Ising Model Selection: A Bayesian Perspective

Michael Price
Department of Statistics
Rice University
Houston, TX 77005
jmp12@rice.edu

Kyle Manning
Department of Statistics
Rice University
Houston, TX 77005
ksm9@rice.edu

Abstract

In this work, we explore applications of Bayesian inference to the Ising model. In particular, a meaningful method of summarizing information contained in the posterior is sought. Commonly, simple statistics like the posterior mean or median are used, but these lack sparsity. To address this, we extend the Decoupling Shrinkage and Selection Loss function - first proposed by Hahn and Carvalho - to Ising models. In order to do so, computational issues due to the normalizing constant in Ising models must be addressed; here, we also investigate the pseudolikelihood as an estimate of the Ising likelihood for MCMC posterior estimation. The results of this method are then demonstrated with a simulated Ising model example.

1 Introduction

Graphical models are a common method of describing relationships between random variables, in which a graph expresses a conditional dependence structure. These models have been used across a variety of fields. Oftentimes, it is desirable to incorporate graphical models within a Bayesian approach to modeling - for example, Stingo et al. (2010) [1] apply a Bayesian graphical modeling approach to miRNA regulatory networks in order to perform feature selection.

However, previous approaches have largely focused on continuous variables. In this paper, we seek to apply Bayesian inference to Ising models, which are one type of discrete variable graphical models. In order to effectively summarize information contained in the posterior, while also maintaining sparsity, we apply the Decoupling Shrinkage and Selection (DSS) loss approach from Hahn and Carvalho (2015) [2], modified with the Ising model likelihood, which is shown below.

$$p(x_1, ...x_p|\theta) = \frac{1}{Z(\theta)} \cdot \exp\Big(\sum_{j=1}^{p} \theta_j x_j + \sum_{\{i,j\} \in E} \theta_{ij} x_i x_j\Big)$$

2 Background

This section offers a brief overview of Bayesian analysis, posterior estimation, and applying the psuedolikelihood.

In general, the Bayesian viewpoint tends to characterize uncertainty with distributions. In contrast to the frequentist view, in which the parameters are not treated as random variables, Bayesians are interested with distributions over parameters of interest. Within this framework, there exist a few key components, which are contained within Bayes' theorem:

$$P(\Theta|D) = \frac{P(D|\Theta) \cdot P(\Theta)}{P(D)}$$

Preprint. Work in progress.

where Θ are parameters of interest, and D is the evidence, or data collected. Then, $P(\Theta)$ represents a prior estimate of the probability of parameters, $P(D|\Theta)$ represents the probability of observing data given Θ, and $P(D)$ is a normalizing constant. With all these components, we can derive $P(\Theta|D)$, the posterior probability of our parameters after seeing data D.

In some cases - for example, conjugate priors - we can derive the posterior exactly; however, in many examples of interest, computing the exact posterior is not feasible and we must turn to posterior approximations. Some of the most popular methods for this are in the Markov Chain Monte Carlo (MCMC) family (Haario et al. (2001) [3]). The idea is to sample from a Markov Chain that converges to a stationary distribution. In this case, we want that distribution to be the posterior.

To perform posterior estimation, many common methods require computation of the likelihood. This presents an issue for the Ising model, which contains a normalizing constant $Z(\theta)$ which is intractable even for a small number of variables. To address this, we apply the pseudolikelihood (Besag, 1977 [4]), defined below. Letting $X_1, ..., X_p$ be a set of random variables, the pseudolikelihood is defined as:

$$PL_\theta(X_1 = x_1, ..., X_p = x_p) = \prod_{j=1}^{p} f_\theta(X_j = x_j | X_{-j} = x_{-j})$$

Note that $-j$ denotes all variables other than j. One simple and effective MCMC method is the Metropolis algorithm. For the Ising model case, we modified the algorithm with the psuedolikelihood instead of the full likelihood, which can provide reasonable estimations of the posterior (Bouranis et al. 2018 [5]).

Algorithm 1: Modified Metropolis Algorithm

```
input  : S, the sample size, P, the burn-in percentage, $q_\theta$, the prior distribution
output : $\theta_p$, the output samples with burn-in removed
1. Let $\theta_0$ be the output from the IsingFit package in R.
for s = 1, 2, ... S do
    2. Sample $\theta_s \sim MVN(\theta_s, c\boldsymbol{I})$
    3. Calculate the acceptance probability, r = $\min(1, \frac{PL(X|\theta_s)q_{\theta_s}}{PL(X|\theta_s - 1)q_{\theta_s - 1}})$
    4. Accept $\theta_s$ with probability r. If rejected, set $\theta_s = \theta_{s-1}$
end
5. Remove the first $\lfloor s \cdot \frac{p}{100} \rfloor$ samples to obtain $\theta_p$.
return $\theta_p$
```

3 DSS Loss Function for Ising Models

3.1 Derivation of DSS Loss Function

However, not only do we want an accurate estimation of the posterior, but also a model that is a meaningful summary of the posterior. Models of this sort can be obtained through simple metrics like the posterior mean and median, but these will almost never have any terms equal to 0, thereby inducing no sparsity into the estimation. To address the shortcomings of these simple techniques, we will utilize the technique introduced by Hahn and Carvalho (2015) [2] in which the posterior distribution can be transformed into a sequence of sparse models.

In general, Hahn and Carvalho's technique attempts to minimize $\mathbb{E}(\mathcal{L}(\hat{Y}, \theta))$, where $\hat{Y}$ represents future observations, θ is the vector of parameters of interest, and the expectation is taken with respect to the posterior predictive distribution. They consider the following loss function

$$\mathcal{L}(\hat{Y}, \theta) = \lambda ||\theta||_0 + \frac{1}{n} \log \left(L(\hat{Y}, \boldsymbol{X}, \theta) \right)$$

where $||\theta||_0 = \sum_{i=1}^{p} \mathbb{I}(\theta_j \neq 0)$, $\lambda > 0$, which is chosen by the user, $\boldsymbol{X}$ being a data matrix of predictors, and L being the log likelihood of the distribution of interest. In [2], Hahn and Carvalho focused on multiple linear regression models with homoscedastic, normal errors, where solving the expectation yields the following loss function, which they call the DSS (Decoupling Shirnkage and Selection) Loss Function:

$$\mathcal{L}(\theta) = \lambda||\theta||_0 + ||\boldsymbol{X}\bar{\theta} - \boldsymbol{X}\theta||_2^2$$

with $\hat{\theta}$ equal to the posterior mean. This function can easily be minimized through the `glmnet` package in R.

To utilize Hahn and Carvalho's technique for Ising Models, we will replace the likelihood with the pseudolikelihood, as the likelihood is intractbale even for a modest number of nodes. In addition, to simplify the minimization procedure, $||\theta||_0$ will be replaced with $||\theta||_1$. Therefore, for Ising models with $n \times p$ data matrix $\boldsymbol{X}$, the loss function we will utilize is defined below

$$\mathcal{L}(\boldsymbol{X}, \theta) = \lambda||\theta||_1 + \tfrac{1}{n}\log\Big(\prod_{i=1}^n \prod_{j=1}^p \frac{\exp(\theta_j x_{ij} + \sum_{k \neq j}\theta_{jk}x_{ij}x_{ik})}{1 + \exp(\theta_j + \sum_{k \neq j}\theta_{jk}x_{ik})}\Big)$$

which can be simplified to

$$\mathcal{L}(\boldsymbol{X}, \theta) = \lambda||\theta||_1 + \tfrac{1}{n}\sum_{i=1}^n \sum_{j=1}^p x_{ij}(\theta_j + \sum_{k \neq j}\theta_{jk}x_{ik}) - \log(1 + \exp(\theta_j + \sum_{k \neq j}\theta_{jk}x_{ik}))$$

When taking the expected with respect to the posterior predictive distribution, the only random variable within the sum will be x_{ij}, thus making

$$\mathbb{E}(\mathcal{L}(\boldsymbol{X}, \theta)) = \lambda||\theta||_1 + \tfrac{1}{n}\sum_{i=1}^n \sum_{j=1}^p \mathbb{E}(x_{ij})(\theta_j + \sum_{k \neq j}\theta_{jk}x_{ik}) - \log(1 + \exp(\theta_j + \sum_{k \neq j}\theta_{jk}x_{ik}))$$

which, after rearranging terms, can be simplified to the final form of the Ising DSS Loss Function:

$$\mathbb{E}(\mathcal{L}(\boldsymbol{X}, \theta)) = \lambda||\theta||_1 + \sum_{j=1}^p [\tfrac{1}{n}\sum_{i=1}^n \hat{\pi}_{ij}(\theta_j + \sum_{k \neq j}\theta_{jk}x_{ik}) - \log(1 + \exp(\theta_j + \sum_{k \neq j}\theta_{jk}x_{ik}))]$$

where $\hat{\pi}_{ij} = \mathbb{E}(x_{ij})$, which is the probability $x_{ij} = 1$ given x_{i-j} with relevant θ's equal to the posterior mean:

$$\hat{\pi}_{ij} = \mathbb{E}(x_{ij}) = \frac{\exp(\bar{\theta}_j + \sum_{k \neq j}\bar{\theta}_{jk}x_{ik})}{1 + \exp(\bar{\theta}_j + \sum_{k \neq j}\bar{\theta}_{jk}x_{ik})}$$

3.2 A Minimization Procedure for the DSS Loss Function

Unfortunately, the Ising DSS Loss Function can not be easily minimized in its entirety. However, holding j constant, the loss function,

$$\mathcal{L}(\boldsymbol{X_j}, \theta_j, \theta_{j.}) = \lambda||\theta_{j.}||_1 + \tfrac{1}{n}\sum_{i=1}^n \hat{\pi}_{ij}(\theta_j + \sum_{k \neq j}\theta_{jk}x_{ik}) - \log(1 + \exp(\theta_j + \sum_{k \neq j}\theta_{jk}x_{ik}))$$
$$\text{where } \theta_{j.} = (\theta_{j1}, \theta_{j2}, ..., \theta_{j(j-1)}, \theta_{j(j+1)}, ..., \theta_{jp})$$

can be easily transformed into a weighted LASSO logistic regression problem, with pseudo-responses $z_{ij} = 1$ and $z_{(i+n)j} = 0$, weights $w_{ij} = \hat{\pi}_{ij}$ and $w_{(i+n)j} = 1 - \hat{\pi}_{ij}$, and associated predictors x_{i-j} for z_{ij} and $z_{(i+n)j}$. This can be easily estimated via the `glmnet` function in R, where a sequence of potential estimates are outputted.

Therefore, to estimate the minimizer of this function, we will follow a similar strategy to that of Ravikumar et al. (2010) [6], in which an Ising model is estimated through with a set of p independent logistic regressions. Their procedure proves to provide consistent results for sufficiently large sample sizes. For our case, we will instead solve the p weighted logistic regression problems described earlier. Therefore, our approach is similar to that of Ravikumar et al.'s (2010), but with inclusion of posterior information, as the logistic regressions are weighted by the posterior means.

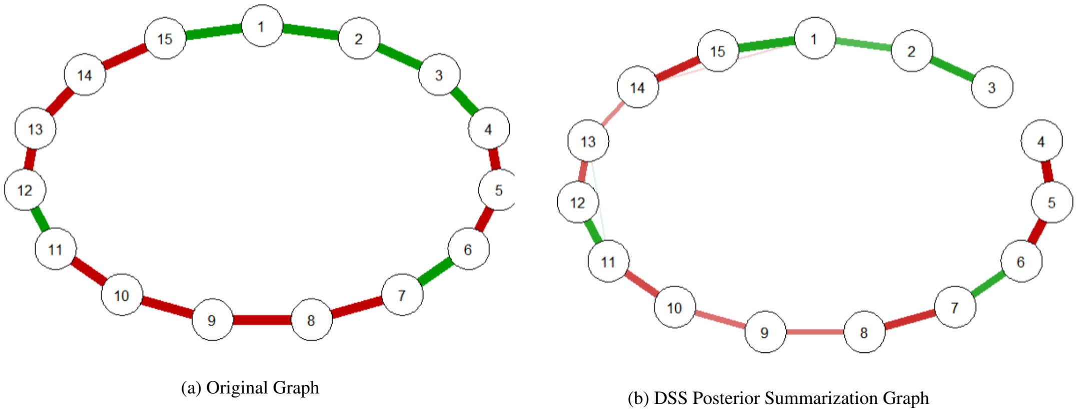
In regards to selecting an appropriate λ for each of the logistic regressions, `glmnet` automatically outputs a sequence of models at varying λ values. We considered the "best" model to be the one that minimizes the Extended Bayesian Information Criterion (Chen et al. 2008) [7], which is the same criterion used by van Borkulo et al. (2014) [8] for the method in [6]. Finally, because each of the interaction effects will be estimated twice in this procedure, the mean of the two estimates will be outputted as the final result. The full algorithm is outlined below:

Algorithm 2: DSS Model Selection

```
input  : $\bar{\theta}$, the posterior mean estimate
         $\boldsymbol{X}$, the observed data matrix
output : $\hat{\theta}$, the selected $\theta$'s from posterior
for $j$ = 1, 2, ... $p$ do
    1. Minimize $\mathcal{L}(\boldsymbol{X_j}, \theta_j, \theta_{j.})$ for a set of $\lambda$'s with the glmnet package in R
    2. Select the model that minimizes the Extended Bayesian Information Criterion
    3. Record results in matrix or list
end
4. Determine final estimate of $\hat{\theta}$ by taking mean of interaction effects
return $\hat{\theta}$
```

4 Example

For our example, we simulated 400 data points from an Ising model with a "ring" structure. The negative interactions have a value of -2, while the positive interactions have a value of 2. Our prior was weakly informative, a multivariate normal distribution with mean 0 and covariance $\sqrt{3} * \boldsymbol{I}$. We utilized the pseudolikelihood Metropolis algorithm for posterior estimation, and our DSS Loss Function for selection. Results are shown in Figure 1.



(a) Original Graph

(b) DSS Posterior Summarization Graph

Figure 1: DSS Selection for Ising Model Results

Our posterior selection technique provides an accurate estimation of the true Ising model. Almost all of the interactions were correctly identified, with only one false negative (interaction between 3 and 4) and two false positives (interaction between 11 and 13 and interaction between 14 and 1). Therefore, utilizing the pseudolikelihood Metropolis algorithm and the Ising DSS Loss Function can prove to be a powerful method to summarize a posterior distribution.

However, it is worth noting that because of the $||\theta||_1$ penalty and the additional prior regularization, the procedure can suffer from overly aggressive shrinkage. In the above example, almost all the estimated interactions have lower absolute values than the their true values, given that most of the edges in the estimated graph are thinner than the true graph. Therefore, one should take this into account if they would like to use this procedure to estimate the true graph. In addition, because of the procedure involves solving logistic regressions weighted by the posterior mean, the pseudolikelihood Metropolis algorithm's posterior mean must be reasonably close to the true parameters in order to obtain an accurate estimation. Because the performance of the Metropolis algorithm begins to significantly deteriorate when the number of nodes in the graph exceeds 20, we recommend utilizing this procedure only when the number of nodes is small.

References

[1] Francesco C Stingo, Yian A Chen, Marina Vannucci, Marianne Barrier, and Philip E Mirkes. A bayesian graphical modeling approach to microrna regulatory network inference. *The annals of applied statistics*, 4(4):2024, 2010.

[2] P Richard Hahn and Carlos M Carvalho. Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110 (509):435–448, 2015.

[3] Heikki Haario, Eero Saksman, Johanna Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[4] Julian Besag. Efficiency of pseudolikelihood estimation for simple gaussian fields. *Biometrika*, pages 616–618, 1977.

[5] Lampros Bouranis, Nial Friel, and Florian Maire. Bayesian model selection for exponential random graph models via adjusted pseudolikelihoods. *Journal of Computational and Graphical Statistics*, 27(3):516–528, 2018.

[6] Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using 1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[7] Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

[8] Claudia D Van Borkulo, Denny Borsboom, Sacha Epskamp, Tessa F Blanken, Lynn Boschloo, Robert A Schoevers, and Lourens J Waldorp. A new method for constructing networks from binary data. *Scientific reports*, 4(1):1–10, 2014.