An Empirical Bayes Approach to Structure Learning of Sparse High-Dimensional Gaussian Graphical Models

Ha Nguyen, Sumanta Basu, Martin T. Wells
Cornell University, Department of Statistics and Data Science

Problem Statement

- In high-dimensional settings where $n < p$, learning the structure of sparse Gaussian graphical models (GGMs) is an important problem with many applications.
- Frequentist methods, such as graphical Lasso (GLASSO), are fast but tend to produce many false positives [4]. Bayesian methods based on Markov chain Monte Carlo (MCMC) are reported to have lower false positive and better edge selection, but are slow [8; 10].
- We propose an empirical Bayes approach based on Expectation Maximization (EM) that is faster than Bayesian methods and retain their high graph selection accuracy.

Significance

- Why Gaussian graphical models?
  - GGMs capture conditional dependencies between random variables in the sparsity structure of the inverse covariance matrix, or the precision matrix $\Omega = \Sigma^{-1}$.
  - Applications of GGMs include cancer gene co-expression networks [2; 8], brain functional connectivity [5], and joint credit risk network [10].
- Current challenges of graph recovery:
  - Full Bayesian methods, such as the G-Wishart and spike and slab priors, have shown favorable comparisons to GLASSO, particularly in terms of reducing the false positive rate [10].
  - However, these approaches face scalability issues in high-dimensional settings due to the computational burden of MCMC.
- Our approach:
  - We employ the scalable empirical Bayes model selection (SEMMS) method, introduced by [1], within the neighborhood selection framework proposed by [6].
  - SEMMS performs nodewise variable selection with high accuracy in high-dimensional settings where $n < p$ and can be efficiently computed with parallelization.

Background & Prior Work

- One can relate a random vector that follows a multivariate normal distribution to an underlying graph structure $\mathcal{G} = (V, E)$, where V is the set of vertices and E is the set of edges. Consider a sample of n i.i.d. observations of a p-dimensional random vector $X = (X_1, \ldots, X_p)$ where $X \overset{iid}{\sim} \mathcal{N}_p(0, \Omega^{-1})$. The sparsity structure of the precision matrix $\Omega = \Sigma^{-1}$ captures the conditional independence relationship among variables:

$$X_j \perp X_k \mid X_{V\setminus\{(j,k)\}} \iff \omega_{jk} = 0 \iff (j,k) \notin E$$

- Our goal is to recover the underlying graph structure $\mathcal{G}$ (structure learning).
- G-Wishart prior: Many popular methods are based on the G-Wishart prior for precision matrices, which is conjugate to the multivariate Gaussian likelihood [3; 9; 7]. However, posterior computation under the G-Wishart prior is a non-trivial task due to the calculation of the normalizing constant of the marginal likelihoods.
- Continuous spike and slab prior: [10] proposed using the continuous spike and slab priors for precision matrices. This prior imposes a two-component Gaussian mixture on the off-diagonal entries of Ω, enabling an efficient block Gibbs sampling procedure known as Stochastic Search Structure Learning (SSSL). However, under high-dimensional settings, its computational cost can still be too long.

Empirical Bayes Methodology

- Nodewise regression: A natural approach to edge selection of the underlying graph $\mathcal{G}$ is to perform nodewise regression [6], where we regress one node on all the other nodes. For a given node j, the associated regression parameter $\beta_j \in \mathbb{R}^p$ defines the edges between node j and the other nodes: $(\beta_j)_k \neq 0 \iff (j,k) \in E$.
- Scalable empirical Bayes model selection (SEMMS): For each node, SEMMS models the continuous response $y \in \mathbb{R}^n$ using an additive combination of K 'putative' variables z_k with coefficients u_k. It assumes that a small but unknown set of candidate predictors have a non-zero effect on the response.

$$\begin{aligned} y &= \sum_{k=1}^{K} z_k \gamma_k u_k + \varepsilon \\ \text{where } u_k &\overset{\text{iid}}{\sim} N(\mu, \sigma^2) \\ \gamma_k &\overset{\text{iid}}{\sim} \text{multinomial}(0, 1, -1; p_0, p_1, p_{-1}) \\ \varepsilon &\sim N(0, \sigma_e^2 I_n) \end{aligned}$$

- Here, z_k are the candidate predictors, u_k are the random effects, and γ_k are treated as latent variables. We are most interested in edge selection, which means identifying which latent variables $\{\gamma_k\}$ are positive, negative, or null (zero).
- In matrix notation, we denote the $n \times (p-1)$ matrix (z_{ik}) by $\mathbf{Z}$, and write $\Gamma \equiv \text{diag}(\gamma_1, \gamma_2, \ldots, \gamma_{p-1})$ and $\mu = \mu \mathbf{1}_{p-1}$. Let $\mathbf{z}_k$ denote the k-th column of $\mathbf{Z}$. Then the model can be rewritten as

$$\begin{aligned} \mathbf{y} &= \mathbf{Z}\Gamma\mathbf{u} + \varepsilon \\ \text{where } \varepsilon &\sim N(\mathbf{0}_n, \sigma_e^2 \mathbf{I}_n) \\ \mathbf{Z}\Gamma\mathbf{u} \mid \Gamma &\sim N(\mathbf{Z}\Gamma\mu, \sigma^2 \mathbf{Z}\Gamma\mathbf{Z}') \end{aligned}$$

Thus, $\mathbf{y} \sim N(\mathbf{Z}\Gamma\mu, \sigma_e^2\mathbf{I}_n + \sigma^2\mathbf{Z}\Gamma\mathbf{Z}')$. Parameter estimation in SEMMS for $\boldsymbol{\theta} = \{\mu, \sigma^2, p_0, p_1, p_2, \sigma_e^2\}$ is carried out using a Generalized Alternating Minimization (GAM) algorithm, which is an efficient and convergent variant of the EM algorithm.

- GAM algorithm: Specifically, in the M step, we plug in the current estimates of the posterior expected values of the latent variables $\{\gamma_k\}$ and maximize with respect to $\boldsymbol{\theta} = \{\mu, \sigma^2, p_0, p_1, p_{-1}, \sigma_e^2\}$. Maximizing with respect to $\{p_0, p_1, p_{-1}\}$ yields the relative frequency $p_s = \frac{\sum_{k=1}^{K} \mathbf{1}_{\{\gamma_k = s\}}}{K}$. In the E step, we update the latent variables $\{\gamma_k\}$ using Bayes' rule

$$Pr(\gamma_k = 0) = \frac{p_0^{(t)} f(y; \gamma_k = 0; \gamma_{-k} = \gamma_{-k}^{(t)})}{\sum_{s\in\{-1,0,1\}} p_{i(s)}^{(t)} f(y; \gamma_k = s; \gamma_{-k} = \gamma_{-k}^{(t)})}$$

- For each node, we select an edge when $Pr(\gamma_k \neq 0) > 0.01$.
- We symmetrize the sparsity pattern for the output graph using standard post-processing techniques, applying OR rule as described in [6].

Numerical Experiments

- Simulated data: We simulate 10 replications, each with a random graph where $p = 300$ and $n \in \{50, 100\}$. The graph density is set to $p \in \{0.002, 0.02\}$. The precision matrix Ω is sampled from the G-Wishart distribution: $\Omega \sim W_{\mathcal{G}}(3, I_p)$.
- For SEMMS, we set a non-informative initial number of non-nulls to be 10 for all nodewise regressions.
- Benchmark: The benchmark methods include SSSL and GLASSO.
  - For SSSL, we use 2000 iterations for burn-in and an additional 5000 iterations for posterior estimation, as recommended by [10]. We initialize the Markov chain with an empty graph. To ensure the starting covariance matrix is positive definite, we use the empirical covariance matrix, adjusting the diagonals by adding the median value of the diagonals when $n < p$. The hyperparameters are set to $\epsilon = 0.02$, $v = 2$, and $\lambda = 2$.
  - For GLASSO, the penalty parameter is chosen using 3-fold cross validation.
- Performance metrics: To evaluate performance of graph structure learning, we report the mean and standard deviation of standard metrics, including true positive rate (TPR), false positive rate (FPR), Matthews's correlation coefficient (MCC) and computing time. All methods are run on 1 core for comparison.

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

- Results: SEMMS outperforms SSSL and GLASSO in structure learning accuracy at a graph density of 0.002, achieving a significantly FPR. At a graph density of 0.02, SEMMS demonstrates a comparable MCC to SSSL with substantially faster runtime.
- Figure 1 demonstrates the importance of a low FPR. In high-dimensional sparse graphs, a FPR of just 0.026 can produce misleading results.

Method	TPR	FPR	MCC	Time
Density 0.002				
SEMMS	0.471 (0.099)	0.001 (0.000)	0.442 (0.082)	23.2 (1.0) s
SSSL	0.627 (0.084)	0.028 (0.002)	0.156 (0.028)	15.7 (0.2) h
GLASSO	0.539 (0.125)	0.012 (0.012)	0.276 (0.114)	67.9 (8.3) s
Density 0.02				
SEMMS	0.200 (0.013)	0.001 (0.000)	0.389 (0.018)	18.6 (0.5) s
SSSL	0.342 (0.029)	0.012 (0.004)	0.342 (0.025)	15.6 (0.3) h
GLASSO	0.534 (0.021)	0.072 (0.013)	0.235 (0.019)	6.6 (3.3) m

Method	TPR	FPR	MCC	Time
Density 0.002				
SEMMS	0.611 (0.094)	0.001 (0.000)	0.555 (0.075)	2.6 (1.8) m
SSSL	0.686 (0.079)	0.007 (0.001)	0.323 (0.047)	16.0 (0.3) h
GLASSO	0.745 (0.094)	0.020 (0.011)	0.250 (0.089)	55.1 (9.5) s
Density 0.02				
SEMMS	0.256 (0.009)	0.001 (0.001)	0.460 (0.027)	53.6 (2.5) s
SSSL	0.452 (0.016)	0.006 (0.001)	0.509 (0.017)	15.9 (0.3) h
GLASSO	0.663 (0.026)	0.095 (0.022)	0.259 (0.027)	12.2 (10.6) m

Table 1: Mean (SD) of structure learning and computing time performance for $\{n, p\} = \{50, 300\}$ (top) and $\{n, p\} = \{100, 300\}$ (bottom).

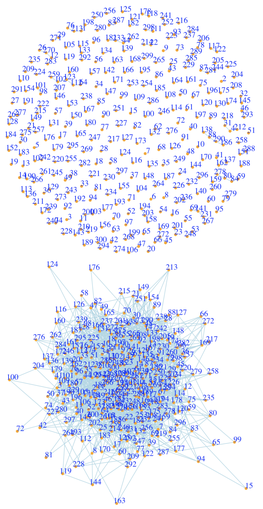

Figure 1: A graph of density 0.002 (top) versus GLASSO output using $n = 50$ (bottom), with TPR of 0.757 and FPR of 0.026.

References

[1] Haim Bar, James Booth, and Martin T. Wells. A scalable empirical bayes approach to variable selection in generalized linear models. 3 2018.

[2] Sagnik Bhadury, Riten Mitra, and Jeremy T Gaskins. Fast bayesian high-dimensional gaussian graphical model estimation, 2023.

[3] Adrian Dobra, Alex Lenkoski, and Abel Rodriguez. Bayesian inference for general gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106(496):1418–1433, 2011. PMID: 26924867.

[4] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 7 2008.

[5] Max Hinne, Ronald J Janssen, Tom Heskes, and Marcel A J Van Gerven. Bayesian estimation of conditional independence graphs improves functional connectivity estimates supporting information s1, 2015.

[6] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 6 2006.

[7] A. Mohammadi and E. C. Wit. Bayesian structure learning in sparse gaussian graphical models. *Bayesian Analysis*, 10:109–138, 2015.

[8] Yang Ni, Veerabhadran Baladandayuthapani, Marina Vannucci, and Francesco C. Stingo. Bayesian graphical models for modern biological applications, 6 2022.

[9] Francesco Stingo and Giovanni M. Marchetti. Efficient local updates for undirected graphical models. *Statistics and Computing*, 25:159–171, 1 2015.

[10] Hao Wang. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10:351–377, 2015.