Learning Bayesian networks given a data set consisting of samples that are not independent and identically distributed

ABNMS 2011

Jessica Kasza, Gary Glonek, Patty Solomon

November 23 2011

THE UNIVERSITY *of* ADELAIDE

The Problem

- Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T \sim N_p(\mathbf{0}, \Sigma)$
- Learning the structure of the Bayesian network of $\boldsymbol{X}$ usually requires *n* **iid** samples
- What if we have a more complex data set?
  - non-independent samples;
  - additional components of variance;
  - data on exogenous variables thought to affect $\boldsymbol{X}$.

Learning Graphical Structure

- A Bayesian network $B = (G, \Theta)$ for a random vector $\boldsymbol{X}$ consists of two components:
    - a directed acyclic graph $G = (V, E)$, $V = \{1, 2, \dots, p\}$,
    - conditional densities for each random variable, $f(x_i|\boldsymbol{x}_{P_i}, \theta_i)$, where P_i is the set of parents of i in G, θ_i the parameters.

  The graph and conditional densities specify a joint density function for $\boldsymbol{X}$:

$$f\left(\boldsymbol{x}|\Theta\right) = \prod_{i=1}^{p} f\left(x_i|\boldsymbol{x}_{P_i}, \theta_i\right).$$

- Want to learn G given a data set $d = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_p\}$, $\boldsymbol{x}_i = (x_{i1}, \dots, x_{in})$
    - A popular approach for learning about genetic regulatory networks.

Score-based Approach

- How well G describes the data is quantified by a score metric, $S(G|d)$.
  - Score we consider is:

$$S(G|d) = p(G)p(d|G) = p(G)\int p(d|G,\Theta)p(\Theta|G)d\Theta.$$

    The BGe score of Geiger and Heckerman (1994).
- Need to specify:
  - $p(G)$: prior on space of DAGs;
  - $p(d|G,\Theta) = \prod_{i=1}^{p} f(\boldsymbol{x}_i|\boldsymbol{x}_{P_i},\theta_i)$;
  - $p(\Theta|G)$: prior for the parameters.

IID samples

When *d* consists of iid samples:

$$\boldsymbol{x}_i | \boldsymbol{x}_{P_i}, \gamma_i, \psi_i \sim N_n(\boldsymbol{x}_{P_i}\gamma_i, \psi_i I_n).$$

Parameter priors:

$$\begin{aligned} \gamma_i | \psi_i & \sim N_{|P_i|}\left(0, \frac{\psi_i}{\tau} I\right), \\ \psi_i^{-1} & \sim Ga\left(\frac{\delta + |P_i|}{2}, \frac{\tau}{2}\right). \end{aligned}$$

These priors chosen to give an **equivalent** score, Geiger and Heckerman (2002).

IID samples

When *d* consists of iid samples:

$$\boldsymbol{x}_i|\boldsymbol{x}_{P_i},\gamma_i,\psi_i \sim N_n(\boldsymbol{x}_{P_i}\gamma_i,\psi_i I_n).$$

Parameter priors:

$$\begin{aligned}\gamma_i|\psi_i &\sim N_{|P_i|}\left(0,\tfrac{\psi_i}{\tau}I\right),\\ \psi_i^{-1} &\sim Ga\left(\tfrac{\delta+|P_i|}{2},\tfrac{\tau}{2}\right).\end{aligned}$$

These priors chosen to give an **equivalent** score, Geiger and Heckerman (2002).

$$\begin{aligned}S(G|d) &= p(G)\prod_{i=1}^{p} f(\boldsymbol{x}_i|\boldsymbol{x}_{P_i}),\\ f(\boldsymbol{x}_i|\boldsymbol{x}_{P_i}) &= \int_{\mathbb{R}^{|P_i|}\times(0,\infty)} f(\boldsymbol{x}_i|\boldsymbol{x}_{P_i},\gamma_i,\psi_i)f(\gamma_i,\psi_i)d\gamma_i d\psi_i.\end{aligned}$$

Algorithms for exploring DAG space

$S(G|d)$ used in conjunction with algorithms for exploring the DAG space.

- Greedy hill climbing,
- High-dimensional Bayesian covariance selection, Dobra *et al.* (2004)

What if samples are not IID?

Grape gene example:

- Learn about the relationships of grape **heat shock** genes
- Grapes sampled from 3 geographically distinct vineyards
- Temperatures at times leading up to picking of grapes available.

What if samples are not IID?

Grape gene example:

- Learn about the relationships of grape **heat shock** genes
- Grapes sampled from 3 geographically distinct vineyards
- Temperatures at times leading up to picking of grapes available.

Must account for effects of exogenous variables!

- Now have

$$\boldsymbol{x}_i|\boldsymbol{x}_{P_i}, \gamma_i, b_i, \psi_i \sim N_n(\boldsymbol{x}_{P_i}\gamma_i + Qb_i, \psi_i I_n)$$

where

$$Q = \left(\boldsymbol{q}_1|\cdots|\boldsymbol{q}_m\right).$$

Dealing with b_i: $\boldsymbol{x}_i | \boldsymbol{x}_{P_i}, \gamma_i, b_i, \psi_i \sim N_n(\boldsymbol{x}_{P_i}\gamma_i + Qb_i, \psi_i I_n)$

- Ignore b_i!

Dealing with b_i: $\boldsymbol{x}_i | \boldsymbol{x}_{P_i}, \gamma_i, b_i, \psi_i \sim N_n(\boldsymbol{x}_{P_i}\gamma_i + Qb_i, \psi_i I_n)$

- Ignore b_i! Not a good idea.

Dealing with b_i: $\boldsymbol{x}_i|\boldsymbol{x}_{P_i}, \gamma_i, b_i, \psi_i \sim N_n(\boldsymbol{x}_{P_i}\gamma_i + Qb_i, \psi_i I_n)$

- Ignore b_i! Not a good idea.
- **Bayesian approach**: place a prior on b_i:

$$b_i|\phi_i \sim N_m(0, \phi_i I)$$

In addition to previously used priors for γ_i and ψ_i.

Dealing with b_i: $\boldsymbol{x}_i | \boldsymbol{x}_{P_i}, \gamma_i, b_i, \psi_i \sim N_n(\boldsymbol{x}_{P_i}\gamma_i + Qb_i, \psi_i I_n)$

- Ignore b_i! Not a good idea.
- **Bayesian approach**: place a prior on b_i:

$$b_i | \phi_i \sim N_m(0, \phi_i I)$$

In addition to previously used priors for γ_i and ψ_i.
Best prior choice is

$$b_i | \psi_i \sim N_m(0, \upsilon^{-1} \psi_i I)$$

$$S_B(G|d) = p(G) \prod_{i=1}^{p} f_{\upsilon}(\boldsymbol{x}_i | \boldsymbol{x}_{P_i}).$$

Dealing with b_i: $\boldsymbol{x}_i|\boldsymbol{x}_{P_i}, \gamma_i, b_i, \psi_i \sim N_n(\boldsymbol{x}_{P_i}\gamma_i + Qb_i, \psi_i I_n)$

- **Residual approach**: remove random effects by analysing residuals: $n \times (n-m)$ matrix P:

$$P^T Q = 0, \quad P^T P = I_{n-m}, \quad PP^T = I_n - Q(Q^T Q)^{-1} Q^T.$$

Then

$$P^T \boldsymbol{x}_i | P^T \boldsymbol{x}_{P_i}, \gamma_i, \psi_i \sim N_{n-m}\left(P^T \boldsymbol{x}_{P_i}\gamma_i, \psi_i I\right).$$

Using previous priors for γ_i and ψ_i, get score metric

$$S_R(G|d) = p(G) \prod_{i=1}^{p} f_R(P^T \boldsymbol{x}_i | P^T \boldsymbol{x}_{P_i}).$$

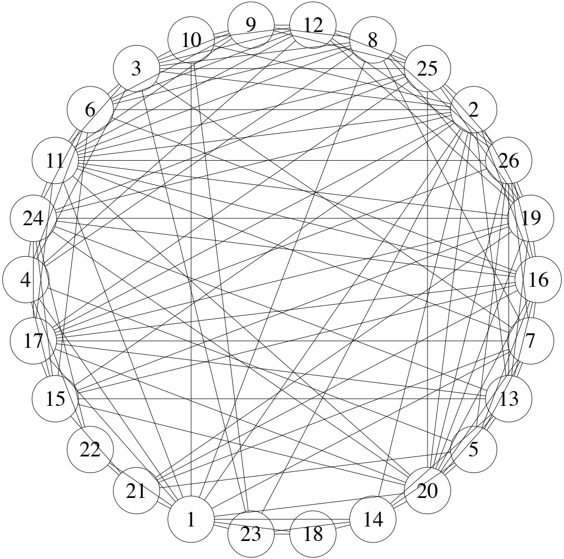
Grape Gene Example

- $n = 50$ samples of $p = 26$ grape berry genes;
- Grape berries sampled from 3 vineyards
- Genes are heat shock genes - and we have temperature measurements.
- Assume the following model for each gene:

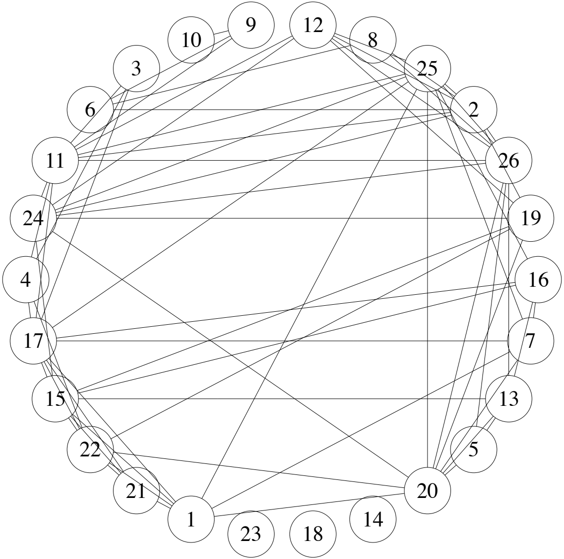
$$\begin{aligned} X_{ij} &= \sum_{l \in P_i} \gamma_{il} X_{lj} + \sum_{r=1}^{m} q_{rj} b_{ir} + \epsilon_{ij}, \;\; \epsilon_{ij} \sim N(0, \psi_i), \\ \gamma_{il} &\sim N(0, \tau^{-1} \psi_i), \\ \psi_i &\sim \text{Inverse Gamma}\left(\frac{\delta + |P_i|}{2}, \frac{\tau}{2} \right). \end{aligned}$$

Use the residual approach to account for b_i. (Is this a good idea?)

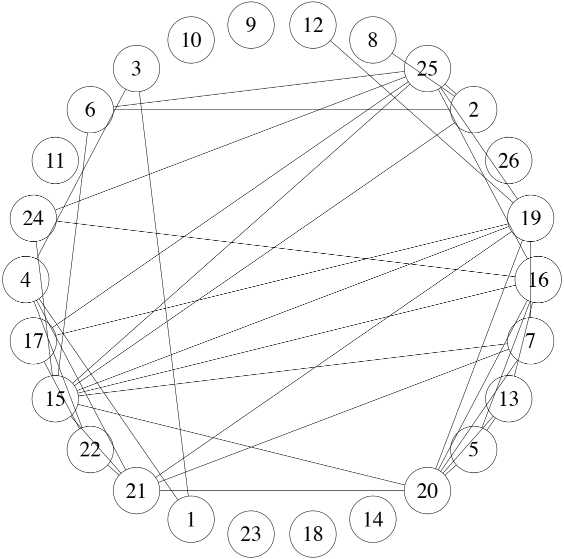
Ignoring effects of vineyard and temperature



Residual approach, vineyard effects



Residual approach, vineyard and temperature effects



Grape Gene Graphs

- As more variation due to exogenous sources is accounted for, graphs become sparser
- Genes 14, 18, 23: disconnected from rest of graph in last two graphs
  - Expressions of these genes have very low ses - no variation to be explained by relationships with other genes!
- Genes 9, 10, 11: correspond to HSP 81, early response to dehydration
  - Role not very well understood, our analysis indicates they are not implicated in heat shock gene network.

Comparing Bayesian and residual approaches

- Should we have used the Bayesian approach in the grape gene example?
  - Residual approach is easier to use;
  - May obtain less information about γ_i, ψ_i:
    - May be important for posterior estimation of parameters.

Comparing Bayesian and residual approaches

- Should we have used the Bayesian approach in the grape gene example?
  - Residual approach is easier to use;
  - May obtain less information about γ_i, ψ_i:
    - May be important for posterior estimation of parameters.
- Full Bayesian approach posterior: $f_B(\gamma_i, \psi_i | \boldsymbol{x}_i, \boldsymbol{x}_{P_i})$.
  Residual approach posterior: $f_R(\gamma_i, \psi_i | \boldsymbol{x}_i, \boldsymbol{x}_{P_i})$.
- We consider the Kullback Leibler divergence:

$$D\left\{f_B, f_R\right\} = \int_{\mathbb{R}^{|P_i|}} \int_0^\infty f_B \log\left(\frac{f_B}{f_R}\right) d\psi_i d\gamma_i.$$

Divergence for marginal covariance matrix Σ

$$var(\boldsymbol{X}|\{\gamma_i, \psi_i\}_{i=1,\dots,p}) = \Sigma$$

- If the true graphical structure of $\boldsymbol{X}$ is known:

$$\begin{aligned} D_\Sigma &\{f_B(\Sigma|\boldsymbol{X}), f_R(\Sigma|\boldsymbol{X})\} \\ &= \sum_{i=1}^{p} D\left\{f_B(\gamma_i, \psi_i|\boldsymbol{x}_i, \boldsymbol{x}_{P_i}), f_R(\gamma_i, \psi_i|\boldsymbol{x}_i, \boldsymbol{x}_{P_i})\right\}. \end{aligned}$$

Divergence for marginal covariance matrix Σ

$$var(\boldsymbol{X}|\{\gamma_i, \psi_i\}_{i=1,\ldots,p}) = \Sigma$$

- If the true graphical structure of $\boldsymbol{X}$ is known:

$$D_\Sigma \left\{ f_B(\Sigma|\boldsymbol{X}), f_R(\Sigma|\boldsymbol{X}) \right\} \\ = \sum_{i=1}^{p} D\left\{ f_B(\gamma_i, \psi_i|\boldsymbol{x}_i, \boldsymbol{x}_{P_i}), f_R(\gamma_i, \psi_i|\boldsymbol{x}_i, \boldsymbol{x}_{P_i}) \right\}.$$

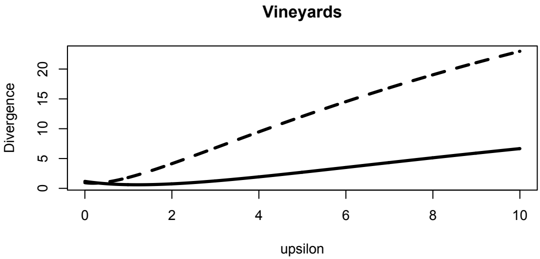
- Divergence for the empty graph:

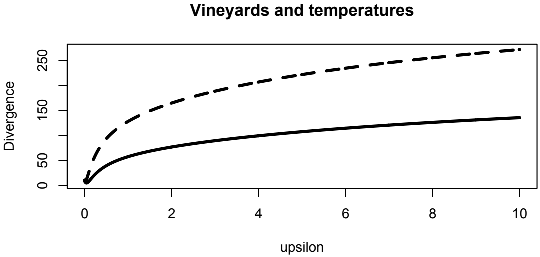
$$D_\Sigma^e = \sum_{i=1}^{p} D\left\{ f_B(\gamma_i, \psi_i|\boldsymbol{x}_i), f_R(\gamma_i, \psi_i|\boldsymbol{x}_i) \right\},$$

- Divergence for an arbitrary full graph:

$$D_\Sigma^f = \sum_{i=1}^{p} D\left\{ f_B(\gamma_i, \psi_i|\boldsymbol{x}_i, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}), f_R(\gamma_i, \psi_i|\boldsymbol{x}_i, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}) \right\}.$$

Divergence in Grape Gene Example





Conclusions and Further Work

- The Bayesian and residual score metrics extend the utility of score-based methods for learning networks to situations where the data does not consist of iid samples.
- Provided sample size is not too small, residual approach is a useful alternative to Bayesian approach
    - Even when the assumptions of the Bayesian approach are valid.

  Some questions:

    - What happens when the chosen prior distribution of the effects of exogenous variables is not suitable?
    - Are there situations where the residual approach performs better than the Bayesian approach?