VARIATIONAL BAYESIAN INFERENCE FOR TENSOR ROBUST PRINCIPAL COMPONENT ANALYSIS *

CHAO WANG†, HUIWEN ZHENG†, RAYMOND CHAN‡, AND YOUWEI WEN§

Abstract. Tensor Robust Principal Component Analysis (TRPCA) holds a crucial position in machine learning and computer vision. It aims to recover underlying low-rank structures and to characterize the sparse structures of noise. Current approaches often encounter difficulties in accurately capturing the low-rank properties of tensors and balancing the trade-off between low-rank and sparse components, especially in a mixed-noise scenario. To address these challenges, we introduce a Bayesian framework for TRPCA, which integrates a low-rank tensor nuclear norm prior and a generalized sparsity-inducing prior. By embedding the priors within the Bayesian framework, our method can automatically determine the optimal tensor nuclear norm and achieve a balance between the nuclear norm and sparse components. Furthermore, our method can be efficiently extended to the weighted tensor nuclear norm model. Experiments conducted on synthetic and real-world datasets demonstrate the effectiveness and superiority of our method compared to state-of-the-art approaches.

Key words. Bayesian inference; tensor recovery; tensor nuclear norm; low rankness

MSC codes. 68Q25, 68R10, 68U05

1. Introduction.

With data becoming ubiquitous from diverse fields and applications, data structures are becoming increasingly complex with higher dimensions. Tensor, a multidimensional array, is an efficient data structure with broad applications, including machine learning [39] and computer vision [40]. Meanwhile, high-dimensional data always lie near a low-dimensional manifold, which can be interpreted by their low rank. In matrix processing, the low-rank assumption allows two-dimensional data recovery from incomplete or corrupted data [11]. However, expanding the low-rank concept from matrices to tensors remains an unresolved challenge. A main challenge in tensor analysis is that the tensor rank is not well defined. Various definitions of tensor rank have been proposed. For example, the CANDECOMP/PARAFAC (CP) rank, as described in [28], is based on the CP decomposition [25] and identifies the smallest number of rank-one tensors needed to represent a tensor. The Tucker rank [14], which stems from the Tucker decomposition [45], consists of a vector where each component corresponds to the rank of a matrix obtained by unfolding the original tensor. Furthermore, developments in tensor singular value decomposition (t-SVD) [27] have led to the tensor multi-rank [14] and tubal rank [26], both of which are analogous to the matrix singular value decomposition (SVD).

Among all these tensor applications, exploring low-rank features in sparse tensor decomposition has become essential, which is called Tensor Robust Principal Component Analysis (TRPCA) [33]. It extends the Robust Principal Component Analysis (RPCA) [24] from matrices to tensors. Specifically, TRPCA seeks to extract the low tubal rank component, $\mathcal{L}$, and eliminate the noise component, $\mathcal{S}$, derived from noisy

*Corresponding author: Youwei Wen.

Funding: This work was funded by the National Natural Science Foundation of China No. 12361089, 12571564, Guangdong Basic and Applied Research Foundation 2024A1515012347, HKRGC Grants No. LU13300125, LU11309922, ITF Grant No. MHP/054/22, and LU BGR 105824.

†Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen 518005, Guangdong Province, China (wangc6@sustech.edu.cn).

‡Lingnan University, Hong Kong SAR, China (raymond.chan@ln.edu.hk).

§Key Laboratory of Computing and Stochastic Mathematics (LCSM), School of Mathematics and Statistics, Hunan Normal University, Changsha, Hunan, China. (wenyouwei@gmail.com)

observations, $\mathcal{X}$, expressed as $\mathcal{X} = \mathcal{L} + \mathcal{S}$. This is achieved through the optimization process [33, 51, 36, 17, 48] described as

$$\min_{\mathcal{X}=\mathcal{L}+\mathcal{S}} \|\mathcal{L}\|_* + \lambda \|\mathcal{S}\|_1 . \tag{1.1}$$

where $\|\mathcal{L}\|_*$ is the tensor nuclear norm as the convex relaxation to a certain tensor rank. Note that minimizing the rank is an NP-hard problem. Various approximations have been proposed to approach different tensor ranks [23, 52, 37]. Here, $\|\mathcal{S}\|_1$ is the ℓ_1 norm of sparsity, and $\lambda > 0$ is the parameter used to balance low-rankedness and sparsity.

In the TRPCA model, we can further reformulate the equality constraint by a penalty term and turn the optimization model (1.1) into

$$\min_{\mathcal{S},\mathcal{L}} \tfrac{\theta_1}{2}\|\mathcal{X} - \mathcal{S} - \mathcal{L}\|_F^2 + \theta_2\|\mathcal{S}\|_1 + \theta_3\|\mathcal{L}\|_*, \tag{1.2}$$

where θ_1, θ_2 and θ_3 are tuning parameters. Note that (1.2) has broadened applications by assuming observation data is constructed not just by low-rank tensor and sparsity but also with certain bias or Gaussian noise, i.e.,

$$\mathcal{X} = \mathcal{L} + \mathcal{S} + \mathcal{E}, \tag{1.3}$$

where $\mathcal{E}$ is the corresponding bias and the Gaussian noise. This model is widely used in mixed noise removal [53, 55] and hyperspectral denoising [41].

The selection of the parameters in the model (1.1) and (1.2) is critical. Under the t-SVD framework, the optimal parameter for λ in (1.1) is suggested in [33] for the tensor nuclear norm. Nevertheless, it cannot be extended to other forms of tensor low-rank regularization, such as the weighted tensor nuclear norm. This issue becomes more serious when dealing with models involving multiple parameters in (1.2). Traditional parameter selection methods, including the discrepancy principle [35], L-curve [19], GCV [18], and RWP [1, 6], are often customized to specific regularization formulations and need iterative minimizations, which makes it inadequate for our tensor recovery problem in (1.2).

In this paper, we address the intricate task of simultaneously estimating tensors $\mathcal{L}$ and $\mathcal{S}$ and their regularization parameters θ_i for TRPCA. We introduce variational Bayesian inference (VBI) [13] as a powerful tool to tackle this challenge, reformulating the optimization problem within a Bayesian framework. By treating regularization parameters θ_i as hyperparameters, we apply the inherent strengths of Bayesian methods, popular for their success in inverse problems [46, 9, 22, 21, 16, 54, 30] and established applications in matrix and tensor problems like matrix completion [50], tensor completion [5, 44], and low-rank tensor approximation [34].

Despite these successes, the adoption of VBI in TRPCA remains limited. To our best knowledge, only [55] has explored VBI for TRPCA, employing a generalized sparsity-inducing prior. However, this method directly expresses the low-rank tensor as a t-product of two smaller factor tensors, presupposing the tubal rank, and models the sparse component $\mathcal{S}$ with independent Gaussian priors, which may not be optimal for sparse data. In contrast, we propose an approach that employs a tensor nuclear norm prior to $\mathcal{L}$, eliminating the need for predefined tensor ranks. For the sparse component $\mathcal{S}$, we adopt a Laplace prior, which better captures sparse structures. This reformulation enhances model flexibility, offering a more principled and less restrictive approach to tensor recovery, thereby mitigating limitations posed by prior assumptions on tensor ranks or sparsity patterns.

In comparison, joint maximum a posteriori (MAP) estimation minimizes the negative log posterior to obtain point estimates for $\mathcal{S}$, $\mathcal{L}$, and $\boldsymbol{\theta}$, simultaneously recovering tensors and parameters. Our VBI framework, however, approximates the full posterior distribution, enabling uncertainty quantification alongside point estimates. For practical applications such as denoising and background subtraction, we use the expectation of the variational distribution as the point estimate for $\mathcal{S}$ and $\mathcal{L}$, offering a robust and versatile approach to tensor recovery.

The primary contributions of this work are succinctly summarized as:

(1) Innovative Variational Bayesian Tensor Recovery Model: This paper proposes a novel variational Bayesian inference model for tensor recovery. It characterizes low-rank tensors using the tensor nuclear norm and sparse tensors via the Laplacian distribution. This approach enables simultaneous inference of both low-rank and sparse components along with the hyperparameters (regularization parameters), eliminating the need for pre-specifying the tensor rank.

(2) Efficient Inference via Laplacian Approximation and MM Framework: We introduce a Laplacian approximation methodology to tackle the computational intricacies associated with non-Gaussian posteriors arising from Laplace priors imposed on sparse tensor $\mathcal{S}$ and low-rank tensor $\mathcal{L}$. This approach directly tackles the ℓ_1 norm minimization and tensor nuclear norm minimization problems in estimating the expectations of sparse tensor $\mathcal{S}$ and low-rank tensor $\mathcal{L}$. For covariance matrix computation, it integrates with the Majorization-Minimization (MM) framework, deriving a tight lower bound for the non-quadratic distributions encountered in the ℓ_1 norm and tensor nuclear norm. This facilitates efficient variance computations, thereby significantly enhancing the efficiency and accuracy of inferring low-rank, sparse tensors as well as their hyperparameters.

The rest of this paper is organized as follows. In Section 2, we introduce the main preliminaries, including tensors and their decomposition. In Section 3, we describe the hierarchical Bayesian model, joint density, and hyperprior. In Section 4, we apply variational Bayesian inference to infer hyperparameters θ_i and solve the tensors $\mathcal{L}, \mathcal{S}$ at the same time. In Section 5, we provide the experimental results and show the superiority of our proposed methods. Finally, in Section 6, some conclusions are drawn.

2. Preliminaries.

This section provides an overview of fundamental notations and definitions that will be utilized throughout the paper.

2.1. Notations.

The set of natural numbers is denoted by $\mathbb{N}$, the set of real numbers by $\mathbb{R}$, and the set of complex numbers by $\mathbb{C}$. In the context of tensors, we adopt the convention of using boldface Euler script letters, exemplified by $\mathcal{A}$, to represent them. Matrices, on the other hand, are indicated with boldface uppercase letters, such as $\mathbf{A}$, with the identity matrix specifically denoted by $\mathbf{I}$. Vectors follow the convention of being written in boldface lowercase letters, like $\mathbf{a}$, whereas single values or scalars are represented by regular lowercase letters, for instance, a. Regarding indexing, for a vector $\mathbf{a}$, the i-th element is denoted by $\mathbf{a}_i$. For a matrix $\mathbf{A}$, $\mathbf{A}_{i:}$ signifies the i-th row, $\mathbf{A}_{:j}$ denotes the j-th column, and the specific element located at the intersection of the i-th row and j-th column is represented by either a_{ij} or, more commonly in matrix notation, $\mathbf{A}_{ij}$. When dealing with a third-order tensor $\mathcal{A}$, each element positioned at the intersection of the i-th, j-th, and k-th dimensions is denoted by a_{ijk} or, more conventionally for tensors, $\mathcal{A}_{ijk}$. This tensor can be dissected into distinct structural components: column fibers are designated as $\mathcal{A}_{:jk}$, row fibers as $\mathcal{A}_{i:k}$, and tube fibers as $\mathcal{A}_{ij:}$. Furthermore, the tensor can be analyzed through

various slices: horizontal slices are noted as $\mathcal{A}_{i::}$, lateral slices as $\mathcal{A}_{:j:}$, and frontal slices as $\mathcal{A}_{::k}$.

We define the inner product of matrices $\mathbf{A}$ and $\mathbf{B}$ as $\langle \mathbf{A}, \mathbf{B} \rangle := \mathrm{Tr}(\mathbf{A}^* \mathbf{B})$, where $\mathbf{A}^*$ is the conjugate transpose of $\mathbf{A}$, and $\mathrm{Tr}(\cdot)$ represents the trace of a matrix. If $\mathbf{A}$ consists only of real numbers, $\mathbf{A}^T$ denotes its transpose. The ℓ_2-norm of a vector $\mathbf{v}$ in the complex number space $\mathbb{C}^n$ is defined by $\|\mathbf{v}\|_2 = (\sum_i |\mathbf{v}_i|^2)^{1/2}$, measuring the vector's magnitude in Euclidean space.

The inner product between two tensors $\mathcal{A}$ and $\mathcal{B}$ in $\mathbb{C}^{n_1 \times n_2 \times n_3}$ is defined as $\langle \mathcal{A}, \mathcal{B} \rangle = \sum_{k=1}^{n_3} \langle \mathcal{A}_{::k}, \mathcal{B}_{::k} \rangle$. The complex conjugate of $\mathcal{A}$, which takes the complex conjugate of each entry of $\mathcal{A}$, is denoted as $\mathrm{conj}(\mathcal{A})$. The conjugate transpose of a tensor $\mathcal{A} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ is a tensor $\mathcal{A}^*$ obtained by conjugate transposing each of the frontal slices and then reversing the order of transposed frontal slices 2 through n_3. The tensor ℓ_1-norm of $\mathcal{A}$ is defined as $\|\mathcal{A}\|_1 = \sum_{ijk} |a_{ijk}|$, and the Frobenius norm as $\|\mathcal{A}\|_F = \sqrt{\sum_{ijl} |a_{ijk}|^2}$.

2.2. T-product and t-SVD. Before introducing the definitions, we define three operators:

$$\mathrm{unfold}(\mathcal{A}) = \begin{bmatrix} \mathcal{A}_{::1} \\ \mathcal{A}_{::2} \\ \vdots \\ \mathcal{A}_{::n_3} \end{bmatrix}, \ \mathrm{fold}\ (\mathrm{unfold}(\mathcal{A})) = \mathcal{A}, \tag{2.1}$$

and

$$\mathrm{bcirc}(\mathcal{A}) := \begin{bmatrix} \mathcal{A}_{::1} & \mathcal{A}_{::n_3} & \cdots & \mathcal{A}_{::2} \\ \mathcal{A}_{::2} & \mathcal{A}_{::1} & \cdots & \mathcal{A}_{::3} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{A}_{::n_3} & \mathcal{A}_{::,n_3-1} & \cdots & \mathcal{A}_{::1} \end{bmatrix} \in \mathbb{R}^{n_1 n_3 \times n_2 n_3}.$$

Here $\mathrm{unfold}(\cdot)$ maps $\mathcal{A}$ to a matrix of size $n_1 n_3 \times n_2$ and $\mathrm{fold}(\cdot)$ is its inverse operator. We introduce the notation $\mathbf{A} := \mathrm{bdiag}(\mathcal{A})$ to concisely represent the block diagonal matrix derived from the tensor $\mathcal{A}$. Here, $\mathrm{bdiag}(\cdot)$ designates the block diagonalization operator, with the i-th block corresponding to $\mathcal{A}_{::i}$.

Now, we focus on applying the Discrete Fourier Transformation (DFT) to tensors. We represent the tensor $\mathcal{A}$ transformed by DFT along its third (tubal) dimension as $\overline{\mathcal{A}}$. This transformation is executed using the MATLAB command fft, specifically performed as $\overline{\mathcal{A}} = \mathrm{fft}(\mathcal{A}, [], 3)$. Conversely, to revert the tensor to its original form from $\overline{\mathcal{A}}$, we use the inverse operation with $\mathcal{A} = \mathrm{ifft}(\overline{\mathcal{A}}, [], 3)$. We also introduce the notation $\overline{\mathbf{A}} := \mathrm{bdiag}(\overline{\mathcal{A}})$ to represent the block diagonal matrix constructed from the tensor $\overline{\mathcal{A}}$. Next, we introduce the definition of t-product.

DEFINITION 2.1. (t-product [27]). *Let $\mathcal{A} \in \mathbb{R}^{n_1 \times l \times n_3}$ and $\mathcal{B} \in \mathbb{R}^{l \times n_2 \times n_3}$, then the t-product $\mathcal{A} * \mathcal{B}$ is defined by*

$$\mathcal{A} * \mathcal{B} = \mathrm{fold}(\mathrm{bcirc}(\mathcal{A}) \cdot \mathrm{unfold}(\mathcal{B})), \tag{2.2}$$

resulting a tensor of size $n_1 \times n_2 \times n_3$. Note that $\mathcal{A} * \mathcal{B} = \mathcal{Z}$ if and only if $\overline{\mathbf{A}}\,\overline{\mathbf{B}} = \overline{\mathbf{Z}}$.

Using the t-product framework, we define the identity tensor $\mathcal{I} \in \mathbb{R}^{n \times n \times n_3}$ as a tensor with its first frontal slice being the $n \times n$ identity matrix, while all subsequent frontal slices consist entirely of zeros. It is clear that $\mathcal{A} * \mathcal{I} = \mathcal{A}$ and $\mathcal{I} * \mathcal{A} = \mathcal{A}$ given the appropriate dimensions. In addition, a tensor $\mathcal{H} \in \mathbb{R}^{n \times n \times n_3}$ is orthogonal if it

satisfies $\mathcal{H}^* * \mathcal{H} = \mathcal{H} * \mathcal{H}^* = \mathcal{I}$. Moreover, we call a tensor f-diagonal if each of its frontal slices is a diagonal matrix. Next, we define the tensor singular value decomposition as below:

DEFINITION 2.2. (tensor singular value decomposition: t-SVD [27]). *The t-SVD of $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is given by*

$$\mathcal{A} = \mathcal{U} * \mathcal{D} * \mathcal{V}^*, \tag{2.3}$$

where $\mathcal{U} \in \mathbb{R}^{n_1 \times n_1 \times n_3}, \mathcal{V} \in \mathbb{R}^{n_2 \times n_2 \times n_3}$ are orthogonal tensors, and $\mathcal{D} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is an f-diagonal tensor.

It follows from Definition 2.1 that $\mathcal{A} = \mathcal{U} * \mathcal{D} * \mathcal{V}^*$ if and only if $\overline{\mathbf{A}} = \overline{\mathbf{U}}\,\overline{\mathbf{D}}\,\overline{\mathbf{V}}^*$. For tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ with tubal rank r, we also have skinny t-SVD similar as matrix. Let r is the tubal rank of $\mathcal{A}$, the skinny t-SVD of $\mathcal{A}$ is $\mathcal{A} = \mathcal{U} * \mathcal{D} * \mathcal{V}^*$, where $\mathcal{U} \in \mathbb{R}^{n_1 \times r \times n_3}, \mathcal{D} \in \mathbb{R}^{r \times r \times n_3}, \mathcal{V} \in \mathbb{R}^{n_2 \times r \times n_3}$, in which $\mathcal{U}^* * \mathcal{U} = \mathcal{I}$ and $\mathcal{V}^* * \mathcal{V} = \mathcal{I}$.

DEFINITION 2.3. (tensor average rank and tubal rank [33]) *The tensor average rank of $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, denoted as $\mathrm{rank}_a(\mathcal{A})$, is defined as*

$$\mathrm{rank}_a(\mathcal{A}) = \frac{1}{n_3}\mathrm{rank}(\mathrm{bcirc}(\mathcal{A})) = \frac{1}{n_3}\sum_{i=1}^{n_3}\mathrm{rank}(\overline{\mathbf{A}}^{(i)}).$$

The tensor tubal rank, denoted as $\mathrm{rank}_t(\mathcal{A})$, is defined as the number of nonzero singular tubes of $\mathcal{S}$, where $\mathcal{S}$ comes from the t-SVD of $\mathcal{A}$, i.e. $\mathcal{A} = \mathcal{U} * \mathcal{S} * \mathcal{V}^*$. In other words, one has

$$\mathrm{rank}_t(\mathcal{A}) = \#\{i, \mathcal{S}(i,i,:) \neq 0\}.$$

For tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ with tubal rank r, we also have skinny t-SVD similar as matrix. Minimizing the tubal rank is an NP-hard problem; we introduce a tensor nuclear norm as a convex relaxation.

DEFINITION 2.4. (tensor nuclear norm [33]). *Let $\mathcal{A} = \mathcal{U} * \mathcal{D} * \mathcal{V}^*$ be the t-SVD of $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. Define $\sigma_{jk}(\mathcal{A})$ is the j-th singular value of $\overline{\mathcal{A}}_{::k}$, or simply σ_{jk} if the context is clear. The tensor nuclear norm (TNN) of $\mathcal{A}$ is defined as*

$$\|\mathcal{A}\|_* = \frac{1}{n_3}\sum_{k=1}^{n_3}\left\|\overline{\mathcal{A}}_{::k}\right\|_* = \frac{1}{n_3}\sum_{k=1}^{n_3}\sum_{j=1}^{\min(n_1,n_2)}\sigma_{jk}.$$

2.3. Probability distribution. Here, we define three kinds of probability distribution: the uniform distribution, the Gamma distribution, and the multivariate Gaussian distribution.

The uniform distribution is a distribution that assigns equal probability mass to a region. For $a, b \in R$ and $a < b$, the uniform distribution for a random variable $x \in R$ is defined as

$$p(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

The Gamma density function is given as

$$p(x) = \mathcal{G}(x|a,b) \propto x^{a-1}\exp(-bx), \tag{2.4}$$

where $a > 0$ and $b > 0$ represent shape and scale parameters respectively. We have its mean and variance of these Gamma distributions:

$$\mathbb{E}(x) = \frac{a}{b}, \quad \mathbb{V}\mathrm{ar}(x) = \frac{a}{b^2}. \tag{2.5}$$

The multivariate Gaussian distribution is fully characterized by a mean vector μ and a covariance matrix $\mathbf{\Sigma}$ and is defined as

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mu, \mathbf{\Sigma}) = (2\pi)^{-\frac{n}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2} \|\boldsymbol{x} - \mu\|^2_{\mathbf{\Sigma}^{-1}}\right),$$

where $\boldsymbol{x} \in \mathbb{R}^n$ is a random variable.

3. Bayesian model.

3.1. The likelihood. In (1.3), we assume the observed data $\mathcal{X}$ can be decomposed into three parts: $\mathcal{E}, \mathcal{S}, \mathcal{L}$. Note that $\mathcal{E}$'s elements as independent and identically distributed (i.i.d.) from a zero-mean normal distribution with precision θ_1. Then we obtain the likelihood function $p(\mathcal{X}|\mathcal{S}, \mathcal{L}, \theta_1)$ characterizes the probability of observing $\mathcal{X}$ conditioned on $\mathcal{S}$, $\mathcal{L}$, and θ_1. By exploiting the properties of the normal distribution, the likelihood function is expressed as:

$$p(\mathcal{X}|\mathcal{S}, \mathcal{L}, \theta_1) \propto \theta_1^{\frac{n}{2}} \exp\left(-\frac{\theta_1}{2} \|\mathcal{X} - \mathcal{S} - \mathcal{L}\|_F^2\right), \tag{3.1}$$

where $\propto$ denotes "proportional to" and $n = n_1 n_2 n_3$ denotes the total flattened dimensionality of $\mathcal{X}$. This formulation captures the probabilistic nature of the constraint violation, enhancing the robustness and applicability of the Bayesian inference process.

To facilitate further analysis and optimization, we consider the log-likelihood function

$$\log p(\mathcal{X}|\mathcal{S}, \mathcal{L}, \theta_1) = -\frac{\theta_1}{2} \|\mathcal{X} - \mathcal{S} - \mathcal{L}\|_F^2 + \frac{n}{2} \log \theta_1 + C_1,$$

where C_1 is a constant term that does not depend on $\mathcal{S}$, $\mathcal{L}$, or θ_1 and can be ignored in inference procedure.

3.2. The prior distributions. In Bayesian inference, the selection of prior distributions is a fundamental step that shapes the posterior beliefs about the unknown parameters. These priors encode our prior knowledge or assumptions about the variables of interest. Here, we choose appropriate prior distributions for $\mathcal{S}$ and $\mathcal{L}$, which represent distinct latent variables with unique characteristics. We remark that the choice of these priors is informed by regularization terms.

3.2.1. Prior distribution for $\mathcal{S}$. For the sparse component $\mathcal{S}$, we employ a Laplace prior distribution that induces ℓ_1-norm regularization. This choice is motivated by the well-established connection between Laplace priors and sparsity promotion in the Bayesian framework [42]. Specifically, the prior density takes the form:

$$p(\mathcal{S}|\theta_2) \propto \theta_2^n \exp\left(-\theta_2 \|\mathcal{S}\|_1\right), \tag{3.2}$$

where $\theta_2 > 0$ is a scale parameter. The ℓ_1-norm arises naturally as the convex envelope of the ℓ_0 pseudo-norm, making it the tightest convex relaxation for sparse recovery problems. From a probabilistic perspective, this corresponds to assuming independent exponentially distributed entries in $\mathcal{S}$, which favors exact zeros in the MAP estimate while maintaining computational tractability through convex optimization.

Taking the logarithm of the prior distribution, we obtain:

$$\log p(\mathcal{S}|\theta_2) = -\theta_2 \|\mathcal{S}\|_1 + n \log \theta_2 + C_2, \tag{3.3}$$

where C_2 represents a constant term that does not depend on $\mathcal{S}$ or θ_2.

3.2.2. Prior distribution for $\mathcal{L}$. For the variable $\mathcal{L}$, we employ a particular Gibbs prior [29] to promote a low-rank structure in $\mathcal{L}$. This prior takes the form of an exponential distribution with a tensor nuclear norm penalty, acting as a convex surrogate for the tensor average rank. It encourages $\mathcal{L}$ to have a low-rank representation, which is often suitable for capturing the underlying low-dimensionality in the data. The prior distribution is given by:

$$p(\mathcal{L}|\theta_3) \propto \theta_3^n \exp\left(-\theta_3 \|\mathcal{L}\|_*\right). \tag{3.4}$$

This characteristic encourages the low rank property in $\mathcal{L}$ and is coherent with the regularization term $\|\mathcal{L}\|_*$ in (1.2). Taking the logarithm of the prior distribution, we have:

$$\log p(\mathcal{L}|\theta_3) = -\theta_3\|\mathcal{L}\|_* + n\log\theta_3 + C_3,$$

where C_3 is a constant term that does not depend on $\mathcal{L}$ or θ_3.

3.3. The hyper-prior distribution. In the field of statistical modeling, the Gamma distribution has obtained significant attention as a versatile prior distribution for hyperparameters, particularly in Bayesian frameworks [2, 3, 4, 38, 43]. The choice of a Gamma distribution as the prior for the hyperparameter θ_i is driven by two key reasons. First, it serves as a conjugate prior for precision parameters in exponential family distributions. For instance, when θ_i controls the precision of a Gaussian likelihood $p(x|\theta_i) \sim \mathcal{N}(0, \theta_i^{-1})$, the Gamma prior ensures the posterior distribution remains a Gamma distribution. This conjugacy simplifies posterior calculations in Bayesian inference, enabling efficient automatic updates of hyperparameters. Second, θ_i typically represents positive physical quantities like precision or rate. The Gamma distribution's support on $(0, +\infty)$ naturally aligns with this positivity constraint, eliminating the need for artificial non-negativity restrictions.

We assign independent Gamma priors to the hyperparameters θ_i, which correspond to the mutually independent components $\mathcal{E}$, $\mathcal{S}$, and $\mathcal{L}$ in the model. This hierarchical structure preserves model consistency while enabling efficient computation. The independence assumption further facilitates automatic feature selection by factorizing the posterior distribution into marginal products over each θ_i. Hence, we have

$$p(\theta_i) = \mathcal{G}(\theta_i|a_{\theta_i}, b_{\theta_i}), i = 1, 2, 3,$$

where a_{θ_i} and b_{θ_i} are the shape and scale parameters for each hyperparameter θ_i. However, a key challenge in adopting the Gamma prior lies in the determination of optimal values for a_{θ_i} and b_{θ_i}. In the absence of strong prior knowledge, researchers often resort to weakly informative or non-informative priors, where the influence of the prior is minimized [4, 2, 38, 3, 43]. This can be achieved by setting extremely small values for a_{θ_i} and b_{θ_i} (e.g., $a_{\theta_i} = b_{\theta_i} = 10^{-4}$), thereby adopting an improper prior [43].

3.4. Joint distribution. The estimation of the unknown tensors $\mathcal{L}$ and $\mathcal{S}$, given the parameters $\theta_i (i = 1, 2, 3)$, can be tackled within the Maximum A Posteriori (MAP) estimation framework. This approach aims to maximize the posterior density $p(\mathcal{S}, \mathcal{L}|\mathcal{X}, \boldsymbol{\theta})$ with respect to $\mathcal{L}$ and $\mathcal{S}$, which is formulated as:

$$(\mathcal{S}^\dagger, \mathcal{L}^\dagger) = \arg\max_{\mathcal{S},\mathcal{L}} p(\mathcal{S}, \mathcal{L}|\mathcal{X}, \boldsymbol{\theta}).$$

Applying Bayes' theorem, the maximization problem can be rewritten in terms of the likelihood function $p(\mathcal{X}|\mathcal{S},\mathcal{L},\boldsymbol{\theta})$ and the prior densities $p(\mathcal{L}|\boldsymbol{\theta})$ and $p(\mathcal{S}|\boldsymbol{\theta})$:

$$\arg\max_{\mathcal{S},\mathcal{L}} p(\mathcal{X}|\mathcal{S},\mathcal{L},\boldsymbol{\theta})p(\mathcal{L}|\boldsymbol{\theta})p(\mathcal{S}|\boldsymbol{\theta}).$$

We remark that, in the MAP framework, the hyperparameters $\boldsymbol{\theta}$ must be either pre-specified or estimated prior to the estimation of $\mathcal{L}$ and $\mathcal{S}$. For a more comprehensive estimation that includes the hyperparameters, the joint maximum a posteriori (JMAP) estimation is employed:

$$(\mathcal{S}^\dagger,\mathcal{L}^\dagger,\boldsymbol{\theta}^\dagger) = \arg\max_{\mathcal{S},\mathcal{L},\boldsymbol{\theta}} p(\mathcal{S},\mathcal{L},\boldsymbol{\theta}|\mathcal{X}) = \operatorname*{argmax}_{\mathcal{S},\mathcal{L},\boldsymbol{\theta}} \frac{p(\mathcal{X},\mathcal{S},\mathcal{L},\boldsymbol{\theta})}{p(\mathcal{X})}. \tag{3.5}$$

For simplicity, we assume independence among the hyperparameters, allowing us to express the joint density function of the variables $\mathcal{X}$, $\mathcal{S}$, $\mathcal{L}$, and $\boldsymbol{\theta}$ as:

$$p(\mathcal{X},\mathcal{S},\mathcal{L},\boldsymbol{\theta}) = p(\mathcal{X}|\mathcal{S},\mathcal{L},\boldsymbol{\theta})p(\mathcal{L}|\theta_3)p(\mathcal{S}|\theta_2)p(\theta_1)p(\theta_2)p(\theta_3).$$

In the literature [4, 2, 38, 3, 43], the Gamma distribution is commonly adopted as a prior for the hyperparameters θ_i $(i = 1,2,3)$ due to its conjugacy with certain likelihood functions, which facilitates analytical tractability in Bayesian inference. However, prior knowledge about the shape and scale parameters (a_{θ_i} and b_{θ_i}) of the Gamma distribution is often lacking. To address this, a non-informative prior can be implemented by adopting an improper uniform prior distribution, defined as $p(x) \propto 1$ for $x \in \{\theta_i \mid i = 1,2,3\}$ over the positive real line [43, 15]. Hence we have

$$p(\mathcal{X},\mathcal{S},\mathcal{L},\boldsymbol{\theta}) \propto \theta_1^{n/2}\theta_2^n\theta_3^n \exp\left(-\frac{\theta_1}{2}\left\|\mathcal{X}-\mathcal{S}-\mathcal{L}\right\|_F^2 - \theta_2\left\|\mathcal{S}\right\|_1 - \theta_3\left\|\mathcal{L}\right\|_*\right). \tag{3.6}$$

4. Variational Bayesian inference. In Bayesian modeling, inference involves conditioning on observed data $\mathcal{X}$ and estimating the posterior density $p(\mathcal{S},\mathcal{L},\boldsymbol{\theta}|\mathcal{X})$. This task can be tackled via Markov Chain Monte Carlo (MCMC) sampling or optimization approaches. However, in this paper, we adopt variational inference as the methodological framework to approximate the latent variables $\mathcal{L}$ and $\mathcal{S}$, along with the parameter vector $\boldsymbol{\theta}$.

4.1. Kullback-Leibler divergence and evidence lower bound. The central goal of variational inference is to identify an optimal variational density $q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})$ that closely approximates the posterior density $p(\mathcal{S},\mathcal{L},\boldsymbol{\theta}|\mathcal{X})$, thereby facilitating efficient inference on the latent variables and parameters [8].

Within this framework, we define a family of densities $\mathcal{Q}$ over the latent variables and parameters. Each candidate $q(\mathcal{S},\mathcal{L},\boldsymbol{\theta}) \in \mathcal{Q}$ represents an approximation to the true posterior. The optimal candidate is chosen by minimizing the Kullback-Leibler (KL) divergence from the true posterior:

$$\begin{aligned}\mathrm{KL}(q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})\|p(\mathcal{S},\mathcal{L},\boldsymbol{\theta}|\mathcal{X})) &= \int_{\mathcal{S},\mathcal{L},\boldsymbol{\theta}} q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})\log\left(\frac{q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})}{p(\mathcal{S},\mathcal{L},\boldsymbol{\theta}|\mathcal{X})}\right) d\mathcal{L}\,d\mathcal{S}\,d\boldsymbol{\theta} \\ &= \mathbb{E}_{q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})}\left[\log\left(\frac{q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})}{p(\mathcal{S},\mathcal{L},\boldsymbol{\theta}|\mathcal{X})}\right)\right].\end{aligned} \tag{4.1}$$

The variational inference task simplifies to finding the variational density $q^\dagger(\mathcal{S},\mathcal{L},\boldsymbol{\theta})$ that minimizes the Kullback-Leibler (KL) divergence from the variational density to

the true posterior:

$$q^\dagger(\mathcal{S},\mathcal{L},\boldsymbol{\theta}) = \underset{q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})\in\mathcal{Q}}{\operatorname{argmin}} \ \mathrm{KL}(q(\mathcal{S},\mathcal{L},\boldsymbol{\theta}) \,\|\, p(\mathcal{S},\mathcal{L},\boldsymbol{\theta}|\mathcal{X})).$$

According to (3.5), the posterior density $p(\mathcal{S},\mathcal{L},\boldsymbol{\theta}|\mathcal{X})$ is the ratio between $p(\mathcal{X},\mathcal{S},\mathcal{L},\boldsymbol{\theta})$ and $p(\mathcal{X})$. The density $p(\mathcal{X})$ involves integrating out the latent variables from the joint density. Unfortunately, this integration is often intractable, rendering direct computation of the posterior challenging. Expand the condition density, we have

$$\mathrm{KL}(q(\mathcal{S},\mathcal{L},\boldsymbol{\theta}) \,\|\, p(\mathcal{S},\mathcal{L},\boldsymbol{\theta}|\mathcal{X})) = -\mathbb{E}_{q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})}\left[\log\left(\frac{p(\mathcal{X},\mathcal{S},\mathcal{L},\boldsymbol{\theta})}{q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})}\right)\right] + \log p(\mathcal{X}).$$

The second term is independent of latent variables and hyperparameters; therefore, it's just a constant in the minimization problem, and we can ignore this term. To circumvent the intractability, we optimize an alternative objective that is equivalent to the KL divergence up to an additive constant. Specifically, we minimize the first term on the right-hand side of the equation, which constitutes the evidence lower bound (ELBO), denoted $\mathcal{J}(q(\mathcal{S},\mathcal{L},\boldsymbol{\theta}))$

$$\mathcal{J}(q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})) \equiv \mathbb{E}_{q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})}\left[\log\left(\frac{p(\mathcal{X},\mathcal{S},\mathcal{L},\boldsymbol{\theta})}{q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})}\right)\right]. \tag{4.2}$$

This is

$$q^\dagger(\mathcal{S},\mathcal{L},\boldsymbol{\theta}) = \underset{q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})\in\mathcal{Q}}{\operatorname{argmax}} \ \mathcal{J}(q(\mathcal{S},\mathcal{L},\boldsymbol{\theta})). \tag{4.3}$$

4.2. Mean-field variational family. To fully specify the optimization problem, we now consider the variational family. The complexity of this family directly impacts the difficulty of the optimization, with more intricate families posing greater challenges.

In this paper, we concentrate on the mean-field variational family, which assumes mutual independence among the latent variables, with each variable being governed by its individual variational factor [8]. This assumption simplifies the variational density into a factorized form:

$$q(\mathcal{S},\mathcal{L},\boldsymbol{\theta}) = q(\mathcal{L})q(\mathcal{S})\prod_{i=1}^{3} q(\theta_i).$$

The selection of variational densities $q(\mathcal{L})$, $q(\mathcal{S})$, and $q(\theta_i)$ is importance. For $q(\mathcal{L})$ and $q(\mathcal{S})$, we adopt normal distributions due to their versatility and analytical convenience. The choice of the variational density $q(\theta_i)$ as a Gamma distribution is motivated by the conjugacy properties derived from the likelihood function (Eq. 3.1) and the prior distributions specified in Eqs. (3.2) and (3.4). These indicate that the posterior distribution of θ_1 and the conditional posteriors of θ_i for $i = 2, 3$ follow Gamma distributions. Since the Gamma distribution is conjugate to itself, selecting $q(\theta_i)$ as a Gamma density ensures compatibility with the posterior, facilitating efficient variational inference.

Let $\mathcal{Q}_\mathcal{G}$ denote the set of Gamma densities for the hyperparameters θ_i ($i = 1, 2, 3$), and $\mathcal{Q}_\mathcal{N}$ denote the set of multivariate normal densities over the tensor space $\mathbb{R}^{n_1\times n_2\times n_3}$. The overall variational family $\mathcal{Q}$ can be expressed as the Cartesian product of these sets: $\mathcal{Q} = \mathcal{Q}_\mathcal{N} \times \mathcal{Q}_\mathcal{N} \times \mathcal{Q}_\mathcal{G}$.

4.3. Laplacian approximation. In (3.6), the non-quadratic properties inherent in both the ℓ_1 norm of $\mathcal{S}$, which represents the sum of the absolute values of all elements, and the tensor nuclear norm of $\mathcal{L}$, which is the weighted sum of its singular values, pose significant obstacles for direct optimization within standard density families. These non-quadraticities complicate direct inference procedures, rendering them computationally intractable. To address this, we utilize the Laplace approximation method, involving mean calculation, variance estimation, and density function construction, to approximate the density with a Gaussian distribution form.

Here, we consider a general density function $q(x)$ with a single random variable x and simplify (4.3) as

$$q^\dagger(x) = \operatorname*{argmax}_{q(x)\in\mathcal{Q}_\mathcal{N}} \int_\Omega q(x) \log \frac{p(x)}{q(x)} dx.$$

where $\mathcal{Q}_\mathcal{N}$ is the set of all the density functions for the Gaussian distribution. According to Gibbs' inequality, for any two probability distributions $q(x)$ and $p(x)$ over a domain Ω, the following holds:

$$\int_\Omega q(x) \log \frac{p(x)}{q(x)} dx \leq 0,$$

with equality achieved if and only if $q(x) = p(x)$, implying identical means and variances. However, the non-quadratic nature of the $\log p(x)$ term complicates the direct estimation of $q(x)$ in practice. To address this, we employ the Laplacian approximation method to estimate $q(x)$. Since $q(x)$ is Gaussian, we have the following properties:

(1) First-Order Condition for the Mean (E_x): The gradient of $\log q(x)$ evaluated at E_x is zero, implying E_x is a maximum of $\log p(x)$.
(2) Second-Order Condition for the Variance (σ_x^2): The negative of the Hessian (second-order derivative) of $\log q(x)$ evaluated at E_x equals the reciprocal of the variance. However, since we directly approximate $p(x)$, we use the Hessian of $\log p(x)$ evaluated at E_x to estimate σ_x^2:

$$\nabla^2 \log p(x)|_{x=\mathsf{E}_x} = -\frac{1}{\sigma_x^2}.$$

We now detail the estimation of E_x and σ_x^2 based on these conditions for some specific density function $p(x)$.

4.3.1. Absolute value function. The ℓ_1 norm of $\mathcal{S}$, as the sum of absolute element values, necessitates approximating the distribution of absolute values to enable effective optimization within Gaussian density families. Given the log-probability density function $\log p(x) \propto -\frac{1}{2}(x-b)^2 - \beta|x|$, the first step of Laplace approximation involves computing the mean E_x of $q(x)$, which corresponds to the maximum of $\log p(x)$:

$$\mathsf{E}_x = \operatorname*{argmax}_x \log p(x) = \operatorname*{argmin}_x \frac{1}{2}(x-b)^2 + \beta|x|.$$

In the paper, we utilize the sans serif font E accompanied by a subscripted variable x to denote the expectation of a random variable x. The solution is given by:

$$\mathsf{E}_x = \begin{cases} b - \beta \text{sign}(b), & \text{if } |b| > \beta \\ 0, & \text{otherwise.} \end{cases}$$

Proceeding to the second stage, we estimate the variance, σ^2. When $\mathsf{E}_x = 0$, we directly set $\sigma_x^2 = 0$. For non-zero E_x, we leverage the inequality $|x| \leq \frac{x^2}{2|y|} + \frac{|y|}{2}$ with equality at $x = y \neq 0$. By setting $y = |\mathsf{E}_x|$, this facilitates a lower bound on $\log p(x)$:

$$\log p(x) \geq -\frac{1}{2}(x-b)^2 - \frac{\beta}{2|\mathsf{E}_x|}x^2 + \text{const.}$$

To simplify the variance estimation process, we exclude the constant term and mean shift from consideration, as they do not impact the variance calculation. Approximating the second-order derivatives of this lower bound around E_x, we derive the variance estimate:

$$\sigma_x^2 \approx \left(1 + \frac{\beta}{|\mathsf{E}_x|}\right)^{-1} = \frac{|\mathsf{E}_x|}{|\mathsf{E}_x| + \beta}.$$

Ultimately, utilizing the estimated mean E_x and variance σ_x^2, we construct the optimal Gaussian density approximation:

$$q(x) = \mathcal{N}\left(\mathsf{E}_x, \frac{|\mathsf{E}_x|}{|\mathsf{E}_x| + \beta}\right). \tag{4.4}$$

Given the approximation $|x| \approx \frac{1}{2|\mathsf{E}_x|}x^2 + \frac{|\mathsf{E}_x|}{2}$ at $x = \mathsf{E}_x$, we derive the expectation of the absolute value of x:

$$\mathsf{E}|x| = |\mathsf{E}_x| + \frac{1}{2(|\mathsf{E}_x| + \beta)}. \tag{4.5}$$

4.3.2. Nuclear norm. The Laplace approximation approach does not directly extend the Gaussian density approximation of absolute functions to the nuclear norm of matrices, given its intrinsic nature as a sum of singular values rather than element-wise absolute values. When considering a density function incorporating the weighted nuclear norm of a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$, we assume $n_1 \leq n_2$ for generality. The targeted density is formulated as:

$$p(\mathbf{X}) \propto \exp\left(-\frac{\alpha}{2}\|\mathbf{X} - \mathbf{A}\|_F^2 - \beta\|\mathbf{X}\|_*\right),$$

where $\mathbf{A}$ is a given matrix, β is a regularization parameter, and $\boldsymbol{w} \in \mathbb{R}^{n_2}$ represents the vector of weights. We find the density $q(\mathbf{X})$ that maximizes:

$$q(\mathbf{X}) = \arg\max_{q(\mathbf{X}) \in \mathcal{Q}_\mathcal{N}} \int q(\mathbf{X}) \log \frac{p(\mathbf{X})}{q(\mathbf{X})}\, d\mathbf{X}.$$

Applying the Laplacian approximation method, the mean $\mathsf{E}_\mathbf{X}$ of $q(\mathbf{X})$ is obtained by solving:

$$\mathsf{E}_\mathbf{X} = \arg\min_\mathbf{X} \left(\frac{\alpha}{2}\|\mathbf{X} - \mathbf{A}\|_F^2 + \beta\|\mathbf{X}\|_*\right).$$

Given the SVD of $\mathbf{A}$ as $\mathbf{A} = \mathbf{U}_\mathbf{A}\mathbf{D}_\mathbf{A}\mathbf{V}_\mathbf{A}^T$, the minimizer $\mathsf{E}_\mathbf{X}$ for the aforementioned problem can be formulated as [10]:

$$\mathsf{E}_\mathbf{X} = \mathbf{U}_\mathbf{A} \max\{\mathbf{D}_\mathbf{A} - \tfrac{\beta}{\alpha}\mathbf{I}, 0\}\mathbf{V}_\mathbf{A}^T,$$

where $\max\{\cdot, 0\}$ denotes an element-wise maximum operation applied to the diagonal matrix. To compute the covariance of $\mathbf{X}$, we introduce an inequality derived from [31]:

$$\|\mathbf{X}\|_* \leq \frac{1}{2}\mathrm{Tr}(\omega(\mathbf{Y})\mathbf{X}\mathbf{X}^T) + \frac{1}{2}\|\mathbf{Y}\|_*,$$

where $\omega(\mathbf{Y}) = (\mathbf{Y}\mathbf{Y}^T)^{-1/2}$, and equality holds when $\mathbf{X} = \mathbf{Y}$. Hence we obtain

$$\|\mathbf{X}\|_* \approx \frac{1}{2}\mathrm{Tr}(\omega(\mathsf{E}_\mathbf{X})\mathbf{X}\mathbf{X}^T) + \frac{1}{2}\|\mathsf{E}_\mathbf{X}\|_*. \tag{4.6}$$

Considering the j-th columns of $\mathbf{X}$ and $\mathbf{A}$ denoted by $\mathbf{X}_{:j}$ and $\mathbf{A}_{:j}$ respectively, we bound the log-likelihood $\log p(\mathbf{X})$ as follows:

$$\log p(\mathbf{X}) \geq -\sum_j \left(\frac{\alpha}{2}\|\mathbf{X}_{:j} - \mathbf{A}_{:j}\|_2^2 + \frac{\beta}{2}\mathbf{X}_{:j}^T\omega(\mathsf{E}_\mathbf{X})\mathbf{X}_{:j} \right) + \text{const.}$$

Evaluating the second-order derivatives of the lower bound with respect to $\mathbf{X}_{:j}$ yields the inverse covariance matrix $\mathbf{\Sigma}_{\mathbf{X}_{:j}}^{-1} = \alpha\mathbf{I} + \beta\omega(\mathsf{E}_\mathbf{X})$. Let $\mathsf{E}_\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be the skinny SVD of $\mathsf{E}_\mathbf{X}$, we have $\omega(\mathsf{E}_\mathbf{X}) = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^T$. Hence we have

$$\mathbf{\Sigma}_{\mathbf{X}_{:j}} = \mathbf{U}\mathbf{D}\left(\alpha\mathbf{D} + \beta\mathbf{I}\right)^{-1}\mathbf{U}^T. \tag{4.7}$$

Finally, the optimal density approximation $q(\mathbf{X})$ is expressed as:

$$q(\mathbf{X}) = \prod_j \mathcal{N}(\mathbf{X}_{:j}|\mathsf{E}_{\mathbf{X}_{:j}}, \mathbf{U}\mathbf{D}\left(\alpha\mathbf{D} + \beta\mathbf{I}\right)^{-1}\mathbf{U}^T). \tag{4.8}$$

We have

$$\mathbb{E}\,\|\mathbf{X}\|_F^2 = \|\mathsf{E}_\mathbf{X}\|_F^2 + \sum_j \mathrm{Tr}\left(\mathbf{\Sigma}_{\mathbf{X}_{:j}}\right) = \|\mathsf{E}_\mathbf{X}\|_F^2 + n_2\sum_i \frac{\mathbf{D}_i}{\alpha\mathbf{D}_i + \beta}. \tag{4.9}$$

According to the approximation of the nuclear norm in (4.6), we have

$$\mathbb{E}\,\|\mathbf{X}\|_* = \frac{1}{2}\sum_j \mathbb{E}\left[\mathbf{X}_{:j}^T\omega(\mathsf{E}_\mathbf{X})\mathbf{X}_{:j}\right] + \frac{1}{2}\|\mathsf{E}_\mathbf{X}\|_*. \tag{4.10}$$

Let $\mathbf{D}_i$ be the i-th singular values of $\mathsf{E}_\mathbf{X}$. Then we have

$$\mathbb{E}\,\|\mathbf{X}\|_* = \|\mathsf{E}_\mathbf{X}\|_* + \frac{n_2}{2}\mathrm{Tr}\left(\left(\alpha\mathbf{D} + \beta\mathbf{I}\right)^{-1}\right) = \|\mathsf{E}_\mathbf{X}\|_* + \frac{n_2}{2}\sum_i \frac{1}{\alpha\mathbf{D}_i + \beta}. \tag{4.11}$$

4.4. Coordinate ascent variational inference. In order to maximize the ELBO $\mathcal{J}(q(\mathcal{S}, \mathcal{L}, \boldsymbol{\theta}))$, we apply coordinate ascent variational inference (CAVI) [8, 47]. Starting from an initial density $(q_0(\boldsymbol{\theta}), q_0(\mathcal{L}), q_0(\mathcal{S}))$, the densities of $\mathcal{S}$, $\mathcal{L}$ and $\boldsymbol{\theta}$ are updated as follows:

$$q_\ell(\mathcal{S}) = \underset{q(\mathcal{S})\in\mathcal{Q}_\mathcal{N}}{\mathrm{argmax}}\,\mathcal{J}\left(q(\mathcal{S})q_{\ell-1}(\mathcal{L})q_{\ell-1}(\boldsymbol{\theta})\right), \tag{4.12}$$

$$q_\ell(\mathcal{L}) = \underset{q(\mathcal{L})\in\mathcal{Q}_\mathcal{N}}{\mathrm{argmax}}\,\mathcal{J}\left(q_\ell(\mathcal{S})q(\mathcal{L})q_{\ell-1}(\boldsymbol{\theta})\right), \tag{4.13}$$

$$q_\ell(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta})\in\mathcal{Q}_\mathcal{G}}{\mathrm{argmax}}\,\mathcal{J}\left(q_\ell(\mathcal{S})q_\ell(\mathcal{L})q(\boldsymbol{\theta})\right), \tag{4.14}$$

where $q_\ell(\boldsymbol{\theta}), q_\ell(\mathcal{L}), q_\ell(\mathcal{S})$ refer to the variational densities obtained in the ℓ-th iteration.

4.4.1. The density $q_\ell(\mathcal{S})$. In accordance with (4.2), we formulate the optimization problem as maximizing the evidence lower bound (ELBO) with respect to the variational distribution $q(\mathcal{S})$:

$$\underset{q(\mathcal{S})\in\mathcal{Q}_{\mathcal{N}}}{\operatorname{argmax}}\ \mathcal{J}(q(\mathcal{S}),q_{\ell-1}(\mathcal{L},\boldsymbol{\theta}))=\underset{q(\mathcal{S})\in\mathcal{Q}_{\mathcal{N}}}{\operatorname{argmax}}\int q(\mathcal{S})\mathbb{E}_{q_{\ell-1}(\mathcal{L},\boldsymbol{\theta})}\log\frac{p(\mathcal{X},\mathcal{S},\mathcal{L},\boldsymbol{\theta})}{q(\mathcal{S})}d\mathcal{S}. \tag{4.15}$$

Given the joint density distribution as defined in (3.6), we can express the expectation term within the ELBO as:

$$\begin{aligned}&\mathbb{E}_{q_{\ell-1}(\mathcal{L},\boldsymbol{\theta})}\left[\log p(\mathcal{X},\mathcal{S},\mathcal{L},\boldsymbol{\theta})\right]\\&=-\sum_{ijk}\left(\frac{\mathsf{E}_{\theta_1}^{\ell-1}}{2}\left(\mathcal{X}_{ijk}-\mathsf{E}_{\mathcal{L}_{ijk}}^{\ell-1}-\mathcal{S}_{ijk}\right)^2+\mathsf{E}_{\theta_2}^{\ell-1}\left|\mathcal{S}_{ijk}\right|\right)+\text{const},\end{aligned}$$

where const is a constant independent of $\mathcal{S}$. According to the discussion in Section 4.3.1, the mean of $\mathcal{S}_{ijk}$ is given by

$$\mathsf{E}_{\mathcal{S}}^{\ell}=\underset{\mathcal{S}}{\operatorname{argmin}}\ \frac{\mathsf{E}_{\theta_1}^{\ell-1}}{2}\left\|\mathcal{X}-\mathsf{E}_{\mathcal{L}}^{\ell-1}-\mathcal{S}\right\|_F^2+\mathsf{E}_{\theta_2}^{\ell-1}\left\|\mathcal{S}\right\|_1.$$

It is known that the minimizer is the well-known soft threshold:

$$\mathsf{E}_{\mathcal{S}_{ijk}}^{\ell}=\begin{cases}\mathcal{X}_{ijk}-\mathsf{E}_{\mathcal{L}_{ijk}}^{\ell-1}-\frac{\mathsf{E}_{\theta_2}^{\ell-1}}{\mathsf{E}_{\theta_1}^{\ell-1}}, & \text{if } \mathcal{X}_{ijk}-\mathsf{E}_{\mathcal{L}_{ijk}}^{\ell-1}\geq\frac{\mathsf{E}_{\theta_2}^{\ell-1}}{\mathsf{E}_{\theta_1}^{\ell-1}},\\ \mathcal{X}_{ijk}-\mathsf{E}_{\mathcal{L}_{ijk}}^{\ell-1}+\frac{\mathsf{E}_{\theta_2}^{\ell-1}}{\mathsf{E}_{\theta_1}^{\ell-1}}, & \text{if } \mathcal{X}_{ijk}-\mathsf{E}_{\mathcal{L}_{ijk}}^{\ell-1}\leq-\frac{\mathsf{E}_{\theta_2}^{\ell-1}}{\mathsf{E}_{\theta_1}^{\ell-1}},\\ 0, & \text{others.}\end{cases} \tag{4.16}$$

Applying (4.4), the variance of $\mathcal{S}_{ijk}$ is given by

$$\boldsymbol{\Sigma}_{\mathcal{S}_{ijk}}^{\ell}=\frac{\mathsf{E}_{\theta_1}^{\ell-1}|\mathsf{E}_{\mathcal{S}_{ijk}}^{\ell}|}{\mathsf{E}_{\theta_1}^{\ell-1}|\mathsf{E}_{\mathcal{S}_{ijk}}^{\ell}|+\mathsf{E}_{\theta_2}^{\ell-1}}.$$

Then the density function of $q(\mathcal{S})$ is given:

$$q_\ell(\mathcal{S}_{ijk})=\mathcal{N}(\mathcal{S}|\mathsf{E}_{\mathcal{S}_{ijk}}^{\ell},\boldsymbol{\Sigma}_{\mathcal{S}_{ijk}}^{\ell}). \tag{4.17}$$

4.4.2. The density $q_\ell(\mathcal{L})$. In accordance with (4.2), we formulate the optimization problem as maximizing the evidence lower bound (ELBO) with respect to the variational distribution $q(\mathcal{L})$:

$$\underset{q(\mathcal{L})\in\mathcal{Q}_{\mathcal{N}}}{\operatorname{argmax}}\ \mathcal{J}(q_\ell(\mathcal{S}),q(\mathcal{L}),q_{\ell-1}(\boldsymbol{\theta}))=\underset{q(\mathcal{L})\in\mathcal{Q}_{\mathcal{N}}}{\operatorname{argmax}}\int q(\mathcal{L})\mathbb{E}_{q_\ell(\mathcal{S})q_{\ell-1}(\boldsymbol{\theta})}\log\frac{p(\mathcal{X},\mathcal{S},\mathcal{L},\boldsymbol{\theta})}{q(\mathcal{L})}d\mathcal{L}. \tag{4.18}$$

Given the joint density distribution as defined in (3.6), we can express the expectation term within the ELBO as:

$$\begin{aligned}\mathbb{E}_{q_\ell(\mathcal{S})q_{\ell-1}(\boldsymbol{\theta})}\left[\log p(\mathcal{X},\mathcal{S},\mathcal{L},\boldsymbol{\theta})\right]&=-\left(\frac{\mathsf{E}_{\theta_1}^{\ell-1}}{2}\left\|\mathcal{L}-(\mathcal{X}-\mathsf{E}_{\mathcal{S}}^{\ell})\right\|_F^2+\mathsf{E}_{\theta_3}^{\ell-1}\left\|\mathcal{L}\right\|_*\right)+\text{const}\\&=-\left(\frac{\mathsf{E}_{\theta_1}^{\ell-1}}{2n_3}\left\|\overline{\mathcal{L}}-(\overline{\mathcal{X}}-\overline{\mathsf{E}_{\mathcal{S}}^{\ell}})\right\|_F^2+\frac{\mathsf{E}_{\theta_3}^{\ell-1}}{n_3}\left\|\overline{\mathcal{L}}\right\|_*\right)+\text{const}\end{aligned}$$

According to the discussion in Section 4.3.2, the mean of $\mathcal{L}$ is given by

$$\mathsf{E}_{\mathcal{L}}^{\ell} = \underset{\mathcal{L}}{\operatorname{argmin}} \frac{\mathsf{E}_{\theta_1}^{\ell-1}}{2} \left\| \mathcal{X} - \mathcal{L} - \mathsf{E}_{\mathcal{S}}^{\ell} \right\|_F^2 + \mathsf{E}_{\theta_3}^{\ell-1} \left\| \mathcal{L} \right\|_* .$$

This subproblem is to solve a proximal operator of the tensor nuclear norm, which has a closed-form solution as tensor singular value thresholding (t-SVT) [33]. Let the SVD of $\mathcal{X} - \mathsf{E}_{\mathcal{S}}^{\ell}$ is given by $\mathcal{X} - \mathsf{E}_{\mathcal{S}}^{\ell} = \mathcal{U}^{\ell} * \mathcal{D}^{\ell} * \mathcal{V}^{\ell^T}$. The update of $\mathsf{E}_{\mathcal{L}}^{\ell}$ is

$$\mathsf{E}_{\mathcal{L}}^{\ell} = \mathcal{U}^{\ell} * \mathcal{D}_{\tau}^{\ell} * \mathcal{V}^{\ell^T}, \tag{4.19}$$

where $\mathcal{D}_{\tau}^{\ell}$ is an $n_1 \times n_2 \times n_3$ tensor that satisfies $\overline{\mathcal{D}}_{\tau}^{\ell} = \max\{\overline{\mathcal{D}^{\ell}} - \tau, 0\}$ with $\tau = \mathsf{E}_{\theta_1}^{\ell-1}/\mathsf{E}_{\theta_3}^{\ell-1}$. Recall that we adopt the notation of an overline $\overline{\mathcal{A}}$ to signify the application of the DFT to the tensor $\mathcal{A}$ specifically along its third dimension.

We apply (4.7) and then obtain the covariance matrix of the vector $\overline{\mathcal{L}}_{:jk}$

$$\mathbf{\Sigma}_{\overline{\mathcal{L}}_{:jk}}^{\ell} = n_3 \overline{\mathcal{U}}_{::k}^{\ell} \overline{\mathcal{D}}_{\tau ::k}^{\ell} \left(\mathsf{E}_{\theta_1}^{\ell-1} \overline{\mathcal{D}}_{\tau ::k}^{\ell} + \mathsf{E}_{\theta_3}^{\ell-1} \mathbf{I} \right)^{-1} \overline{\mathcal{U}}_{::k}^{\ell \, T}.$$

Thus, we construct the density function and parameterize a normal density $q(\overline{\mathcal{L}})$ as:

$$q_{\ell}(\overline{\mathcal{L}}) = \prod_{jk} \mathcal{N}\left(\overline{\mathcal{L}}_{:jk} | \mathsf{E}_{\overline{\mathcal{L}}_{:jk}}^{\ell}, \mathbf{\Sigma}_{\overline{\mathcal{L}}_{:jk}}^{\ell} \right). \tag{4.20}$$

4.4.3. The density $q_{\ell}(\boldsymbol{\theta})$. In accordance with (4.2), we frame the optimization problem as maximizing the evidence lower bound (ELBO) with respect to the variational distribution $q(\boldsymbol{\theta})$,

$$\underset{q(\boldsymbol{\theta}) \in \mathcal{Q}_{\mathcal{G}}}{\operatorname{argmax}} \mathcal{J}(q_{\ell}(\mathcal{S}), q_{\ell}(\mathcal{L}), q(\boldsymbol{\theta})) = \underset{q(\boldsymbol{\theta}) \in \mathcal{Q}_{\mathcal{G}}}{\operatorname{argmax}} \int q(\boldsymbol{\theta}) \mathbb{E}_{q_{\ell}(\mathcal{S},\mathcal{L})} \log \tfrac{p(\mathcal{X},\mathcal{S},\mathcal{L},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}, \tag{4.21}$$

where $\mathcal{Q}_{\mathcal{G}}$ is the set of all the density functions for the Gamma distribution. By taking the partial derivative of the objective function in (4.21) with respect to $q(\boldsymbol{\theta})$, and letting it be equal to 0, we obtain that the optimal $q(\boldsymbol{\theta})$ is proportional to

$$q(\boldsymbol{\theta}) \propto \exp \mathbb{E}_{q_{\ell}(\mathcal{S},\mathcal{L})} \log p(\mathcal{X}, \mathcal{S}, \mathcal{L}, \boldsymbol{\theta})$$

with

$$\begin{aligned} \mathbb{E}_{q_{\ell}(\mathcal{S},\mathcal{L})} \log p(\mathcal{X}, \mathcal{S}, \mathcal{L}, \boldsymbol{\theta}) = & -\tfrac{\theta_1}{2} \mathbb{E}_{q_{\ell}(\mathcal{S},\mathcal{L})} \left\| \mathcal{X} - \mathcal{S} - \mathcal{L} \right\|_F^2 - \theta_2 \mathbb{E}_{q_{\ell}(\mathcal{S})} \left\| \mathcal{S} \right\|_1 \\ & -\theta_3 \mathbb{E}_{q_{\ell}(\mathcal{L})} \left\| \mathcal{L} \right\|_* + \tfrac{n}{2} \log \theta_1 + n \log \theta_2 + n \log \theta_3 + \text{const.} \end{aligned}$$

Since $q(\boldsymbol{\theta}) = \prod_{i=1}^{3} q(\theta_i)$ is assumed to factorize due to the independence of the model components, we derive the form of each $q(\theta_i)$ by comparing the coefficients of θ_i and $\log \theta_i$ with the log-density of a Gamma distribution

$$\log q(\boldsymbol{\theta}) = \sum_{i=1}^{3} \left((a_{\theta_i}^{\ell} - 1) \log \theta_i - b_{\theta_i}^{\ell} \theta_i \right) + \text{const},$$

where $a_{\theta_i}^{\ell}$ and $b_{\theta_i}^{\ell}$ are the shape and rate parameters, respectively. By comparing the coefficients in $\mathbb{E}_{q_{\ell}(\mathcal{S},\mathcal{L})} \log p(\mathcal{X}, \mathcal{S}, \mathcal{L}, \boldsymbol{\theta})$ with those of a Gamma density ($\mathcal{G}(x|a,b) \propto$

$x^{a-1}\exp(-bx)$), we can infer the shape $a^\ell_{\theta_i}$ and scale $b^\ell_{\theta_i}$ parameters for each θ_i. Consequently, the shape parameters are given by:

$$a^\ell_{\theta_1} = \tfrac{n}{2}+1, \quad a^\ell_{\theta_2} = n+1, \quad a^\ell_{\theta_3} = n+1.$$

While the scale parameters are expressed as expectations over the variational distributions $q_\ell(\mathcal{S})$ and $q_\ell(\mathcal{L})$, as defined in the following system of equations:

$$\begin{cases} b^\ell_{\theta_1} = \frac{1}{2}\mathbb{E}_{q_\ell(\mathcal{L})q_\ell(\mathcal{S})}\left[\|\mathcal{X}-\mathcal{S}-\mathcal{L}\|_F^2\right], \\ b^\ell_{\theta_2} = \mathbb{E}_{q_\ell(\mathcal{S})}\|\mathcal{S}\|_1, \\ b^\ell_{\theta_3} = \mathbb{E}_{q_\ell(\mathcal{L})}\|\mathcal{L}\|_*. \end{cases} \tag{4.22}$$

The computation for $b^\ell_{\theta_1}$ involves the expectations of both $\|\mathcal{S}\|_F^2$ and $\|\mathcal{L}\|_F^2$. It is obvious that

$$\mathbb{E}_{q_\ell(\mathcal{S})}\|\mathcal{S}\|_F^2 = \sum_{ijk}\mathbb{E}_{q_\ell}|\mathcal{S}_{ijk}|^2 = \sum_{ijk}\left(\left|\mathsf{E}^\ell_{\mathcal{S}_{ijk}}\right|^2 + \mathbf{\Sigma}^\ell_{\mathcal{S}_{ijk}}\right).$$

According to (4.9), we have

$$\mathbb{E}_{q_\ell(\mathcal{L})}\|\mathcal{L}\|_F^2 = \frac{1}{n_3}\sum_{j,k}\mathbb{E}_{q_\ell(\overline{\mathcal{L}})}\left\|\overline{\mathcal{L}}_{:jk}\right\|_F^2 = \frac{1}{n_3}\sum_{j,k}\left(\left\|\mathsf{E}^\ell_{\overline{\mathcal{L}}_{:jk}}\right\|_F^2 + \mathrm{Tr}\left(\mathbf{\Sigma}^\ell_{\overline{\mathcal{L}}_{:jk}}\right)\right).$$

Hence we have

$$\mathbb{E}_{q_\ell(\mathcal{L})q_\ell(\mathcal{S})}(\|\mathcal{X}-\mathcal{S}-\mathcal{L}\|_F^2) = \left\|\mathcal{X}-\mathsf{E}^\ell_{\mathcal{L}}-\mathsf{E}^\ell_{\mathcal{S}}\right\|_F^2 + \frac{1}{n_3}\sum_{j,k}\mathrm{Tr}\left(\mathbf{\Sigma}_{\overline{\mathcal{L}}_{:jk}}\right) + \sum_{ijk}\mathbf{\Sigma}^\ell_{\mathcal{S}_{ijk}}.$$

For the expectation of $\|\mathcal{S}\|_1$, according to (4.5), we have

$$\mathbb{E}_{q_\ell(\mathcal{S})}\|\mathcal{S}\|_1 = \sum_{ijk}\mathbb{E}_{q_\ell(\mathcal{S})}|\mathcal{S}_{ijk}| = \|\mathsf{E}^\ell_{\mathcal{S}}\|_1 + \frac{1}{2}\sum_{ijk}\left(\mathsf{E}^\ell_{\theta_1}|\mathsf{E}^\ell_{\mathcal{S}_{ijk}}| + \mathsf{E}^\ell_{\theta_2}\right)^{-1}.$$

For the expectation of the nuclear norm $\|\mathcal{L}\|_*$, it is the arithmetic mean of each slice $\overline{\mathcal{L}}_{::k}$ of the tensor $\mathcal{L}$. Hence we need to evaluate $\mathbb{E}_{q_\ell(\overline{\mathcal{L}}_{::k})}\|\overline{\mathcal{L}}_{::k}\|_*$. According to (4.11), we have

$$\mathbb{E}_{q_\ell(\overline{\mathcal{L}}_{::k})}\|\overline{\mathcal{L}}_{::k}\|_* = \left\|\mathsf{E}^\ell_{\overline{\mathcal{L}}_{::k}}\right\|_* + \frac{n_2 n_3}{2}\mathrm{Tr}\left(\left(\mathsf{E}^{\ell-1}_{\theta_1}\overline{\mathcal{D}}^\ell_{\tau::k} + \mathsf{E}^{\ell-1}_{\theta_3}\mathbf{I}\right)^{-1}\right).$$

Hence

$$\mathbb{E}_{q_\ell(\mathcal{L})}\|\mathcal{L}\|_* = \left\|\mathsf{E}^\ell_{\mathcal{L}}\right\|_* + \frac{n_2}{2}\sum_{k}\mathrm{Tr}\left(\left(\mathsf{E}^{\ell-1}_{\theta_1}\overline{\mathcal{D}}^\ell_{\tau::k} + \mathsf{E}^{\ell-1}_{\theta_3}\mathbf{I}\right)^{-1}\right).$$

Now, we focus on the expectation of the nuclear norm $\|\mathcal{L}\|_*$, which requires evaluating $\mathbb{E}_{q_\ell(\overline{\mathcal{L}}_{::k})}\|\overline{\mathcal{L}}_{::k}\|_*$ for each slice $\overline{\mathcal{L}}_{::k}$ of the tensor $\mathcal{L}$.

We summarize the proposed adaptive method in Algorithm 4.1. For simplicity, we refer to our proposed algorithm for solving the tensor nuclear norm model as $\mathrm{VBI}_{\mathrm{TNN}}$.

Remark 4.1. According to [7], a general theoretical treatment of analyzing the convergence of CAVI is missing in the literature. This is due to the lack of tractability of the updating formula involving unwieldy normalization constants and the technical challenge of dealing with optimization over infinite-dimensional distributions. Here, we will empirically show the convergence in Section 5.

Algorithm 4.1 VBI_{TNN} :Variational Bayesian inference for the TNN-based TRPCA.

Initialization: $\mathsf{E}_{\theta_1}, \mathsf{E}_{\theta_2}, \mathsf{E}_{\theta_3}, \mathsf{E}^0_{\mathcal{L}}, \mathsf{E}^0_{\mathcal{S}}, \boldsymbol{\Sigma}^0_{\mathcal{L}}, \boldsymbol{\Sigma}^0_{\mathcal{X}}$
Let $a_{\theta_1} = \frac{n}{2} + 1, \quad a_{\theta_2} = n + 1, \quad a_{\theta_3} = n + 1.$
while $\ell \leq \ell_{\text{Max}}$ or not converged **do**

$$\mathsf{E}^{\ell}_{\mathcal{S}_{ijk}} = \begin{cases} \mathcal{X}_{ijk} - \mathsf{E}^{\ell-1}_{\mathcal{L}_{ijk}} - \frac{\mathsf{E}^{\ell-1}_{\theta_2}}{\mathsf{E}^{\ell-1}_{\theta_1}}, & \text{if } \mathcal{X}_{ijk} - \mathsf{E}^{\ell-1}_{\mathcal{L}_{ijk}} \geq \frac{\mathsf{E}^{\ell-1}_{\theta_2}}{\mathsf{E}^{\ell-1}_{\theta_1}} \\ \mathcal{X}_{ijk} - \mathsf{E}^{\ell-1}_{\mathcal{L}_{ijk}} + \frac{\mathsf{E}^{\ell-1}_{\theta_2}}{\mathsf{E}^{\ell-1}_{\theta_1}}, & \text{if } \mathcal{X}_{ijk} - \mathsf{E}^{\ell-1}_{\mathcal{L}_{ijk}} \leq -\frac{\mathsf{E}^{\ell-1}_{\theta_2}}{\mathsf{E}^{\ell-1}_{\theta_1}} \\ 0, & \text{others} \end{cases}$$

Take the SVD of $\mathcal{X} - \mathsf{E}^{\ell}_{\mathcal{S}}$ as $\mathcal{X} - \mathsf{E}^{\ell}_{\mathcal{S}} = \mathcal{U}^{\ell} * \mathcal{D}^{\ell} * \mathcal{V}^{\ell T}$
$\mathsf{E}^{\ell}_{\mathcal{L}} = \mathcal{U}^{\ell} * \mathcal{D}^{\ell}_{\tau} * \mathcal{V}^{\ell T}$
$\boldsymbol{\Sigma}^{\ell}_{\mathcal{S}_{ijk}} = \frac{\mathsf{E}^{\ell-1}_{\theta_1}|\mathsf{E}^{\ell}_{\mathcal{S}_{ijk}}|}{\mathsf{E}^{\ell-1}_{\theta_1}|\mathsf{E}^{\ell}_{\mathcal{S}_{ijk}}|+\mathsf{E}^{\ell-1}_{\theta_2}}$, and $\boldsymbol{\Sigma}^{\ell}_{\overline{\mathcal{L}}_{:jk}} = n_3 \overline{\mathcal{U}}^{\ell}_{::k} \overline{\mathcal{D}}^{\ell}_{\tau::k} \left(\mathsf{E}^{\ell-1}_{\theta_1} \overline{\mathcal{D}}^{\ell}_{\tau::k} + \mathsf{E}^{\ell-1}_{\theta_3} \mathbf{I} \right)^{-1} \overline{\mathcal{U}}^{\ell\ T}_{::k}$
$q(\mathcal{S}_{ijk}) = \mathcal{N}(\mathcal{S}|\mathsf{E}^{\ell}_{\mathcal{S}_{ijk}}, \boldsymbol{\Sigma}^{\ell}_{\mathcal{S}_{ijk}})$ and $q(\overline{\mathcal{L}}) = \prod_{jk} \mathcal{N}\left(\overline{\mathcal{L}}_{:jk}|\mathsf{E}^{\ell}_{\overline{\mathcal{L}}_{:jk}}, \boldsymbol{\Sigma}^{\ell}_{\overline{\mathcal{L}}_{:jk}}\right).$
$b^{\ell}_{\theta_1} = \left\|\mathcal{X} - \mathsf{E}^{\ell}_{\mathcal{L}} - \mathsf{E}^{\ell}_{\mathcal{S}}\right\|^2_F /2 + \frac{1}{2n_3} \sum_{j,k} \text{Tr}\left(\boldsymbol{\Sigma}_{\overline{\mathcal{L}}_{:jk}}\right) + \sum_{ijk} \boldsymbol{\Sigma}^{\ell}_{\mathcal{S}_{ijk}}/2$
$b^{\ell}_{\theta_2} = \|\mathsf{E}^{\ell}_{\mathcal{S}}\|_1 + \frac{1}{2} \sum_{ijk} \left(\mathsf{E}^{\ell}_{\theta_1}|\mathsf{E}^{\ell}_{\mathcal{S}_{ijk}}| + \mathsf{E}^{\ell}_{\theta_2}\right)^{-1}$
$b^{\ell}_{\theta_3} = \left\|\mathsf{E}^{\ell}_{\mathcal{L}}\right\|_* + \frac{n_2}{2} \sum_k \text{Tr}\left(\left(\mathsf{E}^{\ell-1}_{\theta_1} \overline{\mathcal{D}}^{\ell}_{\tau::k} + \mathsf{E}^{\ell-1}_{\theta_3} \mathbf{I}\right)^{-1}\right)$
$q(\theta_i) = \mathcal{G}(\theta_i|a_{\theta_i}, b^{\ell}_{\theta_i})$, and $\mathsf{E}^{\ell}_{\theta_i} = a_{\theta_i}/b^{\ell}_{\theta_i} i = 1, 2, 3$
end while
return $\mathcal{L} = \mathsf{E}^{\ell}_{\mathcal{L}}, \mathcal{S} = \mathsf{E}^{\ell}_{\mathcal{S}}$

4.5. Variational Bayesian inference for weighted tensor nuclear norm. In this subsection, we consider a variant of the tensor nuclear norm by reweighting the singular values [23, 12]. Note that the standard tensor nuclear norm can be regarded as a special version of the weighted tensor nuclear norm, where the weighting matrix consists of elements that are all equal to one. Formally, for a non-negative matrix $\mathbf{W} \in \mathbb{R}^{\min(n_1,n_2)\times n_3}$ with column vectors $\mathbf{W}_{:k}$, the weighted tensor nuclear norm $\|\mathcal{A}\|_{\mathbf{W}*}$ is defined as:

$$\|\mathcal{A}\|_{\mathbf{W}*} = \frac{1}{n_3} \sum_{k=1}^{n_3} \sum_{j=1}^{\min(n_1,n_2)} \mathbf{W}_{jk} \sigma_{jk},$$

where σ_{jk} denotes the j-th singular value of the k-th frontal slice $\mathcal{A}_{::k}$ of tensor $\mathcal{A}$. To incorporate this weighted norm, we modify the robust principal component model (1.2) as follows:

$$\min_{\mathcal{S},\mathcal{L}} \left\{ \frac{\theta_1}{2} \|\mathcal{X} - \mathcal{L} - \mathcal{S}\|^2_F + \theta_2 \|\mathcal{S}\|_1 + \theta_3 \|\mathcal{L}\|_{\mathbf{W}*} \right\}. \tag{4.23}$$

During the inference of $\mathcal{L}$, we update the expectation of $\overline{\mathcal{L}}_{::k}$ in (4.20) to:

$$\mathsf{E}^{\ell}_{\mathcal{L}} = \mathcal{U}^{\ell} * \mathcal{D}^{\ell}_{\mathbf{W}} * \mathcal{V}^{\ell T}, \tag{4.24}$$

where $\mathcal{D}^{\ell}_{\mathbf{W}}$ is an $n_1 \times n_2 \times n_3$ tensor that satisfies

$$\overline{\mathcal{D}}^{\ell}_{\mathbf{W}::k} = \max\left\{ \overline{\mathcal{D}^{\ell}}_{::k} - \frac{\mathsf{E}^{\ell-1}_{\theta_1}}{\mathsf{E}^{\ell-1}_{\theta_3}} \text{diag}(\mathbf{W}_{:k}), 0 \right\}.$$

Concurrently, the covariance matrix of $\overline{\mathcal{L}}_{::k}$ is adjusted to:

$$\mathbf{\Sigma}^{\ell}_{\overline{\mathcal{L}}_{:jk}} = n_3 \overline{\mathcal{U}}^{\ell}_{::k} \overline{\mathcal{D}}^{\ell}{}_{::k}(\tau^{\ell}) \left(\mathsf{E}^{\ell-1}_{\theta_1} \overline{\mathcal{D}}^{\ell}_{\tau::k} + \mathsf{E}^{\ell-1}_{\theta_3} \operatorname{diag}(\mathbf{W}_{:k}) \right)^{-1} \overline{\mathcal{U}}^{\ell}_{::k}{}^{T}.$$

Given these updates, the computation of $b^{\ell}_{\theta_3} = \mathbb{E}_{q_\ell(\mathcal{L})} \|\mathcal{L}\|_{\mathbf{W}*}$ necessitates a corresponding adjustment:

$$\mathbb{E}_{q_\ell(\mathcal{L})} \|\mathcal{L}\|_{\mathbf{W}*} = \left\|\mathsf{E}^{\ell}_{\mathcal{L}}\right\|_{\mathbf{W}*} + \frac{n_2}{2} \sum_{k=1}^{n_3} \operatorname{Tr}\left(\left(\mathsf{E}^{\ell-1}_{\theta_1} \overline{\mathcal{D}}^{\ell}_{\tau::k} + \mathsf{E}^{\ell-1}_{\theta_3} \operatorname{diag}(\mathbf{W}_{:k})^{-1} \right)^{-1} \right).$$

Note the subtle yet crucial change in the trace term, ensuring consistency with the weighted norm definition.

5. Experiments. In this section, we give experimental results to illustrate the performance of the proposed method. All the experiments are implemented using MATLAB (R2022b) on the Windows 10 platform with Intel Core i5-1135G7 2.40 GHz and 16 GB of RAM.

5.1. Validation on synthetic data. Here, we generate each observation $\mathcal{X}$ in $\mathbb{R}^{n_1 \times n_2 \times n_3}$ by combining a low-rank tensor $\mathcal{L}_0$ and a sparse tensor $\mathcal{S}_0$ with a Gaussian noise $\mathcal{E}_0$ in the the same dimensions. The low-rank tensor $\mathcal{L}_0$ is derived from the t-product of two smaller tensors, namely $\mathcal{P}$ in $\mathbb{R}^{n_1 \times r \times n_3}$ and $\mathcal{H}$ in $\mathbb{R}^{r \times n_2 \times n_3}$, where r is significantly smaller than n_2. The tubal rank of $\mathcal{L}_0$ does not exceed r. The entries of $\mathcal{P}$ are independently and identically distributed according to a Gaussian distribution $\mathcal{N}(0, 1/n_1)$, and those of $\mathcal{H}$ follow $\mathcal{N}(0, 1/n_2)$. The sparse tensor $\mathcal{S}_0$ has entries determined by a Bernoulli process, where each element is either $+1$ or -1 with a probability ρ, and 0 with a probability $1 - 2\rho$. The entries in Gaussian noise $\mathcal{S}_0$ follow $\mathcal{N}(0, \sigma^2)$.

We initiate our analysis by examining the convergence properties using a third-order tensor with dimensions $40 \times 40 \times 30$. The rank parameter r is set to 3, with the parameter ρ at 0.1 and the noise level σ at 10^{-2}. The algorithm is allowed a maximum of 100 iterations, starting with initial guesses for $\mathcal{L}$ and $\mathcal{S}$ as $\mathcal{X}$ and $\mathcal{O}$, respectively. The convergence of the algorithm is monitored using the relative mean square error (RMSE) for $\mathcal{L}$ and $\mathcal{S}$, defined as $\frac{\|\mathsf{E}^{\ell}_{\mathcal{L}} - \mathsf{E}^{\ell-1}_{\mathcal{L}}\|_F}{\|\mathsf{E}^{\ell}_{\mathcal{L}}\|_F}$ and $\frac{\|\mathsf{E}^{\ell}_{\mathcal{S}} - \mathsf{E}^{\ell-1}_{\mathcal{S}}\|_F}{\|\mathsf{E}^{\ell}_{\mathcal{S}}\|_F}$, respectively. The progression of the objective values, RMSE, and parameters $(\theta_1, \theta_2, \theta_3)$ is plotted across iterations in Figure 1. Due to the nonlinear and nonconvex nature of simultaneously optimizing three tensors and their associated parameters, initial fluctuations in the objective values are observed. However, after approximately ten iterations, the objective values begin to decrease steadily and achieve convergence by the 30th iteration. The parameter values similarly stabilize within these iterations. Both RMSE metrics show a sharp decline, reaching as low as 10^{-4} by the 30th iteration. Given these observations, we establish a stopping criterion where the algorithm terminates when RMSE falls below 10^{-4} or when 50 iterations are reached, whichever occurs first. This criterion ensures efficient and effective convergence to an optimal solution within a reasonable number of iterations.

Here, we further evaluate the uncertainty quantification performance of our Variational Bayesian Inference (VBI) algorithm using the same simulated tensor as previously described. Figure 2. presents the mean estimates and 99.73% credible intervals for the recovery of tensor filter $\overline{\mathcal{L}}_{:ij}$ with $i = 20, j = 5, 15, 20$. The mean values consistently align with the ground truth across all fibers, while remarkably narrow

credible intervals (indicated by minimal shading) demonstrate the high precision of our method. This precision is further corroborated by the low parameter standard deviations.

As part of a proof-of-concept study, we employ a partial sum of the tubal nuclear norm [23] as a representative example for a weighted TNN in our numerical experiments. We aim to compare our proposed algorithms, VBI_{TNN} and $\text{VBI}_{\text{PSTNN}}$, against two established methods in tensor rank approximation: TNN [33] and PSTNN [23]. For this comparative analysis, we set the noise levels σ at 10^{-3}, 10^{-2}, and 10^{-1}, the rank r at 3 and 5, and the parameter ρ at 0.01 and 0.1. We assess the performance of these methods by calculating the relative square error between the recovered tensors, $\hat{\mathcal{L}}$ and $\hat{\mathcal{S}}$, and the ground-truth tensors, $\mathcal{L}_{\text{GT}}$ and $\mathcal{S}_{\text{GT}}$. These errors are quantified as follows: $\text{error}_{\mathcal{L}} = \frac{\|\hat{\mathcal{L}}-\mathcal{L}_{\text{GT}}\|_F}{\|\mathcal{L}_{\text{GT}}\|_F}$ for the low-rank component and $\text{error}_{\mathcal{S}} = \frac{\|\hat{\mathcal{S}}-\mathcal{S}_{\text{GT}}\|_F}{\|\mathcal{S}_{\text{GT}}\|_F}$ for the sparse component.

As shown in Table 1, VBI_{TNN} generally outperforms TNN across most tested scenarios, while $\text{VBI}_{\text{PSTNN}}$ is better than PSTNN. Moreover, $\text{VBI}_{\text{PSTNN}}$ consistently delivers the best performance, indicating its superior ability to recover both the low-rank and sparse components of tensors under various noise and rank conditions. This comparative analysis underscores the effectiveness of our proposed methods, particularly $\text{VBI}_{\text{PSTNN}}$, in handling complex tensor decomposition with higher accuracy and robustness against noise.

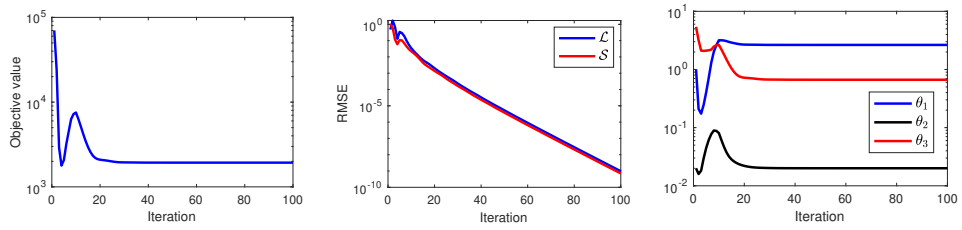

FIG. 1. *Empirical evidence on convergence. Left: objective function, middle: RMSE, right: parameters: θ_1, θ_2, and θ_3, generated by Algorithm 4.1 across iterations.*

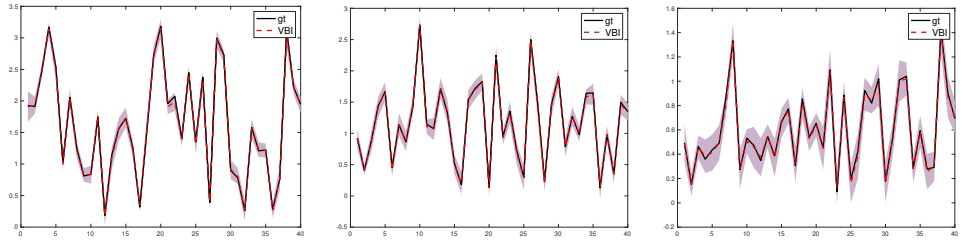

FIG. 2. *Uncertainty quantification: recovery of $\overline{\mathcal{L}}_{:ij}$ with 99.73% credible interval (shaded area) where $i = 20, j = 5, 10, 25$.*

5.2. Image denoising. In this section, we evaluate the performance of the proposed method on image denoising. The peak signal-to-noise ratio (PSNR) [33] and the structural similarity index (SSIM) [49] are used to evaluate the recovery performance quantitatively.

5.2.1. Image with sparse noise. We conduct experiments on four images: "house", "moto", "face", and "hat". In this study, we model the clean images as

TABLE 1
Recovery results on the synthetic datasets with different settings.

Method			TNN		VBI_{TNN}		PSTNN		$\text{VBI}_{\text{PSTNN}}$	
σ	r	ρ	$\text{error}_{\mathcal{L}}$	$\text{error}_{\mathcal{S}}$	$\text{error}_{\mathcal{L}}$	$\text{error}_{\mathcal{S}}$	$\text{error}_{\mathcal{L}}$	$\text{error}_{\mathcal{S}}$	$\text{error}_{\mathcal{L}}$	$\text{error}_{\mathcal{S}}$
10^{-3}	3	0.01	0.0029	0.0075	0.0025	0.0056	0.0028	0.0064	**0.0023**	**0.0052**
		0.1	0.0034	0.0027	0.0032	0.0025	0.0033	0.0024	**0.0029**	**0.0023**
	5	0.01	0.0026	0.0083	0.0025	0.0063	0.0024	0.0070	**0.0022**	**0.0058**
		0.1	0.0033	0.0033	0.0036	0.0032	**0.0030**	**0.0028**	0.0031	0.0029
10^{-2}	3	0.01	0.0286	0.0738	0.0248	0.0556	0.0276	0.0638	**0.0230**	**0.0523**
		0.1	0.0344	0.0274	0.0302	0.0238	0.0325	0.0240	**0.0275**	**0.0223**
	5	0.01	0.0257	0.0820	0.0242	0.0620	0.0240	0.0700	**0.0219**	**0.0576**
		0.1	0.0331	0.0329	0.0322	0.0294	0.0298	0.0281	**0.0281**	**0.0267**
10^{-1}	3	0.01	0.2744	0.7227	0.2317	0.5435	0.2769	0.6398	**0.2255**	**0.5195**
		0.1	0.3222	0.2623	0.2730	0.2262	0.3264	0.2410	**0.2661**	**0.2187**
	5	0.01	0.2392	0.7841	0.2201	0.5921	0.2346	0.6896	**0.2077**	**0.5620**
		0.1	0.2903	0.2961	0.2692	0.2589	0.2864	0.2705	**0.2543**	**0.2484**

the low-rank component and random corruptions as sparse outliers. Each image is corrupted by setting 10 percent of the pixels to random values ranging from 0 to 255, with the locations of these distortions unspecified. We compare our proposed method with several existing techniques, including LRTV [20], $S_{wp}(0.9)$ [51], BTRTF [55], TNN [33], and PSTNN [23], using the original implementations provided by the respective authors. Given the absence of Gaussian noise in this task, the parameter θ_1 is set to a high value of 100 to accommodate this condition, while θ_2 and θ_3 are set to 1. The truncation parameter K for $\text{VBI}_{\text{PSTNN}}$ is consistently set at 50 across all cases.

Quantitative evaluations based on PSNR and SSIM are presented in Table 2, and the corresponding restored images are displayed in Figure 3. Our observations indicate that $\text{VBI}_{\text{PSTNN}}$ consistently outperforms the other methods in terms of PSNR, achieving at least a 0.5 improvement and matching the best-performing methods in SSIM values. Additionally, the restoration of the "hat" image by $\text{VBI}_{\text{PSTNN}}$ and BTRTF shows significantly clearer text compared to other methods. However, some artifacts are noted in the "moto" image restored by BTRTF. In contrast, our method exhibits fewer artifacts across all cases.

TABLE 2
Quantitative comparisons of sparse noise removal results obtained by different methods

Data	Index	LRTV	$S_{wp}(0.9)$	BTRTF	TNN	PSTNN	VBI_{TNN}	$\text{VBI}_{\text{PSTNN}}$
house	PSNR	26.167	28.028	25.930	27.030	27.522	26.878	**28.565**
	SSIM	0.9517	0.9717	0.9374	0.9655	0.9691	0.9596	**0.9741**
moto	PSNR	27.617	28.003	24.871	26.373	27.724	25.945	**28.781**
	SSIM	0.9590	0.9702	0.9130	0.9554	0.9672	0.9440	**0.9719**
face	PSNR	32.524	34.061	32.500	30.770	31.543	30.704	**34.150**
	SSIM	0.9529	**0.9759**	0.9405	0.9509	0.9557	0.9475	0.9694
hat	PSNR	32.626	32.787	32.558	29.453	30.895	29.755	**33.478**
	SSIM	0.9435	**0.9750**	0.9581	0.9473	0.9558	0.9516	0.9735
mean	PSNR	29.733	30.720	28.965	28.407	29.421	28.321	**31.244**
	SSIM	0.9518	**0.9732**	0.9375	0.9548	0.9620	0.9507	0.9722

5.2.2. Image with mixed noise. In this subsection, we perform experiments on four distinct images: "kid", "house", "river", and "hat". Initially, each image is corrupted with sparse noise, following the procedure of our previous experiment. Subsequently, we introduce Gaussian noise to each pixel, modeled by the distribution

TABLE 3
Quantitative comparisons of mixed noise removal results obtained by different methods

Data	Index	3DTNN	$S_{wp}(0.9)$	BTRTF	TNN	PSTNN	VBI_{TNN}	$\text{VBI}_{\text{PSTNN}}$
kid	PSNR	26.670	31.806	32.071	28.691	29.542	29.446	**32.802**
	SSIM	0.9364	**0.9752**	0.9593	0.9487	0.9558	0.9521	0.9720
house	PSNR	27.448	32.302	30.791	29.765	30.459	29.862	**32.496**
	SSIM	0.9292	**0.9708**	0.9370	0.9474	0.9532	0.9414	0.9659
river	PSNR	24.606	26.388	23.818	25.985	26.439	25.367	**26.968**
	SSIM	0.9319	0.9471	0.8606	0.9466	**0.9515**	0.9291	0.9504
hat	PSNR	28.017	32.771	32.553	29.449	30.891	29.753	**33.463**
	SSIM	0.9359	0.9747	0.9581	0.9471	0.9555	0.9514	**0.9733**
mean	PSNR	26.685	30.817	29.808	28.473	29.333	28.607	**31.432**
	SSIM	0.9334	**0.9670**	0.9288	0.9475	0.9540	0.9436	0.9654

$\mathcal{N}(0, 10^{-3})$. The resultant observation, represented mathematically by $\mathcal{X} = \mathcal{L}+\mathcal{S}+\mathcal{E}$, consists of the real image $\mathcal{L}$, augmented by sparse noise $\mathcal{S}$ and Gaussian noise $\mathcal{E}$. To verify that our method's effectiveness is robust to initial conditions, we set the initial values of θ_1 to 100, and θ_2 and θ_3 to 1, as the same as the ones used in the sparse noise-only scenario.

We benchmark our proposed algorithm against several state-of-the-art methods, including 3DTNN [53], $S_{wp}(0.9)$ [51], BTRTF [55], TNN [33], and PSTNN [23]. Performance metrics such as PSNR and SSIM are detailed in Table 3, with visual results presented in Figure 4. Notably, our algorithm outperforms both TNN and PSTNN—methods that utilize similar regularization techniques—across all test cases in terms of PSNR, achieving an average improvement of 0.6 dB over the best-reported results. Qualitatively, the images restored by $\text{VBI}_{\text{PSTNN}}$ exhibit notably sharper boundaries compared to those produced by the other methods, which tend to exhibit some degree of blurring.

FIG. 3. *Comparison of color image Gaussian noise removal performance on four examples.*

5.3. Background modeling. The background modeling problem focuses on distinguishing foreground objects from the background in video sequences. This is commonly achieved by modeling the background as a low-rank tensor, which represents the relatively static scenes across different frames, and treating the moving foreground objects as sparse components. In the context of Tensor Robust Principal

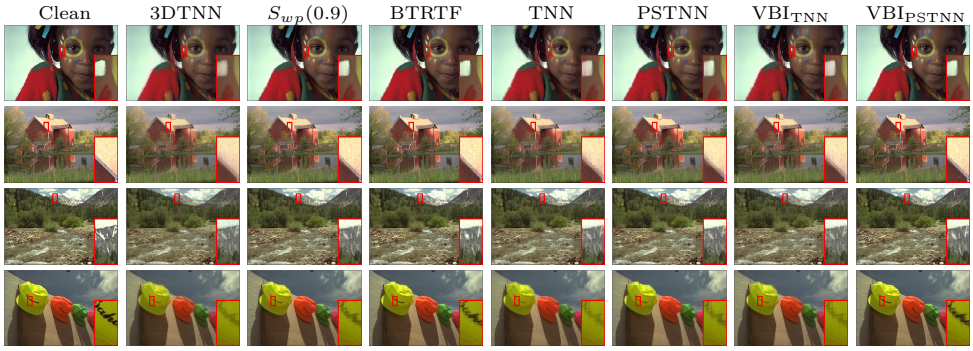

FIG. 4. *Comparison of color image mixed noise removal performance on four examples.*

Component Analysis (TRPCA), these are represented by the low-rank tensor $\mathcal{L}_0$ and the sparse tensor $\mathcal{S}_0$, respectively.

We evaluated our models on sequences from the 12R dataset [32], specifically the "bootstrap" ($120 \times 160 \times 400$), and "sidewalk" ($220 \times 352 \times 400$) videos, all characterized by slow-moving objects against varying backgrounds. Our models were compared with several others, including 3DTNN, TNN, BTRTF, PSTNN, and t-$S_{w,p}(0.9)$. For $\text{VBI}_{\text{PSTNN}}$, the truncated parameter K is set as 5, while the initial values of $\theta_1, \theta_2, \theta_3$ are set as 1, 1, 100, respectively. The results of these comparisons are visually presented in Figure 5. Each video's analysis starts with a frame from the sequence as shown in column (a) of Figure 5, followed by background images generated by the respective methods, from 3DTNN to our approach $\text{VBI}_{\text{PSTNN}}$. Additionally, the motion in each scene is depicted in the second row for each video. In the "bootstrap" video, except for 3DTNN, all the methods achieved superior background separation with fewer ghost silhouettes. In the "sidewalk" videos, all the approaches perform similarly, while 3DTNN has slightly better results.

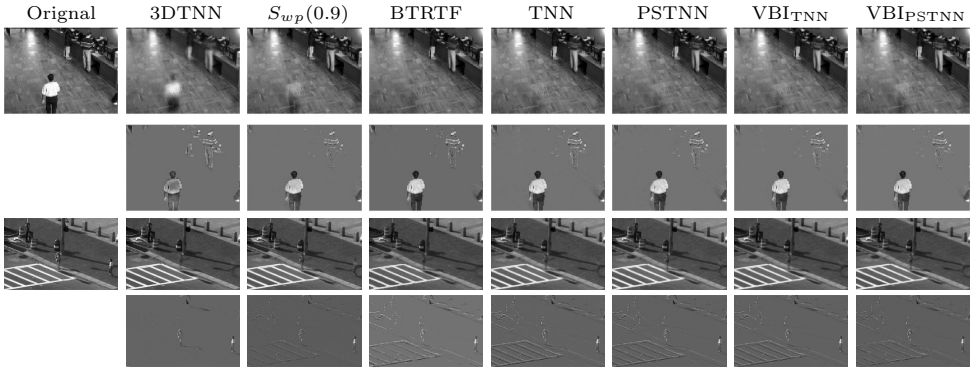

FIG. 5. *Background modeling results of two surveillance video sequences.*

6. Conclusions. In this paper, we presented a method for recovering low-rank tensors from observations contaminated by sparse outliers and Gaussian noise. Utilizing variational Bayesian inference, we effectively resolved the tensors while simultaneously selecting model parameters. Numerical evaluations highlight the advantages and superior performance of our approach compared to existing methods. Currently

limited to linear and convex relaxations, our future work will explore extending this parameter selection technique to nonconvex approximations within tensor recovery models.

REFERENCES

[1] M. Almeida and M. Figueiredo, *Parameter estimation for blind and non-blind deblurring using residual whiteness measures*, IEEE Trans. Image Process., 22 (2013), pp. 2751–2763.

[2] B. Amizic, R. Molina, and A. K. Katsaggelos, *Sparse Bayesian blind image deconvolution with parameter estimation*, EUSIPCO, 2012 (2012), pp. 1–15.

[3] S. D. Babacan, R. Molina, and A. K. Katsaggelos, *Parameter estimation in TV image restoration using variational distribution approximation*, IEEE Trans. Image Process., 17 (2008), pp. 326–339.

[4] J. M. Bardsley, *MCMC-based image reconstruction with uncertainty quantification*, SIAM J. Sci. Comput., 34 (2012), pp. A1316–A1332.

[5] J. A. Bazerque, G. Mateos, and G. B. Giannakis, *Rank regularization and Bayesian inference for tensor completion and extrapolation*, IEEE Trans. Signal Process., 61 (2013), pp. 5689–5703, https://doi.org/10.1109/TSP.2013.2278516, http://ieeexplore.ieee.org/document/6579771/ (accessed 2024-09-01).

[6] F. Bevilacqua, A. Lanza, M. Pragliola, and F. Sgallari, *Whiteness-based parameter selection for poisson data in variational image processing*, Applied Mathematical Modelling, 117 (2023), pp. 197–218.

[7] A. Bhattacharya, D. Pati, and Y. Yang, *On the convergence of coordinate ascent variational inference*, The Annals of Statistics, 53 (2025), pp. 929–962.

[8] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, *Variational inference: A review for statisticians*, Journal of the American statistical Association, 112 (2017), pp. 859–877.

[9] H. Cai, Z. Chao, L. Huang, and D. Needell, *Robust tensor cur decompositions: Rapid low-Tucker-rank tensor recovery with sparse corruptions*, SIAM J. Imag. Sci., 17 (2024), pp. 225–247.

[10] J. Cai, E. Candes, and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM Journal on Optimization, 20 (2010), pp. 1956–1982.

[11] E. Candes and B. Recht, *Exact matrix completion via convex optimization*, Commun. ACM, 55 (2012), pp. 111–119.

[12] Y. Chang, L. Yan, X.-L. Zhao, H. Fang, Z. Zhang, and S. Zhong, *Weighted low-rank tensor recovery for hyperspectral image restoration*, IEEE Trans. Cybern., 50 (2020), pp. 4558–4572.

[13] G. Chantas, N. P. Galatsanos, R. Molina, and A. K. Katsaggelos, *Variational bayesian image restoration with a product of spatially weighted total variation image priors*, IEEE Trans. Image Process., 19 (2009), pp. 351–362.

[14] S. Gandy, B. Recht, and I. Yamada, *Tensor completion and low-n-rank tensor recovery via convex optimization*, Inverse Probl., 27 (2011), p. 025010.

[15] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su, *A weakly informative default prior distribution for logistic and other regression models*, Ann Appl. Stat., 2 (2008), pp. 1360–1383.

[16] J. Glaubitz and A. Gelb, *Leveraging joint sparsity in hierarchical Bayesian learning*, SIAM/ASA J. Uncertainty Quantif., 12 (2024), pp. 442–472.

[17] D. Goldfarb and Z. Qin, *Robust low-rank tensor recovery: Models and algorithms*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 225–253.

[18] G. Golub, M. Heath, and G. Wahba, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, 21 (1979), pp. 215–223.

[19] P. Hansen and D. O'Leary, *The use of the L-curve in the regularization of discrete ill-posed problems*, SIAM J. Sci. Comput, 14 (1993), pp. 1487–1503.

[20] W. He, H. Zhang, L. Zhang, and H. Shen, *Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration*, IEEE Trans. Geosci. Remote Sens., 54 (2015), pp. 178–188.

[21] T. Helin, N. Hyvönen, and J.-P. Puska, *Edge-promoting adaptive Bayesian experimental design for X-ray Imaging*, SIAM J. Sci. Comput., 44 (2022), pp. B506–B530.

[22] J. Jia, Q. Zhao, Z. Xu, D. Meng, and Y. Leung, *Variational Bayes' method for functions with applications to some inverse problems*, SIAM J. Sci. Comput., 43 (2021), pp. A355–A383.

[23] T.-X. Jiang, T.-Z. Huang, X.-L. Zhao, and L.-J. Deng, *Multi-dimensional imaging data*

recovery via minimizing the partial sum of tubal nuclear norm, J. Comput. Appl. Math., 372 (2020), p. 112680.
[24] I. T. Jolliffe and J. Cadima, *Principal component analysis: a review and recent developments*, Phil. Trans. R. Soc. A., 374 (2016), p. 20150202.
[25] H. A. Kiers, *Towards a standardized notation and terminology in multiway analysis*, J. Chemom., 14 (2000), pp. 105–122.
[26] M. E. Kilmer, K. Braman, N. Hao, and R. C. Hoover, *Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 148–172.
[27] M. E. Kilmer and C. D. Martin, *Factorization strategies for third-order tensors*, Linear Algebra Appl., 435 (2011), pp. 641–658.
[28] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.
[29] D. S. Lalush and B. M. Tsui, *Simulation evaluation of Gibbs prior distributions for use in maximum a posteriori SPECT reconstructions*, IEEE Trans. Med. Imaging, 11 (1992), pp. 267–275.
[30] K. J. H. Law and V. Zankin, *Sparse online variational Bayesian regression*, SIAM/ASA J. Uncertainty Quantif., 10 (2022), pp. 1070–1100.
[31] S. Lefkimmiatis and I. Koshelev, *Learning sparse and low-rank priors for image recovery via iterative reweighted least squares minimization*, in Proc. ICLR, 2023.
[32] L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian, *Statistical modeling of complex backgrounds for foreground object detection*, IEEE Trans. Image Process., 13 (2004), pp. 1459–1472.
[33] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, *Tensor robust principal component analysis with a new tensor nuclear norm*, IEEE Trans. Pattern Anal. Mach. Intell., 42 (2020), pp. 925–938.
[34] C. Menzen, M. Kok, and K. Batselier, *Alternating linear scheme in a Bayesian framework for low-rank tensor approximation*, SIAM J. Sci. Comput., 44 (2022), pp. A1116–A1144.
[35] V. Morozov, *Methods for solving incorrectly posed problems*, Springer-Verlag, New York, 1984.
[36] C. Mu, B. Huang, J. Wright, and D. Goldfarb, *Square deal: Lower bounds and improved relaxations for tensor recovery*, in Proc. ICML, PMLR, 2014, pp. 73–81.
[37] Y. Mu, P. Wang, L. Lu, X. Zhang, and L. Qi, *Weighted tensor nuclear norm minimization for tensor completion using tensor-SVD*, Pattern Recognit. Lett., 130 (2020), pp. 4–11.
[38] J. P. Oliveira, J. M. Bioucas-Dias, and M. A. Figueiredo, *Adaptive total variation image deblurring: a majorization–minimization approach*, Signal Process., 89 (2009), pp. 1683–1693.
[39] Y. Panagakis, J. Kossaifi, G. G. Chrysos, J. Oldfield, M. A. Nicolaou, A. Anandkumar, and S. Zafeiriou, *Tensor methods in computer vision and deep learning*, Proc. IEEE, 109 (2021), pp. 863–890.
[40] A.-H. Phan, K. Sobolev, K. Sozykin, D. Ermilov, J. Gusak, P. Tichavskỳ, V. Glukhov, I. Oseledets, and A. Cichocki, *Stable low-rank tensor decomposition for compression of convolutional neural network*, in Proc. ECCV, 2020, pp. 522–539.
[41] Y. Su, H. Zhu, K.-C. Wong, Y. Chang, and X. Li, *Hyperspectral image denoising via weighted multidirectional low-rank tensor recovery*, IEEE Trans. Cybern., 53 (2022), pp. 2753–2766.
[42] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological), (1996), pp. 267–288.
[43] M. E. Tipping, *Sparse bayesian learning and the relevance vector machine*, J. Mach. Learn. Res., 1 (2001), pp. 211–244.
[44] X. Tong, L. Cheng, and Y.-C. Wu, *Bayesian tensor Tucker completion with a flexible core*, IEEE Trans. Signal Process., 71 (2023), pp. 4077–4091.
[45] L. R. Tucker, *Some mathematical notes on three-mode factor analysis*, Psychometrika, 31 (1966), pp. 279–311.
[46] F. Uribe, Y. Dong, and P. C. Hansen, *Horseshoe priors for edge-preserving linear Bayesian inversion*, SIAM J. Sci. Comput., 45 (2023), pp. B337–B365.
[47] C. Wang and D. M. Blei, *Variational inference in nonconjugate models*, J. Mach. Learn. Res., (2013).
[48] H. Wang, J. Peng, W. Qin, J. Wang, and D. Meng, *Guaranteed tensor recovery fused low-rankness and smoothness*, IEEE Trans. Pattern Anal. Mach. Intell., 45 (2023), pp. 10990–11007.
[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, *Image quality assessment: from error visibility to structural similarity*, IEEE Trans. Image Process., 13 (2004), pp. 600–612.
[50] L. Yang, J. Fang, H. Duan, H. Li, and B. Zeng, *Fast low-rank Bayesian matrix completion with hierarchical Gaussian prior models*, IEEE Trans. Signal Process., 66 (2018), pp. 2804–

2817.
[51] M. Yang, Q. Luo, W. Li, and M. Xiao, *Nonconvex 3D array image data recovery and pattern recognition under tensor framework*, Pattern Recognit., 122 (2022), p. 108311.
[52] H. Zheng, Y. Lou, G. Tian, and C. Wang, *A scale-invariant relaxation in low-rank tensor recovery with an application to tensor completion*, SIAM J. Imag. Sci., 17 (2024), pp. 756–783.
[53] Y.-B. Zheng, T.-Z. Huang, X.-L. Zhao, T.-X. Jiang, T.-H. Ma, and T.-Y. Ji, *Mixed noise removal in hyperspectral image via low-fibered-rank regularization*, IEEE Trans. Geosci. Remote Sens., 58 (2019), pp. 734–749.
[54] X. Zhou, Q. Heng, E. C. Chi, and H. Zhou, *Proximal MCMC for Bayesian inference of constrained and regularized estimation*, Am. Stat., (2024), pp. 1–12.
[55] Y. Zhou and Y.-M. Cheung, *Bayesian low-tubal-rank robust tensor factorization with multi-rank determination*, IEEE Trans. Pattern Anal. Mach. Intell., 43 (2019), pp. 62–76.