Parameter Estimation for Ising Model with Variational Bayes

Minwoo Kim[1]

Michigan State University

June 2 2021,
Symposium on Data Science and Statistics

[1] Joint work with Tapabrata Maiti and Shrijita Bhattacharya

Introduction

Ising Model

Consider a sequence of binary random variables

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^\top, \quad X_i \in \{-1, 1\} \text{ for } i = 1, \ldots, n$$

If the binary random variables are dependent, which statistical model is appropriate? Ising model.

- The Ising model, named after the physicist Ernst Ising, is a mathematical model of ferromagnetism in statistical mechanics.
- The model consists of binary random variables (spins).
- Each spin interacts with its neighbors.

Ising Model

Let A_n is a known coupling matrix which represents the dependency structure of $\boldsymbol{X}$.

$$A_n = \begin{pmatrix} 0 & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & 0 & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & 0 \end{pmatrix}$$

$$a_{i,j} \begin{cases} = 0, & \text{if } x_i \text{ and } x_j \text{ are independent} \\ > 0, & \text{if } x_i \text{ and } x_j \text{ are dependent} \end{cases}$$

Common choice of A_n is an adjacency matrix of a graph. In our study, we assume A_n is known.

Ising Model

Then, the likelihood of Ising model is:

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

where $\beta > 0$ and $B \neq 0$.

- β tells us how strongly the dependent variables are interacted.
- B represents overall tendency of the variables,
  - If $B > 0$, x_i's tend to be $+1$,
  - If $B < 0$, x_i's tend to be -1.
- $Z_n(\beta, B)$ is a normalizing constant such that

$$1 = \frac{1}{Z_n(\beta, B)} \sum_{\boldsymbol{x} \in \{-1,1\}^n} \exp\left(\frac{\beta}{2}\boldsymbol{x}^\top A_n \boldsymbol{x} + B\sum_{i=1}^{n} x_i\right)$$

Pseudo-likelihood

Accurate calculation of the normalizing constant involves the sum of 2^n terms, which require high computation costs:

$$Z_n(\beta, B) = \sum_{\boldsymbol{x} \in \{-1,1\}^n} \exp\left(\frac{\beta}{2}\boldsymbol{x}^\top A_n \boldsymbol{x} + B\sum_{i=1}^{n} x_i\right)$$

One can easily calculate the conditional probability of X_i given others to remove the normalizing constant:

$$P(X_i = x_i \mid X_j, j \neq i) = \frac{\exp\left(\beta x_i m_i(\boldsymbol{x}) + B x_i\right)}{\exp\left(\beta m_i(\boldsymbol{x}) + B\right) + \exp\left(-\beta m_i(\boldsymbol{x}) - B\right)}$$

where $m_i(\boldsymbol{x}) = [A_n \boldsymbol{x}]_i = \sum_{j=1}^{n} a_{i,j} \cdot x_j$

Pseudo-likelihood

In this regard, we consider a pseudo-likelihood as the product of the conditional probabilities as follows:

$$\tilde{P}_{\beta,B}(\boldsymbol{X} = \boldsymbol{x}) = \prod_{i=1}^{n} P(X_i = x_i \mid X_j, j \neq i)$$
$$= 2^{-n} \exp\left(\sum_{i=1}^{n} \left[x_i v_i(\beta, B) - \log\cosh\left(v_i(\beta, B)\right)\right]\right) \tag{2}$$

where $v_i(\beta, B) = \beta m_i(\boldsymbol{x}) + B$. We will replace the true likelihood (1) with the pseudo-likelihood (2) in our estimation procedure.

Bayesian Methodology

Prior Distribution

We want to estimate the parameters (β, B) in Bayesian framework. $\theta := (\beta, B)$ is viewed as a random vector with prior distribution:

$$p(\theta) = p_\beta(\beta) p_B(B)$$

We choose normal prior for B.

$$p_B(B) = \frac{1}{\sqrt{2\pi}} e^{-\frac{B^2}{2}}$$

Prior Distribution

We choose log-normal prior for β because it should be positive.

$$p_\beta(\beta) = \frac{1}{\beta\sqrt{2\pi}} e^{-\frac{(\log\beta)^2}{2}}$$

Therefore,

$$p(\theta) = \frac{1}{\beta\sqrt{2\pi}} e^{-\frac{(\log\beta)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{B^2}{2}}$$

Posterior Distribution

With the pseudo-likelihood (2) and prior distributions, we define a (pseudo) posterior as follows:

$$\tilde{\pi}(\theta \mid \boldsymbol{x}) = \frac{\tilde{p}(\theta, \boldsymbol{x})}{\tilde{p}(\boldsymbol{x})} = \frac{\tilde{p}(\theta, \boldsymbol{x})}{\int \tilde{p}(\theta, \boldsymbol{x}) d\theta} \tag{3}$$

where $\tilde{p}(\theta, \boldsymbol{x}) = p(\theta)\tilde{P}_\theta(\boldsymbol{X} = \boldsymbol{x})$. The integral in denominator is intractable.

Variational Bayes

Instead, we use Variational Bayes (VB) to approximate the posterior (3).

1. Choose a family of distributions $\mathcal{Q}$, so-called variational family.
2. Find the optimal variational distribution $q^* \in \mathcal{Q}$ which is closest to $\tilde{\pi}(\theta \mid \boldsymbol{x})$ in terms of Kullback-Leibler (KL) divergence.

$$q^* = \underset{q \in \mathcal{Q}}{\arg\min}\, d_{KL}\left(q, \tilde{\pi}(\theta \mid \boldsymbol{x})\right)$$

Variational Bayes

Note that

$$d_{KL}\left(q, \tilde{\pi}(\theta \mid \boldsymbol{x})\right) = \mathbb{E}_q\left[\log q(\theta) - \log \tilde{p}(\theta, \boldsymbol{x})\right] - \log \tilde{p}(\boldsymbol{x}). \qquad (4)$$

The first term in (4) is negative Evidence Lower BOund (ELBO). We want to find optimal q^* among predetermined $\mathcal{Q}$ which minimizes the negative ELBO:

$$\begin{aligned} q^* &= \underset{q \in \mathcal{Q}}{\arg\min}\ d_{KL}\left(q, \tilde{\pi}(\theta \mid \boldsymbol{x})\right) \\ &= \underset{q \in \mathcal{Q}}{\arg\min}\ \underbrace{\mathbb{E}_q\left[\log q(\theta) - \log \tilde{p}(\theta, \boldsymbol{x})\right]}_{\text{negative ELBO}} \end{aligned}$$

Variational Family

One candidate of our variational family is mean-field family as follows:

$$\mathcal{Q}^{MF} = \{q(\theta) : q(\theta) = q_\beta(\beta) q_B(B)\}$$

where

$$q_\beta(\beta) = \frac{1}{\beta \sigma_\beta \sqrt{2\pi}} e^{-\frac{(\log \beta - \mu_\beta)^2}{2\sigma_\beta^2}},$$

$$q_B(B) = \frac{1}{\sigma_B \sqrt{2\pi}} e^{-\frac{(B - \mu_B)^2}{2\sigma_B^2}}.$$

Note that $q(\theta) \in \mathcal{Q}^{MF}$ is characterized by four variational parameters $(\mu_\beta, \sigma_\beta, \mu_B, \sigma_B)$.

Variational Family

Consider log transformation of β such that $\boldsymbol{z} := (z_1, z_2)^\top = (\log \beta, B)^\top$. Then, another option of variational family is bivariate normal (BN) family with $\boldsymbol{z}$:

$$\mathcal{Q}^{BN} = \{q(\boldsymbol{z}) : q(\boldsymbol{z}) = (2\pi)^{-1} det(\Sigma)^{-1/2} e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu})}\}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ and $\Sigma = \left(\begin{smallmatrix} \sigma_1 & \sigma_{12} \\ \sigma_{12} & \sigma_2 \end{smallmatrix}\right)$. Note that $q(\boldsymbol{z}) \in \mathcal{Q}^{BN}$ is characterized by five variational parameters $(\mu_1, \sigma_1, \mu_2, \sigma_2, \sigma_{12})$.

Stochastic Gradient Method

Let ω denote the set of variational parameters. As a function of ω, we want to optimize the negative ELBO denoted by $\mathcal{L}(\omega)$. In other words, we want to find ω^* such tat

$$
\begin{aligned}
q(\theta; \omega^*) &= \arg\min_{q \in \mathcal{Q}} \mathbb{E}_q \left[\log q(\theta; \omega) - \log p(\theta, \boldsymbol{x}) \right] \\
&= \arg\min_{q \in \mathcal{Q}} \mathcal{L}(\omega)
\end{aligned}
$$

We iteratively update ω until the negative ELBO converges as follows :

$$
\omega^{(t+1)} \leftarrow \omega^{(t)} - \rho_t \nabla_\omega \mathcal{L}
$$

where $\nabla_\omega \mathcal{L}$ is the gradient of $\mathcal{L}(\omega)$ and ρ_t is learning rate.

Stochastic Gradient Method

From Ranganath et al. [2014], the gradient $\nabla_\omega \mathcal{L}$ is:

$$\begin{aligned}\nabla_\omega \mathcal{L} &= \mathbb{E}_q \left[\nabla_\omega \log q(\theta;\omega) \left(\log q(\theta;\omega) - \log p(\theta, \boldsymbol{x})\right)\right] \\ &\simeq \frac{1}{S}\sum_{s=1}^{S} \nabla_\omega \log q(\theta_s;\omega) \left(\log q(\theta_s;\omega) - \log p(\theta_s, \boldsymbol{x})\right) := \widehat{\nabla_\omega \mathcal{L}}\end{aligned}$$

where $\theta_s \sim q(\theta;\omega^{(t)})$. We use $\widehat{\nabla_\omega \mathcal{L}}$ in substitute for $\nabla_\omega \mathcal{L}$ when updating ω:

$$\omega^{(t+1)} \leftarrow \omega^{(t)} - \rho_t \widehat{\nabla_\omega \mathcal{L}}$$

Summary of Algorithm

- Parameters of interest: $\theta = (\beta, B)$
- Given data: $\boldsymbol{x} \sim P_{\beta,B}(\boldsymbol{x})$
- Assumption: A_n is known
- Input:
  - Variational family: $\mathcal{Q}^{MF}$ or $\mathcal{Q}^{BN}$
  - Corresponding initial variational parameters: $\omega^{(0)}$
- Output: optimal variational parameters ω^* obtained by

$$\omega^{(t+1)} \leftarrow \omega^{(t)} - \rho_t \widehat{\nabla_\omega \mathcal{L}}, \quad t = 1, 2, \ldots$$

- Estimate:

$$\hat{\beta} = \frac{1}{M} \sum_{m=1}^{M} \beta_m, \quad \theta_m = (\beta_m, B_m) \sim q(\theta; \omega^*)$$

$$\hat{B} = \frac{1}{M} \sum_{m=1}^{M} B_m$$

Simulation Study

Simulation

To assess applicability of the proposed method, we perform numerical studies as follows:

1. First, we determined the dependency structure of $\boldsymbol{x}$ using an adjacency matrix of a d-regular graph as the known coupling matrix A_n.
2. Then, we generated $\boldsymbol{x}$ from true likelihood (1) with true parameters $\theta_0 = (\beta_0, B_0)$.
3. Given $\boldsymbol{x}$, we implemented our algorithm to get $\hat{\theta} = (\hat{\beta}, \hat{B})$.
4. We repeated the steps 2 and 3 50 times. Then, we have:

$$\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{50}.$$

Simulation

- The measurement of performances is mean squared error (MSE):

$$
\begin{aligned}
MSE(\hat{\theta}) &= \frac{1}{50}\sum_{r=1}^{50}(\hat{\theta}_r - \theta_0)^2 \\
&= \frac{1}{50}\sum_{r=1}^{50}\left((\hat{\beta}_r - \theta_0)^2 + (\hat{B}_r - B_0)^2\right)
\end{aligned}
$$

- We compare our method and Pseduo-MLE in Ghosal et al. [2020]

Performance Comparison

Simulation setup:

1. $(\beta_0, B_0) = (0.2, \pm 0.2)$ and $n = 500$

Degree of graph (d)	Method	Sample size (S)	MSE	Convergence time (sec)
10	PMLE	-	0.051 / 0.022	3.3
	MF family	200	0.060 / 0.031	60.0
		2000	0.052 / 0.021	245.9
	BN family	200	0.047 / 0.019	68.0
		2000	0.045 / 0.016	251.2
50	PMLE	-	0.101 / 0.163	3.6
	MF family	200	0.079 / 0.161	60.2
		2000	0.072 / 0.107	246.6
	BN family	200	0.065 / 0.148	68.1
		2000	0.090 / 0.143	250.8

Performance Comparison

2. $(\beta_0, B_0) = (0.7, \pm 0.5)$ and $n = 500$

Degree of graph (d)	Method	Sample size (S)	MSE	Convergence time (sec)
10	PMLE	-	0.232 / 0.261	3.3
	MF family	200	0.150 /0.146	60.2
		2000	0.151 / 0.132	246.2
	BN family	200	0.144 / 0.136	68.1
		2000	0.140 / 0.133	250.5
50	PMLE	-	0.765 / 1.216	3.5
	MF family	200	0.162 / 0.254	61.0
		2000	0.197 / 0.157	246.2
	BN family	200	0.138 / 0.194	68.0
		2000	0.107 / 0.135	249.9

Performance Comparison

3. $(\beta_0, B_0) = (1.2, \pm 0.5)$ and $n = 500$

Degree of graph (d)	Method	Sample size (S)	MSE	Convergence time (sec)
10	PMLE	-	1.483 / 1.598	3.2
	MF family	200	0.627 / 0.737	60.4
		2000	0.700 / 0.814	246.1
	BN family	200	0.488 / 0.479	67.8
		2000	0.411 / 0.499	250.2
50	PMLE	-	3.190 / 3.526	3.3
	MF family	200	0.836 / 0.792	60.5
		2000	0.972 / 0.947	245.5
	BN family	200	0.336 / 0.294	68.1
		2000	0.272 / 0.208	250.2

Future Work

Theorem (Posterior Consistency)

Consider a neighborhood $\mathcal{U}_\varepsilon = \{|\beta - \beta_0| < \varepsilon, |B - B_0| < \varepsilon\}$*. Let* q^* *is the optimal variational distribution obtained by the VB algorithm with mean-field family. Then,*

$$q^*(\mathcal{U}_\varepsilon^c) \to 0, \quad \varepsilon > 0$$

References

Promit Ghosal, Sumit Mukherjee, et al. Joint estimation of parameters in ising model. *Annals of Statistics*, 48(2): 785–810, 2020.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.