Bayesian structural learning and estimation in Gaussian graphical models and hierarchical log-linear models

Adrian Dobra
University of Washington

Joint work with Héléne Massam and Alex Lenkoski
York University and University of Washington

London Mathematical Society Durham Symposium
Mathematical Aspects of Graphical Models

July 8, 2008

It generalizes the hyper inverse Wishart of Dawid and Lauritzen (1993). Its density is

$$p(K|G) = \frac{1}{I_G(\delta, D)} (\det K)^{(\delta-2)/2} \exp\left\{-\frac{1}{2}\langle K, D\rangle\right\}.$$

wrt the Lebesgue measure on P_G. The posterior of K is $W_G(\delta + n, D + U)$. The marginal likelihood of G is

$$p(x^{(1:n)}|G) = I_G(\delta + n, D + U)/I_G(\delta, D).$$

PROPERTIES OF THE G-WISHART $W_G(\delta, D)$

- When graph is complete, it reduces to the Wishart distribution.
- It is strong hyper-Markov wrt a graph G.
    1. Formulas available for decomposable graphs.
    2. Decompositions in prime components and separators.
- Finding its mode is fast and accurate using the Iterative Proportional Fitting (IPF) algorithm.
- Sampling is possible using the Bayesian IPF of Piccioni (2000).

Center for Statistics and the Social Sciences

Define the operator from P_G into P_G

$$M_{C,A}K = \begin{pmatrix} A^{-1} + K_{C,V\setminus C}(K_{V\setminus C})^{-1}K_{V\setminus C,C} & K_{C,V\setminus C} \\ K_{V\setminus C,C} & K_{V\setminus C} \end{pmatrix}.$$

which is such that $[(M_{C,A}K)^{-1}]_C = A$. To find the mode of $W_G(\delta, D)$, use IPF with $L = D/(\delta - 2)$:

Step a. Set $K^{r+(0/k)} = K^r$.

Step b. For each $j = 1, \ldots, k$, set $K^{r+(j/k)} = M_{C_j, L_{C_j}} K^{r+((j-1)/k)}$.

Step c. Set $K^{r+1} = K^{r+(k/k)}$.

To sample from $W_G(\delta, D)$, use BIPF. Just replace Step b with:

Step b'. Simulate A from $W_{|C_j|}(\delta, D_{C_j})$ and set $K^{r+(j/k)} = M_{C_j, A^{-1}} K^{r+((j-1)/k)}$.

$$\widehat{I_G(\delta, D)} = h_{\delta,D}(\widehat{K})(2\pi)^{|\mathcal{V}|/2}[\det H_{\delta,D}(\widehat{K})]^{-1/2},$$

where $\widehat{K} \in P_G$ is the mode of $W_G(\delta, D)$, $H_{\delta,D}$ is the Hessian and

$$h_{\delta,D}(K) = -\frac{1}{2}\left[\operatorname{tr}(K^T D) - (\delta - 2)\log(\det K)\right].$$

For $(i,j),(l,m) \in \mathcal{V}$, the $((i,j),(l,m))$ entry of $H_{\delta,D}$ is given by

$$\frac{d^2 h_{\delta,D}(K)}{dK_{ij}dK_{lm}} = -\frac{\delta - 2}{2}\operatorname{tr}\left\{K^{-1}(1_{ij})^0 K^{-1}(1_{lm})^0\right\}.$$

Example: Simulating from the C_5-Wishart
C_5 is the cycle with length five

Need to use the Monte Carlo method of Atay-Kayis and Massam (2005) to estimate the prior normalizing constant.

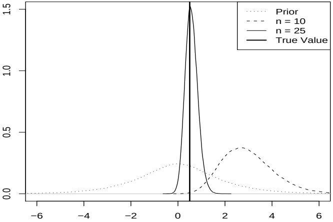


Figure: Marginal distributions of K_{12} based on $10,000$ samples from the G-Wishart prior $W_{C_5}(3, I_5)$ and the G-Wishart posteriors $W_{C_5}(13, D^*_{10})$ (sample size $n = 10$) and $W_{C_5}(28, D^*_{25})$ (sample size $n = 25$). The vertical line $x = 0.5$ shows the true value of K_{12}.

Let $V = \{a, b, c, d\}$, $\mathcal{E}$ all subsets of V and $\mathcal{D}$ all complete subsets of V:

$$\begin{aligned} \mathcal{D} &= \{a, b, c, d, ab, bc, cd, da\}, \\ \mathcal{E} &= \{a, b, c, d, ab, bc, cd, da, ac, bd, abc, bcd, cda, dab, abcd\}. \end{aligned}$$

Take

$$\theta_E = \sum_{F \subseteq E} \log p_F^{(-1)^{|E \setminus F|}} \Leftrightarrow \log p_E = \sum_{F \subseteq E} \theta_F.$$

Distribution of $X = (X_a, X_b, X_c, X_d)$ is Markov wrt to four-cycle means:

$$\theta_E = 0 \text{ for } E \notin \mathcal{D}.$$

which implies:

$$p_{ac} = \frac{p_a p_c}{p_\emptyset}, p_{bd} = \frac{p_b p_d}{p_\emptyset}, p_{abc} = \frac{p_{ab} p_{bc}}{p_b}, p_{bcd} = \frac{p_{bc} p_{cd}}{p_c}, p_{cda} = \frac{p_{cd} p_{da}}{p_d},$$

$$p_{dab} = \frac{p_{da} p_{ab}}{p_a}, p_{abcd} = \frac{p_{ab} p_{bc} p_{cd} p_\emptyset}{p_a p_b p_c p_d}$$

Conjugate Priors for Log-linear Parameters

Diaconis and Ylvisaker, 1979; Massam, Liu and Dobra, 2008

The likelihood for a model G in terms of $(\theta_D, D \in \mathcal{D})$ is:

$$f(y;\theta,G) = \exp\left(\sum_{D\in\mathcal{D}} \theta_D y_D - n\log\left(1+\sum_{E\in\mathcal{E}}\exp\left(\sum_{D\subseteq E, D\in\mathcal{D}}\theta_D\right)\right)\right).$$

The conjugate prior is the generalized hyper Dirichlet which generalizes the hyper Dirichlet of Dawid and Lauritzen (1993):

$$\pi_G(\theta|s,\alpha) = I_G(s,\alpha)^{-1}\exp\left(\sum_{D\in\mathcal{D}} \theta_D s_D - \alpha\log\left(1+\sum_{E\in\mathcal{E}}\exp\left(\sum_{D\subseteq E, D\in\mathcal{D}}\theta_D\right)\right)\right).$$

The posterior of $(\theta_D, D \in \mathcal{D})$ is $\pi_G(y+s, n+\alpha)$. The marginal likelihood of G is:

$$P(Y|G) = I_G(y+s,n+\alpha)/I_G(s,\alpha).$$

Properties of the generalized hyper Dirichlet $\pi_G(\theta|s, \alpha)$

- When model is decomposable, it reduces to the hyper Dirichlet.
- It is strong hyper-Markov wrt a graph G.
  1. Formulas available for decomposable graphs.
  2. Decompositions in prime components and separators.
- Finding its mode is fast and accurate using the Iterative Proportional Fitting (IPF) algorithm.
- Sampling is possible using the Bayesian IPF of Piccioni (2000).

Start with a random choice of $(\theta_D^{(0)}, D \in \mathcal{D})$. For each model generator C_l, $l = 1, 2, \ldots, m$ do:

1. Generate marginals $\tau_{C_l}(D)$, $D \subset C_l$ as independent Gammas with shape $\sum_{D \subseteq F \subseteq C_l} (-1)^{|F \setminus D|}$ and scale $1/\alpha$.
2. Normalize $\tau_{C_l}(D)$, $D \subset C_l$ to obtain marginal tables $p_{C_l}(D)$, $D \subset C_l$.
3. Compute the corresponding $(\theta_l(E), E \subseteq C_l)$:

$$\theta^{k+\frac{l}{m}}(E) = \theta_{k,l}(E \cap C_l) + \sum_{F \subset E, F \in \mathcal{E}_0} (-1)^{|E \setminus F|-1} \log\left(1 + \sum_{L \subseteq C_l^c, L \in \mathcal{E}} \exp\left(\sum_{C \not\subseteq F, C \subseteq F \cup L} \theta^{k+\frac{l-1}{m}}(C)\right)\right).$$

4. Set $\theta^{k+\frac{l}{m}}(E) = 0$ for all $E \notin \mathcal{D}$.

$$\widehat{I_{\mathcal{D}}(s,\alpha)} \approx h_{s,\alpha}(\widehat{\theta}_{\mathcal{D}})(2\pi)^{\frac{d_{\mathcal{D}}}{2}} \det(H_{s,\alpha}(\widehat{\theta}_{\mathcal{D}}))^{-1/2}.$$

The entries of the Hessian are:

$$\frac{d^2 h_{s,\alpha}(\theta_{\mathcal{D}})}{d\theta(i_D)d\theta(l_H)} = -\alpha \sum_{\substack{G\in\mathcal{E}_{\ominus}\\ G\supseteq D}} \sum_{\substack{j_G\in\mathcal{I}_G^*\\ (j_G)_D=i_D}} p(j(G)) \left[\delta_{(j_G)_H}(l_H) - \sum_{\substack{(j_C)_H=l_H\\ C\in\mathcal{E}_{\ominus}, j_C\in\mathcal{I}_C^*}} p(j(C))\right].$$

where

$$\delta_{(j_G)_H}(l_H) = \begin{cases} 1, & \text{if } (j_G)_H = l_H, \\ 0, & \text{otherwise.} \end{cases}$$

Bayesian Model Choice

Candidate models: $\{\mathcal{M}_m, m = 1, \ldots, M\}$. Models are connected through their neighborhoods. Perform model selection using the posterior model probabilities:

$$\{p(\mathcal{M}_m | D), m = 1, \ldots, M\}.$$

Possible decisions:

1. Select the best model $\mathcal{M}_{m*}$ with the highest posterior probability.
2. Average across all models.
3. Average across a reduced set of models:

$$\mathcal{M}(c) = \{\mathcal{M}_m : p(\mathcal{M}_{m*}|D) \geq c \cdot p(\mathcal{M}_m|D)\}.$$

As $n \to \infty$ and M is fixed, $\mathcal{M}(c) \to \{\mathcal{M}_{m*}\}$. However, as $M \to \infty$ and n is fixed, $p(\mathcal{M}(c)|D) \to 0$.

Center for Statistics and the Social Sciences

The Mode Oriented Stochastic Search (MOSS)

The precursor of MOSS is the Shotgun Stochastic Search (SSS) algorithm (Jones et al., 2005; Hans et al., 2007).

MOSS(c)

Let $\mathcal{S}$ be the models visited so far and $\mathcal{L}$ be the unexplored models. Do:
Step (A). Sample a model $\mathcal{M}_j \in \mathcal{L}$ with probabilities proportional with $p(\mathcal{M}_j|D)$. Mark $\mathcal{M}_j$ as explored.
Step (B). Include in $\mathcal{S}$ all the neighbors of $\mathcal{M}_j$.
Step (C). If $\mathcal{L}$ is empty, output $\mathcal{S}(c)$ and STOP. Otherwise go to (A).

Theorem

At each iteration, the probability that MOSS finds $\mathcal{M}_{m*}$ is greater than the probability that any Markov chain algorithm finds $\mathcal{M}_{m*}$.

Example: Efficiency of MOSS

Experiment: Simulate 50 samples from a decomposable graph with 25 vertices. Only 10 vertices are linked with edges (Scott & Carvalho, 2008).

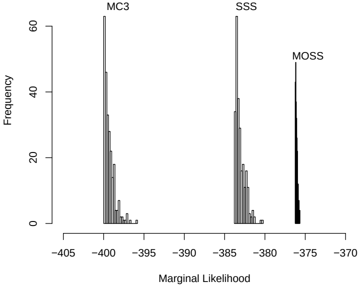


Figure: Distribution of the top 250 marginal likelihoods returned by MOSS, SSS and MC^3 algorithms after evaluating the same number of models and starting at the same randomly generated graph.

GGMs Simulated Study: Yuan and Lin (2007)

SEE PAGE 3 OF THE HANDOUT

- Comparison of MOSS, Yuan and Lin (2007), Meinshausen and Bühlmann (2006), Drton and Perlman (2004).
- Experiment: simulate 25 samples of dimension $p = 5$ and $p = 10$ from eight different models: AR(1), AR(2), AR(3), AR(4), a full graph, a star graph with every vertex conected to the first vertex and a circle graph. Repeat 100 times.
- Assess performance using the average Kullback-Leibler (KL) loss across the replicates; number of false positive and false negative edges.
- Conclusion: MOSS does consistently better than the other three approaches.

Example: Modeling growth determinant uncertainty using GGMs

SEE PAGE 4 OF THE HANDOUT

- Dataset with 41 potential growth determinants from Fernandez et al. (2001).
- Economists hypothesized the existence of seven growth determinants.
- Previous studies based on linear regressions found between 2 and 22 predictors (Theo Eicher, Mark Steel, etc).
- With the same prior specification, our results show:
    1. Linear regressions: 17 growth determinants.
    2. GGMs: seven (relevant) and one (marginally relevant) growth determinants.

Example: Household Study in Rochdale
Source: Whittaker (1990) page 279

Eight dichotomous variables relating women's economic activity and husband's unemployment in Rochdale:

1. A, wife economically active (no,yes)
2. B, age of wife > 38 (no,yes)
3. C, husband unemployed (no,yes)
4. D, child ≤ 4 (no,yes)
5. E, wife's education, high-school$+$ (no,yes)
6. F, husband's education, high-school$+$ (no,yes)
7. G, asian origin (no,yes)
8. H, other household member working (no,yes).

Center for Statistics and the Social Sciences

Example: Household study in Rochdale
Source: Whittaker (1990) page 279

Sparse table with 665 individuals cross-classified in 256 cells, 165 counts of zero, 217 counts ≤ 3 and a few large counts ≥ 30.

5	0	2	1	5	1	0	0	4	1	0	0	6	0	2	0
8	0	11	0	13	0	1	0	3	0	1	0	26	0	1	0
5	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0
4	0	8	2	6	0	1	0	1	0	1	0	0	0	1	0
17	10	1	1	16	7	0	0	0	2	0	0	10	6	0	0
1	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0
4	7	3	1	1	1	2	0	1	0	0	0	1	0	0	0
0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
18	3	2	0	23	4	0	0	22	2	0	0	57	3	0	0
5	1	0	0	11	0	1	0	11	0	0	0	29	2	1	1
3	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	25	0	1	37	26	0	0	15	10	0	0	43	22	0	0
0	0	0	0	2	0	0	0	0	0	0	0	3	0	0	0
2	4	0	0	2	1	0	0	0	1	0	0	2	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

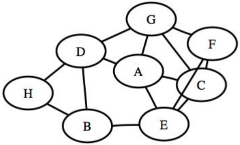
EXAMPLE: HOUSEHOLD STUDY IN ROCHDALE
SOURCE: WHITTAKER (1990)

> *"[...] it is impossible to detect many high order interactions, and one should hesitate to fit the saturated log-linear model [...] However we may fit the all two-way interactions model, because the sufficient statistics are the two-way marginal tables and the entries in these tables are quite respectable. [...] Here, we adopt the quick model selection method of selecting interactions for which the square of the standardized parameter estimate exceeds* 3.84*."*

Based on this heuristic, Joe arrives at the hierarchical model

$$[FG][EF][DH][DG][CG][CF][CE][BH][BE][BD][AG][AE][AD][AC].$$

Total number of possible hierarchical models: 5.6×10^{22}.



Example: Household study in Rochdale

See page 2 of the handout

Joe Whittaker's analysis determined:

$$[FG][EF][DH][DG][CG][CF][CE][BH][BE][BD][AG][AE][AD][AC].$$

Best decomposable graphical model determined by MOSS:

$$[EFG][BEG][BDH][BDG][ADG][ACG].$$

Best graphical model determined by MOSS (out of 2^{28} possible models):

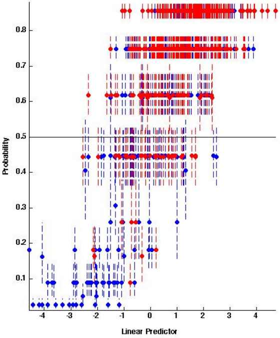
$$[FG][EF][BE][BDH][BDG][ADG][ACG][ACE].$$

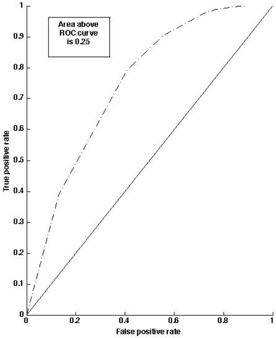
Best hierarchical model determined by MOSS (out of 5.6×10^{22} possible models):

$$[FG][EF][DG][CG][CF][CE][BE][BDH][AG][AE][AD][AC].$$

Markov blanchet of A is C, D, E, G. MOSS determines best hierarchical model:

$$[FG][EF][DG][CG][CF][CE][BE][BDH]\color{red}{[AG][AE][AD][AC]}.$$





Whittaker (1990) estimates logistic regression as:

$$\log \frac{p(a=1|c,d,e,g)}{p(a=0|c,d,e,g)} \quad = \quad \text{const.} - 1.33c - 1.32d + 0.69e - 2.17g,$$

with standard errors 0.3, 0.21, 0.2, 0.47. We estimate the same regression equation to be:

$$\log \frac{p(a=1|c,d,e,g)}{p(a=0|c,d,e,g)} \quad = \quad \text{const.} - 1.30c - 1.26d + 0.70e - 2.31g,$$

with standard errors 0.29, 0.2, 0.19 and 0.47.

Multivariate Regressions

Covariates grouped as responses Y and explanatory X. Possibly X is much bigger than Y. We are interested in learning $p(Y|X)$ and not the joint $p(Y, X)$.

Theorem

(Whittaker, 1990) The conditional independence relationships from $p(Y|X)$ are embedded in graphs having complete subgraphs associated with X.

Example: Genome-wide Analysis of Estrogen Response with Dense SNP Array Data
Source: Dobra et al. (2008)

60 cell lines from NCI used to study resistance to estrogen response (Jarjanazi et al., 2008):

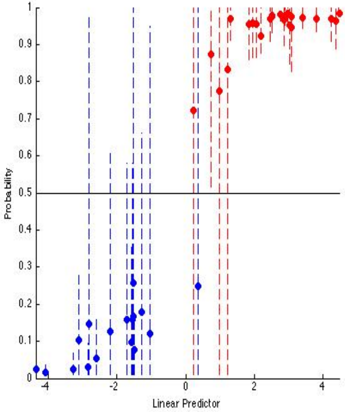
- 25 cell lines were *resistant*.
- 17 cell lines were *sensitive*.

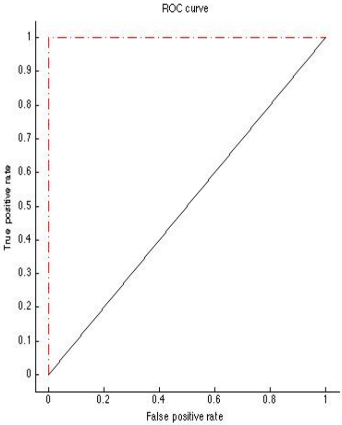
Genotypes of SNPs in these 42 cell lines were obtained from the Affymetrix 125K chip data – only $25,530$ SNPs were retained. A segregating SNP site has three possible genotypes: $0/0$, $0/1$ and $1/1$.

The data is a 2×3^{25530} contigency table with 42 samples.

Example: Genome-wide Analysis of Estrogen Response with Dense SNP Array Data

MOSS selects 17 SNPs that appear in regressions with at most 3 variables. Total number of such regressions: 2.77×10^{12}. Mean number of models evaluated by MOSS: $2,407,299$.





Some Concluding Remarks

Papers and code available from my website:

http://www.stat.washington.edu/adobra/