Model selection for high-dimensional Gaussian Graphical Models

Matt Delhey
Rice University
matt.delhey@rice.edu

1 Introduction

In this paper we assume the general graphical model framework that connects a probability distribution p with a graph $G = \{V, E\}$. *Model selection* corresponds to selecting E, the edges to include in the graph. *Model estimation* corresponds to estimating the inverse covariance matrix Θ for a given E. In practice, E is unknown (and V is known) and therefore estimation of the unknown graph requires both model selection and model estimation. This paper explores four different methods of model selection for Gaussian graphical models (GGMs), supposing the graphical LASSO (GLASSO) method of model estimation.

1.1 Gaussian graphical models

In a Gaussian graphical model, we assume that the data is drawn from a multivariate Gaussian distribution $N_k(0, \Sigma)$, and are solely concerned with inference on Σ. The log-likelihood of the data is then

$$l(\Theta) = logL(\Theta) = \frac{n}{2} \log \det(\Theta) - \frac{1}{2}\Theta\hat{S} \tag{1}$$

For a graphical model in general, an edge between X_i and X_j is absent if and only if X_i and X_j are conditionally independent given all the other variables. For a Gaussian graphical model, the pairwise Markov property such that the conditional independence condition, which removes the edge between X_i and X_j, is upheld if and only if $\Theta_{i,j} = 0$. Therefore estimation of the inverse covariance matrix Θ is equivalent to both model estimation and model selection problems for Gaussian graphical models.

1.2 GLASSO model estimation

The graphical LASSO estimates a sparse inverse covariance matrix Θ using an ℓ_1 (LASSO) penalty given a set of edges, exploiting the fact that estimation of Θ is equivalent to estimation of β in linear regression. An estimate of the corresponding undirected graph can be constructed using either the AND or the OR rule.

The GLASSO estimate for Θ and β, its linear regression equivalent, are given by

$$\hat{\Theta}_{LASSO}(\lambda) = \underset{\Theta}{\operatorname{argmax}} \left[l(\Theta) - \lambda ||\Theta||_1 \right] \tag{2}$$

$$\hat{\beta}_{LASSO}(\lambda) = \underset{\beta}{\operatorname{argmax}} \left[\frac{1}{2} ||Y - X\beta||_2 + \lambda ||\beta||_1 \right] \tag{3}$$

The regularization parameter λ determines the penalty for model non-sparsity, larger values resulting in sparser graphs. Theoretical results show that the GLASSO estimator $\hat{\Theta}_{GLASSO}(\lambda)$ recovers the true graph with high probability for an appropriate λ under regularity conditions, but this does not give any guidance to picking λ in practice.

2 Methods

The goal of this paper to compare four different methods of estimating Θ within the GLASSO framework, either by selecting the best λ or by estimating E through some other means. Additionally, this comparison is made in the *high-dimensional* case, where $k >> n$. The four methods of model selection compared are:

- Bayesian information criterion (**BIC**)
- Cross-validation (**CV**)
- Stability selection (**SS**)
- Bayesian model averaging (**BMA**)

BIC, CV, and SS are perhaps the three most popular methods of model selection for GGMs while BMA is relatively less popular. For this reason, only the BMA approach will be discussed in detail in this paper. However, it is worth noting that BIC and CV specifically select λ, while SS and BMA estimate E using the *regularization path*, or the series of model estimates for a set of candidate values of λ

2.1 Bayesian model averaging for GLASSO

My inclusion of Bayesian model averaging is inspired by a recent paper by Zhe Liu [2] which describes a method for extending BMA methods to the high-dimensional case using the GLASSO framework (BMA-GLASSO) as an extension to the same procedure for the linear regression case (Fraley, Percival [1]). I leave the full technical description of the proposed method to these two papers, and here will only describe it in outline.

The basic approach of BMA methods is to estimate a parameter using a weighted average over posterior distributions of multiple models. The weights correspond to the updated confidence in each model, and thus BMA methods are said to incorporate model uncertainty. Indeed, the advantage of the BMA-GLASSO approach is to join both the advantages of GLASSO (sparsity) with BMA (model uncertainty).

However, in order to average over many models, we first need a model space. Traditional approaches to BMA fail in the high-dimensional setting, where, in the case of model selection for GGMs, there are $2^{d(d-1)/1}$ potential models. The solution proposed by Liu [2] is to use the GLASSO regularization path as the model space or "model dictionary". Posterior probabilities of each of the models are implicitly calculated by averaging estimates of Θ from a Metropolis-Hastings MCMC sampling algorithm, where the sampler visits "neighboring" models uniformly, controlled by a fixed parameter l indicating the number of edges constituting a local neighborhood.

I have implemented the BMA-GLASSO algorithm in an `R` package `bmar` which also includes the code used to run the simulations. Model estimation and graph simulations were conducted using the `huge` package [4]. Code and installation instruction can be found at `github.com/mattdelhey/bmar`.

3 Simulations

Each model selection method was applied to four types of generated graphs (see in Figure 1) where $n = 200$ and $k = 300$. The regularization path consisted of $1,000$ candidate values of λ for each method and each simulation was repeated 5 times.

Successful recovery of the graph was evaluated by comparing the true Θ with its estimate $\hat{\Theta}$ according to sparsity and running time as well as three criteria:

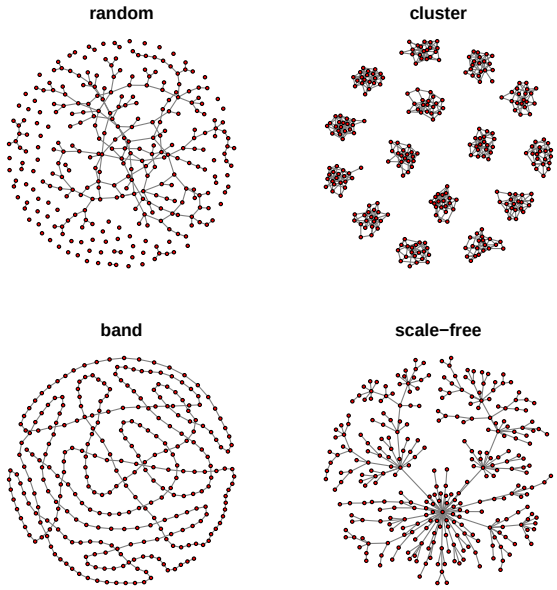

Figure 1: Simulated graphs

- sum of squared errors (**SSE**)

$$SSE = ||\hat{\Theta} - \Theta||_F^2$$

- Kullback-Leibler loss (**KL**)

$$KL = -\log\det\hat{\Theta} + \text{tr}\,\hat{\Theta}\Theta^{-1} - (-\log\det\Theta + d)$$

- F_1-score (weighted average of precision and recall)

$$F_1\text{-score} = 2\,\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Tuning parameters are needed for the model selection methods. For the sake of simplicity, these were fixed to: 5 folds for cross-validation; an inclusion threshold of 0.05, subsample ratio of 0.5, and 10 subsamplings for stability selection; and $10,000$ iterations, $2,000$ burn-in iterations, and a neighborhood of $l = 3$ edges and the flat prior for Bayesian model averaging.

3.1 Results

Table 1 records the average values for the aforementioned metrics and their standard errors.

4 Data: breast cancer

The four methods of model selection were also compared using a real data set consisting of gene expression signatures for p53 mutation status in breast cancer samples (Miller et al [3]). In this case, $n = 250$ and $k = 1,000$. Model estimation was conducted for 100 candidate values of λ before applying each procedure and all other tuning parameters were kept the same as in the simulations. Because the true graph structure is unknown, no error metrics can be given as in the case of the simulations. The resulting estimated graphs are given in Figure 2.

5 Conclusion

The results of the simulations suggest that BMA and SS work the best. BIC in almost every case picked a graph that was too dense. However, the breast cancer data set application demonstrates the

graph	model	SSE	KL	F1-score	Sparsity	Time
random	bic	91.20 (11.0)	41.41 (3.8)	0.69 (0.03)	0.35 (0.02)	0.26 (0.0)
	ss	89.30 (2.6)	21.24 (0.3)	0.01 (0.00)	0.01 (0.00)	1175.48 (20.1)
	cv	170.05 (9.7)	49.22 (2.6)	0.00 (0.00)	0.00 (0.00)	1040.88 (4.5)
	bma	89.67 (10.8)	40.85 (3.8)	0.69 (0.03)	0.35 (0.02)	39.91 (3.0)
cluster	bic	116.69 (9.0)	43.99 (0.23)	0.63 (0.00)	0.33 (0.00)	0.26 (0.0)
	ss	214.96 (40.3)	42.03 (3.8)	0.01 (0.00)	0.02 (0.00)	1176.86 (57.6)
	cv	309.39 (55.0)	71.52 (6.9)	0.00 (0.00)	0.00 (0.00)	1060.44 (0.8)
	bma	115.98 (9.0)	43.57 (0.2)	0.63 (0.00)	0.33 (0.00)	42.72 (9.0)
band	bic	177.45 (1.5)	43.39 (0.5)	0.59 (0.01)	0.30 (0.00)	0.25 (0.0)
	ss	494.47 (3.8)	55.80 (0.5)	0.02 (0.00)	0.02 (0.00)	1250.44 (13.0)
	cv	684.95 (2.2)	96.86 (0.6)	0.00 (0.00)	0.00 (0.00)	104.00 (12.3)
	bma	177.96 (1.6)	43.05 (0.5)	0.59 (0.01)	0.31 (0.01)	43.21 (7.5)
scale-free	bic	135.42 (14.3)	52.67 (4.0)	0.78 (0.02)	0.40 (0.01)	0.28 (0.0)
	ss	59.56 (8.1)	15.01 (2.1)	0.01 (0.00)	0.01 (0.00)	1220.74 (10.9)
	cv	93.97 (15.2)	29.48 (4.2)	0.00 (0.00)	0.00 (0.00)	102.75 (5.3)
	bma	132.48 (14.6)	51.77 (4.1)	0.78 (0.02)	0.40 (0.01)	40.35 (0.5)

Table 1: Results from simulation with 5 replicates. Mean values are given, with standard errors in parenthesis.

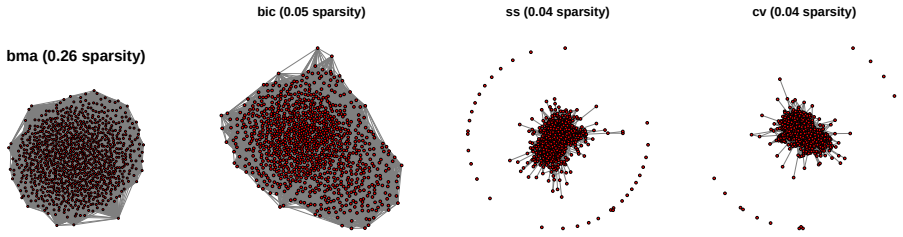

Figure 2: Graphs estimated for the breast cancer dataset.

overselection of BIC and BMA, resulting in graphs likely too dense for interpretative purposes. On the other hand, graphs selected by SS and CV are much sparser and thus are likely more appropriate for the underlying structure. The graphs estimated by BMA tended to be dense. Perhaps this could be mitigated using the sparsity inducing prior mentioned by Liu in [2].

References

[1] Fraley, C., & Percival, D. (2013). Model-averaged ℓ_1 regularization using Markov chain Monte Carlo model composition. Journal of Statistical Computation and Simulation, (ahead-of-print), 1-12.

[2] Liu, Z. (2015). Bayesian Model-Averaged Regularization for Gaussian Graphical Models. *arXiv preprint arXiv:1503.02698*.

[3] Miller et al (2005, PubMed ID:16141321)

[4] Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research, 13*(1), 1059-1062.