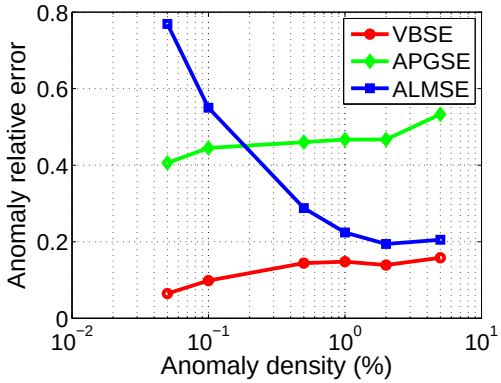
In (A.5), $||\mathbf{Y} - \langle\mathbf{A}\rangle\langle\mathbf{B}\rangle^{\mathrm{T}} - \mathbf{R}\langle\mathbf{E}\rangle||_{\mathrm{F}}^2$ can readily be computed using the most recent $q(\mathbf{A})$, $q(\mathbf{B})$ and $q(\mathbf{E})$. From [21] we have

$$S_1 = T\mathrm{Tr}(\langle\mathbf{A}\rangle^{\mathrm{T}}\langle\mathbf{A}\rangle\boldsymbol{\Sigma}^B) + L\mathrm{Tr}(\langle\mathbf{B}\rangle^{\mathrm{T}}\langle\mathbf{B}\rangle\boldsymbol{\Sigma}^A) + LT\mathrm{Tr}(\boldsymbol{\Sigma}^A\boldsymbol{\Sigma}^B). \tag{A.6}$$

Similarly, S_2 can be computed as

$$\begin{aligned} S_2 &= \mathrm{Tr}(\mathbf{R}\langle\mathbf{E}\mathbf{E}^{\mathrm{T}}\rangle\mathbf{R}^{\mathrm{T}}) - \mathrm{Tr}(\mathbf{R}\langle\mathbf{E}\rangle\langle\mathbf{E}\rangle^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}) \\ &= \mathrm{Tr}(\mathbf{R}(\langle\mathbf{E}\mathbf{E}^{\mathrm{T}}\rangle - \langle\mathbf{E}\rangle\langle\mathbf{E}\rangle^{\mathrm{T}})\mathbf{R}^{\mathrm{T}}) \\ &= \sum_{i=1}^{T} \mathrm{Tr}(\mathbf{R}\boldsymbol{\Sigma}^{E_i}\mathbf{R}^{\mathrm{T}}) \,. \end{aligned} \tag{A.7}$$

Substituting (A.6) and (A.7) into (A.5), we have the estimated noise energy, which appears as the denominator of (46).



(a) Error v.s. anomaly density

APPENDIX B
DERIVATION OF ALMSE

In this section we provide details on the ALMSE algorithm that solves the general data model (1). For the ALM algorithm that solves the more specific RPCA problem in (2), the reader is referred to [13].

Defining

$$f(\mathbf{X},\mathbf{E}) = ||\mathbf{X}||_*^2 + \lambda||\mathbf{E}||_1 \tag{B.1}$$

and

$$h(\mathbf{X},\mathbf{E}) = \mathbf{Y} - \mathbf{X} - \mathbf{R}\mathbf{E}, \tag{B.2}$$

the rank minimization and sparse estimation problem can be formulated as

$$\begin{array}{ll}\underset{\mathbf{X},\mathbf{E}}{\text{minimize}} & f(\mathbf{X},\mathbf{E}) \\ \text{subject to} & h(\mathbf{X},\mathbf{E}) = \mathbf{0},\end{array} \tag{B.3}$$

where λ is a regularization parameter controlling the relative weight placed on the low-rank term and the sparse term.

The augmented Lagrangian function based on (B.3) can be defined as

$$\begin{aligned}&\mathcal{L}(\mathbf{X},\mathbf{E},\mathbf{G},\mu)\\ &= f(\mathbf{X},\mathbf{E}) + \text{Tr}(\mathbf{G}^{\text{T}}h(\mathbf{X},\mathbf{E})) + \tfrac{\mu}{2}||h(\mathbf{X},\mathbf{E})||_{\text{F}}^2\\ &= f(\mathbf{X},\mathbf{E}) + g(\mathbf{X},\mathbf{E}),\end{aligned} \tag{B.4}$$

where $\mathbf{G} \in \mathbb{R}^{L\times T}$ contains the Lagrange multipliers and $g(\mathbf{X},\mathbf{E}) = \text{Tr}(\mathbf{G}^{\text{T}}h(\mathbf{X},\mathbf{E})) + \frac{\mu}{2}||h(\mathbf{X},\mathbf{E})||_{\text{F}}^2$ is a quadratic function in $\mathbf{X}$ and $\mathbf{E}$. The general method for augmented Lagrange multipliers is shown as Algorithm 3 in [13], and we adapt it below (and label it as Algorithm 1 in this paper) to facilitate the derivation that follows.

Algorithm 1 General Method of ALM (adapted from [13])

1: **Input**: $\mathbf{Y}$, λ
2: **while** not converged **do**
3: Solve $(\mathbf{X}_{k+1},\mathbf{E}_{k+1}) = \arg\min_{\mathbf{X},\mathbf{E}} \mathcal{L}(\mathbf{X},\mathbf{E},\mathbf{G}_k,\mu_k)$.
4: $\mathbf{G}_{k+1} = \mathbf{G}_k + \mu_k h(\mathbf{X}_{k+1},\mathbf{E}_{k+1})$.
5: Update μ_k to μ_{k+1}.
6: **end while**(loop-1 indexed by k)
7: **Output:** $\mathbf{X} = \mathbf{X}_{k^*}$, $\mathbf{E} = \mathbf{E}_{k^*}$.

The core of the algorithm is Line 3 in Algorithm 1. To minimize $\mathcal{L}$, we employ a block descent approach, i.e., breaking the joint minimization with respect to $(\mathbf{X},\mathbf{E})$ into cyclic minimizations with respect to $\mathbf{X}$ and $\mathbf{E}$. This cyclic block descent procedure replaces Line 3 in Algorithm 1 with an iterative loop, which we call loop-2 and index by j. The block descent update procedures are summarized as follows

Algorithm 2 Block descent minimization of $\mathcal{L}$

1: **while** not converged **do**
2: Solve $\mathbf{X}_{k+1}^{j+1} = \arg\min_{\mathbf{X}} \mathcal{L}(\mathbf{X},\mathbf{E}_{k+1}^{j},\mathbf{G}_k,\mu_k)$.
3: Solve $\mathbf{E}_{k+1}^{j+1} = \arg\min_{\mathbf{E}} \mathcal{L}(\mathbf{X}_{k+1}^{j+1},\mathbf{E},\mathbf{G}_k,\mu_k)$.
4: **end while**(loop-2 indexed by j)
5: **Output:** $\mathbf{X}_{k+1} = \mathbf{X}_{k+1}^{j^*}$, $\mathbf{E}_{k+1} = \mathbf{E}_{k+1}^{j^*}$.

Line 2 in Algorithm 2 can be solved analytically, following procedures similar to those in the original ALM algorithm. Specifically, by grouping terms independent of $\mathbf{X}$ in (B.4) into constant C, we have

$$\begin{aligned}&\mathcal{L}(\mathbf{X},\mathbf{E}_{k+1}^{j},\mathbf{G}_k,\mu_k) = ||\mathbf{X}||_* + g(\mathbf{X},\mathbf{E}_{k+1}^{j}) + \text{C}\\ &= ||\mathbf{X}||_* + \text{Tr}(\mathbf{G}_k^{\text{T}}(\mathbf{Y}-\mathbf{X})) + \frac{\mu_k}{2}||\mathbf{Y}-\mathbf{X}-\mathbf{R}\mathbf{E}_{k+1}^{j}|| + \text{C}\\ &= ||\mathbf{X}||_* + \frac{\mu_k}{2}||\mathbf{X} - (\mathbf{Y} - \mathbf{R}\mathbf{E}_{k+1}^{j} + \mu_k^{-1}\mathbf{G}_k)||_{\text{F}}^2 + \text{C}\,.\end{aligned} \tag{B.5}$$

It is well known that (B.5) can be minimized using the SVT [11] algorithm as follows

$$\mathbf{X}_{k+1}^{j+1} = \mathbf{U}\mathcal{S}_{\mu_k^{-1}}(\mathbf{\Sigma})\mathbf{V}^{\text{T}}, \tag{B.6}$$

where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\text{T}}$ is a singular value decomposition of $\mathbf{Y} - \mathbf{R}\mathbf{E}_{k+1}^{j} + \mu_k^{-1}\mathbf{G}_k$ and

$$\mathcal{S}_\epsilon(\cdot) = \text{sign}(\cdot)\max\{|\cdot| - \epsilon, 0\} \tag{B.7}$$

denotes the shrinkage thresholding operator with threshold ϵ. Note that $\mathbf{Y} - \mathbf{R}\mathbf{E}_{k+1}^{j} + \mu_k^{-1}\mathbf{G}_k$ is the minimizer of the quadratic function $g(\mathbf{X},\mathbf{E}_{k+1}^{j})$.

Following a similar procedure as above for $\mathbf{E}$, we have

$$\begin{aligned}&\mathcal{L}(\mathbf{X}_{k+1}^{j+1},\mathbf{E},\mathbf{G}_k,\mu_k) = \lambda||\mathbf{E}||_1 + g(\mathbf{X}_{k+1}^{j+1},\mathbf{E}) + \text{C}\\ &= \lambda||\mathbf{E}||_1 + \frac{\mu_k}{2}||\mathbf{R}\mathbf{E} - (\mathbf{Y} - \mathbf{X}_{k+1}^{j+1} + \mu_k^{-1}\mathbf{G}_k)||_{\text{F}}^2 + \text{C},\end{aligned} \tag{B.8}$$

In the original ALM algorithm, when $\mathbf{R} = \mathbf{I}_F$, the optimal $\mathbf{E}_{k+1}^{j+1}$ minimizing (B.8) can be found as

$$\mathbf{E}_{k+1}^{j+1} = \mathcal{S}_{\lambda/\mu_k}(\mathbf{Y} - \mathbf{X}_{k+1}^{j+1} + \mu_k^{-1}\mathbf{G}_k), \tag{B.9}$$

i.e., by determining the minimizer of $g(\mathbf{X}_{k+1}^{j+1},\mathbf{E})$ with respect to $\mathbf{E}$ and shrinkage thresholding the entries. However, for the general problem (1), the minimizer of the quadratic function $g(\mathbf{X}_{k+1}^{j+1},\mathbf{E})$ is not unique, due to the fact that $\mathbf{R}$ is wide and $\mathbf{R}^{\text{T}}\mathbf{R}$ is rank deficient. Instead of looking for an analytical solution, we resort to iterative approaches for this l_1-norm minimization problem. We choose the fast iterative shrinkage thresholding algorithm (FISTA) [40] for this purpose, because of its provably fast convergence and low-computational complexity. Since FISTA is a well known approach, we omit the details. Putting everything together, we have the (exact) ALMSE summarized in Algorithm 3.

In [13], an inexact ALM algorithm was derived, by degenerating the inner loop into one iteration. Here we adopt the same idea and speed up the ALMSE by making loop-2 transparent. The details are straightforward to obtain, and hence are not

Algorithm 3 The exact ALMSE algorithm

Input $\mathbf{Y}$, λ
Initialize $\mathbf{X}_0$, $\mathbf{E}_0$, $\mathbf{G}_0$, μ_0; $k = 0$;
while not converged **do**
  $\mathbf{X}_{k+1}^0 = \mathbf{X}_k$; $\mathbf{E}_{k+1}^0 = \mathbf{E}_k$; $j = 0$;
  while not converged **do**
    $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\text{T}} = \text{svd}(\mathbf{Y} - \mathbf{R}\mathbf{E}_{k+1}^{j} + \mu_k^{-1}\mathbf{G}_k)$;
    $\mathbf{X}_{k+1}^{j+1} = \mathbf{U}\mathcal{S}_{\mu_k^{-1}}(\mathbf{\Sigma})\mathbf{V}^{\text{T}}$;
    ${}^0\mathbf{E}_{k+1}^{j+1} = \mathbf{E}_{k+1}^{j}$; $l = 0$;
    while not converged **do**
      ${}^{l+1}\mathbf{E}_{k+1}^{j+1} = \text{FISTA}(\mathbf{X}_{k+1}^{j+1}, {}^{l}\mathbf{E}_{k+1}^{j}, \mathbf{G}_k, \mu_k)$;
      $l \leftarrow l + 1$;
    end while
    $l^* = l - 1$;
    $\mathbf{E}_{k+1}^{j+1} = {}^{l^*}\mathbf{E}_{k+1}^{j+1}$;
    $j \leftarrow j + 1$;
  end while
  $j^* = j - 1$
  $\mathbf{X}_{k+1} = \mathbf{X}_{k+1}^{j^*}$; $\mathbf{E}_{k+1} = \mathbf{E}_{k+1}^{j^*}$;
  $\mathbf{G}_{k+1} = \mathbf{G}_k + \mu_k h(\mathbf{X}_{k+1},\mathbf{E}_{k+1})$;
  Update μ_k to μ_{k+1};
  $k \leftarrow k + 1$
end while
$k^* = k - 1$
Output: $\mathbf{X} = \mathbf{X}_{k^*}$, $\mathbf{E} = \mathbf{E}_{k^*}$.

shown to save space. Unless otherwise noted, we have used the inexact version of ALMSE in our numerical examples.

REFERENCES

[1] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, May 2011.

[2] N. Chitradevi, V. Palanisamy, K. Baskaran, and U. B. Nisha, "Outlier aware data aggregation in distributed wireless sensor network using robust principal component analysis," in *International Conference on Computing Communication and Networking Technologies*, 2010, pp. 1–9.

[3] S. Lang, B. Zhao, S. Wang, X. Liu, and G. Fang, "Ionospheric ionogram denoising based on robust principal component analysis," in *International Conference on Systems and Informatics*, 2012, pp. 1956–1960.

[4] L. T. Duarte, E. Z. Nadalin, K. N. Filho, R. A. Zanetti, J. Romano, and M. Tygel, "Seismic wave separation by means of robust principal component analysis," in *Proceedings of the 20th European Signal Processing Conference*, 2012, pp. 1494–1498.

[5] P. Huang, S. D. Chen, P. Smaragdis, and M. A. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 57–60.

[6] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in Neural Information Processing Systems 22*, 2009, pp. 2080–2088.

[7] J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, "Towards a practical face recognition system: Robust registration and illumination by sparse representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 597–604.

[8] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: a probabilistic analysis," *Journal of Computer and System Sciences*, vol. 61, no. 2, pp. 217–235, 2000.

[9] Y. Peng, "RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 763–770.

[10] M. Mardani, G. Mateos, and G. Giannakis, "Unveiling anomalies in large-scale networks via sparsity and low rank," in *Proceedings of 45th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, November 2011.

[11] J. Cai, E. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, January 2010.

[12] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," in *International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, December 2009.

[13] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," *University of Illinois at Urbana-Champaign technical report #UILU-ENG-09-2215*, 2009.

[14] L. Wang, "An iterative algorithm for robust kernel principal component analysis," in *International Conference on Machine Learning and Cybernetics*, 2007, pp. 3484–3489.

[15] B. Bao, "Inductive robust principal component analysis," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3794–3800, 2012.

[16] G. Tang and A. Nehorai, "Robust principal component analysis based on low-rank and block-sparse matrix decomposition," in *45th Annual Conference on Information Sciences and Systems*, 2011, pp. 1–5.

[17] ——, "Constrained Cramér-Rao bound on robust principal component analysis," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 5070–5076, 2011.

[18] M. Ayazoglu, M. Sznaier, and O. I. Camps, "Fast algorithms for structured robust principal component analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1704–1711.

[19] J. Gai, Y. Li, and R. L. Stevenson, "Robust Bayesian PCA with Student's t-distribution: the variational inference approach," in *IEEE International Conference on Image Processing*, 2008, pp. 1340–1343.

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

[21] S. Babacan, M. Luessi, R. Molina, and A. Katsaggelos, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 3964–3977, August 2012.

[22] A. M. McIvor, "Background subtraction techniques," in *Proceedings of Image and Vision Computing New Zealand*, Auckland, New Zealand, 2000.

[23] M. Piccardi, "Background subtraction techniques: a review," in *IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, 2004, pp. 3099–3104.

[24] M. Dikmen and T. S. Huang, "Robust estimation of foreground in surveillance videos by sparse error estimation," in *19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[25] A. Patcha and J. Park, "An overview of anomaly detection techniques: existing solutions and latest technological trends," *Elsevier Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[26] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 15:1–15:58, 2009.

[27] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot, "Traffic matrix estimation: existing techniques and new directions," in *Proceedings of the 2002 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, 2002, pp. 161–174.

[28] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," in *Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems*, 2004, pp. 61–72.

[29] A. Abdelkefi, Y. Jiang, W. Wang, A. Aslebo, and O. Kvittem, "Robust traffic anomaly detection with principal component pursuit," in *Proceedings of the ACM CoNEXT Student Workshop*, 2010, pp. 10:1–10:2.

[30] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, 2004, pp. 219–230.

[31] B. I. P. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S. Lau, N. Taft, and J. D. Tygar, "Compromising PCA-based anomaly detectors for network-wide traffic," *University of California at Berkeley Technical Report No. UCB/EECS-2008-73*, 2008.

[32] A. L. Chistov and D. Y. Grigorév, "Complexity of quantifier elimination in the theory of algebraically closed fields," *Mathematical Foundations of Computer Science, Lecture Notes in Computer Science*, vol. 176, pp. 17–31, 1984.

[33] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal of Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[34] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[35] M. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.

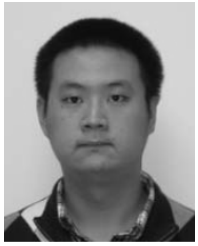
[36] C. Bishop, *Pattern recognition and machine learning*. Springer-Verlag, 2006.

[37] (2004, January) CAVIAR test case scenarios. [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

[38] (2013, January) Internet2 datasets. [Online]. Available: http://internet2.edu/observatory/archive/data-collections.html

[39] M. Thottan and C. Ji, "Anomaly detection in IP networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2191–2204, 2003.

[40] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.



Zhaofu Chen received the B.Sc. degree in information science and engineering from Zhejiang University, Hangzhou, China, in 2008, and the M.Sc. degree in electrical and computer engineering from University of Florida, Gainesville, FL, in 2010. He is currently a Ph.D. candidate in electrical engineering and computer science at Northwestern University, Evanston, IL.

In 2010, he joined the Image and Video Processing Lab at Northwestern University, where he is currently a graduate research assistant. His research interests include statistical analysis of sparse signals with applications in networking and multimedia.

Rafael Molina was born in 1957. He received the degree in mathematics (statistics) in 1979 and the Ph.D. degree in optimal design in linear models in 1983.

He became a Professor of Computer Science and Artificial Intelligence with the University of Granada, Granada, Spain, in 2000. He was a Former Dean of the Computer Engineering School, University of Granada, from 1992 to 2002, and the Head of the Computer Science and Artificial Intelligence Department, University of Granada, from 2005 to 2007. His current research interests include using Bayesian modeling and inference in problems like image restoration (applications to astronomy and medicine), super resolution of images and video, blind deconvolution, computational photography, source recovery in medicine, compressive sensing, low rank matrix decomposition, active learning, and classification.

Dr. Molina serves the IEEE and other professional societies: Applied Signal Processing, an Associate Editor from 2005 to 2007, the IEEE Transactions on Image Processing, an Associate Editor since 2010, Progress in Artificial Intelligence, an Associate Editor since 2011 and Digital Signal Processing, an Area Editor since 2011. He was a recipient of an IEEE International Conference on Image Processing Paper Award in 2007, an ISPA Best Paper Award in 2009 and Co-author of a Paper Awarded the Runner-Up Prize at Reception for early-stage researchers at the House of Commons.

Aggelos K. Katsaggelos received the Diploma degree in electrical and mechanical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, in 1979, and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 1981 and 1985, respectively.

He joined the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA, in 1985, where he is currently a Professor holder of the AT&T Chair. He was the holder of the Ameritech Chair of Information Technology from 1997 to 2003. He is the Director of the Motorola Center for Seamless Communications, a member of the Academic Affiliate Staff, NorthShore University Health System, an Affiliated Faculty with the Department of Linguistics, and he has an appointment with the Argonne National Laboratory. He is the editor of Digital Image Restoration (Springer-Verlag, 1991), co-author of Rate-Distortion Based Video Compression (Kluwer, 1997), co-editor of Recovery Techniques for Image and Video Compression and Transmission (Kluwer, 1998), and co-author of Super-Resolution for Images and Video (Claypool, 2007) and Joint Source-Channel Video Transmission (Claypool, 2007). He has published over five books, 180 journal papers, 450 conference papers, 40 book chapters, and 20 patents.

Dr. Katsaggelos has served the IEEE and other professional societies in many capacities; he was, for example, Editor-in-Chief of the IEEE Signal Processing Magazine from 1997 to 2002, a member of the Board of Governors of the IEEE Signal Processing Society from 1999 to 2001, and a member of the Publication Board of the IEEE Proceedings from 2003 to 2007. He was a recipient of the IEEE Third Millennium Medal in 2000, the IEEE Signal Processing Society Meritorious Service Award in 2001, the IEEE Signal Processing Society Technical Achievement Award in 2010, the IEEE Signal Processing Society Best Paper Award in 2001, the IEEE International Conference on Multimedia and Expo Paper Award in 2006, the IEEE International Conference on Image Processing Paper Award in 2007, and the ISPA Best Paper Award in 2009. He was a Distinguished Lecturer of the IEEE Signal Processing Society from 2007 to 2008 and he has been a Fellow of SPIE since 2009.