Bayesian Inference for Gaussian Graphical Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

The project aims to apply techniques of Gaussian Graphical Models, a powerful probabilistic tool for modeling and representing conditional dependencies among variables in Multivariate Gaussian Distributions. The current project also aims to introduce a Bayesian Inference perspective via estimation of the Precision matrix via a Gibbs sampling algorithm. The project aims to use these techniques with a protein dataset, in an attempt to model relationships among proteins for patients diagnosed with AML. Our implementation of the Gibbs algorithm found a highly interconnected protein network in the data set, and a viable estimate of the Multivariate Gaussian precision matrix.

1 Introduction

1.1 Literature Review

Gaussian Graphical Models are used in a variety of contexts to gain a deeper understanding of features and random variables, in this case variables following a Multivariate Gaussian Distribution. This technique's ability to model conditional dependence among many variables in a high-dimensional modeling scenario.

In particular with proteins and AML, literature suggests a relationship between protein profiling and AML outcomes, particularly response and survival. Given the high-dimensional nature of the data, directly sampling and deriving inference from the posterior distribution will prove difficult. To that end, we utilize a Gibbs sampling algorithm, a special case of the Markov Chain Monte Carlo class of algorithms for posterior distribution estimation via simulation.

Recent literature suggests a viable method for estimating the precision matrix involves Regression, specifically the Pseudo-Likelihood and Joint regression methods. As is the case with regressions involving high-dimensional data, multicollinearity, variance inflation, and overfitting are consistent problems.

A way around this in the Bayesian framework involves the Bayesian lasso as well as various prior choices to control parameter magnitude, similar to how shrinkage occurs in Lasso or Ridge regression.

1.2 Project Guideline

The current project aims to integrate these high-dimensional probabilistic techniques and apply them to the AML data set to gain a more intensive understanding of the relationship between proteins for these patients.

The project will begin with deriving the full conditional distributions to implement the Gibbs sampling algorithm as a way of estimating the precision matrix. The algorithm will be implemented in the usual way, with a burn-in period and specified number of iterations as hyperparameters.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

The estimated Precision matrix will then be used to build the undirected graph from an adjacency matrix, with a 1 or 0 depending on if any two variables (proteins) are conditionally independent or not.

The resulting undirected graph will serve as a way of modeling the interactions between proteins for these AML patients. That is, any proteins that are connected on the graph can be thought of as conditionally dependent, or related in some sense.

2 Methodology

As mentioned before, the project will start with implementing the Gibbs sampler before MCMC diagnostics and finally inference on the proteins and variables themselves.

The project itself makes use of the AML RPPA data set, with 256 observations and 51 proteins specifically for analysis.

Prior to the algorithm and modeling, the data were cleaned so as to isolate the proteins themselves. The distributional nature of the data were also verified, in the Multivariate Gaussian and Gaussian sense.

More specifically, we specify distributions for the prior, marginal, and other distributions. We suggest the following distributions for this particular model and data.

$$Y_n \sim MVN(\mu, \Omega^{-1}) \tag{1}$$

$$\pi(\Omega) \sim Wishart(\mathbf{V}, n) \tag{2}$$

Where V is our matrix of features and n is our degrees of freedom. In this case, we can say **V** is a *51x51* matrix.

We use this information to derive the full conditional distributions, as given below.

$$Y^{\mathit{i}}|Y^{\text{-}\mathit{i}}, \omega^{-ii}, \omega^{ii} \sim N \tag{3}$$

Where here, we have the following.

$$i = 1, 2, \ldots, 256 \tag{4}$$

The *-i* in this case represents all observations except for i in this case, as will be used in the Gibbs sampler.

3 Results and Discussion

With these distributions defined, we now proceed to implementing the Gibbs sampler and discussing results.

The Gibbs sampler was designed in the typical way, and we provide a code sample for implementation below.

```
    gibbs <- function(data,
                  n_iter = 10000,
                  burn_in = 5000,
                  p,
                  n,
                  Sigma)
{  for (iter in 1:n_iter){
    for (i in 1:p){
      not_i <- (1:p)[-1]

      S_11 <- Sigma[not_i, not_i]
      S_12 <- Sigma[not_i, i]

      Omega_11 <- Omega[not_i , not_i]
      beta_mean <- -solve(Omega_11) %*% S_12
      beta_cov <- solve(Omega_11) / n

      beta <- t(rmvnorm(1, mean = as.numeric(beta_mean),
                        sigma = beta_cov))
      omega_ii <- rgamma(1, shape = (n / 2) + 1,
                         rate = (Sigma_obs[i,i] + t(beta) %*% Omega_11 %*% beta) / 2)
      Omega[i,i] <- omega_ii # filling in
      Omega[not_i, i] <- Omega[i,not_i] <- as.numeric(beta) # flattening
    }
    # lasso style shrinkage
    Omega[-diag(p)] <- Omega[-diag(p)] / (1 + lambda)

    # storing
    if (iter > burn_in){
      Omega_samples[,,iter - burn_in] <- Omega
    }

    # progress bar
    if (iter %% 1000 == 0) cat("Iteration", iter, "\n")
  }
  # actually returning stuff
  #return(Omega_post_mean)
  return(Omega_samples)
}
```

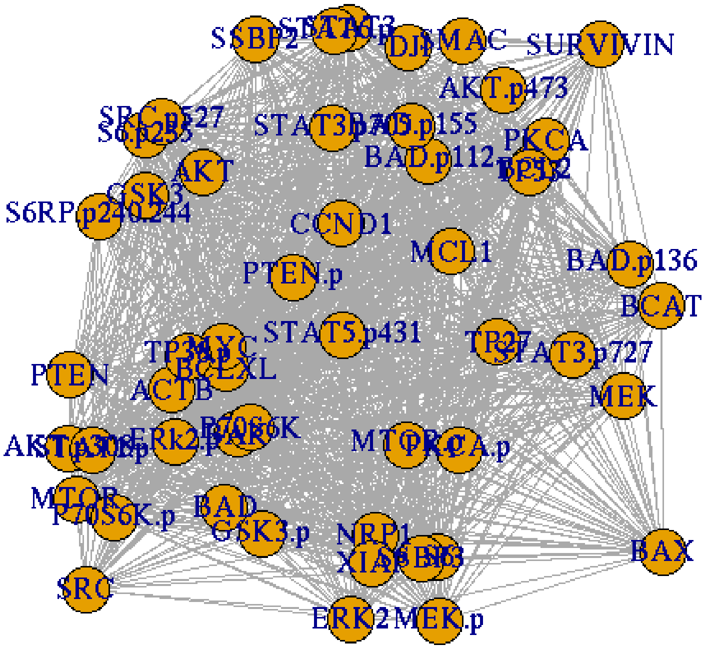

Figure 1: Caption

3.1 Conditional Dependence Protein Graph

We see the default arguments and hyper-parameters included above. Further analysis of this code and Gibbs sampling algorithm revealed the following undirected graph image. We see that this image, the

Figure 2: Caption

nodes labeled with each relevant protein, suggest a highly dependent data set.

In this case, a threshold value of 0.1 was used. That is, if a partial correlation was greater than 0.1, the nodes were connected on the graph.

3.2 Diagnostics and Variational Inference

After completing the graph and establishing the Gibbs sampler, traditional MCMC convergence methods for a single Markov Chain were performed, in particular tests to measure convergence and mixing.

Diagnostic results from the Gibbs sampler can be found detailed by the above plots.

Figure 3: Caption

As evidenced by the plots, we found the mixing and convergence of the Markov Chain to be adequate, however further research and evaluation would be appropriate, for example with differing starting points or varying iterations and a fraction of the time used for the burn-in period. Comparison of various iteration lengths as well as burn-in, and even the necessity of a burn in period at all, in a type of cross-validation, we imagine would be fruitful for further analysis.

Given the convergence of the Markov Chain as well as the diagnostic plots and undirected graph, we decided that variational inference would not be appropriate in this case.

References

References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Griffin, J.E. & Brown P.J. (2010) Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis 5(1)*:171-188.

[2] Kornblau, S.M. et al (2009) Functional proteomic profiling of aml predicts response and survival. *Blood, The Journal of the American Society of Hematology,* 113(1):154-164.

[3] Meinshausen N. & Buhlmann P. (2006) High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics,*34(3):1436-1462.

[4] Mitchell T.J. & Beauchamp J.J. (1988) Bayesian variable selection in linear regression. *Journal of the American Statistical Association,*83(404):1023-1032.

[5] Piironen J. & Vehtari A. (2017) On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. *Artificial Intelligence and Statistics* 905-913.

[6] Wang Z. et al (2022) Bayesian edge regression in undirected graphical models to characterize interpatient heterogeneity in cancer. *Journal of the American Statistical Association*:533-546.