Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Statistically Valid Variational Bayes Algorithm for Ising Model Parameter Estimation

Minwoo Kim, Shrijita Bhattacharya & Tapabrata Maiti

To cite this article: Minwoo Kim, Shrijita Bhattacharya & Tapabrata Maiti (2024) Statistically Valid Variational Bayes Algorithm for Ising Model Parameter Estimation, Journal of Computational and Graphical Statistics, 33:1, 75-84, DOI: 10.1080/10618600.2023.2217869

To link to this article: https://doi.org/10.1080/10618600.2023.2217869

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2024, VOL. 33, NO. 1, 75–84
https://doi.org/10.1080/10618600.2023.2217869

Statistically Valid Variational Bayes Algorithm for Ising Model Parameter Estimation

Minwoo Kim[a], Shrijita Bhattacharya[b], and Tapabrata Maiti[b]

[a]Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia; [b]Department of Statistics and Probability, Michigan State University, East Lansing, MI

ABSTRACT

Ising models originated in statistical physics and are widely used in modeling spatial data and computer vision problems. However, statistical inference of this model remains challenging due to intractable nature of the normalizing constant in the likelihood. Here, we use a pseudo-likelihood instead, to study the Bayesian estimation of two-parameter, inverse temperature and magnetization, Ising model with a fully specified coupling matrix. We develop a computationally efficient variational Bayes procedure for model estimation. Under the Gaussian mean-field variational family, we derive posterior contraction rates of the variational posterior obtained under the pseudo-likelihood. We also discuss the loss incurred due to variational posterior over true posterior for the pseudo-likelihood approach. Extensive simulation studies validate the efficacy of mean-field Gaussian and bivariate Gaussian families as the possible choices of the variational family for inference of Ising model parameters. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2022
Accepted May 2023

KEYWORDS
Black box variational inference; Coupling matrix; ELBO; Kullback-Leibler distance; Posterior contraction rates; Pseudo-likelihood; Stochastic Optimization.

1. Introduction

A popular way of modeling a dependent binary vector $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ is to take advantage of Ising model named after the physicist Ernst Ising (Ising 1924) which has been used in a wide range of applications. For examples, Ising models have been used for voter models in social science (Lipowski, Lipowska, and Ferreira 2017), interactions between genetic markers (Majewski, Li, and Ott 2001), describing complexity class of topological quantum computer (Lahtinen and Pachos 2017), and describing the pairing of electrons (Li et al. 2021). In Statistics, many researchers have considered Ising models for a variety of statistical problems including variable selection and clustering (Smith and Fahrmeir 2007; Li and Zhang 2010; Lee et al. 2014; Li et al. 2015; Fang and Kim 2016; Park, Jin, and Schweinberger 2022).

Many different versions of Ising model have emerged in the literature. In this article, among them, we focus on two-parameter Ising model, which has an inverse temperature (interaction) parameter $\beta > 0$ and a magnetization (threshold) parameter $B \neq 0$, with a symmetric coupling matrix $A_n \in \mathbb{R}^{n\times n}$. An Ising model is often represented by an undirected graph in which each vertex (node) is a binary variable $x_i \in \{-1, 1\}$ and the connections between x_i's are determined by A_n. Here, β characterizes the strength of interactions among x_i's and B represents external influence on $\boldsymbol{x}$. In the first place, Ising model has been introduced for the relations between atom spins Brush (1967) with the domain $\{-1, 1\}^n$. While we work with the domain $\{-1, 1\}^n$, in many current applications, Ising model has been defined with a different domain $\{0, 1\}^n$. One can read Haslbeck et al. (2021) for more details on two different domains.

Estimation of Ising model parameters has received considerable attention in statistics and computer science literature. The existing literature can be broadly divided into two groups. Some literature assume that iid copies of data ($\boldsymbol{x}$ vector) are available for inference, Anandkumar et al. (2012), Bresler (2015), Lokhov et al. (2018), Ravikumar, Wainwright, and Lafferty (2010), and Xue, Zou, and Cai (2012). Another category of literature assumes that only one sample is observable, Bhattacharya and Mukherjee (2018), Chatterjee (2007), Comets (1992), Comets and Gidas (1991), Ghosal and Mukherjee (2020), Gidas (1988), and Guyon and Künsch (1992). Under the assumption of only one observation, Comets and Gidas (1991) showed that the MLE of $\beta > 0$ for Curie-Weiss model is consistent if $B \neq 0$ is known, and vice versa. They also proved that the joint MLE does not exist when neither β nor B is given. In this regard, Ghosal and Mukherjee (2020) addressed joint estimation of (β, B) using pseudo-likelihood and showed that the pseudo-likelihood estimator is consistent under some conditions on coupling matrix A_n. We also assume only one observation of $\boldsymbol{x}$ and provide a variational Bayes algorithm for model parameter estimation with its posterior consistency.

Methodological Contribution: One of the main challenges in the Bayesian estimation of Ising models lies in the intractable nature of the normalizing constant in the likelihood. Following the works of Ghosal and Mukherjee (2020), Bhattacharya and Mukherjee (2018), and Okabayashi, Johnson, and Geyer (2011), we replace the true likelihood of the Ising model by a pseudo-likelihood. As a first contribution, we establish that the posterior based on the pseudo-likelihood is consistent for a suitable choice of the prior distribution. Further, we use

CONTACT Shrijita Bhattacharya bhatta61@msu.edu Department of Statistics and Probability, Michigan State University, East Lansing, MI.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

variational Bayes (VB) approach which has recently become a popular and computationally powerful alternative to MCMC. In order to approximate the unknown posterior distribution using VB, we propose a Gaussian mean field family and general bivariate normal family with transformation of the parameters to $(\log \beta, B)$. For implementation of VB, we consider a black box variational inference (BBVI), Ranganath, Gerrish, and Blei (2014). In BBVI, we need to evaluate the likelihood to compute the gradient estimates, but the existence of an unknown normalizing constant in likelihood of Ising model prevents us using BBVI directly. So, as mentioned above, we use pseudo-likelihood instead of directly using the true likelihood as in Ghosal and Mukherjee (2020). Our VB algorithm based on optimization is computationally more powerful than the sampling based MCMC methods (Møller et al. 2006). Also, by the virtue of PyTorch's automatic differentiation, we do not need to manually compute necessary gradients (See the tutorial: *https://pytorch.org/tutorials/beginner/blitz/autograd_tutorial.html*). Python codes using PyTorch's automatic differentiation and data are available at Github *https://github.com/stat-kim/vb-Ising*.

Theoretical Contribution: The main theoretical contribution of this work lies in establishing the consistency of the variational posterior for the Ising model with the true likelihood replaced by the pseudo-likelihood. In this direction, we first establish the rates at which the true posterior based on the pseudo-likelihood concentrates around the ε_n- shrinking neighborhoods of the true parameters. With a suitable bound on the Kulback-Leibler distance between the true and the variational posterior, we next establish the rate of contraction for the variational posterior and demonstrate that the variational posterior also concentrates around ε_n-shrinking neighborhoods of the true parameter. These results have been derived under three set of assumptions on the coupling matrix A_n (see Section 3 for more details). Indeed, we demonstrate that the variational posterior consistency holds for the same set of assumptions on A_n as those needed for the convergence of the maximum likelihood estimates based on the pseudo-likelihood. One of the main caveats in establishing the posterior contraction rates under the pseudo-likelihood structure is in ensuring that the concentration of the variational posterior occurs in $\mathbb{P}_0^{(n)}$ probability where $\mathbb{P}_0^{(n)}$ is the distribution induced by the true likelihood and not the pseudo-likelihood. Indeed, we could show that in $\mathbb{P}_0^{(n)}$ probability, the contraction of variational posterior happens at the rate $1 - 1/M_n$ in contrast to the faster rate $1 - \exp(-Cn\varepsilon_n^2), C > 0$ for the true posterior. As a final theoretical contribution, we establish that the variational Bayes estimator converges to the true parameters at the rate $1/\varepsilon_n$ where ε_n can be chosen $n^{-\delta}$, $0 < \delta < 1/2$ provided the A_n matrix satisfies certain regularity assumptions.

The rest of the article is organized as follows: Section 2 defines the likelihood and pseudo-likelihood of Ising model with two parameters (β, B) and provides the details of our Bayesian estimation using variational approach. In Section 3 we discuss our main theoretical developments and sketch of the proof. The numerical studies are provided in Section 4. We give a comparison of our variational Bayes estimates to existing maximum likelihood estimators based on pseudo-likelihood Ghosal and Mukherjee (2020) and MCMC based method Møller et al. (2006). Section 5 shows a real-world application using Facebook network data. Technical details of theoretical results are deferred to the supplementary document.

2. Model and Methods

2.1. Ising Model

For a representation of an Ising model with two parameters $\beta > 0$ and $B \neq 0$, we consider an undirected graph which has n vertices x_i, $i = 1, \ldots, n$. Each vertex of the graph takes a value either -1 or 1, that is, $x_i \in \{-1, 1\}$. Then, we define a likelihood of Ising model as the probability of the vector $\boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \{-1, 1\}^n$:

$$\mathbb{P}_{\beta,B}^{(n)}(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{Z_n(\beta, B)} \exp\left(\frac{\beta}{2}\boldsymbol{x}^\top A_n \boldsymbol{x} + B\sum_{i=1}^{n} x_i\right), \quad (1)$$

where $Z_n(\beta, B)$ is a normalizing constant which makes the sum of (1) over the support $\{-1, 1\}^n$ equal to 1 and A_n is a coupling matrix of size $n \times n$ which determines the connections between the coordinates of $\boldsymbol{x}$. More precisely, let $\mathcal{E} = \{(i, j) \mid i \sim j,\ 1 \leq i, j \leq n\}$ be the set of edges in the graph where $i \sim j$ denote that the vertices i and j are connected. Then, we define A_n as a symmetric matrix with $A_n(i, j) = 0$ for all $(i, j) \notin \mathcal{E}$ and $A_n(i, j) > 0$ for all $(i, j) \in \mathcal{E}$.

In our study, with the regard to only one observation of $\boldsymbol{x} \in \{-1, 1\}^n$, estimating all the elements of A_n is impossible because A_n has $n(n-1)/2$ distinct values. In this work, we primarily focus on the problem of estimation of the parameters (β, B) under the assumption of a fully known coupling matrix A_n. The same set up was considered by Bhattacharya and Mukherjee (2018), Ghosal and Mukherjee (2020), and Okabayashi, Johnson, and Geyer (2011).

2.2. Pseudo-Likelihood

It is challenging to use the likelihood (1) directly because of the unknown normalizing constant $Z_n(\beta, B)$. Due to the intractable nature of the likelihood, the standard Bayesian implementation is computationally intractable. We thereby propose the use of the conditional probability of x_i given others. It is easily calculated because x_i is binary:

$$\mathbb{P}_{\beta,B}^{(n)}\left(X_i = 1 \mid X_j, j \neq i\right) = \frac{e^{\beta m_i(\boldsymbol{x}) + B}}{e^{\beta m_i(\boldsymbol{x}) + B} + e^{-\beta m_i(\boldsymbol{x}) - B}},$$

where $m_i(\boldsymbol{x}) = \sum_{j=1}^{n} A_n(i, j)x_j$. The pseudo-likelihood of Ising model corresponding to the likelihood in (1) is defined as the product of one dimensional conditional distributions (see Ghosal and Mukherjee 2020, for further details):

$$\prod_{i=1}^{n} \mathbb{P}_{\beta,B}^{(n)}\left(X_i = x_i \mid X_j, j \neq i\right)$$
$$= 2^{-n} \exp\left(\sum_{i=1}^{n} \left(\beta x_i m_i(\boldsymbol{x}) + B x_i - \log\cosh(\beta m_i(\boldsymbol{x}) + B)\right)\right). \quad (2)$$

Our subsequent Bayesian development will make use of the pseudo-likelihood (2) instead of the true likelihood (1). We shall establish that the variational posterior (7) obtained by the use of the pseudo-likelihood allows for consistent estimation of the model parameters.

2.3. Bayesian Formulation

Let $\theta = (\beta, B)$ be the parameter set of interest. We consider the following independent prior distribution $p(\theta) = p_\beta(\beta)p_B(B)$, with $p_\beta(\beta)$ as a log-normal prior for β and $p_B(B)$ as a normal prior for B as follows:

$$p_\beta(\beta) = \frac{1}{\beta\sqrt{2\pi}} e^{-\frac{(\log\beta)^2}{2}}, \quad p_B(B) = \frac{1}{\sqrt{2\pi}} e^{-\frac{B^2}{2}}. \tag{3}$$

The assumption of log-normal prior on β is to ensure the positivity of β. Based on the prior $p(\theta)$ and the pseudo-likelihood $L(\theta)$ as in (2), we have the following posterior distribution:

$$\Pi(\mathcal{A} \mid X^{(n)}) = \frac{\int_{\mathcal{A}} \pi(\theta, X^{(n)})d\theta}{m\left(X^{(n)}\right)} = \frac{\int_{\mathcal{A}} L(\theta)p(\theta)d\theta}{\int L(\theta)p(\theta)d\theta}, \tag{4}$$

for any set $\mathcal{A} \subseteq \Theta$ where Θ denotes the parameter space of θ. Note, $\pi(\theta, X^{(n)})$ is the joint density of θ and the data $X^{(n)}$ and $m\left(X^{(n)}\right) = \int L(\theta)p(\theta)d\theta$ is the marginal density of $X^{(n)}$ which is free from the parameter set θ.

2.4. Variational Inference

Next, we provide a variational approximation to the posterior distribution (4) considering two choices of the variational family in order to obtain approximated posterior distribution *(variational posterior)*. One candidate of our variational family, for the virtue of simplicity, is a mean-field (MF) Gaussian family as follows:

$$\mathcal{Q}^{\mathrm{MF}} = \Big\{ q(\theta) \mid q(\theta) = q_\beta(\beta)q_B(B), \log\beta \sim N(\mu_1, \sigma_1^2), \\ B \sim N(\mu_2, \sigma_2^2) \Big\}. \tag{5}$$

The above variational family is the same as a lognormal distribution on β and normal distribution on B. Also, we point out that β and B are independent in $\mathcal{Q}^{\mathrm{MF}}$ and each $q(\theta) \in \mathcal{Q}^{\mathrm{MF}}$ is governed by its own parameter set $\nu^{\mathrm{MF}} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^\top$. ν^{MF} denotes the set of variational parameters which will be updated to find the optimal variational distribution closest to the true posterior.

Beyond the mean field family, we suggest a bivariate normal (BN) family to exploit the interdependence among the parameters (β, B) as follows:

$$\mathcal{Q}^{\mathrm{BN}} = \Big\{ q(\theta) \mid q(\theta) = q(\beta, B), (\log\beta, B) \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Big\}, \tag{6}$$

where $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$. Here, $\mathcal{Q}^{\mathrm{MF}} \subset \mathcal{Q}^{\mathrm{BN}}$ since $\mathcal{Q}^{\mathrm{MF}}$ can be obtained from $\mathcal{Q}^{\mathrm{BN}}$ by restricting $\sigma_{12} = 0$. Thus, one may expect the $\mathcal{Q}^{\mathrm{BN}}$ to provide a better approximation to the true posterior over $\mathcal{Q}^{\mathrm{MF}}$. The variational parameters of BN family are $\nu^{\mathrm{BN}} = (\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12})^\top$. Once a variational family is selected, the variational posterior is obtained by minimizing the Kullback-Leibler (KL) divergence between a variational distribution $q \in \mathcal{Q}$ and the true posterior (4). The variational posterior is thus given by

$$Q^* = \arg\min_{Q \in \mathcal{Q}} \mathrm{KL}(Q, \Pi(\mid X^{(n)})), \tag{7}$$

where $\mathrm{KL}(Q, \Pi(\mid X^{(n)}))$ is the KL divergence given by

$$\mathrm{KL}(Q, \Pi(\mid X^{(n)})) = \int \log\left(\frac{q(\theta)}{\pi(\theta \mid X^{(n)})}\right) q(\theta)d\theta,$$

where q and $\pi(\mid X^{(n)})$ are the densities corresponding to Q and $\Pi(\mid X^{(n)})$, respectively.

Based on (4), we rewrite the KL divergence as

$$\begin{aligned} \mathrm{KL}(Q, \Pi(\mid X^{(n)})) &= \int \left(\log q(\theta) - \log \pi(\theta, X^{(n)})\right) q(\theta)d\theta \\ &\quad + \log m\left(X^{(n)}\right) \\ &= -\mathrm{ELBO}(Q, \Pi(, X^{(n)})) + \log m\left(X^{(n)}\right). \end{aligned}$$

The first term is the negative Evidence Lower Bound (ELBO) and observe that the second term does not depend on q. Therefore, minimizing KL divergence is equivalent to maximizing the ELBO. So, we search for an optimal q by maximizing the ELBO:

$$Q^* = \arg\max_{Q \in \mathcal{Q}} \mathrm{ELBO}\left(Q, \Pi(, X^{(n)})\right).$$

To optimize the ELBO, we consider the ELBO as a function of variational parameters ν:

$$\mathcal{L}(\nu) := \mathbb{E}_Q\left(\log \pi(\theta, X^{(n)}) - \log q(\theta; \nu)\right).$$

Ranganath, Gerrish, and Blei (2014) suggested black box variational inference (BBVI) for optimizing the ELBO using gradient descent method. The gradient of $\mathcal{L}(\nu)$ with respect to $\nu \in \nu$ is

$$\begin{aligned} \nabla_\nu \mathcal{L} &= \nabla_\nu \mathbb{E}_Q\left(\log \pi(\theta, X^{(n)}) - \log q(\theta; \nu)\right) \\ &= \int q(\theta; \nu)\nabla_\nu \log q(\theta; \nu)\left(\log \pi(\theta, X^{(n)}) - \log q(\theta; \nu)\right)d\theta \\ &\quad + \int q(\theta; \nu)\nabla_\nu\left(\log \pi(\theta, X^{(n)}) - \log q(\theta; \nu)\right)d\theta \\ &= \mathbb{E}_Q\left(\nabla_\nu \log q(\theta; \nu)\left(\log \pi(\theta, X^{(n)}) - \log q(\theta; \nu)\right)\right). \end{aligned} \tag{8}$$

The last equality holds because $\mathbb{E}_Q(\nabla_\nu \log q(\theta; \nu)) = 0$ and $\nabla_\nu \log \pi(\theta, X^{(n)}) = 0$. Since the expectation in (8) cannot be computed exactly, we use the Monte Carlo estimate:

$$\begin{aligned} \widehat{\nabla}_\nu \mathcal{L} &= \frac{1}{S}\sum_{s=1}^{S} \nabla_\nu \log q(\theta^{(s)}; \nu) \\ &\quad \times \left(\log \pi(\theta^{(s)}, X^{(n)}) - \log q(\theta^{(s)}; \nu)\right), \end{aligned} \tag{9}$$

where $\theta^{(1)}, \ldots, \theta^{(S)}$ are samples generated from $q(\theta^{(s)}; \boldsymbol{\nu})$ and $\log \pi(\theta^{(s)}, X^{(n)}) = \log L(\theta^{(s)}) + \log p(\theta^{(s)})$. The explicit expressions for $\log L$, $\log p$ and $\log q$ are given by

$$\begin{aligned}
\log L(\theta^{(s)}) &= -n \log 2 + \sum_{i=1}^{n} \Big(\beta^{(s)} x_i m_i(\boldsymbol{x}) + B^{(s)} x_i \\
&\qquad - \log \cosh(\beta^{(s)} m_i(\boldsymbol{x}) + B^{(s)}) \Big), \\
\log p(\theta^{(s)}) &= -\log(2\pi) - \log \beta^{(s)} - \frac{1}{2}(\log \beta^{(s)})^2 - \frac{1}{2}(B^{(s)})^2, \\
\log q(\theta^{(s)}; \boldsymbol{\nu}) &= -\log(2\pi) - \log \beta^{(s)} - \frac{1}{2} \log |\boldsymbol{\Sigma}| \\
&\quad -\frac{1}{2} \Big((\log \beta^{(s)}, B^{(s)}) - \boldsymbol{\mu} \Big)^{\top} \\
&\quad \boldsymbol{\Sigma}^{-1} \Big((\log \beta^{(s)}, B^{(s)}) - \boldsymbol{\mu} \Big).
\end{aligned}$$

Although we do not need to manually compute the gradients using PyTorch's automatic differentiation, we provide explicit expressions in Supplement Section A.2. Using the estimate (9), we iteratively update $\boldsymbol{\nu}$ in the direction of increasing the objective function $\mathcal{L}(\boldsymbol{\nu})$. The summary of BBVI algorithm is shown in Algorithm 1.

In Algorithm 1, $\rho_t, t = 1, 2, \ldots$ is a sequence of learning rates which satisfy the Robbin-Monro conditions Robbins and Monro (1951), that is, $\sum_{t=1}^{\infty} \rho_t = \infty$ and $\sum_{t=1}^{\infty} \rho_t^2 < \infty$. Let $\sigma \in \boldsymbol{\nu}$ be a variational parameter which must be positive. However, during the updating procedure, one may obtain a negative value of σ. To preclude this issue, we consider a reparameterization $\sigma = \log(1 + e^{\eta})$ and update the quantity η, as a free parameter instead. We address more details of BBVI algorithm implementation in Supplement Section A.2.

Algorithm 1 Black box variational inference (BBVI)

Initialize: $p(\theta)$, $q(\theta; \boldsymbol{\nu}^1)$ and learning rate sequence ρ_t.

1: **while** ELBO increases **do**
2: Draw $\theta^{(s)} \sim q(\theta; \boldsymbol{\nu}^t), s = 1, \ldots, S$;
3: Get $\widehat{\nabla}_{\nu} \mathcal{L}$ based on the S sample points;
4: Update $\nu^{t+1} \leftarrow \nu^t + \rho_t \widehat{\nabla}_{\nu} \mathcal{L}$;
5: **end while**

Output: Optimal variational parameters $\boldsymbol{\nu}^*$

3. Main Theoretical Results

In this section, we establish the posterior consistency of the variational posterior (7) under the mean-field family, $\mathcal{Q} = \mathcal{Q}^{\mathrm{MF}}$. In this direction, we establish the variational posterior contraction rates to evaluate how well the variational posterior of β and B concentrates around the true values β_0 and B_0. Toward the proof, we make the following assumptions:

Assumption 1 (Bounded row sums of A_n). The row sums of A_n are bounded above

$$\max_{i \in [n]} \sum_{j=1}^{n} A_n(i,j) \leq \gamma,$$

for a constant γ independent of n.

Consider the simple situation where A_n is a 0–1 matrix indicating whether two nodes are connected or not. Then, Assumption 1 implies that even with growing number of edges in a graph, the number of neighbors of each node still remains finite. In the more general case, by Assumption 1, it can be shown that $|m_i(\boldsymbol{x})| \leq \gamma$, $i = 1 \ldots, n$. Due to Assumption 1, $\sup_{x \in [-1,1]^n} \sum_{i=1}^{n} |\sum_{j=1}^{n} A_n(i,j) x_j| = O(n)$. If this does not hold, the log-normalization constant $\log Z_n(\beta, B)$ grows super linearly which implies $\lim_{n \to \infty} (1/n) \log Z_n(\beta, B) = +\infty$. Also, by 1, $||A_n||_2 \leq \gamma$ which is a regularity condition to guarantee that no eigen value of A_n has an unduly large effect on the corresponding Ising model (see the eq. (1.2) and the discussion following it in Ghosal and Mukherjee (2020) for further details).

Assumption 2 (Mean field assumption on A_n). Let $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ such that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} A_n(i,j)^2 = o(n\epsilon_n^2).$$

If A_n is a 0–1 matrix, Assumption 2 implies that even with growing number of edges in a graph, the total number of edges of the graph grow at a rate smaller than the sample size n. More generally, ϵ_n of Assumption 2 controls the contraction rate of the variational posterior (see Theorem 1). Indeed Assumption 2 is used to control the expected L_2 loss between the true likelihood and pseudo likelihood and to bound the normalization constants of the true and pseudo likelihood (see the Supplement Section B.4). Assumption 2 for $\epsilon_n = 1$ was introduced in Definition 1.3 in Basak and Mukherjee (2017) to study the limiting behavior of Ising and Potts models. We direct the reader to Section 1.2 in Ghosal and Mukherjee (2020) for a discussion on matrices which satisfy the mean field assumption.

Assumption 3 (Nonzero limiting variance of row sums). Let $\bar{A}_n = (1/n) \sum_{i=1}^{n} \sum_{j=1}^{n} A_n(i,j)$,

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} A_n(i,j) - \bar{A}_n \right)^2 > 0.$$

If A_n is 0–1 adjacency matrix, then Assumption 3 implies that as the graph grows, all nodes do not have the same number of neighbors in the limit. This is same as assuming that A_n is asymptotically irregular (a graph in which each node has the same degree is said to be regular and irregular otherwise). In the more general case, Assumption 3 ensures that $T_n(\boldsymbol{x}) = (1/n) \sum_{i=1}^{n} (m_i(\boldsymbol{x}) - \bar{m}(\boldsymbol{x}))^2$ is bounded below and above in probability, an essential requirement toward the proof of Theorem 1 (see the Supplement Section B.1). We direct the reader to eq. (1.7) in Ghosal and Mukherjee (2020) for further details.

Although the theoretical results presented in this section are applicable to any class of adjacency matrices with nonnegative entries, the more interesting examples occur when A_n is a scaled adjacency matrix (see also sec. 1.2 in Ghosal and Mukherjee 2020).

Definition 1. A scaled adjacency matrix for a graph G_n with n vertices is defined as

$$A_n(i,j) := \begin{cases} \frac{n}{2|G_n|} & \text{if } (i,j) \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases},$$

where $|G_n|$ denotes the number of edges.

The numerical results in Sections 4 and 5 are based on scaled adjacency matrices. We next provide a simple example of a scaled adjacency matrix used in Section 4 satisfying Assumptions 1, 2, and 3. We use the notation d_n to denote the degree of a regular graph.

Example 1. Consider a graph G_n with $|G_n| = (nH_n)/(2\epsilon_n^2)$ for some sequence $H_n \to \infty$. Then, the number of nonzero entries in A_n is $2|G_n|$ and $\sum_{i=1}^n \sum_{j=1}^n A_n(i,j)^2 = (n^2/(4|G_n|^2)) \cdot 2|G_n| = (n\epsilon_n^2)/H_n = o(n\epsilon_n^2)$. This satisfies Assumption 2. For Assumptions 1 and 3, consider the matrix $A_n \in \mathbb{R}^n$ with a submatrix of size $(n-k_n) \times (n-k_n)$ which is d_n-regular and the other k_n rows only have zero entries. Let us assume $k_n = n/2$ such that only the first $n/2$ rows of A_n have nonzero entries. Then, $nd_n/2 = 2|G_n| = (nH_n)/\epsilon_n^2$, which implies $d_n = (2H_n)/\epsilon_n^2$. Also, for $i = 1, \ldots, n/2$, we get $\sum_{j=1}^n A_n(i,j) = d_n \cdot (n/(2|G_n|)) = (d_n\epsilon_n^2)/H_n = 2$ as a result of which Assumption 1 holds. Finally, Assumption 3 holds since $\bar{A}_n = (d_n\epsilon_n^2)/(2H_n)$ and $(1/n)\sum_{i=1}^n (\sum_{j=1}^n A_n(i,j) - \bar{A}_n)^2 = (d_n\epsilon_n^2)/(2H_n) = 1 > 0$.

We next present the main theorem on the contraction rate for the variational posterior. We also establish the contraction rate of the variational Bayes estimator as a corollary. Let $\theta = (\beta, B)$ be the model parameter and $\theta_0 = (\beta_0, B_0)$ be the true parameter from which the data are generated. Let $L(\theta)$ and $L(\theta_0)$ denote the pseudo-likelihood as in (2) under the model parameters and true parameters, respectively. Let L_0 denote the true probability mass function of the data. Thus, L_0 is as in (1) with $\theta = \theta_0$. We use the notations $\mathbb{E}_0^{(n)}$ and $\mathbb{P}_0^{(n)}$ to denote expectation and probability mass function with respect to L_0.

Theorem 1 (Posterior Contraction). Let $\mathcal{U}_{\varepsilon_n} = \{\theta : \|\theta - \theta_0\|_2 \le \varepsilon_n\}$ be neighborhood of the true parameters. Suppose ϵ_n satisfies Assumption 2, then in $\mathbb{P}_0^{(n)}$ probability

$$Q^*(\mathcal{U}_{\varepsilon_n}^c) \to 0, \quad n \to \infty,$$

where $\varepsilon_n = \epsilon_n\sqrt{M_n \log n}$ for any slowly increasing sequence $M_n \to \infty$ satisfying $\varepsilon_n \to 0$.

The above result establishes that the posterior distribution of β and B concentrates around the true value β_0 and B_0 at a rate slightly larger than ϵ_n. The proof of the above theorem rests on following lemmas, whose proofs have been deferred to Supplement.

Lemma 1. There exists a constant $C_0 > 0$, such that for any $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$,

$$\mathbb{P}_0^{(n)}\left(\log \int_{\mathcal{U}_{\varepsilon_n}^c} \frac{L(\theta)}{L(\theta_0)} p(\theta)d\theta \le -C_0 n\epsilon_n^2\right) \to 1, \quad n \to \infty.$$

Lemma 2. Let ϵ_n be the sequence satisfying the Assumption 2, then for any $C > 0$,

$$\mathbb{P}_0^{(n)}\left(\left|\log \int \frac{L(\theta)}{L(\theta_0)} p(\theta)d\theta\right| \le Cn\epsilon_n^2 \log n\right) \to 1.$$

Lemma 3. Let ϵ_n be the sequence satisfying Assumption 2, then for some $Q \in \mathcal{Q}^{\mathrm{MF}}$ and any $C > 0$,

$$\mathbb{P}_0^{(n)}\left(\int \log \frac{L(\theta_0)}{L(\theta)} q(\theta)d\theta \le Cn\epsilon_n^2 \log n\right) \to 1.$$

Lemma 1 and 2 taken together suffice to establish the posterior consistency of the true posterior based on the pseudo-likelihood $L(\theta)$ as in (4). Lemma 3 on the other hand is the additional condition which needs to ensure the consistency of the variational posterior. We next state an important result which relates the variational posterior to the true posterior.

Formula for KL divergence: By Corollary 4.15 in Boucheron, Lugosi, and Massart (2013),

$$\mathrm{KL}(P_1, P_2) = \sup_f \left[\int f dP_1 - \log \int e^f dP_2\right].$$

Using the above formula in the context of variational distributions, we get

$$\int f dQ^* \le \mathrm{KL}(Q^*, \Pi(\mid X^{(n)})) + \log \int e^f d\Pi(\mid X^{(n)}). \quad (10)$$

The above relation serves as an important tool for the proof of Theorem 1. Next, we give a brief sketch of the proof. Further details have been deferred to the supplement.

We use the term with dominating probability to imply that under $\mathbb{P}_0^{(n)}$, the probability of the event goes 1 as $n \to \infty$.

Sketch of proof of Theorem 1: Let $f = (C_0/2)n\varepsilon_n^2 1[\theta \in \mathcal{U}_{\varepsilon_n}^c]$, then

$$\begin{aligned}(C_0/2)n\varepsilon_n^2 Q^*(\mathcal{U}_{\varepsilon_n}^c) &\le \mathrm{KL}(Q^*, \Pi(\mid X^{(n)})) \\ &+ \log(e^{(C_0/2)n\varepsilon_n^2}\Pi(\mathcal{U}_{\varepsilon_n}^c \mid X^{(n)}) + \Pi(\mathcal{U}_{\varepsilon_n} \mid X^{(n)})) \\ \implies Q^*(\mathcal{U}_{\varepsilon_n}^c) &\le \frac{2}{C_0 n\varepsilon_n^2}\mathrm{KL}(Q^*, \Pi(\mid X^{(n)})) \\ &+ \frac{2}{C_0 n\varepsilon_n^2}\log(1 + e^{(C_0/2)n\varepsilon_n^2}\Pi(\mathcal{U}_{\varepsilon_n}^c \mid X^{(n)})).\end{aligned}$$

By Lemmas 2 and 3, it can be established with dominating probability for any $C > 0$,

$$\mathrm{KL}(Q^*, \Pi(\mid X^{(n)})) \le Cn\epsilon_n^2 \log n, \quad n \to \infty.$$

By Lemmas 1 and 2, it can be established with dominating probability, as $n \to \infty$

$$\Pi(\mathcal{U}_{\varepsilon_n}^c \mid X^{(n)}) \le e^{-C_1 n\varepsilon_n^2}, \quad (11)$$

for any $C_1 > C_0/2$. Therefore, with dominating probability

$$\begin{aligned}Q^*(\mathcal{U}_{\varepsilon_n}^c) &\le \frac{2C}{C_0 M_n} + \frac{2}{C_0 n\varepsilon_n^2}\log\left(1 + e^{-(C_1 - C_0/2)n\varepsilon_n^2}\right) \\ &\sim \frac{2C}{C_0 M_n} + \frac{e^{-(C_1 - C_0/2)n\varepsilon_n^2}}{C_0 n\varepsilon_n^2} \to 0. \quad (12)\end{aligned}$$

This completes the proof.

Note that (11) gives the statement for the contraction of the true posterior. Similarly the contraction rate for the variational posterior follows as a consequence of (12). It is important to note that for both the true posterior and the variational posterior, the size of the Hellinger neighborhood which gets close to 1 probability, is the same as ε_n. Thus, both variational and true posterior have the same contraction rates. However, for the variational posterior, the probability of ε_n-Hellinger neighborhood ($Q^*(\mathcal{U}_{\varepsilon_n})$) of the true density function approaches 1 at the rate $1 - 1/M_n$ for $M_n \to \infty$ and for the true posterior, the probability of ε_n-Hellinger neighborhood ($\Pi(\mathcal{U}_{\varepsilon_n}|X^{(n)})$) of the true density function approaches 1 at the rate $1 - \exp(-Cn\varepsilon_n^2)$. This difference in rate is expected and has also been discussed in the seminal works of Zhang and Gao (2020), Yang, Pati, and Bhattacharya (2020), etc.

Note, Theorem 1 gives the contraction rate of the variational posterior. However, the convergence of the of variational Bayes estimator to the true values of β_0 and B_0 is not immediate. The following corollary gives the convergence rate for the variational Bayes estimate as long as the Assumptions 1, 2, and 3 hold.

Corollary 1 (Variational Bayes Estimator Convergence). Let ε_n be as in Theorem 1, then in $\mathbb{P}_0^{(n)}$ probability,

$$\frac{1}{\varepsilon_n}\mathbb{E}_{Q^*}(\|\theta - \theta_0\|_2) \to 0, \text{ as } n \to \infty.$$

Next, we provide a brief sketch of the proof. Further details of the proof have been deferred to supplement.

Sketch of proof of Corollary 1: Let $f = (C_2/2)n\varepsilon_n\|\theta - \theta_0\|_2$, then

$$(C_2/2)n\varepsilon_n \int \|\theta - \theta_0\|_2 dQ^*(\theta) \le \mathrm{KL}(Q^*, \Pi(|\, X^{(n)})) + \log\left(\int e^{C_2 n\varepsilon_n\|\theta-\theta_0\|_2/2} d\Pi(\theta \mid X^{(n)})\right).$$

By Lemmas 2 and 3, it can be established with dominating probability, for any $C > 0$,

$$\mathrm{KL}(Q^*, \Pi(|\, X^{(n)})) \le Cn\epsilon_n^2 \log n.$$

By Lemmas 1, and 2, it can be established with dominating probability, for some $C_2 > 0$

$$\int e^{(C_2/2)n\varepsilon_n\|\theta-\theta_0\|_2} d\Pi(\theta \mid X^{(n)}) \le \frac{1}{(C_2/2)n\varepsilon_n^2} e^{Cn\epsilon_n^2 \log n}. \quad (13)$$

Therefore, with dominating probability

$$\int \|\theta - \theta_0\|_2 dQ^*(\theta) \le \frac{2C\varepsilon_n}{C_2 M_n} - \frac{2\log(C_2/2)}{C_2 n\varepsilon_n} - \frac{2\varepsilon_n \log(n\varepsilon_n^2)}{C_2 n\varepsilon_n^2} + \frac{2C\varepsilon_n}{C_2 M_n} \le \varepsilon_n o(1).$$

This completes the proof. Note, (13) follows as a consequence of convergence of the true posterior. If $\varepsilon_n = n^{-\delta}$, then the rate of convergence of the variational Bayes estimator is n^{δ}. Since $n\epsilon_n^2 = n\varepsilon_n^2/(M_n \log n) \to \infty$, δ can be chosen anywhere between $0 < \delta < 1/2$. If δ can be chosen very close to 1/2, the rate of convergence will be close to $\sqrt{n}$. However, how close δ can be to 1/2 depends on the extent to which Assumption 2 is satisfied by the adjacency matrix A_n. For example, smaller the Frobenius norm of the adjacency matrix (alternatively the number of total edges for 0–1 adjacency matrix), the closer to $\sqrt{n}$ consistency.

4. Simulation Results

In this section, we compare our VB algorithm with two other methods, PMLE (Ghosal and Mukherjee 2020) and MCMC based method (Møller et al. 2006). We briefly describe the two methods before providing performance comparison.

PMLE: Let $h(\beta, B)$ denote the pseudo-likelihood in (2). Ghosal and Mukherjee (2020) used grid search to find PMLE for Ising parameters which satisfies $\frac{\partial}{\partial\beta}\log h(\beta, B) = 0$ and $\frac{\partial}{\partial B}\log h(\beta, B) = 0$. We create a grid search space so that β contains all values from 0.01 to 2 in increments of 0.01 and the search space for B increases from -1 to 1 by 0.01.

MCMC: Møller et al. (2006) suggested efficient MCMC method employing an auxiliary variable z to deal with an unknown normalizing constant in Ising model. For $\theta = (\beta, B)$, let $g_\theta(x) = \exp\left((\beta/2)x^\top A_n x + B\sum_{i=1}^n x_i\right)$ denote unnormalized density of Ising model in (1). With the initial guess of θ denoted by $\tilde{\theta}$, the *Metropolis-Hastings* ratio is

$$\mathrm{MH}(\theta', z' \mid \theta, z) = \frac{g_{\tilde{\theta}}(z')g_{\theta'}(x_{\text{observed}})g_\theta(z)}{g_{\tilde{\theta}}(z)g_\theta(x_{\text{observed}})g_{\theta'}(z')}, \quad (14)$$

where (θ, z) is the current state and x_{observed} is the observed data. The PMLE is used as $\tilde{\theta}$ and we accept (θ', z') as the next state with probability $\max\left\{1, \mathrm{MH}(\theta', z' \mid \theta, z)\right\}$.

4.1. Performance Comparison

We generate irregular graphs and coupling matrices A_n as in Example 1 with $n \in \{100, 500\}$, $H_n = n^{0.3}$, $\epsilon_n = n^{-0.1}$, and $k_n = n/2$. We compare the performance of the parameter estimation methods for the Ising model in (1) under various combinations of (β_0, B_0). For a given A_n, under each scenario, we repeat the estimation procedures $R = 50$ times. We use $S = 20$ or $S = 200$

Table 1. Mean squared errors and computation times for each pair of (β_0, B_0) when $(n, d_n) = (100, 20)$ (left numbers) and $(n, d_n) = (500, 44)$ (right numbers).

Method	Monte Carlo samples (S)	(0.2, 0.2)	(0.2, −0.2)	Convergence time (sec)
PMLE[1]	–	0.061 / 0.023	0.100 / 0.016	3.2 / 3.5
MCMC[2]	–	0.021 / 0.010	0.027 / 0.006	157.0 / 575.1
MF family[3]	20	0.051 / 0.016	0.056 / 0.011	6.4 / 9.8
	200	0.046 / 0.009	0.052 / 0.008	10.2 / 17.9
BN family[4]	20	0.052 / 0.019	0.062 / 0.014	7.9 / 11.7
	200	0.049 / 0.011	0.056 / 0.009	12.1 / 19.0

[1]PMLE, pseudo maximum likelihood estimate (Ghosal and Mukherjee 2020); [2]MCMC, Markov chain Monte Carlo (Møller et al. 2006); [3]MF, mean-field; [4]BN, bivariate normal.

Table 2. Mean squared errors and computation times for each pair of (β_0, B_0) when $(n, d_n) = (100, 20)$ (left numbers) and $(n, d_n) = (500, 44)$ (right numbers).

Method	Monte Carlo samples (S)	(0.2, 0.5)	(0.2, −0.5)	Convergence time (sec)
PMLE	–	0.060 / 0.011	0.046 / 0.011	3.2 / 3.5
MCMC	–	0.045 / 0.008	0.035 / 0.008	157.3 / 576.1
MF family	20	0.057 / 0.009	0.049 / 0.008	6.5 / 9.9
	200	0.055 / 0.005	0.046 / 0.006	10.0 / 17.3
BN family	20	0.061 / 0.010	0.054 / 0.009	8.0 / 11.5
	200	0.057 / 0.006	0.049 / 0.006	12.3 / 19.0

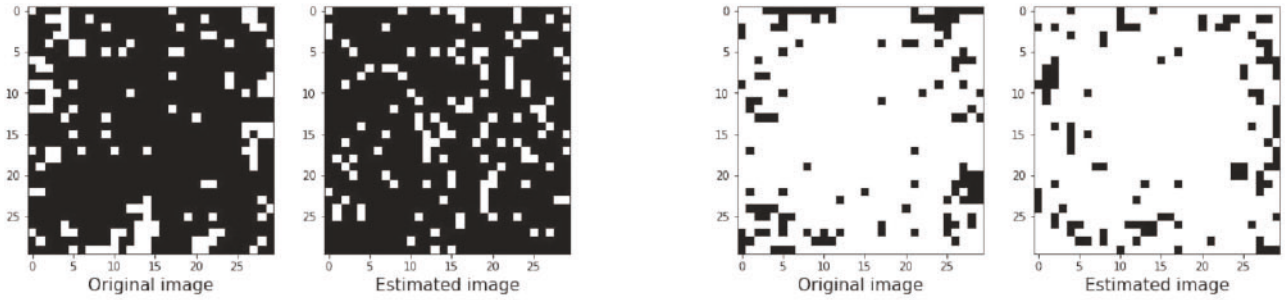

Figure 1. The first two images from left are original (first) and estimated (second), respectively for $(\beta_0, B_0) = (1.2, 0.2)$. The third and fourth images are for $(\beta_0, B_0) = (1.2, -0.2)$.

Table 3. Mean squared errors and computation times for each pair of (β_0, B_0) when $(n, d_n) = (100, 20)$ (left numbers) and $(n, d_n) = (500, 44)$ (right numbers).

Method	Monte Carlo samples (S)	$(0.7, 0.2)$	$(0.7, -0.2)$	Convergence time (sec)
PMLE	–	0.117 / 0.010	0.217 / 0.012	3.1 / 3.5
MCMC	–	1.555 / 0.010	13.120 / 0.014	158.0 / 575.9
MF family	20	0.065 / 0.018	0.066 / 0.018	6.6 / 9.8
	200	0.066 / 0.017	0.067 / 0.017	10.2 / 17.1
BN family	20	0.064 / 0.019	0.062 / 0.019	8.1 / 11.3
	200	0.063 / 0.017	0.063 / 0.018	12.1 / 18.9

Table 4. Mean squared errors and computation times for each pair of (β_0, B_0) when $(n, d_n) = (100, 20)$ (left numbers) and $(n, d_n) = (500, 44)$ (right numbers).

Method	Monte Carlo samples (S)	$(0.7, 0.5)$	$(0.7, -0.5)$	Convergence time (sec)
PMLE	–	0.706 / 0.020	0.709 / 0.016	3.2 / 3.4
MCMC	–	31.089 / 0.035	25.403 / 0.019	157.8 / 580.1
MF family	20	0.067 / 0.028	0.055 / 0.029	6.8 / 9.7
	200	0.074 / 0.024	0.061 / 0.026	10.1 / 16.9
BN family	20	0.067 / 0.026	0.055 / 0.028	8.0 / 11.2
	200	0.071 / 0.024	0.058 / 0.028	12.3 / 19.0

as the Monte Carlo sample size in (9). For fair comparison, we set a fixed learning rate $\rho_t = 0.00002$ and number of iterations (5000) for each scenario although the optimal choice may vary depending on the variational family and (n, S). We report ELBO convergences in Figure 1 of supplement Section A.3. We compute mean squared errors (MSE), $(1/R)\sum_{r=1}^{R}((\hat{\beta}_r - \beta_0)^2 + (\hat{B}_r - B_0)^2)$, for assessing the performances based on R pairs of estimates, $(\hat{\beta}_r, \hat{B}_r)_{r=1,\ldots,R}$. The two numbers in each cell of Tables 1–4, represent MSE and convergence times for $n = 100$ and $n = 500$, respectively.

First, we consider a small value of $\beta_0 = 0.2$ with $B_0 = \pm 0.2, \pm 0.5$. In these cases PMLE is the fastest but less accurate (see Tables 1 and 2). MCMC achieves smaller MSEs but it has the highest runtimes. Our VB methods notably reduce the runtimes without compromising accuracy. Second, the results for higher interaction parameter $\beta_0 = 0.7$ with $B_0 = \pm 0.2, \pm 0.5$ are shown in Tables 3 and 4. The numerical studies validate the superiority of our VB algorithm. For our VB approach, the sensitivity to number of Monte Carlo samples S is not too high. However, smaller values of S can reduce the computation time.

In all the experiments, $\mathcal{Q}^{\mathrm{BN}}$ does not always outperform $\mathcal{Q}^{\mathrm{MF}}$. This may primarily happen because $\mathcal{Q}^{\mathrm{BN}}$ which has more parameters to be estimated does not perform exceedingly well unless the posterior samples of β and B have high covariance. Since across almost all parameter combinations, we observe small covariance values (see Table 1 of Supplement Section A.3), one would only expect minor differences between $\mathcal{Q}^{\mathrm{MF}}$ and $\mathcal{Q}^{\mathrm{BN}}$.

In addition to the numerical experiments in this section, we also performed extra experiments with d_n-regular graphs, although regular graphs do not satisfy the assumptions. The results are provided in Supplement Section A.4. For additional experiments, we used our algorithm to regenerate an image in the next section.

Table 5. Means and standard deviations of F_1 scores for reconstructing the original image 50 times using variational Bayes (VB) algorithm and independent Bernoulli (IB) method.

	$(1.2, 0.2)$	$(1.2, -0.2)$	$(1.2, 0.5)$	$(1.2, -0.5)$
VB	**0.864 ± 0.007**	**0.877 ± 0.009**	0.938 ± 0.004	**0.943 ± 0.010**
IB	0.854 ± 0.009	0.866 ± 0.008	**0.941 ± 0.005**	0.937 ± 0.005
	$(0.7, 0.2)$	$(0.7, -0.2)$	$(0.7, 0.5)$	$(0.7, -0.5)$
VB	0.754 ± 0.011	**0.737 ± 0.015**	**0.882 ± 0.008**	**0.898 ± 0.006**
IB	**0.762 ± 0.010**	0.721 ± 0.012	0.881 ± 0.008	0.898 ± 0.007
	$(0.2, 0.2)$	$(0.2, -0.2)$	$(0.2, 0.5)$	$(0.2, -0.5)$
VB	0.601 ± 0.021	**0.643 ± 0.015**	0.777 ± 0.010	**0.783 ± 0.009**
IB	**0.618 ± 0.015**	0.635 ± 0.014	**0.786 ± 0.009**	0.772 ± 0.010

Larger mean F_1 value is highlighted in bold and it indicates better performance.

4.2. Image Reconstruction

Ising model can be used for constructing an image in computer vision field. In particular, the Bayesian procedure facilitate the reconstruction easily by using the posterior predictive distribution Halim (2007). Consider an image in which each pixel represents either -1(white) or 1(black). For choice of the underlying graph, we generated a two-dimensional grid graph of size 30×30 and added diagonal edges between 400 nodes at the center to make the graph more irregular. Using the scaling of Definition 1 for the coupling matrix A_n and $\epsilon_n = n^{-0.01}$, we get $\max_{i\in[n]}\sum_{j=1}^{n} A_n(i,j) = 1.462$, $(1/(n\epsilon_n^2))\sum_{i=1}^{n}\sum_{j=1}^{n} A_n(i,j)^2 = 0.209$ and $(1/n)\sum_{i=1}^{n}(\sum_{j=1}^{n} A_n(i,j) - \bar{A}_n)^2 = 0.136$. Here, $\epsilon_n = 0.01$ just presents one choice of the sequence ϵ_n which allows Assumption 2 to hold. Nonetheless, one could also choose $\epsilon_n = n^{-\delta}$ such that $0 < \delta < 1/2$ and Assumption 2 holds.

We generate the images using Supplement Section A.1 for a true (β_0, B_0) and use it as our given data $\boldsymbol{x}_{observed}$. With $\boldsymbol{x}_{observed}$ and coupling matrix A_n, we obtain $(\hat{\beta}, \hat{B})$ after implementing the parameter estimation procedure based on the BN family. The estimates $(\hat{\beta}, \hat{B})$ are used for data regeneration using Supplement

Section A.1 again. In Figure 1, we plot two original images and corresponding estimated images for $(\beta_0, B_0) = (1.2, 0.2)$ and $(\beta_0, B_0) = (1.2, -0.2)$ as examples. To quantify the performance of the VB in image reconstruction, we computed F_1-scores (Baddeley 1992). As a baseline, we consider a naive Bernoulli method for regenerating images where pixels are independently black and white with same probability as in the original image. Table 5 shows the means and standard deviations of F_1-scores for variational Bayes (labeled as VB) and independent Bernoulli (labeled as IB) based on 50 replications. As is evident from the table, VB is at par if not better than IB for majority of the cases. Although for this simple experiment of image reconstruction, the improvement provided by VB over IB is only marginal, VB allows for inference on the interaction parameter β and the threshold parameter B. The estimates of β and B allow one to interpret the nature and strength of connectivity among the pixels of the images, something which cannot be obtained by applying a naive technique like IB.

5. Real Data Analysis

In two-parameter Ising model, higher value of β implies stronger interactions between connected nodes and the threshold parameter B controls the model size (number of 1s), where the model size is greater for $B > 0$, smaller for $B < 0$. Here, we apply our methods to a real dataset related to network analysis.

5.1. Data Description

Stanford Network Analysis Project (SNAP) provides a Facebook network dataset (Leskovec and Krevl 2014) available at *http://snap.stanford.edu/data/ego-Facebook.html*.The Facebook network consists of 4039 nodes and 88,234 edges. Each node represents a Facebook user and there is an edge between two nodes if corresponding users are friends. The dataset also contains user features such as birthday, school, gender, and location. The features are fully anonymized. For instance, while the original data may include a feature "location = Michigan," the anonymized data would simply contain "location = anonymized location A." Thus, using the anonymized data, we can determine whether two users stay in the same location, but we do not know where.

Among the 4039 users, we select only users who disclose gender information to create a sub-graph such that there are 3948 nodes and 84,716 edges in the sub-graph. Later we use the gender information for a binary vector observation. Figure 2 shows all the nodes and edges in the Facebook network we used. Each node has different number of neighbors, indicating an irregular graph. The maximum degree of the sub-graph is 1024, and the minimum is 1, with an average degree 42.92. For the matrix A_n, we use scaling of Definition 1 and compute the values with $n = 3948$ and $\epsilon_n = n^{-0.01}$ to check the assumptions in Section 3. We did not find any strong evidence of violation of any of the assumptions.

5.2. Parameter Estimation

We use the selected users ($n = 3948$) as a real dataset with the gender feature as an observed binary vector but the original dataset is anonymized. Specifically, the gender information of each user is "Gender A" or "Gender B" in the original dataset. We encode "Gender A" by 1 and "Gender B" by -1. We do not know which group represents male (or female) but we note that the resulting model size (number of 1s) is 2417.

Since the model size is large, one can expect that B will be positive. The estimates from all the four methods are positive indicating that the estimation is in right direction. All four estimates of the interaction parameter β are less than or equal to 0.25 which suggests that the gender effect of being Facebook friends is not so strong. To analyze features with more than two categories, one could use a Potts model (see Supplement Section C.2).

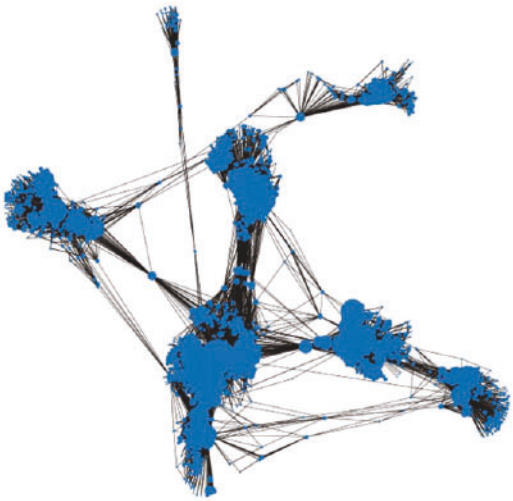

Figure 2. Visualization of Facebook network data (circle sizes denote degrees of the nodes).

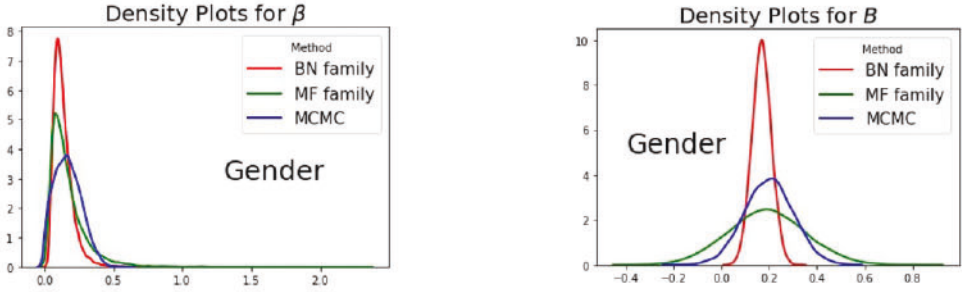

Figure 3. Density plots for the estimated parameters (left: β, right: B) from VB with BN family (red), VB with MF family (green), and MCMC (blue) for the gender feature.

Table 6. The estimated parameters with standard errors and time costs for gender feature.

Method	Monte Carlo samples (S)	$\hat{\beta}$ (SE)	$\hat{B}$ (SE)	Convergence time (sec)
PMLE	–	0.250(—)	0.180(—)	5.9
MCMC	–	0.171(0.097)	0.198(0.102)	27109.2
MF family	200	0.169(0.133)	0.189(0.162)	505.1
BN family	200	0.135(0.068)	0.168(0.040)	515.3

Table 6 summarizes the estimated parameters with standard errors (SE) (for VB and MCMC) and runtimes. SEs of MCMC in Table 6 are calculated based on 10,000 draws after the burn-in period of 10,000 iterations. For our VB methods, to calculate SEs, we draw 10,000 samples of β and B from the optimal variational distributions and calculate sample standard deviations (see the density plots in Figure 3). Table 6 suggests that while estimated parameters are comparable for all the methods, the MCMC implementation takes about 50 times more compared to VB to achieve similar level of accuracy. The PMLE approach does not produce SE and thus limited for statistical inference.

6. Discussion and Conclusion

In this article, we propose a variational Bayes estimation technique for a two-parameter Ising model. The use of pseudo-likelihood avoids the computation of normalizing constants and VB facilitates computation speed. We have mainly worked with mean-field assumption on the adjacency matrix. We concentrated only on irregular graphs to establish theoretical consistency of VB. The promising empirical performance of regular graphs under mean field assumption motivates us to explore them in the future. Finally, exploring the case where mean field assumption breaks is a compelling direction for future research. In Supplement Section C, we discuss possible extensions to multiparameter Ising models and Potts models.

Supplementary Materials

The supplementary file contains implementation details and theoretical details.

Acknowledgments

We are thankful to the Associate Editor and the Reviewers for their comments which helped improve the manuscript significantly.

Disclosure Statement

There is no conflict of interest to report.

Funding

The authors are grateful to P. Ghosal and S. Mukherjee for kindly sharing codes of their work. The research is partially supported by the National Science Foundation grants NSF DMS-1952856 and 1924724.

ORCID

Minwoo Kim http://orcid.org/0000-0003-4240-9878

References

Anandkumar, A., Tan, V. Y., Huang, F., and Willsky, A. S. (2012), "High-Dimensional Structure Estimation in Ising Models: Local Separation Criterion," *The Annals of Statistics*, 40, 1346–1375. [75]

Baddeley, A. J. (1992), "An Error Metric for Binary Images," *Robust Computer Vision*, 5978. [82]

Basak, A., and Mukherjee, S. (2017), "Universality of the Mean-Field for the Potts Model," *Probability Theory and Related Fields*, 168, 557–600. [78]

Bhattacharya, B. B., and Mukherjee, S. (2018), "Inference in Ising Models," *Bernoulli*, 24, 493–525. [75,76]

Boucheron, S., Lugosi, G., and Massart, P. (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford: Oxford University Press. [79]

Bresler, G. (2015), "Efficiently Learning Ising Models on Arbitrary Graphs," in *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*. [75]

Brush, S. G. (1967), "History of the Lenz-Ising Model," *Reviews of Modern Physics*, 39, 883. [75]

Chatterjee, S. (2007), "Estimation in Spin Glasses: A First Step," *The Annals of Statistics*, 35, 1931–1946. [75]

Comets, F. (1992), "On Consistency of a Class of Estimators for Exponential Families of Markov Random Fields on the Lattice," *The Annals of Statistics*, 20, 455–468. [75]

Comets, F., and Gidas, B. (1991), "Asymptotics of Maximum Likelihood Estimators for the Curie-Weiss Model," *The Annals of Statistics*, 19, 557–578. [75]

Fang, Z., and Kim, I. (2016), "Bayesian Ising Graphical Model for Variable Selection," *Journal of Computational and Graphical Statistics*, 25, 589–605. [75]

Ghosal, P., and Mukherjee, S. (2020), "Joint Estimation of Parameters in Ising Model," *The Annals of Statistics*, 48, 785–810. [75,76,78,80]

Gidas, B. (1988), "Consistency of Maximum Likelihood and Pseudo-likelihood Estimators for Gibbs Distributions," in *Stochastic Differential Systems, Stochastic Control Theory and Applications*, eds. W. Fleming, and P.-L. Lions, pp. 129–145, New York: Springer. [75]

Guyon, X., and Künsch, H. R. (1992), "Asymptotic Comparison of Estimators in the Ising Model," in *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis*, eds. P. Barone, A. Frigessi, M. Piccioni, pp. 177–198, New York: Springer. [75]

Halim, S. (2007), "Modified Ising Model for Generating Binary Images," *Jurnal Informatika*, 8, 115–118. [81]

Haslbeck, J. M., Epskamp, S., Marsman, M., and Waldorp, L. J. (2021), "Interpreting the Ising Model: The Input Matters," *Multivariate Behavioral Research*, 56, 303–313. [75]

Ising, E. (1924), "Beitrag zur theorie des ferro-und paramagnetismus," Ph.D. thesis, Grefe & Tiedemann. [75]

Lahtinen, V., and Pachos, J. (2017), "A Short Introduction to Topological Quantum Computation," *SciPost Physics*, 3, 021. [75]

Lee, K.-J., Jones, G. L., Caffo, B. S., and Bassett, S. S. (2014), "Spatial Bayesian Variable Selection Models on Functional Magnetic Resonance Imaging Time-Series Data," *Bayesian Analysis*, 9, 699–732. [75]

Leskovec, J., and Krevl, A. (2014), "SNAP Datasets: Stanford Large Network Dataset Collection," available at *http://snap.stanford.edu/data*. [82]

Li, F., and Zhang, N. R. (2010), "Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces with Applications in Genomics," *Journal of the American Statistical Association*, 105, 1202–1214. [75]

Li, F., Zhang, T., Wang, Q., Gonzalez, M. Z., Maresh, E. L., and Coan, J. A. (2015), "Spatial Bayesian Variable Selection and Grouping for High-Dimensional Scalar-on-Image Regression," *The Annals of Applied Statistics*, 9, 687–713. [75]

Li, W., Huang, J., Li, X., Zhao, S., Lu, J., Han, Z. V., and Wang, H. (2021), "Recent Progresses in Two-Dimensional Ising Superconductivity," *Materials Today Physics*, 21, 100504. [75]

Lipowski, A., Lipowska, D., and Ferreira, A. L. (2017), "Phase Transition and Power-Law Coarsening in an Ising-Doped Voter Model," *Physical Review E*, 96, 032145. [75]

Lokhov, A. Y., Vuffray, M., Misra, S., and Chertkov, M. (2018), "Optimal Structure and Parameter Learning of Ising Models," *Science Advances*, 4, e1700791. [75]

Majewski, J., Li, H., and Ott, J. (2001), "The Ising Model in Physics and Statistical Genetics," *The American Journal of Human Genetics*, 69, 853–862. [75]

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006), "An Efficient Markov Chain Monte Carlo Method for Distributions with Intractable Normalising Constants," *Biometrika*, 93, 451–458. [76,80]

Okabayashi, S., Johnson, L., and Geyer, C. J. (2011), "Extending Pseudo-Likelihood for Potts Models," *Statistica Sinica*, 21, 331–347. [75,76]

Park, J., Jin, I. H., and Schweinberger, M. (2022), "Bayesian Model Selection for High-Dimensional Ising Models, with Applications to Educational Data," *Computational Statistics & Data Analysis*, 165, 107325. [75]

Ranganath, R., Gerrish, S., and Blei, D. (2014), "Black Box Variational Inference," in *Artificial Intelligence and Statistics*, PMLR. [76,77]

Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010), "High-Dimensional Ising Model Selection Using ℓ_1-regularized Logistic Regression," *The Annals of Statistics*, 38, 1287–1319. [75]

Robbins, H., and Monro, S. (1951), "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, 22, 400–407. [78]

Smith, M., and Fahrmeir, L. (2007), "Spatial Bayesian Variable Selection with Application to Functional Magnetic Resonance Imaging," *Journal of the American Statistical Association*, 102, 417–431. [75]

Xue, L., Zou, H., and Cai, T. (2012), "Nonconcave Penalized Composite Conditional Likelihood Estimation of Sparse Ising Models," *The Annals of Statistics*, 40, 1403–1429. [75]

Yang, Y., Pati, D., and Bhattacharya, A. (2020), "α-variational Inference with Statistical Guarantees," *Annals of Statistics*, 48, 886–905. [80]

Zhang, F., and Gao, C. (2020), "Convergence Rates of Variational Posterior Distributions," *Annals of Statistics*, 48, 2180–2207. [80]