Communications for Statistical Applications and Methods
2022, Vol. 29, No. 1, 27–40

https://doi.org/10.29220/CSAM.2022.29.1.027
Print ISSN 2287-7843 / Online ISSN 2383-4757

Variational Bayesian inference for binary image restoration using Ising model

Moonsoo Jang[a], Younshik Chung[1,a]

[a]Department of Statistics, Pusan National University

Abstract

In this paper, the focus on the removal noise in the binary image based on the variational Bayesian method with the Ising model. The observation and the latent variable are the degraded image and the original image, respectively. The posterior distribution is built using the Markov random field and the Ising model. Estimating the posterior distribution is the same as reconstructing a degraded image. MCMC and variational Bayesian inference are two methods for estimating the posterior distribution. However, for the sake of computing efficiency, we adapt the variational technique. When the image is restored, the iterative method is used to solve the recursive problem. Since there are three model parameters in this paper, restoration is implemented using the VECM algorithm to find appropriate parameters in the current state. Finally, the restoration results are shown which have maximum peak signal-to-noise ratio (PSNR) and evidence lower bound (ELBO).

Keywords: variational Bayesian inference, Ising model, expectation maximization algorithm, binary image restoration

1. Introduction

In the past decades, a lot of image analysis methods have been proposed. There have been many attempts to eliminate noise. Geman and Geman (1984) restored a degraded image having white Gaussian noise and blur. Also, Besag *et al.* (1991) introduced the Bayesian image restoration. Geman and Geman (1984) and Besag *et al.* (1991) constructed the posterior distribution considering the degraded image as an observation and the original image as a latent variable which should be estimated. They used the Gibbs sampler to estimate the posterior distribution. Besag (1986) implemented the image restoration using another method, Iterative Conditional Modes (ICM).

The Markov Chain Monte Carlo (MCMC; Smith and Roberts, 1993) such as the Gibbs sampler (Gelfand and Smith, 1990; Casella and George, 1992) and the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970) have been used to estimate the posterior distribution. The MCMC method constructs an ergodic Markov Chain on z whose stationary distribution is the posterior density $p(z|y)$ where z is a latent variable and y is a given data. TIAN *et al.* (2009) suggested the gray scale image restoration algorithm using the adaptive MCMC particle filters to reduce the computational cost. However, the MCMC based methods take time and require some conditions. Also, because the image is quite a big data set, the computational cost is extremely large. Thus, sometimes it is not suitable to use the MCMC methods to estimate the posterior density in the image analysis. Therefore, we use the Variational Inference mainly used in Machine Learning for approximating the posterior density $p(z|y)$.

[1] Corresponding author: Department of Statistics, Pusan National University, 2, Busandaehak-ro 63 beon-gil, Geumjeong-gu, Busan 46241, Korea. E-mail: yschung@pusan.ac.kr

Published 31 January 2022 / journal homepage: http://csam.or.kr

Jordan *et al.* (1999) suggested a different method called variational inference (VI) (also called variational Bayesian inference) based on a graphical model. The variational inference transforms the inference problem to the optimization problem by minimizing the KL-divergence between the true posterior and the variational density. Therefore, VI algorithm solves the optimization problem. Using the VI, we can estimate the true posterior density more cheaply. Also, Blei *et al.* (2017) reviewed the VI for statisticians and Zhang *et al.* (2019) reviewed the advanced version of VI such as stochastic, black box and amortized VI.

In the variational Bayesian inference (VBI), the main goal is to find the variational density, minimizing the KL-divergence between the true posterior distribution and the candidate approximated density which is assumed to satisfy the mean field assumption (Parisi, 1988). In addition, minimizing the KL-divergence is equal to maximizing the evidence lower bound (ELBO), which is the main objective function in VBI.

In this paper, we use the simple Ising model to describe the binary image structure and apply the variational inference to get an approximated posterior distribution. Also, to estimate the model parameters properly, we adapt an empirical Bayes method (Casella, 2001) using the expectation maximization (EM) algorithm (Dempster *et al.*, 1977). That is, maximizing the Q function in the EM algorithm is equal to maximizing ELBO and minimizing KL-divergence. You can see the EM algorithm from a Bayesian perspective in Friedman (2013). However, updating multiple parameters simultaneously is not feasible. Thus, we use the expectation/conditional-maximization (ECM) algorithm suggested by Meng and Rubin (1993). Since we use the VI to approximate the posterior distribution, we adapt the variational version of the EM algorithm and you can see the variational EM algorithm for Gaussian mixture problems in Nasios and Bors (2006).

The remainder of this paper is organized as follows. In Chapter 2, we describe the Ising model and its distribution called Boltzmann-Gibbs distribution. Also, there is an assumption for the original image and the degraded image to construct the posterior density. In Chapter 3, we introduce the variational Bayesian inference (VBI), the mean field approximation and the coordinate ascent variational inference (CAVI) algorithm to find the variational density. In Chapter 4, we calculate the variational density and the ELBO by quoting from Bishop (2006). Since the model parameters affect the result of restoration, we use the VI version of the EM algorithm (VEM) to estimate the parameters properly. Also, for each time a restoration is implemented, the value of optimal parameter is changed. Therefore, the EM algorithm was used for each restoration. Since there are three model parameters, we also use an ECM algorithm in Chapter 5. The entire restoration algorithm and results are shown in Chapter 6 with two measurements, peak signal-to-noise ratio (PSNR) and ELBO.

2. Reviews of Ising model and Markov random field (MRF)

An Ising model is a well-known model in statistical mechanics that consists of binary variables. It is one of the simplest mathematical models of ferromagnetism. The value of each variable represents magnetic dipole moments of atomic spins, which can be up(+1) or down(-1), and each spin interacts with its neighborhood spin status. If the spins have the same state, then the energy is low.

The Ising model uses the nearest neighborhood system, that is, each spin affects other spin right beside it. The total energy of the Ising system is given by

$$E = h\sum_{i} S_i - \beta \sum_{\{i,j\}} S_i S_j, \tag{2.1}$$

where h and β are constants and the bracket sign $\{i, j\}$ represents that S_i and S_j are neighborhoods for each other. S_i means the spin state of site i.

h is the interaction constant between S_i and an external field. Because the spins are interacting with each other in the Ising model, β represents the constant between S_i and S_j for all i, j. Then the probability density of the Ising model is

$$P(S) = \frac{1}{Z}\exp\left(-\frac{1}{kT}E\right), \tag{2.2}$$

which is called Boltzmann-Gibbs distribution where Z is a normalizing constant for the distribution called partition function, k is a Boltzmann constant, E is defined in (2.1) and T is a temperature.

Since the binary image is used, we consider the image as a two-dimensional Ising model. Let S be a set of values of pixels $\{s_1, \dots, s_N\}$ where N is the total number of pixels. Then, the white and black points represent the up (+1) and down (–1) status, respectively.

We suppose that the original image is the Markov random field (MRF) with the nearest neighborhood system. The MRF is one of the graphical models composed with a clique and the neighborhood system. A clique is a set of nodes affecting each other and their functional relationship is called the clique potential. The joint distribution of the MRF equals the total product of clique potentials divided by the partition function. Let X_{ij} denote the intensity value of the original image at (i, j) on $N_1 \times N_2$ lattice L and N_{ij} be the neighborhood of X_{ij}. Besag (1974) suggested the following three conditions for $P(X)$ as conditions becoming an MRF,

1) Positivity : $P(X_{ij}) > 0 \ \forall X_{ij}$.

2) Markovianity : $P(x_{ij}|\text{all points in lattice } L \text{ except } (i, j)) = P(x_{ij}|N_{ij}),\ \forall (i, j) \in\ L.$

3) Homogeneity : $P(x_{ij}|N_{ij})$ depends on the value of the neighborhood, not the location of it.
   That is, $P(x_{ij}|N_{ij}) = P(x_{kl}|N_{kl})$ if the configuration of N_{ij} and N_{kl} is the same.

The MRF is equal to the Gibbs Random Field (GRF) by the Hammersley-Clifford theorem if they have the same neighborhood system. The GRF means that the distribution of the two-dimensional lattice follows the Gibbs distribution in (2.2). Then, the joint distribution of MRF is given by

$$P(X) = \prod_{i=1}^{N_1}\prod_{j=1}^{N_2} P(X_{ij}|N_{ij}) = \prod_{i=1}^{N_1}\prod_{j=1}^{N_2} \frac{1}{Z_{ij}}\exp\left(-\frac{1}{kT}E_{ij}\right), \tag{2.3}$$

where i, j denote the index of the row and column of $N_1 \times N_2$ lattice L and E_{ij} represents the energy of X_{ij} interacting with its neighborhood.

3. Variational Bayesian inference (VBI)

We consider the distribution of X as prior and the clique potential between X and Y as a likelihood function. Then, we calculate the posterior density of X given Y by Bayes rule.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)},$$

, where $P(Y) = \int P(X, Y)dX$, is called evidence (or marginal likelihood). However, for many models, the evidence is unavailable in closed form or requires exponential time to compute. There are two main ways of estimating the true posterior distribution without calculating the evidence. The first one is the MCMC algorithm sampling from the Markov Chain. It is theoretically certain that the generated samples follow the true posterior, but it should satisfy some conditions and is computationally expensive. The second way is the VBI approximating the true posterior density by minimizing the KL-divergence. See Blei *et al.* (2017) for more details of the VBI.

3.1. Kullback-Leibler (KL) divergence

The VBI generalizes the idea behind the Laplace approximation used to find the posterior density (Tierney and Kadane, 1986). In the VBI, we wish to find an approximated density that is closest to the true posterior density with respect to the KL-divergence as follows,

$$KL(q(x)||p(x|y)) = \sum_x q(x) \log \frac{q(x)}{p(x|y)} \tag{3.1}$$

and x is a latent variable and y is an observed data. The KL-divergence is a measurement of the difference between the functions of $q(x)$ and $p(x|y)$.

Let $\mathbf{W}$ be a family of candidates q's which approximate the posterior density. The VBI finds the optimal q^* which minimizes the KL-divergence. That is,

$$q^*(x) = \underset{q(x)}{\arg\min}\ KL\left(q(x)||p(x|y)\right).$$

From equation (3.1),

$$\begin{aligned}\mathrm{KL}(q(x)||p(x|y)) &= \sum_x q(x) \log \frac{q(x)}{p(x|y)} \\ &= \sum_x q(x) \log \frac{q(x)p(y)}{p(x,y)} \\ &= \sum_x q(x) \log q(x) + \sum_x q(x) \log p(y) - \sum_x q(x) \log p(x,y) \\ &= \mathbb{E}_{q(x)}\left[\log q(x)\right] - \mathbb{E}_{q(x)}\left[\log p(x,y)\right] + \log p(y), \end{aligned} \tag{3.2}$$

where $\mathbb{E}_{q(x)}[h(x)]$ denotes the expectation of $h(x)$ with respect to $q(x)$.
Define,

$$\mathrm{ELBO}(q) = \mathbb{E}_{q(x)}[\log p(x,y)] - \mathbb{E}_{q(x)}[\log q(x)]. \tag{3.3}$$

Here, ELBO(q) is called the evidence lower bound (ELBO). Then, the ELBO(q) in (3.3) can be written as,

$$\mathrm{ELBO}(q) = \mathbb{E}_{q(x)}[\log p(y|x)] - \mathrm{KL}(q(x)||p(x)).$$

Therefore, minimizing the KL-divergence is equal to maximizing the ELBO. We can rewrite the KL-divergence with respect to $\log p(y)$.

$$\begin{aligned}\log p(y) &= \mathbb{E}_{q(x)}[\log p(x,y)] - \mathbb{E}_{q(x)}[\log q(x)] + \mathrm{KL}(q(x)||p(x|y)) \\ &= \mathrm{ELBO}(q) + \mathrm{KL}\left(q(x)||p(x|y)\right). \end{aligned} \tag{3.4}$$

Since $\mathrm{KL}(\cdot) \geq 0$, it follows from (3.4) that the ELBO is lower bound for the evidence.

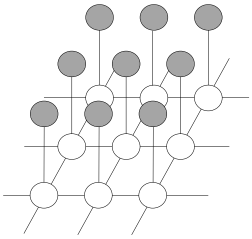

Figure 1: *Graphical Representation of X and Y.*

3.2. Mean field approximation

We described the ELBO in Section 3.1, which is the main objective function of the VBI. The goal of the VBI is to find the best candidate from the variational family $\mathbf{W}$. If the variational family is complex, the optimization problem is more difficult. To simplify the optimization problem, we employ the Mean field approximation suggested by Peterson and Anderson (1987).

The Mean field approximation means that the latent variables are mutually independent.

$$q(x) = \prod_{i=1}^{N} q_i(x_i). \tag{3.5}$$

Each latent variable x_i is governed by its variational factor $q_i(x_i)$. We choose these variational factors to maximize the ELBO.

The mean field approximation is a model for the variational density. The observed data y does not appear in equation (3.5). Thus, by the mean field approximation, it becomes easy to solve the optimization problem.

3.3. Coordinate ascent variational inference

Using the ELBO and the Mean field approximation, simplifies the optimization problem. Bishop (2006) introduced the coordinate ascent variational inference (CAVI) algorithm. The CAVI repeatedly optimizes each factor in the mean field variational density, while holding the other factors constant.

Consider the i^{th} latent variable x_i. We can rewrite the ELBO in equation (3.3) as a function of i^{th} latent factor $q_i(x_i)$ using the Mean field approximation. It follows from Blei *et al.* (2017) that

$$\text{ELBO}(q_i) = \mathbb{E}_i\left[\mathbb{E}_{-i}\left[\log p(x_i, x_{-i}, y)\right]\right] - \mathbb{E}_i\left[\log q_i(x_i)\right] + \text{const}, \tag{3.6}$$

where $\mathbb{E}_{-i}$ denotes the expectation, except for the i^{th} latent variable, and $\mathbb{E}_i$ denotes the expectation for the i^{th} latent variable.

The ELBO in (3.6) takes the form of the negative KL-divergence between $q_i(x_i)$ and $\mathbb{E}_{-i}[\log p(x_i, x_{-i}, y)]$. Therefore, the variational density q_i^* that maximizes the ELBO is given by

$$q_i^*(x_i) \propto \exp\left(\mathbb{E}_{-i}\left[\log p(x, y)\right]\right). \tag{3.7}$$

4. The proposed model and its variational Bayesian inference

4.1. Simple Ising model

We suppose that the original image X is a MRF with the nearest neighborhood and use the Ising model in Chapter 2. Figure 1 shows the relationship between the original image X and observation Y. The white circles represent the original image X and the grey circles represent the degraded image Y. In Figure 1, we see that the original image X has a connection for the nearest nodes and the degraded image Y has a connection only with corresponding X. Thus, the clique potential between X and Y consists of two nodes (x_i, y_i) for $i = 1, \ldots, N$.

We propose the simple Ising model from Bishop (2006) as follows,

$$p(x, y; h, \beta, \eta) \propto \exp\left(\beta \sum_{i=1}^{N} x_i \sum_{j \in N_i} x_j + \eta \sum_{i=1}^{N} x_i y_i - h \sum_{i=1}^{N} x_i\right), \tag{4.1}$$

where x_i and y_i are either -1 or $+1$ and $\eta, \beta > 0$. Here, y is an observed data and x is a latent variable (actually, unknown true value).

Note that the Hammersley-Clifford theorem provides a formula for the joint distribution of MRF in terms of energy function of the Ising model. Thus, the prior distribution of X can be regarded as part of the right-hand side of equation (4.1), which is a function of only x_i's.

$$\begin{aligned} P(X) &= \prod_{i=1}^{N_1} \prod_{j=1}^{N_2} P(X_{ij} | N_{ij}) \\ &\propto \prod_{i=1}^{N_1} \prod_{j=1}^{N_2} \exp(-E_{ij}) \\ &\propto \exp\left(-h \sum_{i=1}^{N} x_i + \beta \sum_{i=1}^{N} x_i \sum_{j \in N_i} x_j\right), \end{aligned} \tag{4.2}$$

where E_{ij} represents the energy function at X_{ij} with its neighborhood N_{ij} in $N_1 \times N_2$ lattice L and for the simplicity of notation, we combine the index of row and column. In addition, N represents the total number of nodes. Therefore, from (4.1), the likelihood function can be expressed as

$$p(y|x; \eta) \propto \exp\left(\eta \sum_{i=1}^{N} x_i y_i\right). \tag{4.3}$$

4.2. Variational Bayesian inference for simple Ising model

Using equations (4.2) and (4.3), we compute the variational density q^* by substituting equation (4.1) for equation (3.7). From equation (3.7) and (4.1),

$$\begin{aligned} \log q_i^*(x_i) &= \mathbb{E}_{-i}[\log p(x, y; \theta)] \\ &\propto \mathbb{E}_{-i}\left[\eta \sum_{i=1}^{N} x_i y_i + \beta \sum_{i=1}^{N} x_i \sum_{j \in N_i} x_j - h \sum_{i=1}^{N} x_i\right] \\ &\propto \mathbb{E}_{-i}\left[\eta x_i y_i + \beta x_i \sum_{j \in N_i} x_j - h x_i\right] \\ &\propto \eta x_i y_i + \beta x_i \sum_{j \in N_i} \mu_j - h x_i, \end{aligned} \tag{4.4}$$

where $\theta = (h, \beta, \eta)$ and $\mu_j = \mathbb{E}_{q_i^*}[x_i]$. Following (3.7), the approximated variational distribution is

$$\begin{aligned} q_i^*(x_i) &= \frac{\exp\left(x_i\left(\beta\sum_{j\in N_i}\mu_j + \eta y_i - h\right)\right)}{\exp\left(-\left(\beta\sum_{j\in N_i}\mu_j + \eta y_i - h\right)\right) + \exp\left(\beta\sum_{j\in N_i}\mu_j + \eta y_i - h\right)} \\ &= \frac{\exp\left(x_i(m_i + L_i)\right)}{\exp(-(m_i + L_i) + \exp(m_i + L_i)}, \end{aligned} \tag{4.5}$$

where $m_i = \beta\sum_{j\in N_i}\mu_j$ and $L_i = \eta y_i - h$.

Since x_i is a binary variable, which can have -1 or +1, the expectation of x_i is calculated by

$$\begin{aligned} \mu_i = \mathbb{E}_{q_i^*}[x_i] &= q_i^*(x_i = +1) - q_i^*(x_i = -1) \\ &= \frac{\exp\left(m_i + L_i\right) - \exp\left(-\left(m_i + L_i\right)\right)}{\exp\left(m_i + L_i\right) + \exp\left(-\left(m_i + L_i\right)\right)} \\ &= \tanh\left(m_i + L_i\right) \\ &= \tanh\left(\eta y_i + \beta\sum_{j\in N_i}\mu_j - h\right). \end{aligned} \tag{4.6}$$

Let $\theta = (h, \beta, \eta)$ and $\phi = (m_i, L_i)$ be model parameters and variational parameters, respectively. To get an expectation of x_i with respect to q^*, we need to know the mean value of neighborhoods, μ_j. By iterating the above formulation (4.6), we can solve this problem,

$$\mu_i^{(k+1)} = \tanh\left(\beta\sum_{j\in N_i}\mu_j^{(k)} + \eta y_i - h\right). \tag{4.7}$$

Also, the ELBO is calculated by

$$\begin{aligned} \mathrm{ELBO}(q) &= \mathbb{E}_{q^*}[\log p(x, y; \theta) - \log q^*(x; \phi)] \\ &= \sum_{i=1}^{N}\mathbb{E}_{q_i^*}\left[x_i\left(\eta y_i + \beta\sum_{j\in N_i}x_j - h\right) - \log\left(e^{-\left(\eta y_i + \beta\sum_{j\in N_i}x_j - h\right)} + e^{\eta y_i + \beta\sum_{j\in N_i}x_j - h}\right)\right] \\ &- \sum_{i=1}^{N}\mathbb{E}_{q_i^*}\left[x_i(m_i + L_i) - \log\left(e^{-(m_i + L_i)} + e^{m_i + L_i}\right)\right] \\ &= \sum_{i=1}^{N}\mathbb{E}_{q_i^*}\left[x_i\left(\eta y_i + \beta\sum_{j\in N_i}x_j - h\right) - \log\left(2\cosh\left(\eta y_i + \beta\sum_{j\in N_i}x_j - h\right)\right)\right] \\ &- \sum_{i=1}^{N}\mathbb{E}_{q_i^*}\left[x_i(m_i + L_i) - \log\left(2\cosh(m_i + L_i)\right)\right]. \end{aligned} \tag{4.8}$$

5. Choosing the parameters

Following the empirical Bayes method (Casella, 2001), we find the model parameter θ maximizing the evidence (marginal likelihood). We construct the variational density q^* and calculate the ELBO.

$$\log p(y; \theta) = \mathrm{ELBO}(q) + \mathrm{KL}(q(x)||p(x|y)).$$

Variational density q^* is calculated by $\exp(\mathbb{E}_{-i}[\log p(x, y; \theta)])$ in equation (3.7). Now, we want to make the ELBO in equation (4.8) larger by adjusting the parameters. We consider the EM algorithm since the form of ELBO is similar to the Q function of the EM. The main goal of the EM is to find the maximum likelihood estimates or posterior mode by iterating **E-step** and **M-step**.

5.1. Expectation maximization (EM) and the variational expectation maximization (VEM) algorithm

The main purpose of the EM algorithm (Dempster *et al.* 1977) is to find the maximum likelihood estimates,

$$\underset{\theta}{\arg\max}\ L(\theta; y) \propto p(y; \theta).$$

In the case of the marginal distribution of y being complex, we can introduce the latent variable x as below,

$$\begin{aligned}\log p(y; \theta) &= \log \sum_x p(x, y; \theta)\\ &= \log \sum_x q(x) \frac{p(x, y; \theta)}{q(x)}\\ &\geq \sum_x q(x) \log \frac{p(x, y; \theta)}{q(x)} = L(q, \theta).\end{aligned}$$

$L(q, \theta)$ is a lower bound of marginal log likelihood. Instead of maximizing $\log p(y; \theta)$, we maximize $L(q, \theta)$, which is called the Q function of the EM. To compute the Q function, we give the initial value to the model parameter, and then find the MLE of it. The EM algorithm iterates those steps for a fixed number of times or until the parameter converges. Thus, we iterate **E-step** and **M-step** described below to find the MLE of θ.

E-step Compute Q function,

$$\begin{aligned}Q(\theta; \theta^{(t)}) &= \mathbb{E}_{p(x)}[\log p(x, y; \theta)]\\ &= \sum_x p(x|y; \theta^{(t)}) \log p(x, y; \theta).\end{aligned}$$

M-step Find a new θ maximizing the Q function,

$$\theta^{(t+1)} = \underset{\theta}{\arg\max}\ Q\left(\theta; \theta^{(t)}\right),$$

where y is an observed data, x is a latent variable and its density is $p(y|x; \theta^{(t)})$. A general EM algorithm takes the simple form of $p(x|y; \theta)$ for computational convenience. In the VBI, we approximate the true posterior as q^*. Thus, we use it instead of $p(x|y; \theta)$. Then, $L(q, \theta)$ is equal to the ELBO in the VBI. The detailed process of variational EM (VEM) is as follows,

VE-step

$$\begin{aligned}Q(\theta; \phi^{(t)}) &= \mathbb{E}_{q^*}[\log p(x, y; \theta)]\\ &= \sum_x q^*\left(x; \phi^{(t)}\right) \log p(x, y; \theta),\end{aligned}$$

where $\theta = (h, \beta, \eta)$ and $\phi = (m_i, L_i)$. Also $\phi^{(t)}$ is computed by $(h^{(t)}, \beta^{(t)}, \eta^{(t)})$. That is, at iteration t, $m_i^{(t)} = \beta^{(t)} \sum_{j \in N_i} \mu_j^{(t)}$ and $L_i^{(t)} = \eta^{(t)} y_i - h^{(t)}$. In other words, in the **E-step**, we specify the variational density $q^*(x; \phi^{(t)})$ by computing $(m_i^{(t)}, L_i^{(t)})$.

$$Q(\theta; \phi^{(t)}) = \sum_x q^* \left(x; \phi^{(t)}\right) \log p(x, y; \theta).$$

M-step

$$\theta^{(t+1)} = \underset{\theta}{\arg\max}\ Q(\theta; \phi^{(t)}).$$

The VEM algorithm is the same as the EM algorithm except for the computation of the Q function.

5.2. Expectation/conditional-maximization algorithm

The VEM algorithm can be used to update the model parameters. However, it is complex to update model parameters simultaneously because there are three model parameters. Thus, we employ the expectation/conditional-maximization (ECM) algorithm (Meng and Rubin, 1993).

Suppose that the model parameter θ is $(\theta_1, \ldots, \theta_n)$ and implement the EM algorithm. Then, in the **CM-step**, one parameter θ_k is updated while holding the other parameters constant. In other words, at iteration t, update $\theta_k^{(t+1)}$ using $(\theta_1^{(t+1)}, \ldots, \theta_{k-1}^{(t+1)}, \theta_{k+1}^{(t)}, \ldots, \theta_n^{(t)})$. Since our model parameters are h, β and η, we update the parameters sequentially using the ECM algorithm. Now, we call it VECM by combining VEM and ECM.

VE-step Compute $Q(\theta; \phi)$

$$Q(\theta; \phi) = \sum_{i=1}^{N} Q_i(\theta; \phi), \tag{5.1}$$

where

$$\begin{aligned} Q_i(\theta; \phi) &= \mathbb{E}_{q_i^*}[\log p(x_i, y_i; \theta)] = \sum_{x_i} q_i^*(x_i; \phi) \log p(x_i, y_i; \theta) \\ &= q_i^*(x_i = +1; \phi) \times \log p(x_i = +1, y_i; \theta) + q_i^*(x_i = -1; \phi) \times \log p(x_i = -1, y_i; \theta). \end{aligned}$$

Then, $Q(\theta; \phi)$ is computed by

$$\begin{aligned} Q(\theta; \phi) &= \sum_{i=1}^{N} \sigma(2(m_i + L_i)) \log p(x_i = +1, y_i; \theta) + \sum_{i=1}^{N} \sigma(-2(m_i + L_i)) \log p(x_i = -1, y_i; \theta) \\ &= \sum_{i=1}^{N} \tanh(m_i + L_i) \left(\beta \sum_{j \in N_i} x_j + \eta y_i - h\right) - \sum_{i=1}^{N} \log\left(2\cosh\left(\beta \sum_{j \in N_i} x_j + \eta y_i - h\right)\right), \end{aligned} \tag{5.2}$$

where $\sigma(\cdot)$ is a sigmoid function.

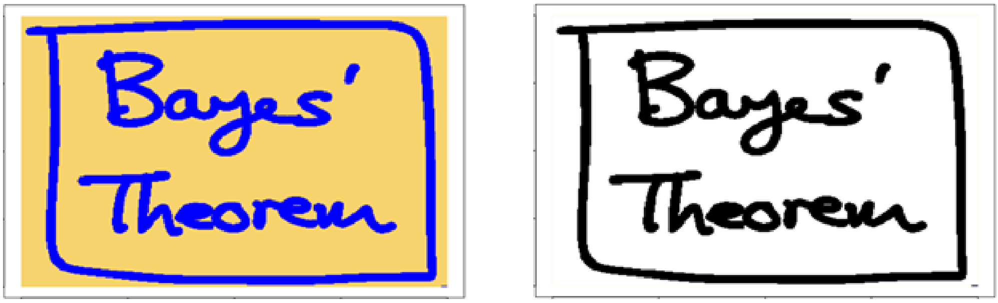

Figure 2: *Original images.*

The expression in (5.2) is summarized for each parameter as follows,

$$Q\left(h;h^{(t)}\right) = -h\sum_{i=1}^{N}\tanh(m_i+L_i) - \sum_{i=1}^{N}\log\left(2\cosh\left(\beta\sum_{j\in N_i}x_j+\eta y_i-h\right)\right), \tag{5.3}$$

$$Q\left(\beta;\beta^{(t)}\right) = \beta\sum_{i=1}^{N}\left(\tanh(m_i+L_i)\sum_{j\in N_i}x_j\right) - \sum_{i=1}^{N}\log\left(2\cosh\left(\beta\sum_{j\in N_i}x_j+\eta y_i-h\right)\right), \tag{5.4}$$

$$Q\left(\eta;\eta^{(t)}\right) = \eta\sum_{i=1}^{N}\tanh(m_i+L_i)y_i - \sum_{i=1}^{N}\log\left(2\cosh\left(\beta\sum_{j\in N_i}x_j+\eta y_i-h\right)\right). \tag{5.5}$$

CM-step Maximize Q function with respect to each parameter, sequentially.

$$h^{(t+1)} = \underset{h}{\arg\max}\, Q\left(h;h^{(t)}\right), \tag{5.6}$$

$$\beta^{(t+1)} = \underset{\beta}{\arg\max}\, Q\left(\beta;\beta^{(t)}\right), \tag{5.7}$$

$$\eta^{(t+1)} = \underset{\eta}{\arg\max}\, Q\left(\eta;\eta^{(t)}\right). \tag{5.8}$$

Each Q function in equations (5.3)-(5.5) has a logarithm hyperbolic cosine term that becomes a hyperbolic tangent function if they are differentiated. Since the hyperbolic tangent has no maximum value, we cannot find the $\theta^{(t+1)}$ using differentiation. Thus, we find the maximum value by substituting numbers within a given range for each parameter.

6. Simulation results

Using the VECM algorithm, we can implement the binary image restoration.

The left figure in Figure 2 shows the original color image and the right one is the binary image. To get the degraded image, we add some noise to the binary image with a probability of 0.1.

So far, we have discussed the structure of posterior and the variational density in Chapter 4 and the EM algorithm to find the maximum likelihood estimates in Chapter 5. By summarizing this content, the entire algorithm for restoration can be made.

```
Algorithm 1: Binary image restoration
    Step 1. Initialize the hyperparameters $h^{(1)}$, $\beta^{(1)}$, $\eta^{(1)}$ and calculate the initial values of $\mu$, $\phi$.
    Step 2. Update the hyperparameters using the VECM algorithm.
        $h^{(k+1)} = \arg\max_{h} \; Q\left(h; h^{(k)}\right)$
        $\beta^{(k+1)} = \arg\max_{\beta} \; Q\left(\beta; \beta^{(k)}\right)$
        $\eta^{(k+1)} = \arg\max_{\eta} \; Q\left(\eta; \eta^{(k)}\right)$
    Step 3. Update $\mu^{(k+1)}$ using the updated hyperparameters.
        $L_i^{(k+1)} = \eta^{(k+1)} y_i - h^{(k+1)}$
        $m_i^{(k+1)} = \beta^{(k+1)} \sum \mu_j^{(k)}$
        $\mu^{(k+1)} = \tanh\left(m_i^{(k+1)} + L_i^{(k+1)}\right)$
        $x_i^{(k+1)} = \begin{cases} +1, & \text{if } \mu_i^{(k+1)} \geq 0 \\ -1, & \text{if } \mu_i^{(k+1)} < 0 \end{cases}$
```

The left column of Figure 3 is the degraded image with 10% noise and the figures in the middle column are the results that have the highest Peak Signal-to-Noise Ratio (PSNR) and those figures on the right are the maximum ELBO during 20 iterations. PSNR is a measurement of restoration.

$$\mathrm{PSNR} = 10 \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right),$$

where MAX_I is the difference between the maximum and minimum values of pixels and MSE is the mean squared error. The highest PSNR is 22.34, 22.31, 23.05, and 22.40 for each simulation.

Figure 4 shows the change of parameters and ELBO for each iteration. First, two simulations start from the same initial value of $h = 1, \beta = 1, \eta = 1$. The results of 3rd and 4th start from $h = -1, \beta = 1.5, \eta = 3$ and $h = 2, \beta = 2, \eta = 1.8$, respectively.

7. Conclusion

In this manuscript, the degraded binary images were restored using VBI with the Ising model. We apply the VECM algorithm to maximize the ELBO and update the model parameters. If the model parameters are not updated, the restoration is very restricted. Without parameter estimating, the restoration is greatly influenced by the initial value of the hyperparameters. If we set $h = 0$, then the restoration result without parameter estimating is quite good. On the other hand, for example, if $h = 1, \beta = 1, \eta = 1$, then the restored image becomes a black screen after a few iterations. Likewise, if $h = -1, \beta = 1, \eta = 1$, then the restored image becomes a white screen. Because the restoration process is highly sensitive to the initial values of the hyperparameters, especially h in this simulation data, it is important to find the appropriate model parameters for each iteration.

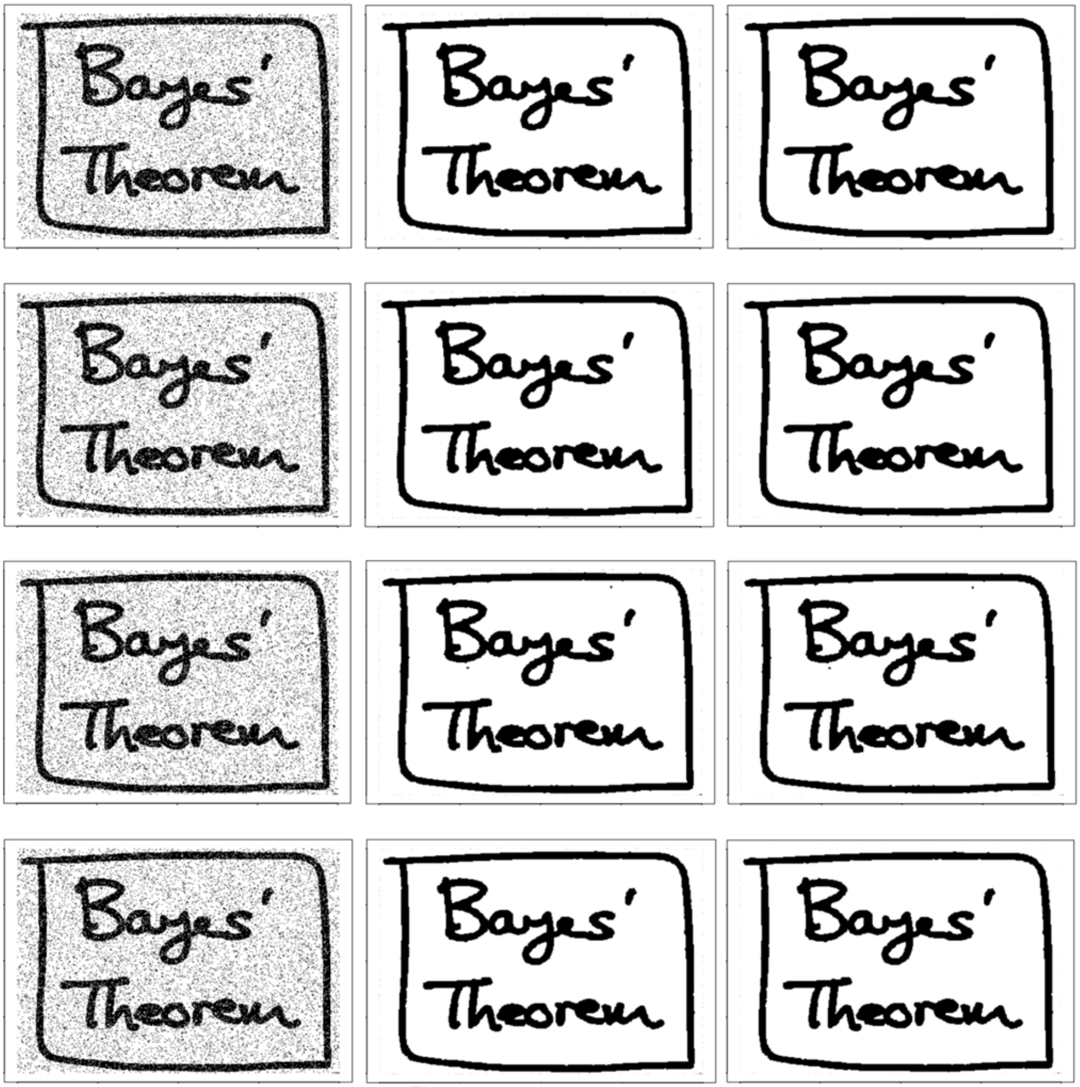

Figure 3: *Restored images.*

When the Q functions are computed in equations (5.6)-(5.8), we have to solve the numerical equations to find a maxima. In this case, we substitute the numbers within a restricted range for each set of parameters. In addition, the simple Ising model and prior distribution are used. Therefore, this paper is applied only to the binary images. These two main problems can be solved by changing the likelihood and the prior distributions. If the distribution of noise is changed, for example with Gaussian or Student's t-distribution, and another model is used instead of Ising, then it is possible to expand to gray scale or color images. The gray scale and color image analysis using the VI method will be our next thesis.

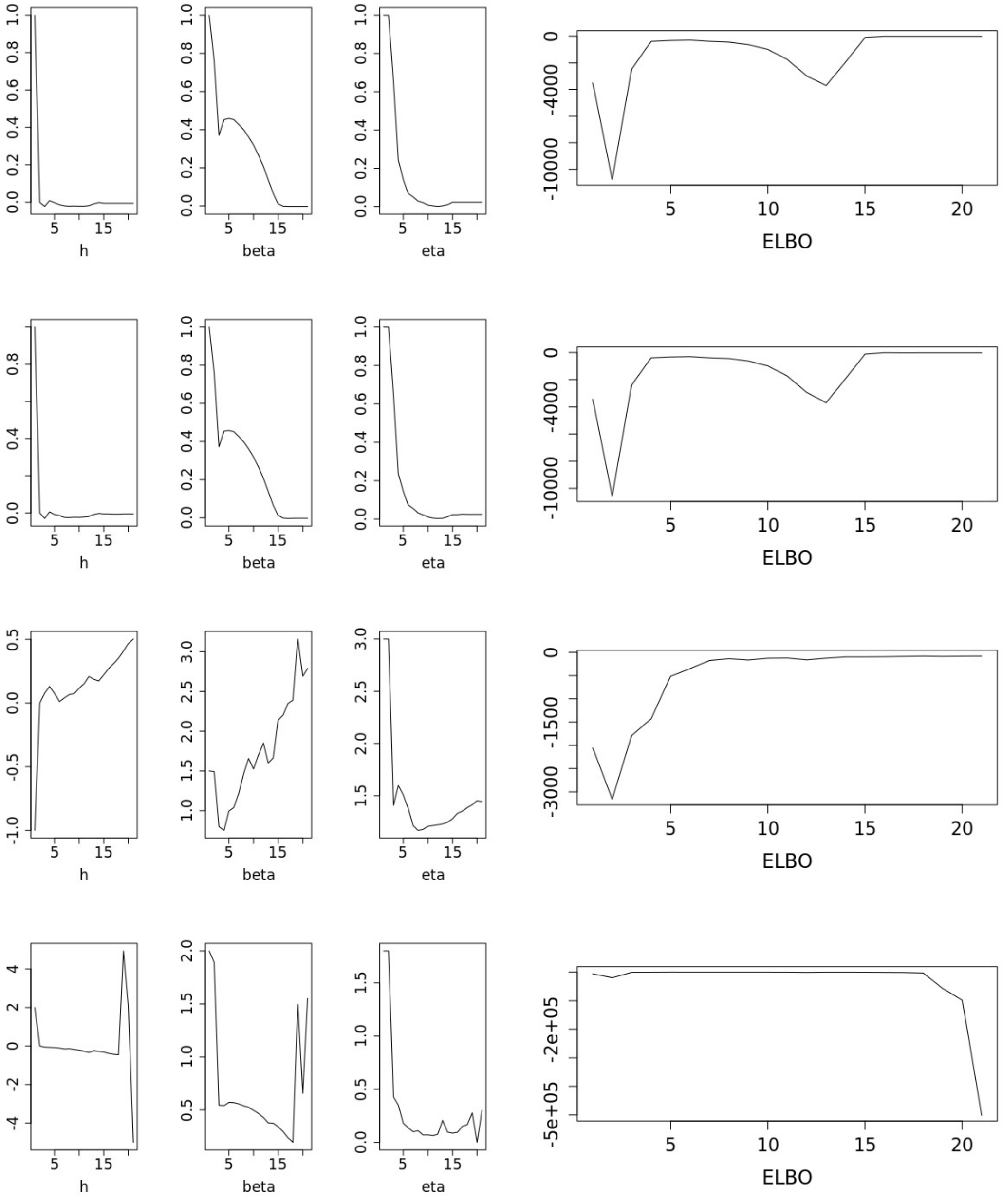

Figure 4: *Simulation results of* h, β, η *and ELBO.*

References

Besag J (1974). Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society Series B*, **48**, 259–279.

Besag J (1986). On the statistical analysis of dirty pictures, *Journal of the Royal Statistical Society Series B*, **48**, 259–302.

Besag J, York J, and Mollie A (1991). Bayesian image restoration, with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

Bishop C (2006). *Pattern Recognition and Machine Learning*, Springer-Verlag, Berlin, Heidelberg.

Blei DM, Kucukelbir A, and McAuliffe JD (2017). Variational inference: A review for Statisticians, *Journal of the American Statistical Association*, **112**, 859–877.

Casella G and George EI (1992). Explaining the Gibbs Sampler, *The American Statistician*, **46**, 167–174.

Casella G (2001). Emperical Bayes Gibbs sampling, *Biostatistics*, **2**, 485–500.

Dempster AP, Laird NM, and Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Series B*, **39**, 1–38.

Friedman N (2013). *The Bayesian Structural EM Algorithm*, arXiv preprint arXiv:1301.7373.

Gelfand AE and Smith FM (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398–409.

Geman S and Geman D (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Hastings WK (1970). Monte Carlo sampling methods using Markov Chains and their applications, *Biometrika*, **57**, 97–109.

Jordan MI, Ghahramani Z, Jaakkola TS, and Saul LK (1999). An introduction to variational methods for graphical models, *Machine Learning, Kluwer Academic Publishers*, **37**, 183-–233.

Meng X and Rubin DB (1993). Maximum likelihood estimation via the ECM algorithm: A general framework, *Biometrika*, **80**, 267–278.

Metropolis N, Rosenbluth AW, Rosenbluth MN, and Teller AH (1953). Equation of state calculations by fast computing machines, *The Journal of Chemical Physics*, **21**.

Nasios N and Bors AG (2006). Variational learning for Gaussian mixture models, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **36**, 849–862.

Parisi G (1988). *Statistical Field Theory*, Addison-Wesley, Redwood City.

Peterson C and Anderson JR (1987). A mean field theory learning algorithm for neural networks, *Complex Systems*, **1**, 995–1019.

Smith AFM and Roberts GO (1993). Bayesian computation via the Gibbs sampler and Related Markov Chain Monte Carlo Methods, *Journal of the Royal Statistical Society Series B*, **55**, 3–23.

Tian H, Shen T, Hao B, Hu Y, and Yang N (2009) Image restoration based on adaptive MCMC particle filter. In *2009 2nd International Congress on Image and Signal Processing, IEEE*, 1–5.

Tierney L and Kadane JB (1986). Accurate approximations for posterior moments and marginal densities, *Journal of American Statistical Association*, **81**, 82–86.

Zhang C, Bütepage J, Kjellstr¨öm H, and Mandt S (2019). Advances in variational inference, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**, 2008–2026.

Received February 12, 2021; Revised December 10, 2021; Accepted December 21, 2021