Bayesian Search Algorithms for Decomposable Gaussian Graphical Models

Emt and KPM
CS,UBC

1 A Heirarchical Model

Consider d variables denoted by $y_1, y_2, \ldots, y_d$. Let us say that we observe n measurements i.e. $y_i \in \mathbb{R}^n$. We assume that these measurements follow a Gassian distribution:

$$y_i \sim \mathcal{N}(\mu, \Sigma) \tag{1}$$

For simplicity, we fix $\mu = 0$, and consider the problem of covariance estimation (or covariance selection).

1.1 Priors

In this report, we are interested in the cases when the dependence among variables is sparse. The independence among the variables is reflected in *structural zeros* in precision matrix Σ^{-1} [Dem72]. We now describe a prior which forces sparsity in precision matrix, while maintaining positive-definiteness of covariance matrix.

A positive definite matrix A has an inverse Wishart (IW) density, if the density is of the following form,

$$p(A|\delta', \Phi') \quad \propto \quad |A|^{-\frac{\delta'+m+1}{2}} \text{etr}\left(-\frac{1}{2}A^{-1}\Phi'\right) \tag{2}$$

where etr(A) = exp trace(A), δ' is a positive real number and Φ' is a positive definite matrix of same size as A. Note that this is a conjugate prior to multivariate Gaussian and hence is a convenient choice for prior distribution. However this distribution does not enforce any structural sparsity on the matrix A. We now describe a special case of decomposable graph where we can use IW distribution to build a distribution which maintains the sparsity patterns along with positive definite constraints(please see [Lau96] for details on decomposable graphs).

Let G be a decomposable graph with cliques $C_{1:k}$ and separators $S_{1:k}$. Decomposibility of a graph implies that the cliques $C_{1:k}$ of this graph can be ordered in a perfect sequence (and vice versa). In what follows, for any matrix A, we use A_{C_i,C_j} to denote the submatrix referenced by the vertices in clique C_i and C_j. A hyper inverse Wishart (HIW) distribution on Σ is defined as follows,

$$p(\Sigma|\Phi, \delta, G) \quad = \quad \frac{\prod_{i=1}^{k} p(\Sigma_{C_i,C_i}|\delta, \Phi_{C_i,C_i})}{\prod_{i=2}^{k} p(\Sigma_{S_i,S_i}|\delta, \Phi_{S_i,S_i})} \tag{3}$$

where Φ is a $d \times d$ positive-definite matrix, and δ is a positive real number. The above expression says that the distribution over Σ can be constructed using separate distributions over cliques and separators, which themselves follow an inverse Wishart distribution given by Eq. (2). It is easy to see that the normalizing constant of this distribution is equal to the product of normalizing constants of IW distribution for cliques divided by the product of normalizing constants of separators.

There are various choices for δ and Φ. As discussed in [GG99], δ expresses the relative weight of the prior, and it is reasonable to assume a gamma distribution with certain mean and variance. In this work, we fix it to a constant value. For Φ, there are at least 3 choices (discussed in [Arm05]):

1. Set $\Phi = \tau I$ where $\tau > 0$.
2. Set $\Phi = \tau(\rho J + (1-\rho)I)$, where $\tau > 0$, J is a $d \times d$ matrix of ones and ρ is a correlation coefficient in the open interval $(-(p-1)^{-1}, 1)$. This is called *equicorrelated version* in [GG99].
3. Set $\Phi = \tau S_y/(n-1)$, where $\tau > 0$, where $S_y = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$, and $\bar{y}$ is sample mean. This is refered to as *scaled sum of squares* (described in [Arm05]).

Finally for prior on graph, we have different choices discussed in [Arm05]. For now, we assume a uniform distribution over graph in this work.

1.2 Posterior distribution

As shown in [Arm05], the posterior distribution of Σ is the following,

$$p(\Sigma|y_{1:n}, \delta, \Phi, G) \sim \text{HIW}(G, \delta^*, \Phi^*) \tag{4}$$

where $\delta^* = \delta + n - 1$ and $\Phi^* = \Phi + S_y$, where , where $S_y = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$, and $\bar{y}$ is sample mean. We can also integrate out Σ to obtain the following marginal likelihood,

$$p(y_{1:n}|\delta, \Phi, G) = (2\pi)^{-d(n-1)/2} \frac{h(G, \delta, \Phi)}{h(G, \delta^*, \Phi^*)} \tag{5}$$

where $h(G, \delta, \Phi)$ is the normalizing constant of a HIW wishart distribution, and as described earlier it can be derived from the normalizing constants of cliques and separators.

2 The Bayesian Search Algorithms

We are interested in finding posterior distribution over graphs,

$$p(G|y_{1:n}, \delta, \Phi) \propto p(y_{1:n}|G, \delta, \Phi)p(G) \tag{6}$$

Given a graph, we can compute the log posterior using the above equation. We can use this as a score associated with a graph, and search for decomposable graphs with most significant score. We may use the following search algorithms,

Breadth first search We start with a null graph, and add an edge at all possible locations (such that we get decopomsable graphs) and compute their score. Next we add second edge, find all possible decomposable graph with 2 edges, and compute scores. We keep on going till we have added some e number of edges. This is definitely very computationaly intensive, and may not be feasible as the number of candidate edges grows exponentially.

Depth first search We start with a null graph, and find all decomposable graph with one edge. We find the graph with maximum score, and then add a second edge to it at all possible locations. We find the graph with maximum score, and continue till we have added some e number of edges. This will result in a single graph. To find another graph, we repeat this procedure while avoiding all the previous edges already added.

Beam Search This is same as depth first search, except instead of keeping one graph, we keep B graph at ever stage, where B is the beam width.

For all these search algorithms, we need to know whether an edge addition leads to a decomposable graph or not. The following result from [Arm05] ensures the *leagality* of an edge addition,

Theorem 1. *Let $G = (V, E)$ be a decomposable graph with cliques $C_{1:k}$. Consider two vertices a and b such that (a, b) is not an edge in G. Let C_i and C_j are two cliuqes such that a is in C_i, b is in C_j. An edge addition (a, b) is legal iff intersection $C_i \cap C_j$ separates a from b.*

This gives us following algorithm to find the legal addition: First, find all the cliques containing a and b (there is no single clique containing both a and b as this edge is not present in G). Next, find C_i and C_j such that their intersection is maximal among all the cliques containing a and b, and remove it from G. If there is no path from a to b in this new graph then this edge is a legal addition.

There is a similar result for edge deletion,

Theorem 2. *Deletion of an edge (a, b) is legal iff this edge is a member of only one clique.*

Finally, the following result shows how to update the score when an edge is added (or deleted),

Theorem 3. *Let G' be a graph obtained by a (legal) addition of edge (a, b) to a decomposable graph G. Let $S_{q2} = C_i \cap C_j$ is the maximal intersection in G as described in Theorem 1. Let $\delta^* = \delta + n - 1$ and $\Phi^* = \Phi + S_y$. We have the following update for marginal likelihood,*

$$\log \frac{p(y_{1:n}|G', \delta, \Phi)}{p(y_{1:n}|G, \delta, \Phi)} = \log \frac{h(G', \delta, \Phi)h(G, \delta^*, \Phi^*)}{h(G', \delta^*, \Phi^*)h(G, \delta, \Phi)} = \log \frac{r(\Phi, \delta)}{r(\Phi^*, \delta^*)} \tag{7}$$

where

$$\log r(\Phi, \delta) = \frac{\delta + |S_{q2}|}{2} \log \frac{|\Phi_{DD|S_{q2}}|}{|\Phi_{aa|S_{q2}} \Phi_{bb|S_{q2}}|} + \frac{1}{2} \log |\Phi_{DD|S_{q2}}| + \log \frac{\Gamma\left(\frac{\delta+|S_{q2}|}{2}\right)}{\Gamma\left(\frac{\delta+|S_{q2}|+1}{2}\right)} \tag{8}$$

with $D = \{a, b\}, \Phi_{DD|S_{q2}} = \Phi_{DD} - \Phi_{DS_{q2}}(\Phi_{S_{q2}S_{q2}})^{-1}\Phi_{DS_{q2}}^T$ and $\Phi_{aa|S_{q2}}, \Phi_{bb|S_{q2}}$ are defined similarly.

Computational complexity of above update depends on the size of set S_{q2}, and the most expensive step is inversion of $\Phi_{S_{q2}S_{q2}}$. For sparse graphs we expect this matrix to be of considerable smaller size than d. One can use Cholesky update for fast computation of this matrix, and also the fast update tricks described in [PS08].

2.1 Estimate Σ

Having found a graph (say for example MAP estimate), we can estimate the covariance matrix as follows:

$$\mathbb{E}(\Omega|y_{1:n}, G, \Phi, \delta)^{-1} = \sum_{i=1}^{k} \left[(\delta^* + |C_i| - 1)(\Phi^*_{C_iC_i})^{-1}\right]^0 - \sum_{i=2}^{k} \left[(\delta^* + |S_i| - 1)(\Phi^*_{S_iS_i})^{-1}\right]^0 \tag{9}$$

where $[A]^0$ means that a matrix A with zeros filled according to G, and C_i, S_i are cliques and separators of G.

3 Preliminary results

In this section, we report a small experiment comparing Bayes search with L1MB (see [MB06]) for a simple AR1 model (described in [YL07]). We set Φ to be an identity matrix and $\delta = 5$. We use beam search with search depth equal to d, i.e. number of variables. This is because the truth is an AR1 model containing d edges (yes, we are cheating!). We run 10 simulations for 2 cases: $n = 40, d = 10$ and $n = 10, d = 10$. Figure 1 shows KL divergence, false positives and false negatives. We see that bayes search performs much better than L1MB for the case when $n \leq d$. Similar results are found for AR2 model and 'circle' model (Fig. 2 and 3).

References

[Arm05] H. Armstrong. *Bayesian estimation of decomposable Gaussian graphical models*. PhD thesis, The University of New South Wales, 2005.

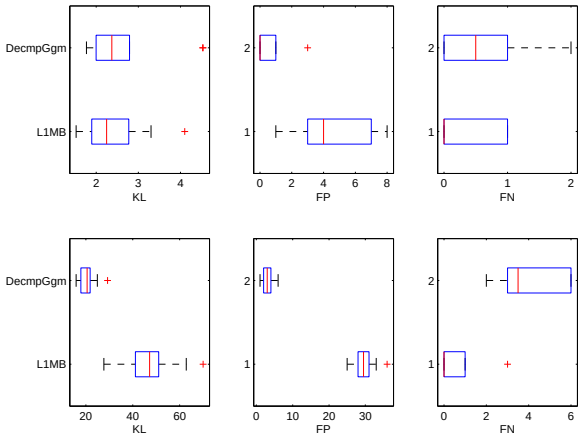

Figure 1: Comparison of Bayesian search with L1MB for AR1. Top row shows $n > d$, and bottom row shows $n < d$.

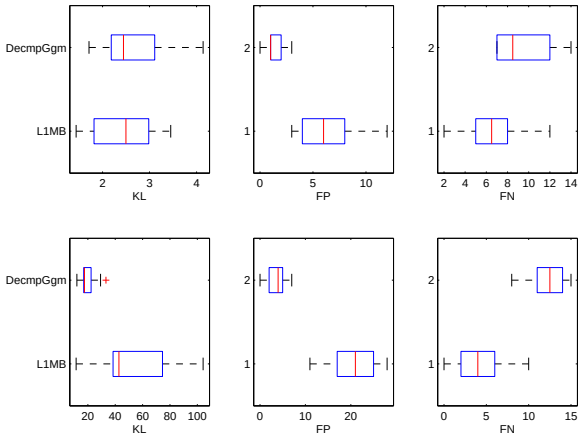

Figure 2: Comparison of Bayesian search with L1MB for AR2. Top row shows $n > d$, and bottom row shows $n < d$.

[Dem72] A.P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

[GG99] P. Giudici and PJ Green. Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801, 1999.

[Lau96] S.L. Lauritzen. *Graphical Models*. Oxford Science Publications, 1996.

[MB06] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of statistics*, 34(3):1436, 2006.

[PS08] Justin Ziniel Philip Schniter, Lee C. Potter. Fast Bayesian Matching Pursuit. *Compressed Sensing Workshop, UCSD*, 2008.

[YL07] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19, 2007.

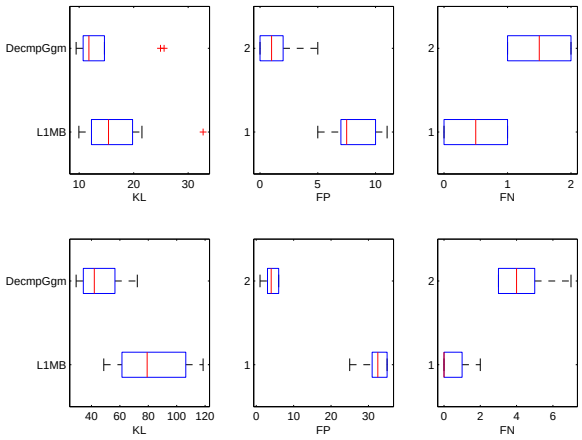

Figure 3: Comparison of Bayesian search with L1MB for 'circle'. Top row shows $n > d$, and bottom row shows $n < d$.