SIAM J. IMAGING SCIENCES
Vol. 17, No. 1, pp. 248–272

Image Segmentation Using Bayesian Inference for Convex Variant Mumford–Shah Variational Model*

Xu Xiao†, Youwei Wen‡, Raymond Chan§, and Tieyong Zeng¶

Abstract. The Mumford–Shah model is a classical segmentation model, but its objective function is nonconvex. The smoothing and thresholding (SaT) approach is a convex variant of the Mumford–Shah model, which seeks a smoothed approximation solution to the Mumford–Shah model. The SaT approach separates the segmentation into two stages: first, a convex energy function is minimized to obtain a smoothed image; then, a thresholding technique is applied to segment the smoothed image. The energy function consists of three weighted terms and the weights are called the regularization parameters. Selecting appropriate regularization parameters is crucial to achieving effective segmentation results. Traditionally, the regularization parameters are chosen by trial-and-error, which is a very time-consuming procedure and is not practical in real applications. In this paper, we apply a Bayesian inference approach to infer the regularization parameters and estimate the smoothed image. We analyze the convex variant Mumford–Shah variational model from a statistical perspective and then construct a hierarchical Bayesian model. A mean field variational family is used to approximate the posterior distribution. The variational density of the smoothed image is assumed to have a Gaussian density, and the hyperparameters are assumed to have Gamma variational densities. All the parameters in the Gaussian density and Gamma densities are iteratively updated. Experimental results show that the proposed approach is capable of generating high-quality segmentation results. Although the proposed approach contains an inference step to estimate the regularization parameters, it requires less CPU running time to obtain the smoothed image than previous methods.

Key words. image segmentation, Mumford–Shah model, Bayesian inference, mean field variational approximation, regularization parameters

MSC codes. 68-U10, 94-A08, 90-C99, 62-F99

DOI. 10.1137/23M1545379

1. Introduction.

Image segmentation is an important task in image processing with applications in various fields such as science, engineering, medicine, and commerce. Its aim is to

*Received by the editors January 5, 2023; accepted for publication (in revised form) August 25, 2023; published electronically January 30, 2024.
https://doi.org/10.1137/23M1545379
Funding: This work is supported in part by NSFC grant 12361089; the National Key R&D Program of China under grants 2021YFE0203700, NSFC/RGC N_CUHK 415/19, ITF MHP/038/20, CRF 8730063, RGC 14300219, 14302920, 14301121, and CUHK Direct Grant for Research, HKRGC grants CityU11301120, CityU11309922, C1013-21GF, CityU grant 9380101; HKRGC-NSFC grant N_CityU214/19.

†Key Laboratory of Computing and Stochastic Mathematics (LCSM), School of Mathematics and Statistics, Hunan Normal University, Changsha, Hunan, China (xiaoxu@hunnu.edu.cn).
‡Corresponding author and cofirst author. Key Laboratory of Computing and Stochastic Mathematics (LCSM), School of Mathematics and Statistics, Hunan Normal University, Changsha, Hunan, China (wenyouwei@gmail.com).
§Department of Mathematics, City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, Hong Kong SAR, China; Hong Kong Centre for Cerebro-Cardiovascular Health Engineering, Hong Kong (raymond.chan@cityu.edu.hk).
¶The Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong, China (zeng@math.cuhk.edu.hk).

partition the image into multiple parts or regions based on the characteristics of the pixels in the image, allowing for subsequent analysis such as object recognition [37, 60, 73], image compression/editing [9, 55], and occlusion boundary estimation within motion or stereo systems [43, 75, 86]. There are many efficient approaches for image segmentation, e.g., the mathematical model-based approaches [72, 91], pattern recognition techniques [87], tracking-based approaches [42], statistical approaches [29], and artificial intelligence approaches [58], etc.

Image segmentation is intrinsically difficult because the images are usually degraded by some blurs and corrupted by noise. In general, the standard image degradation model is

$$\mathbf{g} = \mathcal{H}\mathbf{u} + \mathbf{n}, \tag{1.1}$$

where $\mathbf{g} \in \mathbb{R}^n$ is the degraded image, $\mathbf{u} \in \mathbb{R}^m$ is the clean image, $\mathbf{n} \in \mathbb{R}^n$ is the noise or the texture, and the matrix $\mathcal{H} \in \mathbb{R}^{n\times m}$ can be the identity operator for the denoising problem or a blurring operator for the image restoration problem. The purpose of image segmentation is to obtain a smoothed image $\mathbf{u}$ from the given image $\mathbf{g}$ such that $\mathbf{u}$ has the piecewise continuous regions with sharp boundaries and contains the global structural information. Mumford and Shah [65, 66] proposed a classical image segmentation model, which has been studied extensively [24, 26, 39, 52, 56, 72, 78, 88]. The model is to seek an optimal piecewise smooth approximation $\mathbf{u}$ by minimizing the following function:

$$\min_{\mathbf{u},\Gamma} \left\{ \frac{\beta}{2} \int_\Omega (\mathbf{g} - \mathcal{H}\mathbf{u})^2 dx + \frac{\lambda_1}{2} \int_{\Omega\setminus\Gamma} |\nabla \mathbf{u}|^2 dx + \lambda_2 \cdot \text{Length}(\Gamma) \right\}. \tag{1.2}$$

Here Ω is a bounded open connected set where $\mathbf{u}$ is defined, Γ is a compact curve in Ω. The first term in (1.2) represents the data-fitting term and the next two terms represent regularization terms. The positive parameters β, λ_1, and λ_2 measure the trade-off among the fidelity between $\mathbf{g}$ and $\mathcal{H}\mathbf{u}$, the smoothness of $\mathbf{u}$ in the region $\Omega \setminus \Gamma$, and the regularity of Γ, respectively.

Since the energy function in (1.2) is nonconvex, it is difficult to find its minimizer [39]. In order to deal with this numerical difficulty, many simplified models have been proposed, such as restricting $\mathbf{u}$ to be a piecewise constant function [52, 78]. Instead of seeking a piecewise constant solution, a convex approximation approach of the Mumford–Shah model (1.2) was proposed to obtain a smooth approximate solution in [21, 23]. This approach consists of two stages. In the first stage, a smoothing procedure is applied by minimizing the convex function

$$\min_{\mathbf{u}} \mathcal{J}(\mathbf{u}; \beta, \lambda_1, \lambda_2) \equiv \left\{ \frac{\beta}{2} \int_\Omega (\mathbf{g} - \mathcal{H}\mathbf{u})^2 dx + \frac{\lambda_1}{2} \int_\Omega |\nabla \mathbf{u}|^2 dx + \lambda_2 \int_\Omega |\nabla \mathbf{u}| dx \right\}. \tag{1.3}$$

The second stage involves employing a thresholding technique, such as the K-means method, to segment the minimizer $\mathbf{u}$ of (1.3). The convex approximation objective function can produce exact solutions for the Mumford–Shah model in (1.2) for two-phase disk images [21, 23].

The selection of the regularization parameters in the model (1.3) affects the performance of segmentation. Although there are three parameters in the minimization problem (1.3), we can fix the parameter β and then tune the other two parameters. If the parameter λ_1 is set to 0, the middle term in (1.3) vanishes and the problem becomes the classical total-variation (TV)

image restoration problem. There are papers that propose methods to choose the optimal λ_2; see [5, 80]. However, the middle term in (1.3) comes from the Mumford–Shah model (1.2) and it is crucial because it allows small local variations in the minimizer $\mathbf{u}$. Without this term, there will be staircase artifacts in the solution, which would lead to a spurious boundary in the thresholded segmented image [21, 23].

In [21], the regularization parameters $\{\lambda_i\}_{i=1}^2$ in (1.3) are selected by the trial-and-error method. This means that we solve the minimization problem (1.3) with various values of $\{\lambda_i\}_{i=1}^2$. Then the optimal parameters $\{\lambda_i\}_{i=1}^2$ are those that give the best segmentation quality in the solution $\mathbf{u}$ when compared with the ground truth. However, the trial-and-error approach is computationally expensive because it involves trying all combinations of regularization parameters, and it is time consuming. Moreover, this approach may not be practical because we may not have access to the ground truth in practice.

Several approaches have been proposed to automatically select regularization parameters for some classical variational problems, which include the generalized cross validation method [41], discrepancy principle method [64], joint maximum a posterior method (JMAP) [14], Bayesian inference method [27], and so on. In this paper, we apply the variational Bayesian inference to select the regularization parameters for problem (1.3). Variational Bayesian inference has been successfully applied to TV-regularized image restoration and blind image deconvolution problem [27, 34, 54]. Similarly to the JMAP method, the variational Bayesian inference method can simultaneously estimate the image and hyperparameter (i.e., regularization parameters). However, while the JMAP method is to find the image $\mathbf{u}$ and the hyperparameters $\beta, \lambda_1, \lambda_2$ that maximize the posterior density $p(\mathbf{u}, \theta|\mathbf{g})$ (here $\theta = (\beta, \lambda_1, \lambda_2)$), the variational Bayesian inference method is to seek a probability density function $q(\mathbf{u}, \theta)$ from a set of tractable distribution families that approximates the posterior density $p(\mathbf{u}, \theta|\mathbf{g})$ and then to infer the image $\mathbf{u}$ and the hyperparameters $\beta, \lambda_1, \lambda_2$ from $q(\mathbf{u}, \theta)$.

The Kullback–Leibler divergence is used to measure the closeness of the density $q(\mathbf{u}, \theta)$ and the posterior density $p(\mathbf{u}, \theta|\mathbf{g})$. Then we reformulate the inference as an optimization problem. Under the mean field approximation assumption [13, 71], the variational density of the image $\mathbf{u}$ and the hyperparameter θ can be evaluated by the approach of coordinate ascent inference. However, obtaining the density of $\mathbf{u}$ given the density of θ is difficult because (1.3) includes a TV norm term and the TV norm prior is not conjugate to the Gaussian likelihood, which makes the inference intractable. In [27, 34, 54], the majorization-minimization approach is applied to obtain a quadratic upper bound of the TV norm. In order to avoid a zero denominator, a smoothed parameter is introduced in the quadratic function. The additional smoothed parameter is a trade-off between the quality of the restored image edges and the speed of convergence. Therefore, instead of the smoothed approach, we apply the Laplace approximation approach to infer the variational distribution parameters of the image $\mathbf{u}$. The obtained $\mathbf{u}$ achieves the minimum of a TV regularization optimization problem. This means that the $\mathbf{u}$ we get is more accurate and our speed of convergence is not affected by the extra smoothed parameter. The posterior density of the hyperparameter θ has the form of Gamma density; thus we assume that the variational density of θ is Gamma. The parameters in the densities of $\mathbf{u}$ and θ are iteratively updated. Numerical experiments show that our method competes well with other methods that use trial-and-error to determine the best parameters in terms of accuracy and is much faster in terms of time.

The rest of this paper is organized as follows. In section 2, we describe the hierarchical Bayesian model, joint density, and hyperprior. In section 3, we apply variational inference to infer the image $\mathbf{u}$ and the regularization parameters β and $\{\lambda_i\}_i^2$. Next, we show how to use appropriate thresholds to segment $\mathbf{u}$ by K-means. In section 4, we provide the experimental results on grayscale and color images. Finally, the conclusions are given in section 5.

2. Bayesian model. In order to obtain the numerical solution of the minimization problem (1.3), we formulate the objective function $\mathcal{J}(\mathbf{u};\beta,\lambda_1,\lambda_2)$ into a discretization form as follows:

$$\mathcal{J}(\mathbf{u};\beta,\lambda_1,\lambda_2) \equiv \frac{\beta}{2}\left\|\mathbf{g}-\mathcal{H}\mathbf{u}\right\|_2^2 + \frac{\lambda_1}{2}\left\|\nabla\mathbf{u}\right\|_2^2 + \lambda_2\left\|\nabla\mathbf{u}\right\|_1. \tag{2.1}$$

To simplify, we assume that all images are vectors reshaped from two-dimensional matrices of size $M\times N$. The vectors $\mathbf{g},\mathbf{u}\in X$, where X is an Euclidean subspace of $\mathbb{R}^{MN}$ with an index set Ω. The gradient $\nabla\mathbf{u}$ is in $X\times X$ given by $(\nabla\mathbf{u})_i=((\nabla_h\mathbf{u})_i,(\nabla_v\mathbf{u})_i)$, where ∇_h,∇_v represent the discrete version of the horizontal and vertical gradient operators respectively. The backward difference is used to compute the discretization of the gradient, and a periodic boundary condition is applied to extend the value of $\mathbf{u}$. We denote ∇^T being an adjoint of ∇. The TV term in (2.1) is defined as $||\nabla\mathbf{u}||_1 := \sum_{i\in\Omega}\sqrt{(\nabla_h\mathbf{u})_i^2+(\nabla_v\mathbf{u})_i^2}$.

For convenience's sake, we summarize the symbols used in this paper.

- $\mathcal{N}(\mu,\Sigma)$: Gaussian density function with mean μ and variance Σ.
- $\mathcal{G}(a,b)$: Gamma density function with shape parameter a and scale parameter b.
- $p(x)=\mathcal{N}(\mu_x,\Sigma_x)$: The variable x follows a Gaussian distribution with mean μ_x and variance Σ_x.
- $p(\cdot)$: the true density function.
- $q(\cdot)$: the variational density function .
- $q_k(\cdot)$: the variational density function in the kth iteration.

2.1. Maximum a posteriori (MAP) interpretation. In this subsection, we describe the variational problem (1.3) from the statistical perspective. In the MAP approach [2, 67], we assume that the parameters β,λ_1, and λ_2 in (1.3) are known. We model the observed noise $\mathbf{n}$ as a zero mean white Gaussian vector. Given the image $\mathbf{u}$ and the noise variance σ^2, according to the observation model (1.1), the conditional probability density of the random variable $\mathbf{g}$ is

$$p(\mathbf{g}|\mathbf{u},\beta)=\mathcal{N}(\mathbf{g}|\mathcal{H}\mathbf{u},\frac{1}{\beta}I)\propto\beta^{MN/2}\exp\left[-\frac{\beta}{2}||\mathbf{g}-\mathcal{H}\mathbf{u}||_2^2\right]. \tag{2.2}$$

Here $\mathcal{N}$ denotes the Gaussian density function, and the parameter β is related to the noise variance σ^2 by $\beta=1/\sigma^2$.

The Bayesian approach requires choosing a prior distribution on the image $\mathbf{u}$. The choice of prior can significantly affect the quality of the resulting image reconstructions. The prior corresponding to the objective function $\mathcal{J}(\mathbf{u};\beta,\lambda_1,\lambda_2)$ in (2.1) is an unnormalized hybrid Gaussian–Laplacian distribution prior for the image $\mathbf{u}$ with parameters λ_1 and λ_2,

$$p(\mathbf{u}|\lambda_1,\lambda_2)=\frac{1}{Z(\lambda_1,\lambda_2)}\exp\left(-\frac{\lambda_1}{2}\left\|\nabla\mathbf{u}\right\|_2^2-\lambda_2\left\|\nabla\mathbf{u}\right\|_1\right), \tag{2.3}$$

where $Z(\lambda_1, \lambda_2)$ is a normalizing factor known as the partition function, and it is defined by

$$Z(\lambda_1, \lambda_2) = \int \exp\left(-\frac{\lambda_1}{2}||\nabla \mathbf{u}||_2^2 - \lambda_2||\nabla \mathbf{u}||_1\right) d\mathbf{u}. \tag{2.4}$$

We remark that we can choose other priors on the image $\mathbf{u}$. These priors include the Laplace prior [7, 81, 89], Gaussian prior [28, 31, 45], simultaneous autoregression prior [63], TV prior [4, 6] and its variants [47, 48, 83, 84, 85], and so on. If we change the prior on the image $\mathbf{u}$, the objective function in the minimization problem should be modified accordingly.

We remark that the hybrid Gaussian–Laplacian mixture model (HGLMM) of the image $\mathbf{u}$ is a geometric means of the Gaussian distribution with the mean 0, the variance $\frac{1}{\lambda_1}$, the Laplacian distribution with the location parameter 0, and the scale parameter λ_2. The HGLMM has been used for various applications including image annotation [51], annealed importance sampling [69], and the averaged predictions of multiple neural networks employing a softmax layer [44]. We remark that there exist small local variations in the image; the hybrid distribution can prevent the staircase artifacts in the image $\mathbf{u}$, which would lead to spurious boundary in the threshold segmented image.

When the parameters β, λ_1, and λ_2 are given, we just need to estimate the unknown image $\mathbf{u}$. We can derive the posterior density $p(\mathbf{u}|\mathbf{g}, \beta, \lambda_1, \lambda_2)$, and the solution $\mathbf{u}$ can be obtained by a MAP approach,

$$\widehat{\mathbf{u}} = \arg\max_{\mathbf{u}} p(\mathbf{u}|\mathbf{g}, \beta, \lambda_1, \lambda_2) = \arg\max_{\mathbf{u}} \frac{p(\mathbf{g}|\mathbf{u}, \beta)p(\mathbf{u}|\lambda_1, \lambda_2)}{p(\mathbf{g})}.$$

We want to maximize the right side where the denominator $p(\mathbf{g})$ has no direct functional dependence with $\mathbf{u}$. Therefore, the denominator $p(\mathbf{g})$ can be removed. Taking the negative logarithm to the posterior density, it is straightforward to observe that the MAP approach is equivalent to the minimization problem in (2.1). The first term in (2.1) is derived from the likelihood function while the other two terms are derived from the prior of the image.

2.2. Joint density and hyperprior. In the MAP approach of the convex variant image segmentation model, we need to know the values of the regularization parameters β, λ_1, and λ_2 in order to obtain $\mathbf{u}$. However, they are unknown in practical applications and we need to choose them. In the framework of Bayesian estimation, the parameters $\beta, \lambda_1, \lambda_2$ are treated as the latent variables, which are called hyperparameters. In this paper, we assume that $\beta, \lambda_1, \lambda_2$ are independent of each other, where β is the hyperparameter of the data density $p(\mathbf{g}|\mathbf{u}, \beta)$, λ_1 and λ_2 are the hyperparameters of the image prior density $p(\mathbf{u}|\lambda_1, \lambda_2)$. For simplification, we set $\theta = (\beta, \lambda_1, \lambda_2)$. The joint density of the variable $\mathbf{g}, \mathbf{u}, \theta$ is given by

$$p(\mathbf{g}, \mathbf{u}, \theta) = p(\mathbf{g}|\mathbf{u}, \beta)p(\mathbf{u}|\lambda_1, \lambda_2)p(\lambda_1)p(\lambda_2)p(\beta). \tag{2.5}$$

The functions $p(\lambda_1)$, $p(\lambda_2)$, and $p(\beta)$ are the prior densities assigned to the hyperparameters, which are also known as hyperpriors [11]. There are three common types of hyperparametric priors: uniform prior, Jeffreys prior, and Gamma prior. The probability density function (pdf) of the uniform prior is constant, the Gamma distribution is the conjugate distribution of the Gaussian distribution [32, 67], and the prior pdf based on the Jeffreys' prior is invariant under a transformation of parameter [47].

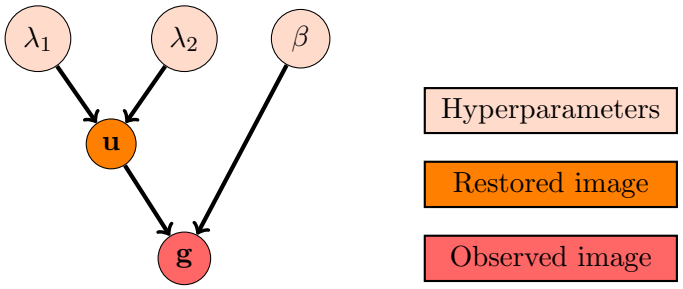


Figure 1. *Graphical model to represent the dependencies among the random variables. The edges represent the dependencies among the variables.*

We plot the dependencies among the random variables and the parameters by a graphical model shown in Figure 1. The circular nodes correspond to the random variables. The edges represent the dependencies among the variables. Each circular node represents a conditional probability density that defines the distribution of the random variable associated with that node given the values of its parent variables.

Applying the joint MAP (JMAP) estimation [14], we can estimate the image $\mathbf{u}$ and the regularization parameters $\theta = (\beta, \lambda_1, \lambda_2)$ together:

$$\begin{aligned}(\widehat{\mathbf{u}}, \widehat{\theta}) &= \underset{\mathbf{u},\theta}{\operatorname{argmax}}\, p(\mathbf{u}, \theta|\mathbf{g}) \\ &= \underset{\mathbf{u},\theta}{\operatorname{argmin}} - \log p(\mathbf{g}|\mathbf{u}, \beta) - \log p(\mathbf{u}|\lambda_1, \lambda_2) - \log p(\lambda_1) - \log p(\lambda_2) - \log p(\beta). \end{aligned} \tag{2.6}$$

We remark that the objective function of the minimization problem in (2.6) is nonconvex, which makes it difficult to find an optimal solution numerically.

3. Variational Bayesian inference.

In this section, we apply variational inference to infer $\mathbf{u}$ and θ. Unlike JMAP, the aim of variational inference does not maximize the posterior density $p(\mathbf{u}, \theta|\mathbf{g})$, but seeks a variational density $q(\mathbf{u}, \theta)$ to approximate the posterior density $p(\mathbf{u}, \theta|\mathbf{g})$, i.e., $q(\mathbf{u}, \theta) \approx p(\mathbf{u}, \theta|\mathbf{g})$, and then uses $q(\mathbf{u}, \theta)$ to infer on $\mathbf{u}$ and θ by evaluating their expectations [13]. We remark that both $q(\mathbf{u}, \theta)$ and $p(\mathbf{u}, \theta|\mathbf{g})$ are PDFs. However, $q(\mathbf{u}, \theta)$ is an approximation of the true density $p(\mathbf{u}, \theta|\mathbf{g})$ and is obtained using variational inference.

3.1. Kullback–Leibler divergence.

In variational inference [13], we specify a set of densities $\mathcal{Q}$ to approximate the posterior density $p(\mathbf{u}, \theta|\mathbf{g})$. Then we choose an optimal variational density $q(\mathbf{u}, \theta) \in \mathcal{Q}$ which is closest to $p(\mathbf{u}, \theta|\mathbf{g})$. The closeness of the two densities is measured by the Kullback–Leibler (KL) divergence, which is defined as follows:

$$\begin{aligned} KL(q(\mathbf{u}, \theta)|p(\mathbf{u}, \theta|\mathbf{g})) &= \int_{\mathbf{u},\theta} q(\mathbf{u}, \theta) \log\left(\frac{q(\mathbf{u}, \theta)}{p(\mathbf{u}, \theta|\mathbf{g})}\right) \mathrm{d}\mathbf{u}\mathrm{d}\theta \\ &= \mathbb{E}_{q(\mathbf{u},\theta)} \log\left(\frac{q(\mathbf{u}, \theta)}{p(\mathbf{u}, \theta|\mathbf{g})}\right). \end{aligned} \tag{3.1}$$

The variational inference is to find a variational density $q(\mathbf{u},\theta) \in \mathcal{Q}$ by minimizing the KL divergence, i.e.,

$$q^*(\mathbf{u},\theta) = \underset{q(\mathbf{u},\theta)\in\mathcal{Q}}{\operatorname{argmin}} KL(q(\mathbf{u},\theta)|p(\mathbf{u},\theta|\mathbf{g})).$$

According to the Bayesian rule, the conditional density of the latent variables $\mathbf{u}$ and θ given the observed image $\mathbf{g}$ can be formulated by [13]:

$$p(\mathbf{u},\theta|\mathbf{g}) = \frac{p(\mathbf{g},\mathbf{u},\theta)}{p(\mathbf{g})}.$$

Notice that the joint density $p(\mathbf{g},\mathbf{u},\theta)$ is given by (2.5). The marginal density of $\mathbf{g}$ is given by

$$\begin{aligned} p(\mathbf{g}) &= \int_{\mathbf{u},\theta} p(\mathbf{g},\mathbf{u},\theta) d\mathbf{u} d\theta \\ &= \int_{\mathbf{u},\beta,\lambda_1,\lambda_2} p(\mathbf{g}|\mathbf{u},\beta) p(\mathbf{u}|\lambda_1,\lambda_2) p(\beta) p(\lambda_1) p(\lambda_2) d\mathbf{u} d\beta d\lambda_1 d\lambda_2. \end{aligned}$$

The density $p(\mathbf{g})$ is called the evidence which is often analytically intractable and is difficult to obtain [67]. It is because we need to compute the integral in very high dimension which is generally not straightforward.

Some approximation methods have been proposed to compute the posterior density, such as the Markov chain Monte Carlo (MCMC) method [61], variational approximation [27, 50], and expectation propagation [38]. Here we focus our attention on variational approximation, because it is a deterministic optimization algorithm that has guaranteed convergence, and similarly to MCMC, it estimates the posterior without an additional step to perform inference. According to the numerical results shown in the literature [48, 49], the difference in accuracy between the Gibbs sampling and the variational inference is not significant. Therefore we compared our approach with the variational inference only.

3.2. Evidence lower bound. In this section, we review the evidence lower bound; more details can be found in [13]. Using $p(\mathbf{g},\mathbf{u},\theta) = p(\mathbf{u},\theta|\mathbf{g})p(\mathbf{g})$, we have

$$KL(q(\mathbf{u},\theta)|p(\mathbf{u},\theta|\mathbf{g})) = \log p(\mathbf{g}) - \mathcal{L}(q(\mathbf{u},\theta)),$$

where $\mathcal{L}(q(\mathbf{u},\theta))$ is the evidence lower bound (ELBO) defined as

$$\begin{aligned} \mathcal{L}(q(\mathbf{u},\theta)) &= \mathbb{E}_{q(\mathbf{u},\theta)}(\log p(\mathbf{g},\mathbf{u},\theta)) - \mathbb{E}_{q(\mathbf{u},\theta)}(\log q(\mathbf{u},\theta)) \\ &= \mathbb{E}_{q(\mathbf{u},\theta)} \log\left(\frac{p(\mathbf{g},\mathbf{u},\theta)}{q(\mathbf{u},\theta)}\right). \end{aligned} \tag{3.2}$$

According to Jensen's inequality, we have [13]

$$\log p(\mathbf{g}) = \log \int_{\mathbf{u},\theta} q(\mathbf{u},\theta) \frac{p(\mathbf{g},\mathbf{u},\theta)}{q(\mathbf{u},\theta)} \mathrm{d}\mathbf{u}\mathrm{d}\theta = \log \mathbb{E}_{q(\mathbf{u},\theta)}\left(\frac{p(\mathbf{g},\mathbf{u},\theta)}{q(\mathbf{u},\theta)}\right) \geq \mathcal{L}(q(\mathbf{u},\theta)).$$

Hence the KL divergence is nonnegative and it is equal to zero when $q(\mathbf{u},\theta) = p(\mathbf{u},\theta|\mathbf{g})$. Because the KL divergence contains the logarithm of the evidence $\log p(\mathbf{g})$, we cannot minimize KL divergence exactly. Instead we consider maximizing the ELBO

$$q^*(\mathbf{u},\theta) = \underset{q(\mathbf{u},\theta)\in\mathcal{Q}}{\operatorname{argmax}}\, \mathcal{L}(q(\mathbf{u},\theta)). \tag{3.3}$$

The reason is that maximizing the ELBO is mathematically equivalent to minimizing the KL divergence and the ELBO only requires the joint probability density $p(\mathbf{g},\mathbf{u},\theta)$ and an approximation density $q(\mathbf{u},\theta)$. The complexity of the density $q(\mathbf{u},\theta)$ determines the complexity of the maximization problem. In the following, we apply the mean field approximation method to represent $q(\mathbf{u},\theta)$.

3.3. Mean field variational approximation method. We specify the mean field variational family to approximate the posterior density. In this approach, the latent variables are mutually independent and each variable has its own variational factor [13, 71]. Therefore, we can express the variational density as

$$q(\mathbf{u},\theta) = q(\mathbf{u})q(\theta) \quad \text{and} \quad q(\theta) = q(\beta)q(\lambda_1)q(\lambda_2).$$

It is important to choose the variational densities $q(\mathbf{u})$ and $q(\theta)$. If the variational density is very complex, it will be difficult to solve the minimization problem. The image $\mathbf{u}$ has a Gaussian–Laplacian prior density and we cannot find a closed form $q(\mathbf{u})$ to approximate the posterior density of $\mathbf{u}$. One can consider employing a Gibbs sampling method [36, 68] or a Laplace approximation method [74, 76] to obtain $q(\mathbf{u})$ with closed form. The Gibbs sampling is an algorithm in MCMC while the Laplace approximation refers to approximating the complex distribution with a Gaussian distribution.

In this paper, we apply the Laplace approximation to compute the variational density $q(\mathbf{u})$ in section 3.4.2. For the parameter θ, the likelihood function of θ is the form of the Gamma density (see (2.2), (2.3)), and the posterior density of θ is Gamma. Hence $q(\theta)$ should also be a Gamma density. In fact, the Gamma density is often used as a density with nonnegative parameters; see [27, 34, 63]. The density function of Gamma is defined as

$$p(x) = \mathcal{G}(x|a_x,b_x) = \frac{(b_x)^{a_x}}{\Gamma(a_x)} x^{a_x-1}\exp[-xb_x], \tag{3.4}$$

where $\Gamma(\cdot)$ is the Gamma function, $a_x > 0, b_x > 0$ represent shape and scale parameters respectively [10]. Then, the density functions of the hyperparameters β, λ_1, and λ_2 are

$$p(\beta) = \mathcal{G}(\beta|a_\beta,b_\beta),\ p(\lambda_1) = \mathcal{G}(\lambda_1|a_{\lambda_1},b_{\lambda_1}),\ p(\lambda_2) = \mathcal{G}(\lambda_2|a_{\lambda_2},b_{\lambda_2}).$$

The mean and variance of the Gamma distribution [10] are

$$\mathbb{E}(x) = \frac{a_x}{b_x}, \quad \mathbb{V}\mathrm{ar}(x) = \frac{a_x}{(b_x)^2}. \tag{3.5}$$

Let $\mathcal{Q}_\mathcal{G}$ be the set of Gamma densities and $\mathcal{Q}_\mathcal{N}$ be the set of Gaussian densities. Then we have $\mathcal{Q} = \mathcal{Q}_\mathcal{N} \times \mathcal{Q}_\mathcal{G}$. We choose the variational density $q(\mathbf{u}) \in \mathcal{Q}_\mathcal{N}$ and $q(x) \in \mathcal{Q}_\mathcal{G}$

($x \in \{\beta, \lambda_1, \lambda_2\}$) in (3.3). More precisely, the image $\mathbf{u}$ has a Gaussian density with the mean μ and the variance Σ and the hyperparameter x ($x \in \{\beta, \lambda_1, \lambda_2\}$) has a Gamma density with the shape parameter $\widetilde{a}_x$ and the scale parameter $\widetilde{b}_x$:

$$q(\mathbf{u}) = \mathcal{N}(\mu, \Sigma) \in \mathcal{Q}_{\mathcal{N}} \quad \text{and} \quad q(\beta) = \mathcal{G}(\widetilde{a}_\beta, \widetilde{b}_\beta), q(\lambda_i) = \mathcal{G}(\widetilde{a}_{\lambda_i}, \widetilde{b}_{\lambda_i}) \in \mathcal{Q}_{\mathcal{G}}, i = 1, 2.$$

It is worth noting that although (a_x, b_x) and $(\widetilde{a}_x, \widetilde{b}_x)$ are the shape and scale parameters of the Gamma distribution, the former is the parameter of the prior $p(\theta)$ while the latter is the parameter of the approximate posterior $q(\theta)$. In this paper, we will infer the parameters μ, Σ and $\widetilde{a}_x, \widetilde{b}_x$ ($x \in \{\beta, \lambda_1, \lambda_2\}$) by using coordinate ascent variational inference.

3.4. Coordinate ascent variational inference. Coordinate ascent variational inference [13, 79] is widely applied to maximize the ELBO $\mathcal{L}(q(\mathbf{u}), q(\theta))$. Starting from an initial density $(q_0(\theta), q_0(\mathbf{u}))$ with $q_0(\theta) \in \mathcal{Q}_{\mathcal{G}}, q_0(\mathbf{u}) \in \mathcal{Q}_{\mathcal{N}}$, the densities of $\mathbf{u}$ and θ are updated as follows:

$$q_k(\theta) = \operatorname*{argmax}_{q(\theta) \in \mathcal{Q}_{\mathcal{G}}} \mathcal{L}(q_{k-1}(\mathbf{u}), q(\theta)), \tag{3.6}$$

$$q_k(\mathbf{u}) = \operatorname*{argmax}_{q(\mathbf{u}) \in \mathcal{Q}_{\mathcal{N}}} \mathcal{L}(q(\mathbf{u}), q_k(\theta)). \tag{3.7}$$

Here, $q_k(\theta), q_k(\mathbf{u})$ refer to the variational densities obtained in the kth iteration. We will discuss how to maximize (3.6) and (3.7) in the following subsections.

3.4.1. The density $q_k(\theta)$. We first consider how to obtain the density $q_k(\theta)$. Assume that we have obtained $q_{k-1}(\boldsymbol{u}) = \mathcal{N}(\mu_{\mathbf{u}}^{k-1}, \Sigma_{\mathbf{u}}^{k-1})$. The necessary condition for optimality of the unconstrained optimization problem

$$\max_{q(\theta)} \mathcal{L}(q_{k-1}(\mathbf{u}), q(\theta))$$

is the partial derivative with respect to $q(\theta)$ being equal to 0, i.e.,

$$\frac{\partial \mathcal{L}(q_{k-1}(\mathbf{u}), q(\theta))}{\partial q(\theta)} = \mathbb{E}_{q_{k-1}(\mathbf{u})}(\log p(\mathbf{g}, \mathbf{u}, \theta)) - \log q(\theta) - C_1 = 0,$$

where $C_1 = 1 + \int_{\mathbf{u}} q_{k-1}(\mathbf{u}) \log q_{k-1}(\mathbf{u})$. Therefore, we obtain

$$q_k(\theta) \propto \exp(\mathbb{E}_{q_{k-1}(\mathbf{u})}(\log p(\mathbf{g}, \mathbf{u}, \theta))). \tag{3.8}$$

By using $p(\mathbf{g}|\mathbf{u}, \beta)$ given in (2.2) and $p(\mathbf{u}|\lambda_1, \lambda_2)$ given in (2.3) and substituting them into (2.5), we obtain

$$p(\mathbf{g}, \mathbf{u}, \theta) \propto \frac{\beta^{\frac{MN}{2}}}{Z(\lambda_1, \lambda_2)} \exp\left(-\mathcal{J}(\mathbf{u}; \beta, \lambda_1, \lambda_2)\right) p(\theta). \tag{3.9}$$

Here $\mathcal{J}(\mathbf{u}; \beta, \lambda_1, \lambda_2)$ is given by (2.1) and $p(\theta) = p(\beta)p(\lambda_1)p(\lambda_2)$. In the literature [1, 5, 8, 70, 77], the Gamma distribution is widely used as the prior of the hyperparameters β, λ_1, and λ_2. However, when we choose the Gamma prior, we do not have any prior knowledge about the shape

and scale parameters in the Gamma density. It is common to fix the shape and scale parameters to very small values (for example, $a_x = b_x = 10^{-4}$ in (3.4)) such that the effect on the sampled values for the hyperparameters are negligible [1, 5, 8, 70], or they are set to zero, i.e., $a_x = b_x = 0$ [77]. In the latter case, the improper noninformative prior distributions are placed on β and λ_1, λ_2 so that $p(x) \propto 1/x$ for $x \in \{\beta, \lambda_1, \lambda_2\}$. The improper noninformative prior is also called a vague prior or weakly informative prior [35].

In this paper, the shape and scale parameters of the Gamma prior are set to zero. Consequently, the jointed density is given by

$$p(\mathbf{g}, \mathbf{u}, \theta) \propto \frac{\beta^{\frac{MN}{2}-1}}{\lambda_1 \lambda_2 Z(\lambda_1, \lambda_2)} \exp\left(-\mathcal{J}(\mathbf{u}; \beta, \lambda_1, \lambda_2)\right). \tag{3.10}$$

According to the mean field approximation method, we have $q_k(\theta) = q_k(\beta) q_k(\lambda_1) q_k(\lambda_2)$ with

$$q_k(\beta) \propto \beta^{\widetilde{a}^k_\beta - 1} \exp(-\beta \widetilde{b}^k_\beta) \quad \text{and} \quad q_k(\lambda_i) \propto \beta^{\widetilde{a}^k_{\lambda_i} - 1} \exp(-\lambda_i \widetilde{b}^k_{\lambda_i}), i = 1, 2.$$

Once we obtain the densities of the hyperparameters β, λ_i $(i = 1, 2)$, we can infer the regularization parameters by their mean as follows (see (3.5)):

$$\mu^k_\beta = \frac{\widetilde{a}^k_\beta}{\widetilde{b}^k_\beta}, \quad \mu^k_{\lambda_1} = \frac{\widetilde{a}^k_{\lambda_1}}{\widetilde{b}^k_{\lambda_1}}, \quad \text{and} \quad \mu^k_{\lambda_2} = \frac{\widetilde{a}^k_{\lambda_2}}{\widetilde{b}^k_{\lambda_2}}. \tag{3.11}$$

Now we consider how to compute $\widetilde{a}^k_x$ and $\widetilde{b}^k_x$, $x = \beta, \lambda_1, \lambda_2$. Comparing these density functions with (3.10), we can easily identify that

$$\widetilde{a}^k_\beta = \frac{MN}{2}$$

and

$$\begin{cases} \widetilde{b}^k_\beta &= \frac{1}{2}\mathbb{E}_{q_{k-1}(\mathbf{u})}(||\mathbf{g} - \mathcal{H}\mathbf{u}||_2^2), \\ \widetilde{b}^k_{\lambda_1} &= \frac{1}{2}\mathbb{E}_{q_{k-1}(\mathbf{u})}(||\nabla \mathbf{u}||_2^2), \\ \widetilde{b}^k_{\lambda_2} &= \mathbb{E}_{q_{k-1}(\mathbf{u})}(||\nabla \mathbf{u}||_1). \end{cases} \tag{3.12}$$

We observe that there exist closed-form formulas for the shape and scale parameters of the variable β, which imply that $q_k(\beta)$ in (3.8) is a feasible solution of the optimization problem in (3.6) and the constraint with respect to β is inactive. However, there are no closed-form formulas for the shape parameters of the variables λ_i, $i = 1, 2$, because the partition $Z(\lambda_1, \lambda_2)$ in (2.4) does not have a closed form. Hence, the right side of (3.10) cannot be expressed as a Gamma density with respect to λ_i, $i = 1, 2$, and the constraint with respect to λ_i, $i = 1, 2$, is active. We are required to determine the values of $\widetilde{a}^k_{\lambda_i}$, $i = 1, 2$. Applying the mean field variational method, we know that the latent variables are mutually independent and each variable has its own variational factor [13, 71]. Hence $Z(\lambda_1, \lambda_2)$ can be decomposed into the product of two independent partition functions, i.e., $Z(\lambda_1, \lambda_2) = Z_1(\lambda_1) Z_2(\lambda_2)$. In general, the partition function $Z_i(x), i = 1, 2$, can be approximated by the power function $Z_i(x) \propto x^{-\alpha MN}$, where α is a positive real number and the coefficient of MN comes from the size of the image $\mathbf{u}$

[62]. For example, $Z(\lambda) \propto \lambda^{MN/2}$ for the prior $p(\mathbf{u}) \propto \exp(-\frac{\lambda}{2}\|\nabla \mathbf{u}\|_2^2)$ in [70] and $Z(\lambda) \propto \lambda^{MN}$ for the prior $p(\mathbf{u}) \propto \exp(-\lambda\|\nabla \mathbf{u}\|_1)$ in [7]. Hence, we set

$$\widetilde{a}_{\lambda_i}^k = \alpha_i^k MN, i = 1, 2. \tag{3.13}$$

The parameters α_1 and α_2 in (3.13) will be chosen by heuristically experimental evidence [1, 5, 70] or the sequential imputation method [12, 53]. In the former approach, the parameters are fixed to constants. In the latter approach, we use an iteratively updated procedure to obtain estimates of α_1 and α_2. During the iterative procedure, we have a sequence of shape parameter $\{\widetilde{a}_{\lambda_i}^k\}$ and scale parameter $\{\widetilde{b}_{\lambda_i}^k\}$ (see (3.12) and (3.13)), and then the mean of the parameter λ_i can be obtained by $\mu_{\lambda_i}^k = \widetilde{a}_{\lambda_i}^k / \widetilde{b}_{\lambda_i}^k$ $(i = 1, 2)$. Assuming that we have obtained d pairs of $(\widetilde{a}_{\lambda_i}^j, \widetilde{b}_{\lambda_i}^j)$ $(i = 1, 2,\ j = 0, 1, \ldots, d)$ and the parameters $\widetilde{a}_{\lambda_i}^j, \widetilde{b}_{\lambda_i}^j$ are independent (see [53] for details), then the posterior mean of λ_i can be estimated by the sequential imputations

$$\widetilde{\mu}_{\lambda_i} = \frac{1}{d+1}\sum_{j=0}^{d} \mathbb{E}(\lambda_i | \widetilde{a}_{\lambda_i}^j, \widetilde{b}_{\lambda_i}^j), \quad i = 1, 2, \tag{3.14}$$

where $\widetilde{\mu}_{\lambda_i} (i = 1, 2)$ is a natural unbiased estimate of $\mu_{\lambda_i} (i = 1, 2)$ [12, 53]. By expectation of the Gamma distribution, we have

$$\mathbb{E}(\lambda_i | \widetilde{a}_{\lambda_i}^k, \widetilde{b}_{\lambda_i}^k) = \mu_{\lambda_i}^k, \quad i = 1, 2.$$

By the relationship among the shape parameter a, the scale parameter b, and the expectation μ, we have $\mu = a/b$. It is natural to update the shape parameter $\widetilde{a}_{\lambda_i}^k$ as $\widetilde{a}_{\lambda_i}^k = \widetilde{b}_{\lambda_i}^k \widetilde{\mu}_{\lambda_i}^k$; here

$$\widetilde{\mu}_{\lambda_i}^k = \frac{1}{k}\sum_{j=0}^{k-1} \mathbb{E}(\lambda_i | \widetilde{a}_{\lambda_i}^j, \widetilde{b}_{\lambda_i}^j).$$

Then we obtain

$$\widetilde{a}_{\lambda_i}^k = \frac{\widetilde{b}_{\lambda_i}^k}{k}\sum_{j=0}^{k-1} \mu_{\lambda_i}^j. \tag{3.15}$$

In (3.12), to get the scale parameters in the Gamma density, one needs to compute the expectation over $\mathbf{u}$. We now consider how to compute these expectations. Let $\mathrm{Tr}(A)$ be the trace of the matrix A. It is easy to check that, if $q(\mathbf{x}) = \mathcal{N}(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$, we have

$$\mathbb{E}_{q(\mathbf{x})}(\|A\mathbf{x} - \mathbf{b}\|_2^2) = \|A\mu_{\mathbf{x}} - \mathbf{b}\|_2^2 + \mathrm{Tr}(A^T A \Sigma_{\mathbf{x}}).$$

Hence, using $q_{k-1}(\mathbf{u}) = \mathcal{N}(\mu_{\mathbf{u}}^{k-1}, \Sigma_{\mathbf{u}}^{k-1})$, we have

$$\begin{cases} \mathbb{E}_{q_{k-1}(\mathbf{u})}(||\mathbf{g} - \mathcal{H}\mathbf{u}||_2^2) = ||\mathbf{g} - \mathcal{H}\mu_{\mathbf{u}}^{k-1}||_2^2 + \mathrm{Tr}(\mathcal{H}^T \mathcal{H} \Sigma_{\mathbf{u}}^{k-1}), \\ \mathbb{E}_{q_{k-1}(\mathbf{u})}(||\nabla \mathbf{u}||_2^2) = ||\nabla \mu_{\mathbf{u}}^{k-1}||_2^2 + \mathrm{Tr}(\nabla^T \nabla \Sigma_{\mathbf{u}}^{k-1}), \end{cases} \tag{3.16}$$

where $\mathcal{H}^T, \nabla^T$ are the conjugate transpose of $\mathcal{H}$ and ∇, respectively. The expectation of $\|\nabla \mathbf{u}\|_1$ is difficult to evaluate due to the form of the TV prior. One approach is to apply the

method of iteratively reweighted least squares (IRLS) to approximate $\|\nabla \mathbf{u}\|_1$; see [30]. In the IRLS approach, the L_1-norm $\|\mathbf{x}\|_1$ is represented by the weighted least square norm $\|\mathbf{x}\|^2_{W_k}$ with iteratively updated weights W_k. Here W_k is a diagonal matrix with diagonal entries $1/(\mathbf{x}_k^2+\epsilon)^{-1/2}$ with a given parameter $\epsilon>0$, and $\mathbf{x}_k$ is the solution at the kth iteration step. Here we apply the weighted norm $\|\nabla \mathbf{u}\|^2_{W_{k-1}}$ to approximate $\|\nabla \mathbf{u}\|_1$. The diagonal entry of the weighted matrix W_{k-1} is given by $1/(\left|\nabla \mu_{\mathbf{u}}^{k-1}\right|+\epsilon)$, where ϵ is a positive number set to 10^{-3} in the paper. Hence we have

$$\mathbb{E}_{q_{k-1}(\mathbf{u})}(\|\nabla \mathbf{u}\|_1) \approx \left\|\nabla \mu_{\mathbf{u}}^{k-1}\right\|^2_{W_{k-1}} + \mathrm{Tr}(\nabla^T W_{k-1} \nabla \Sigma_{\mathbf{u}}^{k-1}). \tag{3.17}$$

3.4.2. The density $q_k(\mathbf{u})$. Next we find the density $q_k(\mathbf{u})$ in (3.7). The necessary condition for optimality of the unconstrained optimization problem

$$\max_{q(\mathbf{u})\in\mathcal{Q}_{\mathcal{N}}} \mathcal{L}(q(\mathbf{u}), q_k(\theta))$$

is the partial derivative with respect to $q(\mathbf{u})$ being equal to 0, i.e.,

$$\frac{\partial \mathcal{L}(q(\mathbf{u}), q_k(\theta))}{\partial q(\mathbf{u})} = \mathbb{E}_{q_k(\theta)}(\log p(\mathbf{g},\mathbf{u},\theta)) - \log q(\mathbf{u}) - C_2 = 0,$$

where $C_2 = 1 + \int_{\mathbf{u}} q_k(\theta)\log q_k(\theta)$. Substituting the joint density $p(\mathbf{g},\mathbf{u},\theta)$ in (2.5) and the mean of hyperparameter β,λ_i $(i=1,2)$ in (3.11) into the above formulation, the density $q_k(\mathbf{u})$ can be rewritten as follows:

$$\widehat{q}_k(\mathbf{u}) \propto \exp\left(-\mathcal{J}(\mathbf{u};\mu_\beta^k,\mu_{\lambda_1}^k,\mu_{\lambda_2}^k)\right),$$

where $\widehat{q}_k(\mathbf{u})$ denote the variational density satisfying the above necessary condition. Recalling the definition of the function $\mathcal{J}(\mathbf{u};\beta,\lambda_1,\lambda_2)$ in (2.1), we know that the function $\mathcal{J}(\mathbf{u};\beta,\lambda_1,\lambda_2)$ is not quadratic, which implies that $\widehat{q}_k(\mathbf{u})$ is an unnormalized density and the constraint in the optimization problem (3.7) is active.

In order to obtain a feasible solution of (3.7), we apply the Laplace approximation scheme [76] to approximate $\widehat{q}_k(\mathbf{u})$ such that $q_k(\mathbf{u})$ is a Gaussian density. The main idea is to find a Gaussian approximation of the unnormalized density centered at its maximum. The Laplace approximation consists of the following three steps.

(1) *Estimation of the mean.* We find a maximum $\mu_{\mathbf{u}}$ of the density $\widehat{q}_k(\mathbf{u})$, which is achieved at the minimum of the function $\mathcal{J}(\mathbf{u};\mu_\beta^k,\mu_{\lambda_1}^k,\mu_{\lambda_2}^k)$. Thus we have

$$\begin{aligned}\mu_{\mathbf{u}}^k &= \operatorname*{argmin}_{\mathbf{u}} \mathcal{J}(\mathbf{u};\mu_\beta^k,\mu_{\lambda_1}^k,\mu_{\lambda_2}^k)\\ &= \operatorname*{argmin}_{\mathbf{u}} \frac{\mu_\beta^k}{2}\|\mathbf{g}-\mathcal{H}\mathbf{u}\|_2^2 + \frac{\mu_{\lambda_1}^k}{2}\|\nabla\mathbf{u}\|_2^2 + \mu_{\lambda_2}^k\|\nabla\mathbf{u}\|_1.\end{aligned} \tag{3.18}$$

The above minimization problem can be solved by the alternating direction method of multipliers algorithm [15, 46], split-Bregman algorithm [40, 18], and primal-dual method [22, 25, 33]. Here we apply the primal-dual method to obtain the minimizer $\mathbf{u}$.

(2) *Estimation of the covariance.* After calculating $\mu_{\mathbf{u}}^k$, we approximate the term $\|\nabla \mathbf{u}\|_1$ with a weighted least square term $\|\nabla \mathbf{u}\|_{W_k}^2$; see IRLS in [30]. Here W_k is the diagonal matrix with diagonal entries $1/(\left|(\nabla \mu_{\mathbf{u}}^k)\right| + \epsilon)$, where $\epsilon > 0$ is a given parameter (also $\epsilon = 10^{-3}$). Thus we have the following approximation:

$$\mathcal{J}(\mathbf{u}; \mu_\beta^k, \mu_{\lambda_1}^k, \mu_{\lambda_2}^k) \approx \frac{\mu_\beta^k}{2} \|\mathbf{g} - \mathcal{H}\mathbf{u}\|_2^2 + \frac{\mu_{\lambda_1}^k}{2} \|\nabla \mathbf{u}\|_2^2 + \mu_{\lambda_2}^k \|\nabla \mathbf{u}\|_{W_k}^2 .$$

Noticing that $\widehat{q}_k(\mathbf{u}) \propto \exp\left(-\mathcal{J}(\mathbf{u}; \mu_\beta^k, \mu_{\lambda_1}^k, \mu_{\lambda_2}^k)\right)$, we obtain

$$\log \widehat{q}_k(\mathbf{u}) \propto -\frac{1}{2} \left(\mathbf{u} - \mu_{\mathbf{u}}^k\right)^T \left(\Sigma_{\mathbf{u}}^k\right)^{-1} \left(\mathbf{u} - \mu_{\mathbf{u}}^k\right).$$

Here $\Sigma_{\mathbf{u}}^k$ is a symmetric matrix defined by

$$\Sigma_{\mathbf{u}}^k = \left(\mu_\beta^k \mathcal{H}^T \mathcal{H} + \mu_{\lambda_1}^k \nabla^T \nabla + \mu_{\lambda_2}^k \nabla^T W_k \nabla\right)^{-1}. \tag{3.19}$$

(3) *Construction of the density function.* We define the approximation density as

$$q_k(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{u}}^k, \Sigma_{\mathbf{u}}^k). \tag{3.20}$$

3.5. Determination of the segmentation threshold. Once we have obtained the smoothed image $\mathbf{u}$ by the variational Bayesian inference, we can apply a thresholding procedure to get the segmented solution in the next stage; see the 2-stage segmentation method in [21]. In this subsection, we review this thresholding procedure which is to apply the K-means method [82] to automatically select the threshold. The thresholding procedure consists of the following steps.

(1) *Image normalization.* The image $\mathbf{u}$ is normalized to $[0, 1]$ using the linear transformation formula as follows:

$$\bar{\mathbf{u}} = \frac{1}{\mathbf{u}_{\max} - \mathbf{u}_{\min}} (\mathbf{u} - \mathbf{u}_{\min} \cdot \mathbf{1}). \tag{3.21}$$

Here $\mathbf{u}_{\max}, \mathbf{u}_{\min}$ correspond to the maximum and minimum of $\mathbf{u}$, respectively; $\mathbf{1}$ is a vector of all ones with the same dimension as $\mathbf{u}$. We remark that the above linear transformation formula is for a gray image. In order to extend to the color image, the normalization scheme can be applied for each channel.

(2) *Pixel cluster.* The K-means method is applied to divide the image into K phases $T_1, \ldots, T_K$, where K is the number of phases, and we have $T_1 \cup T_2 \cup \cdots \cup T_K = \Omega$.

(3) *Threshold vector estimation.* We compute the mean ρ_i of each category T_i $(i = 1, \ldots, K)$ by

$$\rho_i = \frac{\int_{T_i} \bar{\mathbf{u}} dx}{\int_{T_i} dx}, \quad i = 1, 2, \ldots, K, \tag{3.22}$$

where ρ_i is a scalar if $\bar{\mathbf{u}}$ is a gray image or is an l-vector if $\bar{\mathbf{u}}$ is an l-channel image.

Algorithm 3.1 Segmentation algorithm using variational Bayesian inference.

Inputs: $\mathbf{g}, \mathcal{H}, K$.

Outputs: $\mathbf{u}, \beta, \lambda_1, \lambda_2$.

Initialize $\alpha_1^0 = 0.1, \alpha_2^0 = 0.1$, $\mu_{\mathbf{u}}^0 = \mathbf{g}/2$, $\Sigma_{\mathbf{u}}^0 = \mathbf{0}$.
$\widetilde{a}_\beta = MN/2$.
while convergence criterion not met **do**
  Compute $\widetilde{a}_{\lambda_i}^k$ by (3.15).
  Compute $\widetilde{b}_\beta^k, \widetilde{b}_{\lambda_i}^k (i = 1, 2)$ using (3.12).
  $q_k(\beta) = \mathcal{G}(\widetilde{a}_\beta, \widetilde{b}_\beta^k), q_k(\lambda_i) = \mathcal{G}(\widetilde{a}_{\lambda_i}^k, \widetilde{b}_{\lambda_i}^k)$.
  Compute $\mu_\beta^k, \mu_{\lambda_i}^k (i = 1, 2)$ using (3.11).
  Compute $\mu_{\mathbf{u}}^k$ and $\Sigma_{\mathbf{u}}^k$ using (3.18) and (3.19).
  $q_k(\mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{u}}^k, \Sigma_{\mathbf{u}}^k)$.
end while
return $\mathbf{u} = \mu_{\mathbf{u}}^{k+1}, \beta = \mu_\beta^{k+1}, \lambda_1 = \mu_{\lambda_1}^{k+1}$ and $\lambda_2 = \mu_{\lambda_2}^{k+1}$.
Apply K-means algorithm to segment the image.

(4) *Image segmentation.* The ith phase of $\mathbf{u}$ is given by

$$\Omega_i := \left\{ \ell \in \Omega : ||(\bar{\mathbf{u}})_\ell - \rho_i||_2 = \min_{1 \le j \le K} ||(\bar{\mathbf{u}})_\ell - \rho_j||_2 \right\}, \quad i = 1, \ldots, K,$$

where $(\bar{\mathbf{u}})_\ell$ denotes the value of $\bar{\mathbf{u}}$ at the ℓth pixel.

Finally, we summarize the proposed adaptive segmentation method in Algorithm 3.1.

The initial value of $\alpha_1, \alpha_2, \mu_{\mathbf{u}}^0$, and $\Sigma_{\mathbf{u}}^0$ are set to $\alpha_1^0 = \alpha_2^0 = 0.1$ and $\mu_{\mathbf{u}}^0 = \mathbf{g}/2, \Sigma_{\mathbf{u}}^0 = \mathbf{0}$. We remark that when $\Sigma_{\mathbf{u}}^0 = \mathbf{0}$, we have $\widetilde{b}_\beta^1 = \frac{1}{2}||\mathbf{g} - \mathcal{H}\mu_{\mathbf{u}}^0||_2^2$, $\widetilde{b}_{\lambda_1}^1 = \frac{1}{2}||\nabla \mu_{\mathbf{u}}^0||_2^2$, and $\widetilde{b}_{\lambda_2}^1 = \left\|\nabla \mu_{\mathbf{u}}^0\right\|_{W_{k-1}}^2$ according to (3.12), (3.16), and (3.17).

4. Numerical experiments.

In this section, we give experimental results to illustrate the performance of the proposed method. All results were obtained under Windows and MATLAB R2018b on a PC with a 3.4 GHz CPU processor and 4 GB of RAM. We apply the *segmentation accuracy* (SA) to measure the quality of the segmentation results. The SA is defined as

$$SA := \frac{\#\text{correctly classified pixels}}{\#\text{all pixels}}. \tag{4.1}$$

Accordingly, we know that $SA \le 1$, and the larger the SA, the better the segmentation result.

4.1. Choice of the parameters.

The parameters α_1 and α_2 in (3.13) will be chosen by heuristically experimental evidence [1, 5, 70] or the sequential imputation method [12, 53]. The test images are the three-phase sky (3Ps) image with size 125×150, the 61×58 three-phase ball (3Pb) image, the 256×256 four-phase synthetical (4Ps) image, and the five-phase synthetical (5Ps) image with size 91×96. The original images are shown in Figure 2. We segment the clean 3Ps, 3Pb, 4Ps, and 5Ps image and the noisy 4Ps image with the noise variances of 0.05, 0.1, 0.2, 0.3, respectively. We apply the sequential imputation method to iteratively update α_1, α_2. The initial values are set to $\alpha_1^0 = \alpha_2^0 = 0.1$. The evolutions of the parameters α_i

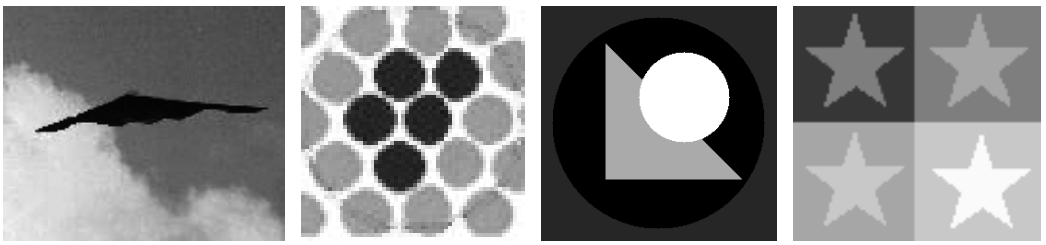
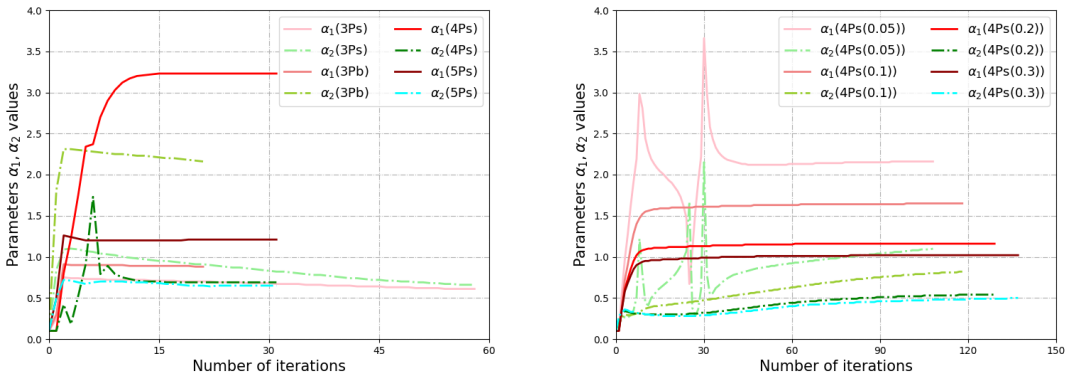


Figure 2. *The three-phase sky (3Ps) image (left), the three-phase ball (3Pb) images (center left), the four-phase synthetical (4Ps) image (center right), and the five-phase synthetical (5Ps) image (right).*



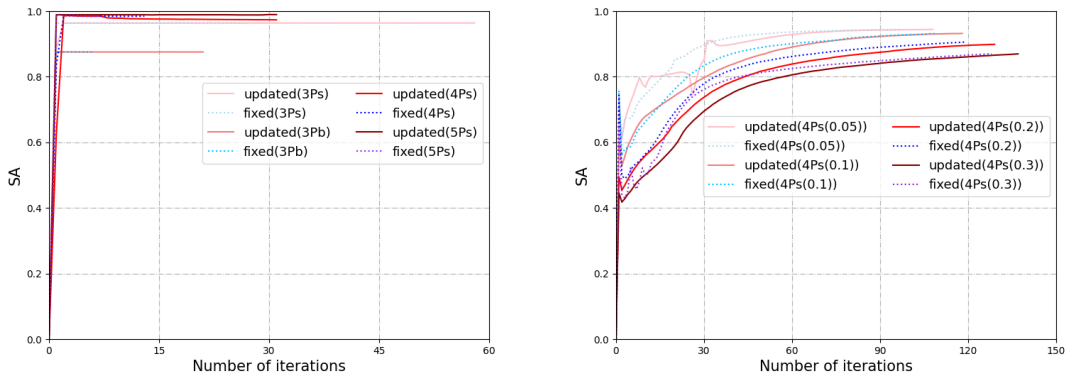
(a) non-noisy ($\alpha_i (i = 1, 2)$) (b) noisy ($\alpha_i (i = 1, 2)$)

Figure 3. *The evolution of the parameters α_1, α_2 along the iteration number for the clean and the noisy image. Here the sequential imputation method is applied to iteratively update α_1, α_2, and "4Ps(0.05)" in the legend denotes the noisy 4Ps image with noise variance of* 0.05.

along the number of iterations are shown in Figure 3. We observe that the parameters α_1 and α_2 fluctuate greatly at the beginning of iterations, and they gradually become stable as the number of iterations increases. It means that the parameters will converge to constants.

We also compare the SA values obtained by the sequential imputation method [12, 53] and heuristically experimental evidence [1, 5, 70]. In the heuristically experimental evidence, we set the parameters to $\alpha_1 = 2, \alpha_2 = 1.5$. The evolutions of the SA values along the number of iterations are shown in Figure 4. It can be observed that the SA values obtained by the sequential imputation method [12, 53] and heuristic experimental evidence [1, 5, 70] are almost the same, but the sequential imputation method (more details, see [12, 53]) requires more iterations to converge.

4.2. Grayscale image. Here we segment the noisy and blurred synthetic images. The synthetic images are the two-phase (2P) image with size 128×128 and the four-phase (4P) image with size 256×256. The original clean images, the noisy images, and the blurred images are shown in Figure 5. Both images are corrupted by Gaussian noise with mean of 0 and variance of 0.25. We also compare the segmentation results on the two-phase and four-phase images that are corrupted by the motion blur and noise. The motion blur is generated by the MATLAB command `fspecial('motion',15,90)`, and then the blurred images are corrupted



(a) non-noisy (SA values) (b) noisy (SA values)

Figure 4. *The evolution of the SA values along the iteration number for the clean and the noisy image. Here "4Ps(0.05)" in the legend denotes the noisy 4Ps image with noise variance of* 0.05.

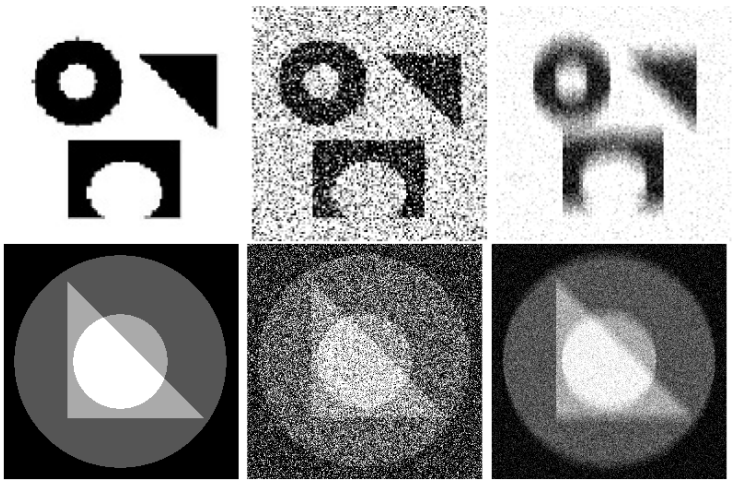
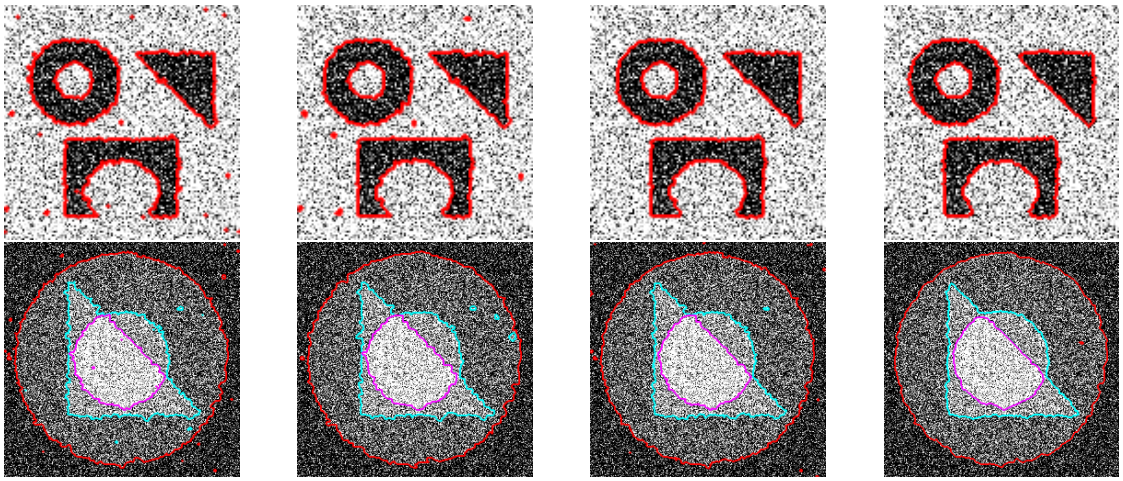


Figure 5. *The original clean images (left), the noisy images (middle), and the blurred images (right) of the two-phase (2P) image (top) and the four-phase (4P) image (bottom).*

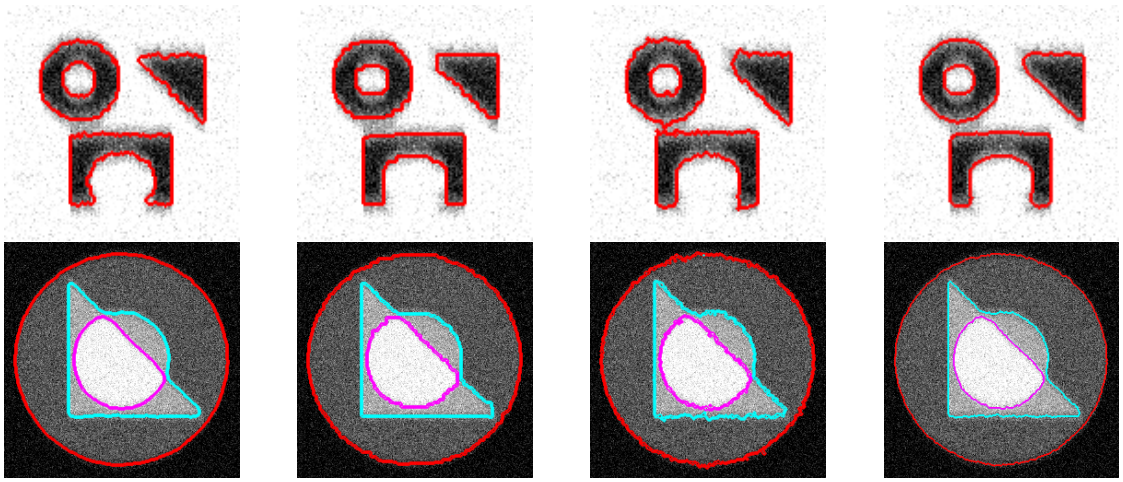
by Gaussian white noise with variance 0.01. We compare the proposed method with the SaT method [21], and the thresholded-Rudin–Osher–Fatemi (T-ROF) method [20]. They are two-stage methods which are regarded as restoration by solving a minimization problem in the first stage and then thresholding in the second stage. We also compare the results obtained by a different regularization function. In [59], the difference between the L_1 norm and L_2 norm was applied as a prior of the smoothed image, this is $p(\mathbf{u}|\lambda) = \frac{1}{Z(\lambda)} \exp(-\frac{\lambda}{2}(\|\nabla \mathbf{u}\|_1 - \|\nabla \mathbf{u}\|_2))$, while in our paper, we apply the hybrid Gaussian–Laplace distribution as a prior of the smoothed image; see (2.3). The regularization parameters in the methods of [20, 21, 59] are chosen by a trial-and-error method such that the highest SA values can be achieved.

We segment the noisy images in the middle of Figure 5 and the blurred images at the right of Figure 5. The segmentation results for noisy images are shown in Figure 6, while



(a) Cai et al. (b) Lou et al. (c) Cai et al. (d) Proposed

Figure 6. *Noisy synthetic images segmentation results. The images are corrupted by Gaussian noise.*



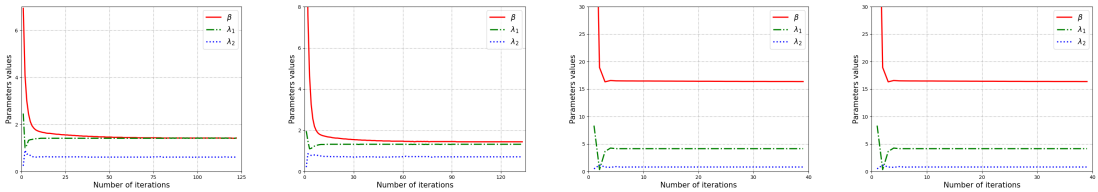
(a) Cai et al. (b) Lou et al. (c) Cai et al. (d) Proposed

Figure 7. *Blurred synthetic images segmentation results. The images are corrupted by motion blur and Gaussian noise.*

the segmentation results for the blurred and noisy images are shown in Figure 7. The first column to the third column in Figures 6 and 7 show the results of methods in [21, 59, 20], and the last column shows the results of our method. The boundaries of the segmentation results are shown with red color and superimposed on the given images. The boundaries of the 4-phase image in the second row are shown with red, cyan, and magenta colors. For the segment results of the noisy images, we can observe that our method and the method in [20] have the best results with no isolated misclassified points in the two-phase image, and it is clear that our method obtains smoother edges of circles and lines in the four-phase image, while the other methods give more sawtooth artifacts on the circles and lines.

Table 1
CPU time in seconds and SA for the synthetical images segmentation.

Type	Image	SA				Time			
		[21]	[59]	[20]	Proposed	[21]	[59]	[20]	Proposed
Noisy	2P	0.9863	0.9858	**0.9893**	0.9874	**0.20**	0.21	0.53	0.27(0.08)
	4P	0.9746	0.9728	0.9756	**0.9780**	**0.49**	0.91	1.12	0.89(0.36)
Blurred	2P	0.9058	0.9076	0.9070	**0.9089**	0.24	0.97	0.23	**0.19**(0.09)
	4P	0.9842	0.9826	0.9752	**0.9844**	**0.42**	4.42	1.70	0.82(0.38)
average		0.9627	0.9622	0.9618	**0.9647**	**0.34**	1.63	0.65	0.54(0.23)



(a) 2P/noisy (b) 4P/noisy (c) 2P/motion (d) 4P/motion

Figure 8. *The parameters $\beta, \lambda_1, \lambda_2$ versus iteration number for two-phase (2P) image and four-Phase (4P) image.*

The CPU running time and the SA obtained by different methods are shown in Table 1. The best results among all the methods are shown in boldface and the second best results are marked in underline. The average SAs are 0.9627 [21], 0.9622 [59], 0.9618 [20], and 0.9647 (proposed method). We observe that the proposed method achieves either the highest or the second highest SA value in all tests. In the average sense, the proposed method achieves the highest SA value. For the CPU running time, we provide not only the total running time of the proposed method but also the time required for parameter inference, indicated by the brackets in Table 1. Although parameter inference accounts for about half of the total runtime, it is still faster than the methods presented in [20, 59].

We also plot the curve of the regularization parameters $\beta, \lambda_1, \lambda_2$ versus iteration number in Figure 8. For the cases of noisy images, the regularization parameters fluctuate greatly in the first 10 iterations, and then gradually tend to be stable. For the cases of the blurred images, the regularization parameters converge at a constant value after several iterations. Thus we suggest stopping the iteration after a fixed number of iterations, say 100, in real applications. We remark that although the parameters thus obtained already give accurate segmentation, the convergence analysis of the regularization parameters is still an open problem.

4.3. Color image. We extend the image segmentation algorithm from grayscale images to color images. Color images are mainly represented by red, green, and blue (RGB) color models. It is shown that the lab (perceived lightness, red-green and yellow-blue) color space is better adapted for color image segmentation than the RGB space. Cai et al. [19] proposed a three-stage approach for segmenting color images. The first stage is to smooth the image in each channel independently using the convex variant of the Mumford–Shah model (1.3).

The second stage is to lift the smoothed image into a six-dimensional space where a new vector-valued image is composed of the smoothed image and its transformation into the lab color space. The last stage is to threshold the resultant image using a multichannel approach to obtain a segmentation. Here we extend the variational Bayesian inference to segment the color images by the same three-stage approach.

In addition, here we also utilize the Sørensen–Dice similarity coefficient (DICE) score to quantify the segmentation accuracy. And the DICE is defined as

$$DICE := \frac{2|S_m \cap S_t|}{|S_m| + |S_t|}, \tag{4.2}$$

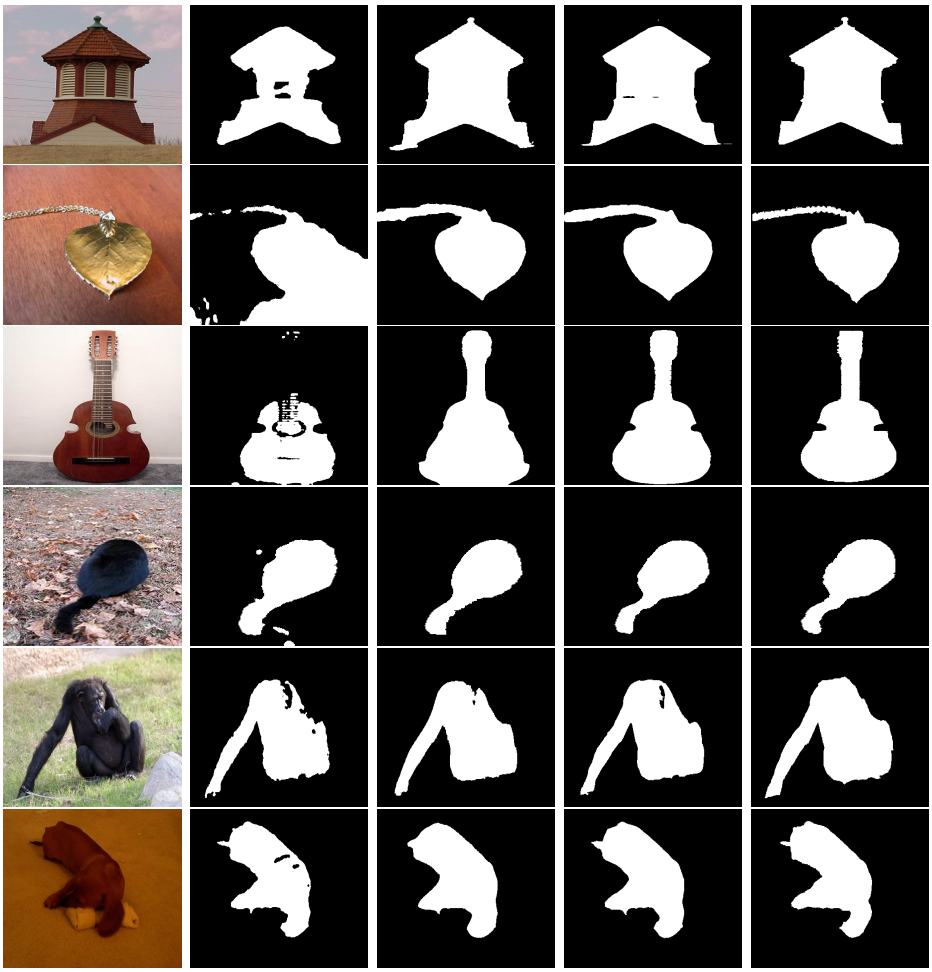
where S_t is the ground truth segmentation, S_m is the segmentation outcome generated by a given method, and $|\cdot|$ denotes the number of pixels. Similarly, a higher DICE score indicates better segmentation performance.

4.3.1. Segmentation for noise-free color image. We consider segmenting the noise-free color images. The color images tested are downloaded from the MSRA data set.[1] They are the tower, golden leaf, guitar, cat, orangutan, and dog shown in the first column of Figure 9. We compare our method with two segmentation methods for noise-free color image proposed in [90] and [57]. We denote them as "SDRE" [90] and "TSVS" [57], respectively. "TSVS" [57] is also a three-stage approach to segment the color image, but it first lifts the color image into a high dimension space and then smoothes the image. The "SDRE" method incorporates the saliency map into the level set framework. The regularization parameters in [57, 90] are chosen by a trial-and-error method such that the highest SA value is achieved. The segmentation results obtained by "SDRE," "TSVS," and the proposed method are shown in the second column to fourth column of Figure 9. We also show the hand-drawn ground-truth segmentation results in the last column of Figure 9. We can observe that the proposed method gives better segmentation results. We also show the SA, DICE, and the CPU running time obtained by difference methods in Table 2. Also, the brackets in Table 2 are the parameter inference time. We observe that the proposed method achieves the highest SA values and DICE scores in all tested images. The average CPU running times of "SDRE" [90] and "TSVS" [57] are more than twice that of the proposed method.

4.3.2. Segmentation for noisy color image. Here we compare the proposed method on noisy color images with the methods in [19] and [20], which we denote as "SLaT" and "T-ROF," respectively. The regularization parameter in [19, 20] were also chosen by the trial-and-error method such that the highest SA values are achieved. The clean images are buffalo, goshawk, horse, and pyramid with size 481×321, and were downloaded from the BSDS500 dataset.[2] They are shown in Figure 10(a), and the noisy versions of these images are shown in Figure 10(b). The variance of the Gaussian noise added is 0.01 for the buffalo and goshawk images and 0.1 for horse and pyramid images. The segmentation results obtained by SLaT, T-ROF, and our method are shown in the third to sixth columns of Figure 10, respectively. The last column shows the hand-drawn ground-truth segmentation results. We observe from Figure 10 that the segmentation results obtained by our method are closer to the real

[1] https://mmcheng.net/msra10k/

[2] https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/



(a) Original (b) SDRE (c) TSVS (d) Proposed (e) Truth

Figure 9. *The segmentation results for the color images of tower, golden leaf, guitar, cat, orangutan, and dog. Download from the MSRA dataset color image segmentation results.*

segmentation results on the wings of the goshawk and the tail of the horse. We list the SA, DICE, and the CPU running time obtained by different methods in Table 3, and the parameter inference time is indicated in parentheses. The average SA (DICE) for SLaT [19] and T-ROF [20] and our method are 0.9821 (0.9453), 0.9842 (0.9502), and 0.9847 (0.9523), respectively. When we compare the average CPU running time, our method only needs half the CPU running time of SLaT [19] and one third of that of T-ROF [20].

5. Conclusion. In this paper, we have developed a method to segment images and select the regularization parameters simultaneously for the convex variant Mumford–Shah variational model. We described the variational model from the statistical perspective. The regularization parameters are treated as random variables and variational Bayesian inference was applied to estimate the smoothed image and the regularization parameters. The mean field variational family is used to approximate the posterior density. We assumed that the image

Table 2
CPU time in seconds, SA, and DICE for the MSRA dataset images segmentation.

Image	SA			DICE			Time		
	[90]	[57]	Proposed	[90]	[57]	Proposed	[90]	[57]	Proposed
Tower	0.9546	0.9777	**0.9883**	0.9230	0.9657	**0.9815**	**4.89**	10.06	5.57(2.14)
Golden leaf	0.7951	0.9698	**0.9746**	0.6390	0.9274	**0.9382**	15.77	8.49	**3.61**(1.51)
Guitar	0.9199	0.9583	**0.9763**	0.8228	0.9240	**0.9551**	9.5	12.88	**5.01**(1.90)
Cat	0.9751	0.9874	**0.9915**	0.9176	0.9551	**0.9691**	6.22	2.38	**1.34**(0.43)
Orangutan	0.9707	0.9738	**0.9779**	0.9403	0.9474	**0.9556**	4.47	8.61	**2.88**(1.14)
Dog	**0.9869**	0.9811	0.9852	**0.9662**	0.9527	0.9629	**2.65**	6.73	2.98(1.23)
average	0.9337	0.9747	**0.9823**	0.8682	0.9454	**0.9604**	7.25	8.19	**3.57**(1.39)

(a) Clean (b) Noisy (c) SLaT (d) TROF (e) Proposed (f) Truth

Figure 10. *Comparison of segmentation results for color noisy images.*

Table 3
SA, DICE, and CPU time in seconds for the noisy color images segmentation.

Noise	Image	SA			DICE			Time		
		[19]	[20]	Proposed	[19]	[20]	Proposed	[19]	[20]	Proposed
$\mathcal{N}(0, 0.01)$	Buffalo	0.9959	0.9951	**0.9965**	0.9658	0.9605	**0.9712**	6.56	13.63	**4.29**(1.56)
	Goshawk	0.9903	**0.9914**	0.9907	0.8988	**0.9126**	0.9077	6.61	12.93	**3.62**(1.35)
$\mathcal{N}(0, 0.1)$	Horse	0.9540	**0.9590**	0.9583	0.9344	**0.9411**	0.9405	7.65	21.46	**5.98**(2.19)
	Pyramid	0.9883	0.9913	**0.9933**	0.9820	0.9865	**0.9897**	6.75	14.89	**4.28**(1.58)
	average	0.9821	0.9842	**0.9847**	0.9453	0.9502	**0.9523**	6.89	15.73	**4.54**(1.67)

has a Gaussian distribution and the regularization parameters have a Gamma distribution. A coordinate ascent approach was applied to obtain the density functions. The segmentation results for both grayscale images and color images has shown that our approach is competitive with the other variants of the Mumford–Shah variational model in terms of segmentation accuracy, and is faster in terms of CPU running time.

REFERENCES

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

[2] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*, Elsevier, Amsterdam, 2018.

[3] S. D. Babacan, R. Molina, M. N. Do, and A. K. Katsaggelos, *Bayesian blind deconvolution with general sparse image priors*, in Proceedings of the European Conference on Computer Vision, Springer, Berlin, 2012, pp. 341–355.

[4] S. D. Babacan, R. Molina, and A. K. Katsaggelos, *Total variation image restoration and parameter estimation using variational posterior distribution approximation*, in Proceedings of the 2007 IEEE International Conference on Image Processing, IEEE, Piscataway, NJ, 2007, pp. 97–100.

[5] 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.

[6] S. D. Babacan, R. Molina, and A. K. Katsaggelos, *Variational Bayesian blind deconvolution using a total variation prior*, IEEE Trans. Image Process., 18 (2008), pp. 12–26.

[7] S. D. Babacan, R. Molina, and A. K. Katsaggelos, *Bayesian compressive sensing using Laplace priors*, IEEE Trans. Image Process., 19 (2009), pp. 53–63.

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

[9] W. A. Barrett and A. S. Cheney, *Object-based image editing*, in Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, ACM, New York, 2002, pp. 777–784.

[10] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer, New York, 2011.

[11] J. M. Bernardo and A. F. Smith, *Bayesian Theory*, Wiley, New York, 2009.

[12] C. Berzuini, N. G. Best, W. R. Gilks, and C. Larizza, *Dynamic conditional independence models and Markov chain Monte Carlo methods*, J. Amer. Statist. Assoc., 92 (1997), pp. 1403–1412.

[13] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, *Variational inference: A review for statisticians*, J. Amer. Statist. Assoc., 112 (2017), pp. 859–877.

[14] C. Bouman and K. Sauer, *A generalized Gaussian image model for edge-preserving MAP estimation*, IEEE Trans. Image Process., 2 (1993), pp. 296–310.

[15] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), pp. 1–122.

[16] K. Bui, F. Park, Y. Lou, and J. Xin, *A weighted difference of anisotropic and isotropic total variation for relaxed Mumford–Shah color and multiphase image segmentation*, SIAM J. Imaging Sci., 14 (2021), pp. 1078–1113.

[17] K. Bui, F. Park, Y. Lou, and J. Xin, *Efficient image segmentation framework with difference of anisotropic and isotropic total variation for blur and Poisson noise removal*, Front. Comput. Sci, 5 (2023), 1131317.

[18] J.-F. Cai, S. Osher, and Z. Shen, *Split Bregman methods and frame based image restoration*, Multiscale Model. Simul., 8 (2009), pp. 337–369.

[19] X. Cai, R. Chan, M. Nikolova, and T. Zeng, *A three-stage approach for segmenting degraded color images: Smoothing, lifting and thresholding (SLaT)*, J. Sci. Comput., 72 (2017), pp. 1313–1332.

[20] X. Cai, R. Chan, C.-B. Schonlieb, G. Steidl, and T. Zeng, *Linkage between piecewise constant Mumford–Shah model and Rudin–Osher–Fatemi model and its virtue in image segmentation*, SIAM J. Sci. Comput., 41 (2019), pp. B1310–B1340.

[21] X. Cai, R. Chan, and T. Zeng, *A two-stage image segmentation method using a convex variant of the Mumford–Shah model and thresholding*, SIAM J. Imaging Sci., 6 (2013), pp. 368–390.
[22] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145.
[23] R. Chan, H. Yang, and T. Zeng, *A two-stage image segmentation method for blurry images with Poisson or multiplicative gamma noise*, SIAM J. Imaging Sci., 7 (2014), pp. 98–127.
[24] T. F. Chan, S. Esedoglu, and M. Nikolova, *Algorithms for finding global minimizers of image segmentation and denoising models*, SIAM J. Appl. Math., 66 (2006), pp. 1632–1648.
[25] T. F. Chan, G. H. Golub, and P. Mulet, *A nonlinear primal-dual method for total variation-based image restoration*, SIAM J. Sci. Comput., 20 (1999), pp. 1964–1977.
[26] T. F. Chan and L. A. Vese, *Active contours without edges*, IEEE Trans. Image Process., 10 (2001), pp. 266–277.
[27] 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.
[28] S. F. Chen and R. Rosenfeld, *A Gaussian Prior for Smoothing Maximum Entropy Models*, Technical report, Carnegie Mellon University School of Computer Science, Pittsburgh, PA, 1999.
[29] D. Cremers, M. Rousson, and R. Deriche, *A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape*, Int. J. Comput. Vis., 72 (2007), pp. 195–215.
[30] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, *Iteratively reweighted least squares minimization for sparse recovery*, Comm. Pure Appl. Math., 63 (2010), pp. 1–38.
[31] E. De Carvalho and D. T. Slock, *Maximum-likelihood blind FIR multi-channel estimation with Gaussian prior for the symbols*, in Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE Computer Society, Los Alamitos, CA, 1997, pp. 3593–3596.
[32] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for Machine Learning*, Cambridge University Press, Cambridge, 2020.
[33] A. V. Fiacco and G. P. McCormick, *The sequential unconstrained minimization technique for nonlinear programing, a primal-dual method*, Manag. Sci., 10 (1964), pp. 360–366.
[34] N. P. Galatsanos, V. Z. Mesarovic, R. Molina, A. K. Katsaggelos, and J. Mateos, *Hyperparameter estimation in image restoration problems with partially-known blurs*, Opt. Eng., 41 (2002), pp. 1845–1854.
[35] 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.
[36] E. I. George and R. E. McCulloch, *Variable selection via Gibbs sampling*, J. Amer. Statist. Assoc., 88 (1993), pp. 881–889.
[37] T. Gevers and A. W. Smeulders, *Color-based object recognition*, Pattern Recognit., 32 (1999), pp. 453–464.
[38] Z. Ghahramani and M. J. Beal, *Propagation algorithms for variational Bayesian learning*, in Proceedings of the Advances in Neural Information Processing Systems, Vol. 13, MIT Press, Cambridge, MA, 2000, pp. 1–7.
[39] M. Gobbino, *Finite difference approximation of the Mumford-Shah functional*, Comm. Pure Appl. Math., 51 (1998), pp. 197–228.
[40] T. Goldstein and S. Osher, *The split Bregman method for L1-regularized problems*, SIAM J. Imaging Sci., 2 (2009), pp. 323–343.
[41] G. H. Golub, M. Heath, and G. Wahba, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, 21 (1979), pp. 215–223.
[42] K. Hausman, F. Balint Benczedi, D. Pangercic, Z. C. Marton, R. Ueda, K. Okada, and M. Beetz, *Tracking-based interactive segmentation of textureless objects*, in Proceedings of the 2013 IEEE International Conference on Robotics and Automation, IEEE, Piscataway, NJ, 2013, pp. 1122–1129.
[43] X. He and A. Yuille, *Occlusion boundary detection using pseudo-depth*, in Proceedings of the European Conference on Computer Vision, Springer, Berlin, 2010, pp. 539–552.
[44] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, *Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors*, preprint, arXiv:1207.0580, 2012.

[45] M. Hirsch, C. J. Schuler, S. Harmeling, and B. Schölkopf, *Fast removal of non-uniform camera shake*, in Proceedings of the 2011 International Conference on Computer Vision, IEEE, Piscataway, NJ, 2011, pp. 463–470.
[46] M. Hong and Z.-Q. Luo, *On the linear convergence of the alternating direction method of multipliers*, Math. Program., 162 (2017), pp. 165–199.
[47] H. Jeffreys, *The Theory of Probability*, Oxford University Press, Oxford, 1998.
[48] B. Jin and J. Zou, *A Bayesian inference approach to the ill-posed Cauchy problem of steady-state heat conduction*, Internat. J. Numer. Methods Engrg., 76 (2008), pp. 521–544.
[49] B. Jin and J. Zou, *Hierarchical Bayesian inference for ill-posed problems via variational method*, J. Comput. Phys., 19 (2010), pp. 7317–7343.
[50] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, *An introduction to variational methods for graphical models*, Mach. Learn., 37 (1999), pp. 183–233.
[51] B. Klein, G. Lev, G. Sadeh, and L. Wolf, *Fisher Vectors Derived From Hybrid Gaussian-Laplacian Mixture Models for Image Annotation*, preprint, arXiv:1411.7399, 2014.
[52] G. Koepfler, C. Lopez, and J. M. Morel, *A multiscale algorithm for image segmentation by variational method*, SIAM J. Numer. Anal., 31 (1994), pp. 282–299.
[53] A. Kong, J. S. Liu, and W. H. Wong, *Sequential imputations and Bayesian missing data problems*, J. Amer. Statist. Assoc., 89 (1994), pp. 278–288.
[54] N. Lermé, F. Malgouyres, D. Hamoir, and E. Thouin, *Bayesian image restoration for mosaic active imaging*, Inverse Probl. Imaging, 8 (2014), pp. 733–760.
[55] A. S. Lewis and G. Knowles, *Image compression using the 2-D wavelet transform*, IEEE Trans. Image Process., 1 (1992), pp. 244–250.
[56] F. Li, M. K. Ng, T. Y, Zeng, and C. Shen, *A multiphase image segmentation method based on fuzzy region competition*, SIAM J. Imaging Sci., 3 (2010), pp. 277–299.
[57] X. Li, X. Yang, and T. Zeng, *A three-stage variational image segmentation framework incorporating intensity inhomogeneity information*, SIAM J. Imaging Sci., 13 (2020), pp. 1692–1715.
[58] J. Long, E. Shelhamer, and T. Darrell, *Fully convolutional networks for semantic segmentation*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Piscataway, NJ, 2015, pp. 3431–3440.
[59] Y. Lou, T. Zeng, S. Osher, and J. Xin, *A weighted difference of anisotropic and isotropic total variation model for image processing*, SIAM J. Imaging Sci., 8 (2015), pp. 1798–1823.
[60] D. G. Lowe, *Object recognition from local scale-invariant features*, in Proceedings of the Seventh IEEE International Conference on Computer Vision, IEEE Computer Society, Los Alamitos, CA, 1999, pp. 1150–1157.
[61] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré, *Markov chain Monte Carlo without likelihoods*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 15324–15328.
[62] A. Mohammad-Djafari, *Maximum entropy and Bayesian methods*, in A Full Bayesian Approach for Inverse Problems, Springer, Dordrecht, The Netherlands, 1996, pp. 135–144.
[63] R. Molina, J. Mateos, and A. K. Katsaggelos, *Blind deconvolution using a variational approach to parameter, image, and blur estimation*, IEEE Trans. Image Process., 15 (2006), pp. 3715–3727.
[64] V. A. Morozov, *Methods for Solving Incorrectly Posed Problems*, Springer, New York, 2012.
[65] D. Mumford and J. Shah, *Boundary detection by minimizing functionals*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Silver Springs, MD, 1985, pp. 137–154.
[66] D. B. Mumford and J. Shah, *Optimal approximations by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–685.
[67] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, New York, 2012.
[68] R. M. Neal, *Markov chain sampling methods for Dirichlet process mixture models*, J. Comput. Graph. Statist., 9 (2000), pp. 249–265.
[69] R. M. Neal, *Annealed importance sampling*, Stat. Comput., 11 (2001), pp. 125–139.
[70] 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.
[71] M. Opper and D. Saad, *Advanced Mean Field Methods: Theory and Practice*, MIT Press, Cambridge, 2001.

[72] T. Pock, A. Chambolle, D. Cremers, and H. Bischof, *A convex relaxation approach for computing minimal partitions*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Piscataway, NJ, 2009, pp. 810–817.
[73] M. Riesenhuber and T. Poggio, *Models of object recognition*, Nature Neurosci., 3 (2000), pp. 1199–1204.
[74] H. Rue, S. Martino, and N. Chopin, *Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 71 (2009), pp. 319–392.
[75] P. Sand and S. Teller, *Particle video: Long-range motion estimation using point trajectories*, Int. J. Comput. Vis., 80 (2008), pp. 72–91.
[76] L. Tierney and J. B. Kadane, *Accurate approximations for posterior moments and marginal densities*, J. Amer. Statist. Assoc., 81 (1986), pp. 82–86.
[77] M. E. Tipping, *Sparse Bayesian learning and the relevance vector machine*, J. Mach. Learn. Res., 1 (2001), pp. 211–244.
[78] L. A. Vese and T. F. Chan, *A multiphase level set framework for image segmentation using the Mumford and Shah model*, Int. J. Comput. Vis., 50 (2002), pp. 271–293.
[79] C. Wang and D. M. Blei, *Variational inference in nonconjugate models*, J. Mach. Learn. Res., 14 (2013), pp. 1005–1031.
[80] Y. Wen and R. Chan, *Parameter selection for total-variation-based image restoration using discrepancy principle*, IEEE Trans. Image Process., 21 (2012), pp. 1770–1781.
[81] P. M. Williams, *Bayesian regularization and pruning using a Laplace prior*, Neural Comput., 7 (1995), pp. 117–143.
[82] M. A. Wong and J. Hartigan, *Algorithm AS* 136: *A k-means clustering algorithm*, J. R. Stat. Soc. Ser. C. Appl. Stat., 28 (1979), pp. 100–108.
[83] T. Wu, J. Shao, X. Gu, M. K. Ng, and T. Zeng, *Two-stage image segmentation based on nonconvex L2-Lp approximation and thresholding*, Appl. Math. Comput., 403 (2021), 126168.
[84] T. Wu, X. Gu, Y. Wang, and T. Zeng, *Adaptive total variation based image segmentation with semi-proximal alternating minimization*, Signal Process., 183 (2021), 108017.
[85] T. Wu, Z. Mao, Z. Li, Y. Zeng, and T. Zeng, *Efficient color image segmentation via quaternion-based L_1/L_2 regularization*, J. Sci. Comput., 93 (2022), 9.
[86] K. Yamaguchi, D. McAllester, and R. Urtasun, *Efficient joint segmentation, occlusion labeling, stereo and flow estimation*, in Proceedings of the European Conference on Computer Vision, Springer, Cham, Switzerland, 2014, pp. 756–771.
[87] P. J. Yim, P. L. Choyke, and R. M. Summers, *Gray-scale skeletonization of small vessels in magnetic resonance angiography*, IEEE Trans. Med. Imaging, 19 (2000), pp. 568–576.
[88] J. Yuan, E. Bae, X. Tai, and Y. Boykov, *A continuous max-flow approach to Potts model*, in Proceedings of the European Conference on Computer Vision, Springer, Berlin, 2010, pp. 379–392.
[89] L. Zhang, W. Wei, C. Tian, F. Li, and Y. Zhang, *Exploring structured sparsity by a reweighted Laplace prior for hyperspectral compressive sensing*, IEEE Trans. Image Process., 25 (2016), pp. 4974–4988.
[90] X. Zhi and H. Shen, *Saliency driven region-edge-based top down level set evolution reveals the asynchronous focus in image segmentation*, Pattern Recognit., 80 (2018), pp. 241–255.
[91] W. Zhu, X. Tai, and T. Chan, *Image segmentation using Euler's elastica as the regularization*, J. Sci. Comput., 57 (2013), pp. 414–438.