Learning Markov Equivalence Classes of Directed Acyclic Graphs: an Objective Bayesian Approach

Guido Consonni

Department of Statistical Sciences
Università Cattolica del Sacro Cuore, Milan

(Joint work with Federico Castelletti, Marco Della Vedova and Stefano Peluso)

University of Helsinki
30 October 2017
Outline

1. Graphical models
2. Objective Bayes model comparison
3. Gaussian essential graphs
4. MCMC Algorithm
5. Experiments
6. Summary
Problem

- $q$ variables
Graphical models

Problem

- $q$ variables
- Aim: discover conditional independence relationships among such variables.
Problem

- $q$ variables
- Aim: discover conditional independence relationships among such variables.
- Data: $(y_1, \ldots, y_n)$
Problem

- $q$ variables
- Aim: discover conditional independence relationships among such variables.
- Data: $(y_1, \ldots, y_n)$
- $y_i = (y_{i,1}, \ldots, y_{i,q})^\top$ $i = 1, \ldots, n$
Problem

- $q$ variables
- Aim: discover conditional independence relationships among such variables.
- Data: $(y_1, \ldots, y_n)$
- $y_i = (y_{i,1}, \ldots, y_{i,q})^\top$ $i = 1, \ldots, n$
- $y_i$s are i.i.d. from a parametric family of sampling distributions
Graphical models

Background and notation

- Graph $\mathcal{G} = (V, E)$
  - $V = \{1, \ldots, q\}$ set of nodes
  - $E \subseteq V \times V$ set of edges
Background and notation

- Graph \( G = (V, E) \)
  - \( V = \{1, \ldots, q\} \) set of nodes
  - \( E \subseteq V \times V \) set of edges
- \( (u, v) \in E \iff u \rightarrow v \)
Background and notation

- Graph \( \mathcal{G} = (V, E) \)
  - \( V = \{1, \ldots, q\} \) set of nodes
  - \( E \subseteq V \times V \) set of edges
- \((u, v) \in E \iff u \rightarrow v\)
- \(\{(u, v), (v, u)\} \in E \iff u - v\)
Background and notation

- **Graph** \( \mathcal{G} = (V, E) \)
  - \( V = \{1, \ldots, q\} \) set of nodes
  - \( E \subseteq V \times V \) set of edges
- \( (u, v) \in E \iff u \rightarrow v \)
- \( \{(u, v), (v, u)\} \in E \iff u - v \)
- Undirected graph (UG) graph with only directed edges
- Directed Acyclic Graph (DAG) graph with only directed edges (no cycles)
  - If the distributions of \( y_i \) exhibits conditional independencies determined by \( \mathcal{D} \) (Markov wrt \( \mathcal{D} \)), then
    \[
    f_{\mathcal{D}}(y_i) = \prod_{j \in V} f(y_{i,j} \mid y_{i,\text{pa}_\mathcal{D}(j)})
    \]
  - \( \text{pa}_\mathcal{D}(j) \) parents of node \( j \) in \( \mathcal{D} \)
Background and notation

- DAGs encoding the same conditional independencies are called Markov equivalent.
- Theorem (Verma & Pearl, 1991)
  - Two DAGs $D_1$ and $D_2$ are Markov equivalent if and only if they have the same skeleton and the same v-structures.
- **skeleton**: the underlying undirected graph
- **v-structure**: $u \rightarrow w \leftarrow z$

Example: three Markov equivalent DAGs

```
1 3 4
2 1 3 4
2 1 3 4

1 3
2 1
3 4
```

Guido Consonni

OBAYES Learning of MECs of DAGs

Helsinki, 30 October 2017
Background and notation

- DAGs encoding the same conditional independencies are called *Markov equivalent*.

Theorem (Verma & Pearl, 1991)

Two DAGs $D_1$ and $D_2$ are Markov equivalent if and only if they have the same skeleton and the same v-structures.

- **skeleton**: the underlying undirected graph
- **v-structure**: $u \rightarrow w \leftarrow z$
Background and notation

- DAGs encoding the same conditional independencies are called Markov equivalent

Theorem (Verma & Pearl, 1991)

Two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent if and only if they have the same skeleton and the same $v$-structures.
Background and notation

- DAGs encoding the same conditional independencies are called Markov equivalent.

Theorem (Verma & Pearl, 1991)

Two DAGs $D_1$ and $D_2$ are Markov equivalent if and only if they have the same skeleton and the same $v$-structures.

- skeleton: the underlying undirected graph
Background and notation

- DAGs encoding the same conditional independencies are called *Markov equivalent*

**Theorem (Verma & Pearl, 1991)**

*Two DAGs $\mathcal{D}_1$ and $\mathcal{D}_2$ are Markov equivalent if and only if they have the same skeleton and the same v-structures.*

- *skeleton*: the underlying undirected graph
- *v-structure*: $u \rightarrow w \leftarrow z$
Graphical models

Background and notation

- DAGs encoding the same conditional independencies are called *Markov equivalent*

**Theorem (Verma & Pearl, 1991)**

*Two DAGs $D_1$ and $D_2$ are Markov equivalent if and only if they have the same skeleton and the same $v$-structures.*

- *skeleton*: the underlying undirected graph
- *$v$-structure*: $u \rightarrow w \leftarrow z$
- Example: three Markov equivalent DAGs

![DAGs $D_1$, $D_2$, and $D_3$](image-url)
Essential graph

- *Essential Graph* (EG) $\mathcal{G}$, also called *Completed Partially Directed Graph* (CPDAG) is the union (over the edge sets) of Markov equivalent DAGs and represents a Markov equivalence class.

\[ \mathcal{G} \]

\[ \mathcal{D}_1 \]

\[ \mathcal{D}_2 \]

\[ \mathcal{D}_3 \]
- Essential Graph (EG) $G$, also called Completed Partially Directed Graph (CPDAG) is the union (over the edge sets) of Markov equivalent DAGs and represents a Markov equivalence class.

$G$ is a chain graph:
- both directed and undirected edges
- chain component: a set of nodes joined by an undirected path
- $\mathcal{T}$: set of chain components of $G$
- $\mathcal{T} = \{\{1, 2, 3\}, \{4\}\}$
Essential graph (Characterization Theorem)

Theorem (Andersson et al., 1997)

A graph $\mathcal{G} = (V, E)$ is the EG for some DAG $\mathcal{D}$ with vertex set $V$ if and only if $\mathcal{G}$ satisfies the following four conditions:

- $\mathcal{G}$ is a CG;
- for each chain component $\tau \in \mathcal{T}$ the subgraph $\mathcal{G}_\tau$ is a decomposable UG;
- $\mathcal{G}$ has no flags (no induced subgraphs of the form $u \rightarrow v \leftarrow z$);
- each directed edge $u \rightarrow v$ contained in $\mathcal{G}$ is strongly protected.
Bayesian model comparison

- Bayesian model $\mathcal{M}_k$
Bayesian model comparison

- Bayesian model $M_k$
  - $f_{M_k}(Y | \theta_k)$, a family of sampling densities indexed by a model specific parameter $\theta_k$
Bayesian model comparison

- Bayesian model $\mathcal{M}_k$
  - $f_{\mathcal{M}_k}(Y | \theta_k)$, a family of sampling densities indexed by a model specific parameter $\theta_k$
  - $p(\theta_k)$, a prior density on $\theta_k$, assumed to be proper
Bayesian model comparison

- Bayesian model $\mathcal{M}_k$
  - $f_{\mathcal{M}_k}(Y | \theta_k)$, a family of sampling densities indexed by a model specific parameter $\theta_k$
  - $p(\theta_k)$, a prior density on $\theta_k$, assumed to be \textit{proper}
- $\mathcal{M}_1, \ldots, \mathcal{M}_K$ a collection of Bayesian models for the data matrix $Y$
Bayesian model comparison

- Bayesian model $\mathcal{M}_k$
  - $f_{\mathcal{M}_k}(Y \mid \theta_k)$, a family of sampling densities indexed by a model specific parameter $\theta_k$
  - $p(\theta_k)$, a prior density on $\theta_k$, assumed to be *proper*
- $\mathcal{M}_1, \ldots, \mathcal{M}_K$ a collection of Bayesian models for the data matrix $Y$
- Goal: comparing models
Bayesian model comparison

- Bayesian model \( \mathcal{M}_k \)
  - \( f_{\mathcal{M}_k}(Y | \theta_k) \), a family of sampling densities indexed by a model specific parameter \( \theta_k \)
  - \( p(\theta_k) \), a prior density on \( \theta_k \), assumed to be \textit{proper}

- \( \mathcal{M}_1, \ldots, \mathcal{M}_K \) a collection of Bayesian models for the data matrix \( Y \)

- Goal: comparing models

- Tool: marginal likelihood of \( \mathcal{M}_k \) (and Bayes factor)

\[
m_{\mathcal{M}_k}(Y) = \int f_{\mathcal{M}_k}(Y | \theta_k)p(\theta_k)d\theta_k
\]
Bayesian model comparison

- Bayesian model $\mathcal{M}_k$
  - $f_{\mathcal{M}_k}(Y | \theta_k)$, a family of sampling densities indexed by a model specific parameter $\theta_k$
  - $p(\theta_k)$, a prior density on $\theta_k$, assumed to be proper

- $\mathcal{M}_1, \ldots, \mathcal{M}_K$ a collection of Bayesian models for the data matrix $Y$

- Goal: comparing models

- Tool: marginal likelihood of $\mathcal{M}_k$ (and Bayes factor)

$$m_{\mathcal{M}_k}(Y) = \int f_{\mathcal{M}_k}(Y | \theta_k) p(\theta_k) d\theta_k$$

- If no/weak prior information set $p(\theta_k) = p^D(\theta_k)$ (default objective parameter prior)

- Problems with objective priors
Bayesian model comparison

- Bayesian model $\mathcal{M}_k$
  - $f_{\mathcal{M}_k}(Y | \theta_k)$, a family of sampling densities indexed by a model specific parameter $\theta_k$
  - $p(\theta_k)$, a prior density on $\theta_k$, assumed to be *proper*
- $\mathcal{M}_1, \ldots, \mathcal{M}_K$ a collection of Bayesian models for the data matrix $Y$
- Goal: comparing models
- Tool: marginal likelihood of $\mathcal{M}_k$ (and Bayes factor)

$$m_{\mathcal{M}_k}(Y) = \int f_{\mathcal{M}_k}(Y | \theta_k)p(\theta_k)d\theta_k$$

- If no/weak prior information set $p(\theta_k) = p^D(\theta_k)$ (default *objective parameter prior*)
- Problems with objective priors
  - often improper
Bayesian model comparison

- Bayesian model $\mathcal{M}_k$
  - $f_{\mathcal{M}_k}(Y | \theta_k)$, a family of sampling densities indexed by a model specific parameter $\theta_k$
  - $p(\theta_k)$, a prior density on $\theta_k$, assumed to be proper

- $\mathcal{M}_1, \ldots, \mathcal{M}_K$ a collection of Bayesian models for the data matrix $Y$

- Goal: comparing models

- Tool: marginal likelihood of $\mathcal{M}_k$ (and Bayes factor)

$$m_{\mathcal{M}_k}(Y) = \int f_{\mathcal{M}_k}(Y | \theta_k)p(\theta_k)d\theta_k$$

- If no/weak prior information set $p(\theta_k) = p^D(\theta_k)$ (default objective parameter prior)

- Problems with objective priors
  - often improper
  - defined up to an arbitrary constant

Guido Consonni

OBAYES Learning of MECs of DAGs

Helsinki, 30 October 2017
Bayesian model comparison

- Bayesian model $\mathcal{M}_k$
  - $f_{\mathcal{M}_k}(Y | \theta_k)$, a family of sampling densities indexed by a model specific parameter $\theta_k$
  - $p(\theta_k)$, a prior density on $\theta_k$, assumed to be proper
- $\mathcal{M}_1, \ldots, \mathcal{M}_K$ a collection of Bayesian models for the data matrix $Y$
- Goal: comparing models
- Tool: marginal likelihood of $\mathcal{M}_k$ (and Bayes factor)

\[
m_{\mathcal{M}_k}(Y) = \int f_{\mathcal{M}_k}(Y | \theta_k)p(\theta_k)d\theta_k
\]

- If no/weak prior information set $p(\theta_k) = p^D(\theta_k)$ (default objective parameter prior)
- Problems with objective priors
  - often improper
  - defined up to an arbitrary constant
  - cannot be naively used to compute marginal likelihoods
Fractional marginal likelihood

- Fractional Bayes factor (O’Hagan, 1995)

\[ b = \frac{b}{n}, \quad 0 < b < 1, \]  

a fraction of the sample size \( n \).

\[ f_{\text{marginal likelihood}}(\mathcal{M}_k) = \int f_{\text{model } \mathcal{M}_k}(Y|\theta_k) p_D(\theta_k) d\theta_k \int f_{\text{model } \mathcal{M}_k}(Y|\theta_k) p_D(\theta_k) d\theta_k - f_{\text{model } \mathcal{M}_k}(Y|\theta_k) = \left[ f_{\text{model } \mathcal{M}_k}(Y|\theta_k) \right]^{1-b}_{b}. \]

- Implied fractional prior:

\[ p_F(\theta_k|b, Y) \propto f_{\text{model } \mathcal{M}_k}(Y|\theta_k) p_D(\theta_k) \]

- Fractional likelihood is a discounted full likelihood

- Default choice: \( b = \frac{n_0}{n}, n_0 \) minimal (integer) training sample size which makes the induced fractional prior proper
Fractional marginal likelihood

- Fractional Bayes factor (O’Hagan, 1995)
- \( b = b(n), \) 0 < \( b < 1 \), a fraction of the sample size \( n \).
Fractional marginal likelihood

- Fractional Bayes factor (O’Hagan, 1995)
- \( b = b(n), \ 0 < b < 1, \) a fraction of the sample size \( n. \)
- fractional marginal likelihood of model \( M_k: \)

\[
m_{M_k}(Y; b) = \frac{\int f_{M_k}(Y | \theta_k)p^D(\theta_k)d\theta_k}{\int f^b_{M_k}(Y | \theta_k)p^D(\theta_k)d\theta_k}
\]
Fractional marginal likelihood

- Fractional Bayes factor (O’Hagan, 1995)
- \( b = b(n), 0 < b < 1 \), a fraction of the sample size \( n \).
- *Fractional marginal likelihood* of model \( \mathcal{M}_k \):

\[
m_{\mathcal{M}_k}(Y; b) = \frac{\int f_{\mathcal{M}_k}(Y | \theta_k)p_D(\theta_k)d\theta_k}{\int f_{\mathcal{M}_k}^b(Y | \theta_k)p_D(\theta_k)d\theta_k}
\]

- \( f^b_{\mathcal{M}_k}(Y | \theta_k) = \left[f_{\mathcal{M}_k}(Y | \theta_k)\right]^b \)
Fractional marginal likelihood

- Fractional Bayes factor (O’Hagan, 1995)
- \( b = b(n), \ 0 < b < 1, \) a fraction of the sample size \( n. \)
- *fractional marginal likelihood* of model \( \mathcal{M}_k:*

\[
m_{\mathcal{M}_k}(Y; b) = \frac{\int f_{\mathcal{M}_k}(Y | \theta_k) p_D(\theta_k) d\theta_k}{\int f_{\mathcal{M}_k}^b(Y | \theta_k) p_D(\theta_k) d\theta_k}
\]

- \( f_{\mathcal{M}_k}^b(Y | \theta_k) = \left[f_{\mathcal{M}_k}(Y | \theta_k)\right]^b \)

\[
m_{\mathcal{M}_k}(Y; b) = \int f_{\mathcal{M}_k}^{1-b}(Y | \theta_k) p^F(\theta_k | b, Y) d\theta_k
\]
Fractional marginal likelihood

- Fractional Bayes factor (O’Hagan, 1995)
- \( b = b(n), \ 0 < b < 1, \) a fraction of the sample size \( n.\)
- fractional marginal likelihood of model \( \mathcal{M}_k:\)

\[
m_{\mathcal{M}_k}(Y; b) = \frac{\int f_{\mathcal{M}_k}(Y \mid \theta_k)p^D(\theta_k)d\theta_k}{\int f_{\mathcal{M}_k}^b(Y \mid \theta_k)p^D(\theta_k)d\theta_k}
\]

- \( f_{\mathcal{M}_k}^b(Y \mid \theta_k) = \left[f_{\mathcal{M}_k}(Y \mid \theta_k)\right]^b \)

\[
m_{\mathcal{M}_k}(Y; b) = \int f_{\mathcal{M}_k}^{1-b}(Y \mid \theta_k)p^F(\theta_k \mid b, Y)d\theta_k
\]

- Implied fractional prior:

\[
p^F(\theta_k \mid b, Y) \propto f_{\mathcal{M}_k}^b(Y \mid \theta_k)p^D(\theta_k)
\]
Fractional marginal likelihood

- Fractional Bayes factor (O’Hagan, 1995)
- \( b = b(n), 0 < b < 1 \), a fraction of the sample size \( n \).
- Fractional marginal likelihood of model \( M_k \):

\[
m_{M_k}(Y; b) = \frac{\int f_{M_k}(Y | \theta_k) p^D(\theta_k) d\theta_k}{\int f_{M_k}^b(Y | \theta_k) p^D(\theta_k) d\theta_k}
\]

- \( f_{M_k}^b(Y | \theta_k) = \left[ f_{M_k}(Y | \theta_k) \right]^b \)

\[
m_{M_k}(Y; b) = \int f_{M_k}^{1-b}(Y | \theta_k) p^F(\theta_k | b, Y) d\theta_k
\]

- Implied fractional prior:

\[
p^F(\theta_k | b, Y) \propto f_{M_k}^b(Y | \theta_k) p^D(\theta_k)
\]

- Fractional likelihood is a discounted full likelihood
Fractional marginal likelihood

- Fractional Bayes factor (O’Hagan, 1995)
- \( b = b(n), \, 0 < b < 1, \) a fraction of the sample size \( n. \)
- Fractional marginal likelihood of model \( \mathcal{M}_k: \)
  \[
  m_{\mathcal{M}_k}(Y; b) = \frac{\int f_{\mathcal{M}_k}(Y \mid \theta_k) p^D(\theta_k) d\theta_k}{\int f^b_{\mathcal{M}_k}(Y \mid \theta_k) p^D(\theta_k) d\theta_k}
  \]
- \( f^b_{\mathcal{M}_k}(Y \mid \theta_k) = \left[ f_{\mathcal{M}_k}(Y \mid \theta_k) \right]^b \)
  \[
  m_{\mathcal{M}_k}(Y; b) = \int f^{1-b}_{\mathcal{M}_k}(Y \mid \theta_k) p^F(\theta_k \mid b, Y) d\theta_k
  \]
- Implied fractional prior:
  \[
  p^F(\theta_k \mid b, Y) \propto f^b_{\mathcal{M}_k}(Y \mid \theta_k)p^D(\theta_k)
  \]
- Fractional likelihood is a discounted full likelihood
- Default choice: \( b = n_0/n, \) \( n_0 \) minimal (integer) training sample size which makes the induced fractional prior proper
Gaussian essential graphs

- Given $\Omega_D$, $y_1, \ldots, y_n$ i.i.d. $\mathcal{N}_q(\mu, \Omega_D^{-1})$
Gaussian essential graphs

- Given $\Omega_D$, $y_1, \ldots, y_n$ i.i.d. $\mathcal{N}_q(\mu, \Omega_D^{-1})$
- $\Omega_D$ Markov with respect to a DAG $\mathcal{D}$
Gaussian essential graphs

- Given $\Omega_D$, $y_1, \ldots, y_n$ i.i.d. $\mathcal{N}_q(\mu, \Omega_D^{-1})$
- $\Omega_D$ Markov with respect to a DAG $D$
- no directed edge $u \rightarrow v$ $\iff \rho_{uv} \cdot \{1, \ldots, v\} \setminus \{u, v\} = 0$
  (assuming a well-numbering of the vertices in $V$)
Gaussian essential graphs

- Given $\Omega_D$, $y_1, \ldots, y_n$ i.i.d. $\mathcal{N}_q(\mu, \Omega_D^{-1})$
- $\Omega_D$ Markov with respect to a DAG $D$
- no directed edge $u \rightarrow v \implies \rho_{uv \cdot \{1, \ldots, v\} \setminus \{u, v\}} = 0$
  (assuming a well-numbering of the vertices in $V$)

- equivalence class of $D \rightarrow \text{EG } G$ with chain components $\tau \in \mathcal{T}$
Gaussian essential graphs

- Given $\Omega_D, \ y_1, \ldots, y_n$ i.i.d. $\mathcal{N}_q(\mu, \Omega_D^{-1})$
- $\Omega_D$ Markov with respect to a DAG $D$
- no directed edge $u \rightarrow v \implies \rho_{uv} \cdot \{1,\ldots,v\}\backslash\{u,v\} = 0$
  (assuming a well-numbering of the vertices in $V$)

- equivalence class of $D \rightarrow$ EG $G$ with chain components $\tau \in \mathcal{T}$
- Conditional density of variables in chain component $\tau$

$$f_{G_\tau}(y_{i,\tau} | y_{i,\text{pa}_G(\tau)}, \theta_{G_\tau}) = \mathcal{N}_{|\tau|}(y_{i,\tau} | \mu_\tau + \Gamma_\tau y_{i,\text{pa}_G(\tau)}, \Omega_{G_\tau}^{-1})$$
Gaussian essential graphs

- Given $\Omega_D$, $y_1, \ldots, y_n$ i.i.d. $\mathcal{N}_q(\mu, \Omega_D^{-1})$
- $\Omega_D$ Markov with respect to a DAG $D$
- no directed edge $u \rightarrow v \implies \rho_{uv}.\{1,\ldots,v\}\{u,v\} = 0$
  (assuming a well-numbering of the vertices in $V$)
- equivalence class of $D \rightarrow \text{EG } G$ with chain components $\tau \in \mathcal{T}$
- Conditional density of variables in chain component $\tau$
  \[ f_{G_\tau}(y_{i,\tau} | y_{i,\text{pa}_G(\tau)}, \theta_{G_\tau}) = \mathcal{N}_{|\tau|}(y_{i,\tau} | \mu_\tau + \Gamma_{\tau} y_{i,\text{pa}_G(\tau)}, \Omega_{G_\tau}^{-1}) \]
- $\mu_\tau = \mathbb{E}(y_{i,\tau} | \mu, \Omega_D)$
Gaussian essential graphs

- Given $\Omega_D, \ y_1, \ldots, y_n$ i.i.d. $\mathcal{N}_q(\mu, \Omega_D^{-1})$
- $\Omega_D$ Markov with respect to a DAG $D$
- no directed edge $u \rightarrow v \implies \rho_{uv} \cdot \{1,\ldots,v\}\{u,v\} = 0$
  (assuming a well-numbering of the vertices in $V$)

- equivalence class of $D \rightarrow \text{EG } \mathcal{G}$ with chain components $\tau \in \mathcal{T}$
- Conditional density of variables in chain component $\tau$

  $$f_{\mathcal{G}_\tau}(y_{i,\tau} | y_{i,\text{pa}_{\mathcal{G}}(\tau)}, \theta_{\mathcal{G}_\tau}) = \mathcal{N}_{|\tau|}(y_{i,\tau} | \mu_\tau + \Gamma_\tau y_{i,\text{pa}_{\mathcal{G}}(\tau)}, \Omega_{\mathcal{G}_\tau}^{-1})$$

- $\mu_\tau = \mathbb{E}(y_{i,\tau} | \mu, \Omega_D)$
- $\Gamma_\tau$ matrix of regression parameters
Gaussian essential graphs

- Given $\Omega_D$, $y_1, \ldots, y_n$ i.i.d. $\mathcal{N}_q(\mu, \Omega_D^{-1})$
- $\Omega_D$ Markov with respect to a DAG $D$
- no directed edge $u \to v \implies \rho_{uv} \cdot \{1, \ldots, v\} \setminus \{u, v\} = 0$
  (assuming a well-numbering of the vertices in $V$)

- equivalence class of $D \rightarrow \text{EG} \ G$ with chain components $\tau \in \mathcal{T}$
- Conditional density of variables in chain component $\tau$
  
  $$f_{G,\tau}(y_{i,\tau} \mid y_{i,\text{pa}_G(\tau)}, \theta_{G,\tau}) = \mathcal{N}_{|\tau|}(y_{i,\tau} \mid \mu_{\tau} + \Gamma_{\tau} y_{i,\text{pa}_G(\tau)}, \Omega_{G,\tau}^{-1})$$

- $\mu_{\tau} = \mathbb{E}(y_{i,\tau} \mid \mu, \Omega_D)$
- $\Gamma_{\tau}$ matrix of regression parameters
- $\Omega_{G,\tau}$ (conditional) precision matrix
  Recall (Theorem 1) that $\Omega_{G,\tau}$ is Markov with respect to the decomposable graph $G_{\tau}$. 
Likelihood and prior factorization for an EG

\[ f_G(y_{i,\tau} \mid y_{i,\text{pa}_G(\tau)}, \theta_G) = \mathcal{N}|_{\tau}(y_{i,\tau} \mid B^\top_{\tau} x_{i,\tau}, \Omega^{-1}_G) \]

\[ x_{i,\tau} = \begin{bmatrix} 1 \\ y_{i,\text{pa}_G(\tau)} \end{bmatrix}; \quad B_{\tau} = \begin{bmatrix} \mu_{\tau}^\top \\ \Gamma_{\tau}^\top \end{bmatrix} \]
Likelihood and prior factorization for an EG

\[ f_{\mathcal{G}_\tau}(\mathbf{y}_{i,\tau} \mid \mathbf{y}_{i,\text{pa}_\mathcal{G}(\tau)}, \theta_{\mathcal{G}_\tau}) = \mathcal{N}_{|\tau|}(\mathbf{y}_{i,\tau} \mid \mathbf{B}_\tau^\top \mathbf{x}_{i,\tau}, \Omega_{\mathcal{G}_\tau}^{-1}) \]

\[ \mathbf{x}_{i,\tau} = \begin{bmatrix} 1 \\ \mathbf{y}_{i,\text{pa}_\mathcal{G}(\tau)} \end{bmatrix}; \quad \mathbf{B}_\tau = \begin{bmatrix} \mu_{\tau}^\top \\ \Gamma_{\tau}^\top \end{bmatrix} \]

- \( \mathbf{B}_\tau \) unconstrained (\( \mathcal{G} \) has no flags)
Likelihood and prior factorization for an EG

\[ f_{\mathcal{G}_\tau}(y_{i,\tau} \mid y_{i,\text{pa}_G(\tau)}, \theta_{\mathcal{G}_\tau}) = \mathcal{N}_{|\tau|}(y_{i,\tau} \mid B_\tau^\top x_{i,\tau}, \Omega_{\mathcal{G}_\tau}^{-1}) \]

\[ x_{i,\tau} = \begin{bmatrix} 1 \\ y_{i,\text{pa}_G(\tau)} \end{bmatrix}; \quad B_\tau = \begin{bmatrix} \mu_\tau^\top \\ \Gamma_\tau^\top \end{bmatrix} \]

- \( B_\tau \) unconstrained (\( G \) has no flags)
Likelihood and prior factorization for an EG

\[ f_{G_{\tau}}(y_{i,\tau} | y_{i,\text{pa}_G(\tau)}, \theta_{G_{\tau}}) = \mathcal{N}_{|\tau|}(y_{i,\tau} | B_{\tau}^T x_{i,\tau}, \Omega_{G_{\tau}}^{-1}) \]

\[ x_{i,\tau} = \begin{bmatrix} 1 \\ y_{i,\text{pa}_G(\tau)} \end{bmatrix}; \quad B_{\tau} = \begin{bmatrix} \mu_{\tau}^T \\ \Gamma_{\tau}^T \end{bmatrix} \]

- \( B_{\tau} \) unconstrained (\( G \) has no flags)

\[ Y_{\tau} | X_{\tau}, B_{\tau}, \Omega_{G_{\tau}} \sim \mathcal{N}_{n,|\tau|}(X_{\tau} B_{\tau}, I_n, \Omega_{G_{\tau}}^{-1}) \]

Lik factorization under a chain graph
Likelihood and prior factorization for an EG

\[
f_{G_\tau}(y_{i,\tau} | y_{i,pa_G(\tau)}, \theta_{G_\tau}) = \mathcal{N}_{|\tau|}(y_{i,\tau} | B_\tau^T x_{i,\tau}, \Omega_{G_\tau}^{-1})
\]

\[
x_{i,\tau} = \begin{bmatrix} 1 \\ y_{i,pa_G(\tau)} \end{bmatrix} \quad \text{;} \quad B_\tau = \begin{bmatrix} \mu_\tau^T \\ \Gamma_\tau^T \end{bmatrix}
\]

- \(B_\tau\) unconstrained (\(G\) has no flags)

\[
Y_\tau | X_\tau, B_\tau, \Omega_{G_\tau} \sim \mathcal{N}_{n,|\tau|}(X_\tau B_\tau, I_n, \Omega_{G_\tau}^{-1})
\]

Lik factorization under a chain graph

\[
f_{G}(Y | \mu, \Omega_G) = \prod_{\tau \in \mathcal{T}} \mathcal{N}_{n,|\tau|}(Y_\tau | X_\tau B_\tau, I_n, \Omega_{G_\tau}^{-1})
\]
Likelihood and prior factorization for an EG

\[ f_{G_{\tau}}(y_{i,\tau} \mid y_{i,\text{pa}_G(\tau)}, \theta_{G_{\tau}}) = \mathcal{N}_{|\tau|}(y_{i,\tau} \mid B_{\tau}^\top x_{i,\tau}, \Omega_{G_{\tau}}^{-1}) \]

\[ x_{i,\tau} = \begin{bmatrix} 1 \\ y_{i,\text{pa}_G(\tau)} \end{bmatrix} ; \quad B_{\tau} = \begin{bmatrix} \mu_{\tau}^\top \\ \Gamma_{\tau}^\top \end{bmatrix} \]

- \( B_{\tau} \) unconstrained (\( G \) has no flags)

\[ Y_{\tau} \mid X_{\tau}, B_{\tau}, \Omega_{G_{\tau}} \sim \mathcal{N}_{n,|\tau|}(X_{\tau}B_{\tau}, I_n, \Omega_{G_{\tau}}^{-1}) \]

Lik factorization under a chain graph

\[ f_{G}(Y \mid \mu, \Omega_{G}) = \prod_{\tau \in T} \mathcal{N}_{n,|\tau|}(Y_{\tau} \mid X_{\tau}B_{\tau}, I_n, \Omega_{G_{\tau}}^{-1}) \]

If \( \theta_{G_{\tau}} \) a priori independent (global parameter independence)

\[ m_{G}(Y) = \prod_{\tau \in T} m_{G_{\tau}}(Y_{\tau} \mid X_{\tau}) \]
Marginal likelihood of the essential graph

- Use theory developed in Consonni et al (2017) to perform objective Bayes learning for Gaussian multivariate regression DAG models
Marginal likelihood of the essential graph

- Use theory developed in Consonni et al (2017) to perform objective Bayes learning for Gaussian multivariate regression DAG models
- Default prior

\[ p^D(B_\tau, \Omega_\tau) \propto |\Omega_\tau|^{\frac{a_D - |\tau| - 1}{2}} \]

- \( G_\tau \) is decomposable
- \( C_\tau \) set of (maximal) cliques
- \( S_\tau \) set of separators
- Then

\[
m_{G_\tau}(Y_\tau | X_\tau) = \prod_{C \in C_\tau} m_{\tau}(Y_C, \tau | X_\tau) \prod_{S \in S_\tau} m_{\tau}(Y_S, \tau | X_\tau) \]

- \( m_{\tau}(Y_C, \tau | X_\tau) \) and \( m_{\tau}(Y_S, \tau | X_\tau) \) computed under the complete graph using standard priors (no need for hyper-Inverse Wishart of the like)
- Finally obtain the marginal likelihood of the essential graph

\[ m_G(Y) = \prod_{\tau \in T} m_{G_\tau}(Y_\tau | X_\tau) \]
Marginal likelihood of the essential graph

- Use theory developed in Consonni et al (2017) to perform objective Bayes learning for Gaussian multivariate regression DAG models
- Default prior
  \[ p^D(B_\tau, \Omega_\tau) \propto |\Omega_\tau|^{a_D-|\tau|-1} \]
- \( G_\tau \) is decomposable

- \( G_\tau \) is decomposable
Marginal likelihood of the essential graph

- Use theory developed in Consonni et al (2017) to perform objective Bayes learning for Gaussian multivariate regression DAG models

- Default prior

\[ p^D(B_\tau, \Omega_\tau) \propto |\Omega_\tau|^{a_D - |\tau| - 1/2} \]

- \( G_\tau \) is decomposable

- \( C_\tau \) set of (maximal) cliques

- \( S_\tau \) set of separators
Marginal likelihood of the essential graph

- Use theory developed in Consonni et al (2017) to perform objective Bayes learning for Gaussian multivariate regression DAG models
- Default prior
  \[ p_D(B_\tau, \Omega_\tau) \propto |\Omega_\tau|^{\frac{a_D-|\tau|-1}{2}} \]
- \( \mathcal{G}_\tau \) is decomposable
- \( \mathcal{C}_\tau \) set of (maximal) cliques
- \( \mathcal{S}_\tau \) set of separators
- Then
  \[ m_{\mathcal{G}_\tau}(Y_{\tau} | X_{\tau}) = \frac{\prod_{C \in \mathcal{C}_\tau} m_\tau(Y_{C,\tau} | X_{\tau})}{\prod_{S \in \mathcal{S}_\tau} m_\tau(Y_{S,\tau} | X_{\tau})} \]
Marginal likelihood of the essential graph

- Use theory developed in Consonni et al (2017) to perform objective Bayes learning for Gaussian multivariate regression DAG models
- Default prior
  \[ p^D(B_\tau, \Omega_\tau) \propto |\Omega_\tau|^{a_D - |\tau| - 1} \]
- \( \mathcal{G}_\tau \) is decomposable
- \( \mathcal{C}_\tau \) set of (maximal) cliques
- \( \mathcal{S}_\tau \) set of separators
- Then
  \[ m_{\mathcal{G}_\tau}(Y_\tau | X_\tau) = \frac{\prod_{C \in \mathcal{C}_\tau} m_\tau(Y_{C,\tau} | X_\tau)}{\prod_{S \in \mathcal{S}_\tau} m_\tau(Y_{S,\tau} | X_\tau)} \]
- \( m_\tau(Y_{C,\tau} | X_\tau) \) and \( m_\tau(Y_{S,\tau} | X_\tau) \) computed under the complete graph using standard priors
  (no need for hyper-Inverse Wishart of the like)
Marginal likelihood of the essential graph

- Use theory developed in Consonni et al (2017) to perform objective Bayes learning for Gaussian multivariate regression DAG models
- Default prior
  \[ p^D(B_\tau, \Omega_\tau) \propto |\Omega_\tau|^{a_D-|\tau|-1/2} \]
- \( G_\tau \) is decomposable
- \( C_\tau \) set of (maximal) cliques
- \( S_\tau \) set of separators
- Then
  \[ m_{G_\tau}(Y_\tau | X_\tau) = \frac{\prod_{C \in C_\tau} m_\tau(Y_{C,\tau} | X_\tau)}{\prod_{S \in S_\tau} m_\tau(Y_{S,\tau} | X_\tau)} \]
- \( m_\tau(Y_{C,\tau} | X_\tau) \) and \( m_\tau(Y_{S,\tau} | X_\tau) \) computed under the complete graph using standard priors (no need for hyper-Inverse Wishart of the like)
- Finally obtain the marginal likelihood of the essential graph \( G \) is
  \[ m_G(Y) = \prod_{\tau \in \mathcal{T}} m_{G_\tau}(Y_\tau | X_\tau) \]
A Markov chain on EG space

- $S_q$, the set of all EGs on $q$ nodes
A Markov chain on EG space

- $S_q$, the set of all EGs on $q$ nodes
- $S \subseteq S_q$, e.g. the set of all EGs on $q$ nodes having fewer edges than a threshold $M$ (sparsity constraint)
A Markov chain on EG space

- $S_q$, the set of all EGs on $q$ nodes
- $S \subset S_q$, e.g. the set of all EGs on $q$ nodes having fewer edges than a threshold $M$ (sparsity constraint)

- Define a set of operators (based on insertion or removal of directed edges, undirected edges and v-structures) that determine the transition from $\mathcal{G} \in S$ to $\mathcal{G}' \in S$ (He et al., 2013)
A Markov chain on EG space

- \( S_q \), the set of all EGs on \( q \) nodes
- \( S \subset S_q \), e.g. the set of all EGs on \( q \) nodes having fewer edges than a threshold \( M \) (sparsity constraint)

Define a set of operators (based on insertion or removal of directed edges, undirected edges and \( \nu \)-structures) that determine the transition from \( G \in S \) to \( G' \in S \) (He et al., 2013)

For each \( G \in S \) obtain a *perfect* set of operators \( O_G \) inducing a Markov chain \( \{G_t\} \) on \( S \) satisfying *optimal* properties: *validity*; *distinguishability*; *irreducibility*; *reversibility*
A Markov chain on EG space

- $S_q$, the set of all EGs on $q$ nodes
- $S \subset S_q$, e.g. the set of all EGs on $q$ nodes having fewer edges than a threshold $M$ (sparsity constraint)

Define a set of operators (based on insertion or removal of directed edges, undirected edges and $v$-structures) that determine the transition from $G \in S$ to $G' \in S$ (He et al., 2013)

For each $G \in S$ obtain a perfect set of operators $O_G$ inducing a Markov chain $\{G_t\}$ on $S$ satisfying optimal properties: validity; distinguishability; irreducibility; reversibility

Then the probability of transition from $G \in S$ to $G' \in S$ is

$$p_{G,G'} = \begin{cases} 
1/|O_G| & \text{if } G' \in S, \\
0 & \text{otherwise.}
\end{cases}$$
MCMC algorithm on Markov equivalence classes of DAGs

The transition from $\mathcal{G} \in S$ to $\mathcal{G}' \in S$ is accepted with probability

$$\alpha = \min \left\{ 1; \frac{m_{\mathcal{G}'}(\mathbf{Y})}{m_{\mathcal{G}}(\mathbf{Y})} \cdot \frac{p(\mathcal{G}')}{p(\mathcal{G})} \cdot \frac{q(\mathcal{G} | \mathcal{G}')}{q(\mathcal{G}' | \mathcal{G})} \right\}$$
The transition from $\mathcal{G} \in \mathcal{S}$ to $\mathcal{G}' \in \mathcal{S}$ is accepted with probability

$$\alpha = \min \left\{ 1; \frac{m_{\mathcal{G}'}(\mathbf{Y})}{m_{\mathcal{G}}(\mathbf{Y})} \cdot \frac{p(\mathcal{G}')}{p(\mathcal{G})} \cdot \frac{q(\mathcal{G} | \mathcal{G}')}{q(\mathcal{G}' | \mathcal{G})} \right\}$$

- $m_{\mathcal{G}}(\mathbf{Y})$ marginal likelihood of EG $\mathcal{G}$
The transition from $\mathcal{G} \in \mathcal{S}$ to $\mathcal{G}' \in \mathcal{S}$ is accepted with probability

$$\alpha = \min \left\{ 1; \frac{m_{\mathcal{G}'}(Y)}{m_{\mathcal{G}}(Y)} \cdot \frac{p(\mathcal{G}')}{p(\mathcal{G})} \cdot \frac{q(\mathcal{G} \mid \mathcal{G}')}{{q(\mathcal{G}' \mid \mathcal{G})}} \right\}$$

- $m_{\mathcal{G}}(Y)$ marginal likelihood of EG $\mathcal{G}$
- $q(\mathcal{G}' \mid \mathcal{G}) = p_{\mathcal{G},\mathcal{G}'}$ proposal distribution
MCMC algorithm on Markov equivalence classes of DAGs

The transition from $G \in S$ to $G' \in S$ is accepted with probability

$$\alpha = \min \left\{ 1; \frac{m_{G'}(Y)}{m_G(Y)} \cdot \frac{p(G')}{p(G)} \cdot \frac{q(G | G')}{q(G' | G)} \right\}$$

- $m_G(Y)$ marginal likelihood of EG $G$
- $q(G' | G) = p_{G,G'}$ proposal distribution
- $p(G)$ prior on $G$
MCMC algorithm on Markov equivalence classes of DAGs

- The transition from $G \in \mathcal{S}$ to $G' \in \mathcal{S}$ is accepted with probability

$$\alpha = \min \left\{ 1; \frac{m_{G'}(Y)}{m_G(Y)} \cdot \frac{p(G')}{p(G)} \cdot \frac{q(G \mid G')}{q(G' \mid G)} \right\}$$

- $m_G(Y)$ marginal likelihood of EG $G$
- $q(G' \mid G) = p_{G,G'}$ proposal distribution
- $p(G)$ prior on $G$
  - Assume a beta-binomial prior on the skeleton $G''$ of $G$
MCMC algorithm on Markov equivalence classes of DAGs

- The transition from $\mathcal{G} \in \mathcal{S}$ to $\mathcal{G}' \in \mathcal{S}$ is accepted with probability

$$
\alpha = \min \left\{ 1; \frac{m_{\mathcal{G}'}(\mathbf{Y})}{m_{\mathcal{G}}(\mathbf{Y})} \cdot \frac{p(\mathcal{G}')}{p(\mathcal{G})} \cdot \frac{q(\mathcal{G} | \mathcal{G}')}{q(\mathcal{G}' | \mathcal{G})} \right\}
$$

- $m_{\mathcal{G}}(\mathbf{Y})$ marginal likelihood of EG $\mathcal{G}$
- $q(\mathcal{G}' | \mathcal{G}) = p_{\mathcal{G}, \mathcal{G}'}$ proposal distribution
- $p(\mathcal{G})$ prior on $\mathcal{G}$
  - Assume a beta-binomial prior on the skeleton $\mathcal{G}''$ of $\mathcal{G}$
  - $\mathcal{G}_{(j)}$ $j$-th element of the vectorized lower triangular part of the adjacency matrix of $\mathcal{G}''$
MCMC algorithm on Markov equivalence classes of DAGs

- The transition from $G \in S$ to $G' \in S$ is accepted with probability

\[
\alpha = \min \left\{ 1; \frac{m_{G'}(Y)}{m_G(Y)} \cdot \frac{p(G')}{p(G)} \cdot \frac{q(G | G')}{q(G' | G)} \right\}
\]

- $m_G(Y)$ marginal likelihood of EG $G$
- $q(G' | G) = p_{G,G'}$ proposal distribution
- $p(G)$ prior on $G$
  - Assume a beta-binomial prior on the skeleton $G^u$ of $G$
  - $G(j)$ $j$-th element of the vectorized lower triangular part of the adjacency matrix of $G^u$
  - $G(j) | \pi \sim \text{Ber}(\pi)$, $j = 1, \ldots, q(q - 1)/2$, independently
MCMC algorithm on Markov equivalence classes of DAGs

The transition from $\mathcal{G} \in \mathcal{S}$ to $\mathcal{G}' \in \mathcal{S}$ is accepted with probability

$$\alpha = \min \left\{ 1; \frac{m_{\mathcal{G}'}(Y)}{m_{\mathcal{G}}(Y)} \cdot \frac{p(\mathcal{G}')}{p(\mathcal{G})} \cdot \frac{q(\mathcal{G} | \mathcal{G}')} {q(\mathcal{G}' | \mathcal{G})} \right\}$$

- $m_{\mathcal{G}}(Y)$ marginal likelihood of $\text{EG } \mathcal{G}$
- $q(\mathcal{G}' | \mathcal{G}) = p_{\mathcal{G},\mathcal{G}'}$ proposal distribution
- $p(\mathcal{G})$ prior on $\mathcal{G}$
  - Assume a beta-binomial prior on the skeleton $\mathcal{G}^u$ of $\mathcal{G}$
  - $\mathcal{G}_{(j)}$ $j$-th element of the vectorized lower triangular part of the adjacency matrix of $\mathcal{G}^u$
  - $\mathcal{G}_{(j)} | \pi \sim \text{Ber}(\pi), \ j = 1, \ldots, q(q-1)/2$, independently
  - $\pi \sim \text{Beta}(a, b)$
Simulation study

- We compare our Objective Bayes Essential graph Search (OBES) with three benchmark methods:
Simulation study

- We compare our Objective Bayes Essential graph Search (OBES) with three benchmark methods:
  - Greedy DAG Search algorithm
    (Chickering, 2002)
  - Greedy Equivalence Search with tuning parameter $\gamma \in \{0, 0.5, 1\}$
    (Chickering, 2002; Hauser and Bühlmann, 2012)
  - PC algorithm with confidence level $\alpha \in \{0.01, 0.05, 0.1\}$
    (Spirtes et al., 2000; Colombo and Maathuis, 2014)
Simulation study

- We compare our Objective Bayes Essential graph Search (OBES) with three benchmark methods:
  - Greedy DAG Search algorithm (Chickering, 2002)
  - Greedy Equivalence Search with tuning parameter $\gamma \in \{0, 0.5, 1\}$ (Chickering, 2002; Hauser and Bühlmann, 2012)
  - PC algorithm with confidence level $\alpha \in \{0.01, 0.05, 0.1\}$ (Spirtes et al., 2000; Colombo and Maathuis, 2014)
- OBES is *fully* Bayes, and produces a posterior distribution on $S$. To obtain a summary of this distribution we define first the marginal probability of edge inclusion

$$p_{u \rightarrow v}(Y) = \sum_{G \in S_{u \rightarrow v}} p(G | Y)$$
Simulation study

- We compare our Objective Bayes Essential graph Search (OBES) with three benchmark methods:
  - Greedy DAG Search algorithm (Chickering, 2002)
  - Greedy Equivalence Search with tuning parameter $\gamma \in \{0, 0.5, 1\}$ (Chickering, 2002; Hauser and Bühlmann, 2012)
  - PC algorithm with confidence level $\alpha \in \{0.01, 0.05, 0.1\}$ (Spirtes et al., 2000; Colombo and Maathuis, 2014)
- OBES is fully Bayes, and produces a posterior distribution on $\mathcal{S}$. To obtain a summary of this distribution we define first the marginal probability of edge inclusion
  \[
  p_{u \rightarrow v}(\mathbf{Y}) = \sum_{\mathcal{G} \in \mathcal{S}_{u \rightarrow v}} p(\mathcal{G} | \mathbf{Y})
  \]
- Next we take the median probability (graph) model, namely the graph containing only those directed edges $u \rightarrow v$ such that $p_{u \rightarrow v}(\mathbf{Y}) \geq 0.5$ (not an EG but a Partial DAG)
We compare our Objective Bayes Essential graph Search (OBES) with three benchmark methods:

- Greedy DAG Search algorithm (Chickering, 2002)
- Greedy Equivalence Search with tuning parameter $\gamma \in \{0, 0.5, 1\}$ (Chickering, 2002; Hauser and Bühlmann, 2012)
- PC algorithm with confidence level $\alpha \in \{0.01, 0.05, 0.1\}$ (Spirtes et al., 2000; Colombo and Maathuis, 2014)

OBES is *fully* Bayes, and produces a posterior distribution on $S$. To obtain a summary of this distribution we define first the marginal probability of edge inclusion

$$p_{u \rightarrow v}(Y) = \sum_{G \in S_{u \rightarrow v}} p(G | Y)$$

Next we take the *median probability (graph) model*, namely the graph containing only those directed edges $u \rightarrow v$ such that $p_{u \rightarrow v}(Y) \geq 0.5$ (not an EG but a Partial DAG)

We make a consistent extension of the PDAG and finally we obtain its EG named *projected median probability model*
Simulation setting

- Consider $q \in \{10, 20\}$, $n \in \{50, 100, 200\}$
- For each scenario $(n, q)$ generate 50 datasets
- To obtain each dataset, randomly generate a DAG $\mathcal{D}$ on $q$ nodes and simulate $n$ i.i.d observations from $\mathcal{N}(\mathbf{0}, \Omega_D^{-1})$
- For each method (OBES and benchmarks) compute the Structural Hamming Distances between estimates and true EG over the 50 simulations
Experiments

Results

- OB Med.
- OB Proj.
- GDS
- GES 0
- GES 0.5
- GES 1
- PC 0.01
- PC 0.05
- PC 0.10

- q = 10
- n = 50
- n = 100
- n = 200

Shd

Guido Consonni

OBAYES Learning of MECs of DAGs

Helsinki, 30 October 2017
Results

- **n = 50**
  - q = 20
  - Shd
  - OB Med.
  - OB Proj.
  - GDS
  - GES 0
  - GES 0.5
  - GES 1
  - PC 0.01
  - PC 0.05
  - PC 0.10

- **n = 100**
  - q = 20
  - Shd
  - OB Med.
  - OB Proj.
  - GDS
  - GES 0
  - GES 0.5
  - GES 1
  - PC 0.01
  - PC 0.05
  - PC 0.10

- **n = 200**
  - q = 20
  - Shd
  - OB Med.
  - OB Proj.
  - GDS
  - GES 0
  - GES 0.5
  - GES 1
  - PC 0.01
  - PC 0.05
  - PC 0.10
Protein-signaling data

- Nine datasets from Sachs et al. (2005)
Protein-signaling data

- Nine datasets from Sachs et al. (2005)
- Data include the levels of $q = 11$ phosphorylated proteins and phospholipids quantified using flow cytometry under different experimental conditions, each with sample size in the range 700-1000
Protein-signaling data

- Nine datasets from Sachs et al. (2005)
- Data include the levels of $q = 11$ phosphorylated proteins and phospholipids quantified using flow cytometry under different experimental conditions, each with sample size in the range 700-1000
- We implement OBES on each dataset setting the maximum number of edges at 22) and running $10^5$ MCMC iterations
Results (Dataset I)

- Heat map with posterior probabilities of edge inclusion $p_{u \rightarrow v}(Y)$

We observe $p_{u \rightarrow v}(Y) \geq 0.5$ for 6 undirected edges and 2 directed edges: $10 \rightarrow 9 \leftarrow 11$ (av-structure).

For comparison with benchmarks we construct the median probability graph model including all edges $u \rightarrow v$ such that $p_{u \rightarrow v} \geq 0.5$. 
Results (Dataset I)

- Heat map with posterior probabilities of edge inclusion $p_{u \rightarrow v}(Y)$

- We observe $p_{u \rightarrow v}(Y) \geq 0.5$ for
Results (Dataset I)

- Heat map with posterior probabilities of edge inclusion $p_{u \rightarrow v}(Y)$

We observe $p_{u \rightarrow v}(Y) \geq 0.5$ for

- 6 undirected edges
Results (Dataset I)

- Heat map with posterior probabilities of edge inclusion $p_{u \rightarrow v}(Y)$

We observe $p_{u \rightarrow v}(Y) \geq 0.5$ for
- 6 undirected edges
- 2 directed edges: $10 \rightarrow 9 \leftarrow 11$ (a $v$-structure)
Results (Dataset I)

- Heat map with posterior probabilities of edge inclusion $p_{u \rightarrow v}(Y)$

We observe $p_{u \rightarrow v}(Y) \geq 0.5$ for
- 6 undirected edges
- 2 directed edges: $10 \rightarrow 9 \leftarrow 11$ (a $v$-structure)

For comparison with benchmarks we construct the median probability graph model including all edges $u \rightarrow v$ such that $p_{u \rightarrow v} \geq 0.5$
Results (Dataset I)

OBES median probability model
Results (Dataset I)

- Define the *quantile probability graph model* of order $p^*$ as the graph including all edges $u \rightarrow v$ such that $p_{u\rightarrow v}(Y) \geq p^*$

GDS (●), GES 0.5 (●), PC 0.01 (○)
Results (Dataset I)

- Define the *quantile probability graph model* of order $p^*$ as the graph including all edges $u \rightarrow v$ such that $p_{u \rightarrow v}(Y) \geq p^*$
- For a grid of thresholds $p^* \in [0.1, 0.9]$ obtain a collection of quantile probability graph models

**GDS (●), GES 0.5 (○), PC 0.01 (●)**
Results (Dataset I)

- Define the *quantile probability graph model* of order $p^*$ as the graph including all edges $u \rightarrow v$ such that $p_{u\rightarrow v}(Y) \geq p^*$
- For a grid of thresholds $p^* \in [0.1, 0.9]$ obtain a collection of quantile probability graph models
- For each graph choose a feature of interest (e.g. number of directed edges) and obtain its frequency distribution by varying $p^*$

![Frequency distributions for different features](image)

GDS (●), GES 0.5 (●), PC 0.01 (●)
We presented an objective Bayes method to obtain the marginal likelihood of an EG. We constructed an MCMC sampler to explore the space of EGs under sparsity constraints. We compared our Objective Bayes Essential graph Search (OBES) method with current benchmarks. OBES is fully Bayes, and thus can provide an uncertainty evaluation of any feature of interest (e.g. the probability of inclusion of a particular edge) and not only a single estimate of the EG. Being objective, it is virtually free from prior specifications.
We presented an objective Bayes method to obtain the marginal likelihood of an EG. We constructed an MCMC sampler to explore the space of EGs under sparsity constraints. We compared our Objective Bayes Essential graph Search (OBES) method with current benchmarks. OBES is fully Bayes, and thus can provide an uncertainty evaluation of any feature of interest (e.g., the probability of inclusion of a particular edge) and not only a single estimate of the EG. Being objective, it is virtually free from prior specifications.
We presented an objective Bayes method to obtain the marginal likelihood of an EG.
We presented an objective Bayes method to obtain the marginal likelihood of an EG.

We constructed an MCMC sampler to explore the space of EGs under sparsity constraints.
We presented an objective Bayes method to obtain the marginal likelihood of an EG

We constructed an MCMC sampler to explore the space of EGs under sparsity constraints

We compared our Objective Bayes Essential graph Search (OBES) method with current benchmarks
We presented an objective Bayes method to obtain the marginal likelihood of an EG.

We constructed an MCMC sampler to explore the space of EGs under sparsity constraints.

We compared our Objective Bayes Essential graph Search (OBES) method with current benchmarks.

OBES is fully Bayes, and thus can provide an uncertainty evaluation of any feature of interest (e.g. the probability of inclusion of a particular edge) and not only a single estimate of the EG.
We presented an objective Bayes method to obtain the marginal likelihood of an EG.

We constructed an MCMC sampler to explore the space of EGs under sparsity constraints.

We compared our Objective Bayes Essential graph Search (OBES) method with current benchmarks.

OBES is fully Bayes, and thus can provide an uncertainty evaluation of any feature of interest (e.g. the probability of inclusion of a particular edge) and not only a single estimate of the EG.

Being objective, it is virtually free from prior specifications.
On going research

- The Sachs data were collected under different experimental conditions
On going research

- The Sachs data were collected under different experimental conditions
- We analyzed them separately
  Extension to structural learning of *multiple* essential graphs with features potentially shared across graphs (Peterson et al., 2015)
On going research

- The Sachs data were collected under different experimental conditions
- We analyzed them separately
  - Extension to structural learning of *multiple* essential graphs with features potentially shared across graphs (Peterson et al., 2015)
- Extension of the methodology to jointly model *observational* and *interventional* data
  (produced under different exogenous perturbations of variables or by randomized intervention experiments) (Hauser & Bülmann, 2015)
Main references


Main references

Objective Bayes factors for Gaussian directed acyclic graphical models.

Objective Bayes Covariate-Adjusted Sparse Graphical Model Selection.

Hyper Markov laws in the statistical analysis of decomposable graphical models.

Parameter priors for directed acyclic graphical models and the characterization of several probability distributions.

Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs.


**Main references**

**He, Y., Jia, J. & Yu, B. (2013).**
Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs.

Fractional Bayes factors for model comparison.

**Peterson, C., Stingo, F. C. & Vannucci, M. (2015).**
Bayesian inference of multiple Gaussian graphical models.

**Sachs, K., Perez, O., Pe’er, D., Lauffenburger, D. A. & Nolan, G. P. (2005).**
Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data.
*Science,*** **308**, 523-529.

**Spirtes, P. and Glymour, C. and Scheines, R. (2000).**
Causation, Prediction and Search (2nd edition).
*Cambridge, MA: The MIT Press.,* 1-16.
Supplementary slides
Gaussian multivariate regression DAGs

- Use theory developed in Consonni et al (2017) to perform objective Bayes learning for Gaussian multivariate regression DAG models
- They extend the methodology of Geiger and Heckerman (2002) for constructing compatible priors across DAG models
  - to the regression setting with covariate $X$
  - to the objective Bayes setting
    (using the fractional Bayes factor)
- Key ideas
  - Assign a prior on the unconstrained parameters for the complete DAG model (easy; can use standard conjugate matrix normal Wishart)
  - Show that the fractional prior belongs to the above conjugate family
  - Deduce the marginal distribution of $Y_J$ with $J \subseteq V$ under the complete model (easy because of conjugacy)
  - If $G$ is decomposable, marginal likelihood is

$$m_G(Y | X) = \frac{\prod_{C \in \mathcal{C}} m(Y_C | X)}{\prod_{S \in \mathcal{S}} m(Y_S | X)}$$

- Important: both $m(Y_C)$ and $m(Y_S)$ are in closed form and computed under the complete graph using standard priors (no need for hyper-Inverse Wishart of the like)
Properties of the Markov chain on EG space

- **validity**: the transition probability from $G$ to $G'$ that is not an EG is zero and the transition probability of an EG to itself is null
Properties of the Markov chain on EG space

- **validity**: the transition probability from $G$ to $G'$ that is not an EG is zero and the transition probability of an EG to itself is null

- **distinguishability**: for any $G'$ direct successor of $G$, there is a unique operator that transforms $G$ in $G'$
Properties of the Markov chain on EG space

- **validity**: the transition probability from $G$ to $G'$ that is not an EG is zero and the transition probability of an EG to itself is null
- **distinguishability**: for any $G'$ direct successor of $G$, there is a unique operator that transforms $G$ in $G'$
- **irreducibility**: starting from $G$, there is a positive probability of reaching any other EG in $S$ via a sequence of operators
Properties of the Markov chain on EG space

- **validity**: the transition probability from $\mathcal{G}$ to $\mathcal{G}'$ that is not an EG is zero and the transition probability of an EG to itself is null
- **distinguishability**: for any $\mathcal{G}'$ direct successor of $\mathcal{G}$, there is a unique operator that transforms $\mathcal{G}$ in $\mathcal{G}'$
- **irreducibility**: starting from $\mathcal{G}$, there is a positive probability of reaching any other EG in $\mathcal{S}$ via a sequence of operators
- **reversibility**: if $\mathcal{G}'$ is a direct successor of $\mathcal{G}$, then $\mathcal{G}$ is also a direct successor of $\mathcal{G}'$
MCMC algorithm

- Start from an arbitrary $G_0$ (e.g. the null graph)
- For $t = 1, \ldots, T$
MCMC algorithm

- Start from an arbitrary $G_0$ (e.g. the null graph)
  For $t = 1, \ldots, T$
    - set $G = G_{t-1}$
MCMC algorithm

- Start from an arbitrary $G_0$ (e.g. the null graph)
- For $t = 1, \ldots, T$
  - set $G = G_{t-1}$
  - generate $G'$ from the proposal $q(G' \mid G)$

Guido Consonni

OBAYES Learning of MECs of DAGs

Helsinki, 30 October 2017
MCMC algorithm

- Start from an arbitrary $G_0$ (e.g. the null graph)
- For $t = 1, \ldots, T$
  - set $G = G_{t-1}$
  - generate $G'$ from the proposal $q(G' | G)$
  - compute the probability of acceptance $\alpha$
MCMC algorithm

- Start from an arbitrary $G_0$ (e.g. the null graph)
  For $t = 1, \ldots, T$
  - set $G = G_{t-1}$
  - generate $G'$ from the proposal $q(G' | G)$
  - compute the probability of acceptance $\alpha$
  - update $G_t = G'$ with probability $\alpha$, $G_t = G_{t-1}$ with probability $1 - \alpha$
Data generation for simulations

- Fix \( q \), number of nodes
Data generation for simulations

- Fix $q$, number of nodes
- Randomly generate a topologically ordered DAG with probability of edge inclusion $p_{\text{edge}} = 3/(2q - 2)$

The DAG thus obtained implies the set of linear equations

\[
Y_{i,j} = \mu_j + \sum_{k \in \text{pa}(j)} \beta_{k,j} Y_{i,k} + \varepsilon_{i,j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, q
\]

with $\varepsilon_{i,j} \sim N(0, \sigma_{j}^2)$ independently.

$\mu_j \leftarrow 0$

$\sigma_{j}^2 \leftarrow \text{runif}(0, 2)$

$\beta_{k,j} \leftarrow \text{runif}(-1, -0.1) \cup [0.1, 1]$. Generate datasets of size $n$ accordingly.
Data generation for simulations

- Fix $q$, number of nodes
- Randomly generate a topologically ordered DAG with probability of edge inclusion $p_{edge} = 3/(2q - 2)$
- The DAG thus obtained implies the set of linear equations

$$Y_{i,j} = \mu_j + \sum_{k \in pa(j)} \beta_{k,j}Y_{i,k} + \varepsilon_{i,j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, q$$
Data generation for simulations

- Fix $q$, number of nodes
- Randomly generate a topologically ordered DAG with probability of edge inclusion $p_{edge} = 3/(2q - 2)$
- The DAG thus obtained implies the set of linear equations

$$Y_{i,j} = \mu_j + \sum_{k \in pa(j)} \beta_{k,j}Y_{i,k} + \varepsilon_{i,j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, q$$

- with $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$ independently
Data generation for simulations

- Fix $q$, number of nodes
- Randomly generate a topologically ordered DAG with probability of edge inclusion $p_{edge} = 3/(2q - 2)$
- The DAG thus obtained implies the set of linear equations

$$Y_{i,j} = \mu_j + \sum_{k \in \text{pa}(j)} \beta_{k,j} Y_{i,k} + \varepsilon_{i,j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, q$$

- with $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$ independently
- $\mu_j \leftarrow 0$
Data generation for simulations

- Fix \( q \), number of nodes
- Randomly generate a topologically ordered DAG with probability of edge inclusion \( p_{\text{edge}} = \frac{3}{2q - 2} \)
- The DAG thus obtained implies the set of linear equations

\[
Y_{i,j} = \mu_j + \sum_{k \in \text{pa}(j)} \beta_{k,j} Y_{i,k} + \varepsilon_{i,j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, q
\]

- with \( \varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2) \) independently
- \( \mu_j \leftarrow 0 \)
- \( \sigma_j^2 \leftarrow \text{runif}(0, 2) \)
Data generation for simulations

- Fix $q$, number of nodes
- Randomly generate a topologically ordered DAG with probability of edge inclusion $p_{\text{edge}} = 3/(2q - 2)$
- The DAG thus obtained implies the set of linear equations

$$Y_{i,j} = \mu_j + \sum_{k \in \text{pa}(j)} \beta_{k,j} Y_{i,k} + \varepsilon_{i,j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, q$$

- with $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$ independently
- $\mu_j \leftarrow 0$
- $\sigma_j^2 \leftarrow \text{runif}(0, 2)$
- $\beta_{k,j} \leftarrow \text{runif}[-1, -0.1] \cup [0.1, 1]$.
Data generation for simulations

- Fix $q$, number of nodes
- Randomly generate a topologically ordered DAG with probability of edge inclusion $p_{edge} = 3/(2q - 2)$
- The DAG thus obtained implies the set of linear equations

$$Y_{i,j} = \mu_j + \sum_{k \in pa(j)} \beta_{k,j} Y_{i,k} + \varepsilon_{i,j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, q$$

- with $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2)$ independently
- $\mu_j \leftarrow 0$
- $\sigma_j^2 \leftarrow \text{runif}(0, 2)$
- $\beta_{k,j} \leftarrow \text{runif}[-1, -0.1] \cup [0.1, 1]$.
- Generate datasets of size $n$ accordingly
Diagnostics (Dataset I)

- MCMC traceplots of some graph features for the visited graphs

![Traceplot of undirected edges](image1)
![Traceplot of directed edges](image2)
![Traceplot of v-structures](image3)
![Traceplot of chain components](image4)
Further slides
Possibly add condition on $n$ and $n_0$
Gaussian multivariate regression DAGs

- Use theory developed in Consonni et al (2017) to perform objective Bayes learning for Gaussian multivariate regression DAG models
- They extend the methodology of Geiger and Heckerman (2002) for constructing compatible priors across DAG models
  - to the regression setting with covariate \( X \)
  - to the objective Bayes setting
    - (using the fractional Bayes factor)
- **Key ideas**
  - Assign a prior on the *unconstrained* parameters for the *complete* DAG model
    - (easy; can use standard conjugate matrix normal Wishart)
  - Show that the fractional prior belongs to the above conjugate family
  - Deduce the marginal distribution of \( Y_J \) with \( J \subseteq V \) under the *complete* model
    - (easy because of conjugacy)
  - If \( G \) is decomposable, marginal likelihood is

\[
m_G(Y \mid X) = \frac{\prod_{C \in C} m(Y_C \mid X)}{\prod_{S \in S} m(Y_S \mid X)}
\]

- Important: both \( m(Y_C) \) and \( m(Y_S) \) are in closed form and computed under the *complete* graph using standard priors
  - (no need for hyper-Inverse Wishart of the like)
Gaussian multivariate regression DAG models

\[ Y \mid M, \Phi, \Sigma \sim \mathcal{N}_{n,q}(M, \Phi, \Sigma) \iff \text{vec}(Y) \sim \mathcal{N}_q(\text{vec}(M), \Sigma \otimes \Phi) \]
Gaussian multivariate regression DAG models

- \( Y | M, \Phi, \Sigma \sim \mathcal{N}_{n \times q}(M, \Phi, \Sigma) \iff \text{vec}(Y) \sim \mathcal{N}_q(\text{vec}(M), \Sigma \otimes \Phi) \)
  
- \( M \) mean matrix
Gaussian multivariate regression DAG models

- \( Y \mid M, \Phi, \Sigma \sim \mathcal{N}_{n,q}(M, \Phi, \Sigma) \iff \text{vec}(Y) \sim \mathcal{N}_q(\text{vec}(M), \Sigma \otimes \Phi) \)

- \( M \) mean matrix
- \( \Phi \) row covariance matrix
Gaussian multivariate regression DAG models

\[
Y | M, \Phi, \Sigma \sim \mathcal{N}_{n,q}(M, \Phi, \Sigma) \iff \text{vec}(Y) \sim \mathcal{N}_q(\text{vec}(M), \Sigma \otimes \Phi)
\]

- \( M \) mean matrix
- \( \Phi \) row covariance matrix
- \( \Sigma \) column covariance matrix
Gaussian multivariate regression DAG models

- $Y \mid M, \Phi, \Sigma \sim \mathcal{N}_{n,q}(M, \Phi, \Sigma) \iff \text{vec}(Y) \sim \mathcal{N}_q(\text{vec}(M), \Sigma \otimes \Phi)$
  - $M$ mean matrix
  - $\Phi$ row covariance matrix
  - $\Sigma$ column covariance matrix

- $Y \mid B, \Omega_{\mathcal{D}} \sim \mathcal{N}_{n,q}(XB, I_n, \Omega_{\mathcal{D}}^{-1})$
Gaussian multivariate regression DAG models

- \( Y \mid M, \Phi, \Sigma \sim \mathcal{N}_{n,q}(M, \Phi, \Sigma) \iff \text{vec}(Y) \sim \mathcal{N}_q(\text{vec}(M), \Sigma \otimes \Phi) \)
  - \( M \) mean matrix
  - \( \Phi \) row covariance matrix
  - \( \Sigma \) column covariance matrix

- \( Y \mid B, \Omega_D \sim \mathcal{N}_{n,q}(XB, I_n, \Omega_D^{-1}) \)
  - \( Y \) \( n \times q \) matrix of observations on the responses
Gaussian multivariate regression DAG models

- $Y | M, \Phi, \Sigma \sim \mathcal{N}_{n,q}(M, \Phi, \Sigma) \iff \text{vec}(Y) \sim \mathcal{N}_q(\text{vec}(M), \Sigma \otimes \Phi)$
  - $M$ mean matrix
  - $\Phi$ row covariance matrix
  - $\Sigma$ column covariance matrix

- $Y | B, \Omega_D \sim \mathcal{N}_{n,q}(XB, I_n, \Omega_D^{-1})$
  - $Y$ $n \times q$ matrix of observations on the responses
  - $X$ $n \times (p + 1)$ design matrix of observations on the $p$ exogenous variables (including a column vector with all entries equal to 1 for the intercept term)
Gaussian multivariate regression DAG models

- \( Y \mid M, \Phi, \Sigma \sim \mathcal{N}_{n,q}(M, \Phi, \Sigma) \iff \text{vec}(Y) \sim \mathcal{N}_q(\text{vec}(M), \Sigma \otimes \Phi) \)
  - \( M \) mean matrix
  - \( \Phi \) row covariance matrix
  - \( \Sigma \) column covariance matrix

- \( Y \mid B, \Omega_D \sim \mathcal{N}_{n,q}(XB, I_n, \Omega_D^{-1}) \)
  - \( Y \) \( n \times q \) matrix of observations on the responses
  - \( X \) \( n \times (p + 1) \) design matrix of observations on the \( p \) exogenous variables (including a column vector with all entries equal to 1 for the intercept term)
  - \( B \) \( (p + 1) \times q \) matrix of unconstrained regression parameters
Gaussian multivariate regression DAG models

- $Y \mid M, \Phi, \Sigma \sim N_{n,q}(M, \Phi, \Sigma) \iff \text{vec}(Y) \sim N_q(\text{vec}(M), \Sigma \otimes \Phi)$
  - $M$ mean matrix
  - $\Phi$ row covariance matrix
  - $\Sigma$ column covariance matrix

- $Y \mid B, \Omega_D \sim N_{n,q}(XB, I_n, \Omega_D^{-1})$
  - $Y$ $n \times q$ matrix of observations on the responses
  - $X$ $n \times (p + 1)$ design matrix of observations on the $p$ exogenous variables (including a column vector with all entries equal to 1 for the intercept term)
  - $B$ $(p + 1) \times q$ matrix of unconstrained regression parameters
  - $\Omega_D$ $q \times q$ precision matrix, Markov with respect to a DAG $D$
Case 1) $\mathcal{D}$ complete

- $\mathcal{D}$ complete $\implies$ all pairs of edges present (no conditional independencies among the $q$ responses)
Case 1) $\mathcal{D}$ complete

- $\mathcal{D}$ complete $\Rightarrow$ all pairs of edges present (no conditional independencies among the $q$ responses)
- $\Omega$ (precision matrix) *unconstrained* (only s.p.d.)
Case 1) $\mathcal{D}$ complete

- $\mathcal{D}$ complete $\implies$ all pairs of edges present (no conditional independencies among the $q$ responses)
- $\Omega$ (precision matrix) *unconstrained* (only s.p.d.)
- Consonni et al. (2017) extended the methodology of Geiger & Heckerman (2002) for the construction of parameter priors to Gaussian regression DAG models using the fractional Bayes factor
Case 1) $\mathcal{D}$ complete

- $\mathcal{D}$ complete $\implies$ all pairs of edges present (no conditional independencies among the $q$ responses)
- $\Omega$ (precision matrix) *unconstrained* (only s.p.d.)
- Consonni et al. (2017) extended the methodology of Geiger & Heckerman (2002) for the construction of parameter priors to Gaussian regression DAG models using the fractional Bayes factor
- Start from the default prior $p^D(B, \Omega) \propto |\Omega|^{\frac{a_D + q - 1}{2}}$
Case 1) $\mathcal{D}$ complete

- $\mathcal{D}$ complete $\implies$ all pairs of edges present (no conditional independencies among the $q$ responses)
- $\Omega$ (precision matrix) *unconstrained* (only s.p.d.)
- Consonni et al. (2017) extended the methodology of Geiger & Heckerman (2002) for the construction of parameter priors to Gaussian regression DAG models using the fractional Bayes factor
- Start from the default prior $p^D(B, \Omega) \propto |\Omega|^{\frac{a_D+q-1}{2}}$
- Obtain the fractional prior $p^F(B, \Omega)$
Case 1) $\mathcal{D}$ complete

- $\mathcal{D}$ complete $\implies$ all pairs of edges present (no conditional independencies among the $q$ responses)
- $\Omega$ (precision matrix) *unconstrained* (only s.p.d.)
- Consonni et al. (2017) extended the methodology of Geiger & Heckerman (2002) for the construction of parameter priors to Gaussian regression DAG models using the fractional Bayes factor
- Start from the default prior $p^D(B, \Omega) \propto |\Omega|^{aD+q-1/2}$
- Obtain the fractional prior $p^F(B, \Omega)$
- $p^F(B, \Omega)$ is a matrix normal Wishart, which satisfies Geiger & Heckerman assumptions and produces a closed form expression for the marginal density of any (column) submatrix of $Y$
Case 2) $\mathcal{D}$ not complete

- $\Omega_\mathcal{G}$ Markov with respect to a decomposable UG $\mathcal{G}$

\[
m_\mathcal{G}(Y \mid X) = \frac{\prod_{C \in C} m(Y_C \mid X)}{\prod_{S \in S} m(Y_S \mid X)}
\]
Case 2) $\mathcal{D}$ not complete

- $\Omega_\mathcal{G}$ Markov with respect to a decomposable UG $\mathcal{G}$
- Marginal likelihood of $\mathcal{G}$

$$m_\mathcal{G}(Y | X) = \frac{\prod_{C \in \mathcal{C}} m(Y_C | X)}{\prod_{S \in \mathcal{S}} m(Y_S | X)}$$

- $\mathcal{C}$ set of cliques
- $\mathcal{S}$ set of separators
- $Y_C$ submatrix of responses belonging to $C \in \mathcal{C}$
- $Y_S$ submatrix of responses belonging to $S \in \mathcal{S}$

- Important: $m(Y_C)$ is computed under any complete graph using the matrix normal Wishart fractional prior (and similarly for $m(Y_S)$)
- Use (1) to obtain the marginal likelihood of an EG.
Case 2) $\mathcal{D}$ not complete

- $\Omega_G$ Markov with respect to a decomposable UG $G$
- Marginal likelihood of $G$

$$m_G(Y \mid X) = \frac{\prod_{C \in C} m(Y_C \mid X)}{\prod_{S \in S} m(Y_S \mid X)}$$

- $C$ set of cliques
Case 2) $\mathcal{D}$ not complete

- $\Omega_G$ Markov with respect to a decomposable UG $G$
- Marginal likelihood of $G$

$$m_G(Y \mid X) = \frac{\prod_{C \in C} m(Y_C \mid X)}{\prod_{S \in S} m(Y_S \mid X)}$$

- $C$ set of cliques
- $S$ set of separators
Case 2) $\mathcal{D}$ not complete

- $\Omega_G$ Markov with respect to a decomposable UG $G$
- Marginal likelihood of $G$

\[
m_G(Y \mid X) = \frac{\prod_{C \in C} m(Y_C \mid X)}{\prod_{S \in S} m(Y_S \mid X)}
\]

- $C$ set of cliques
- $S$ set of separators
- $Y_C$ submatrix of responses belonging to $C \in C$
Case 2) $\mathcal{D}$ not complete

- $\Omega_G$ Markov with respect to a decomposable UG $G$
- Marginal likelihood of $G$

$$m_G(Y \mid X) = \frac{\prod_{C \in \mathcal{C}} m(Y_C \mid X)}{\prod_{S \in \mathcal{S}} m(Y_S \mid X)}$$

- $\mathcal{C}$ set of cliques
- $\mathcal{S}$ set of separators
- $Y_C$ submatrix of responses belonging to $C \in \mathcal{C}$
- $Y_S$ submatrix of responses belonging to $S \in \mathcal{S}$
Case 2) $\mathcal{D}$ not complete

- $\Omega_\mathcal{G}$ Markov with respect to a decomposable UG $\mathcal{G}$
- Marginal likelihood of $\mathcal{G}$

$$m_\mathcal{G}(Y \mid X) = \frac{\prod_{C \in C} m(Y_C \mid X)}{\prod_{S \in S} m(Y_S \mid X)}$$

- $C$ set of cliques
- $S$ set of separators
- $Y_C$ submatrix of responses belonging to $C \in C$
- $Y_S$ submatrix of responses belonging to $S \in S$

- Important: $m(Y_C)$ is computed under any complete graph using the matrix normal Wishart fractional prior (and similarly for $m(Y_S)$)
Case 2) $\mathcal{D}$ not complete

- $\Omega_G$ Markov with respect to a decomposable UG $G$
- Marginal likelihood of $G$

$$m_G(Y \mid X) = \frac{\prod_{C \in \mathcal{C}} m(Y_C \mid X)}{\prod_{S \in \mathcal{S}} m(Y_S \mid X)}$$

- $\mathcal{C}$ set of cliques
- $\mathcal{S}$ set of separators
- $Y_C$ submatrix of responses belonging to $C \in \mathcal{C}$
- $Y_S$ submatrix of responses belonging to $S \in \mathcal{S}$

- Important: $m(Y_C)$ is computed under any complete graph using the matrix normal Wishart fractional prior (and similarly for $m(Y_S)$)
- Use (1) to obtain the marginal likelihood of an EG.
EG Chain component $\tau$

Fractional prior

$$p^F(B_\tau, \Omega_\tau) \propto \left| \Omega_\tau \right|^\frac{aD+n_0-|p_a \mathcal{G}(\tau)|-|\tau|-2}{2} \cdot e^{-\frac{n_0}{2} \text{tr} \left( \Omega_\tau \left\{ (B_\tau - \hat{B}_\tau)^\top \hat{C}_\tau (B_\tau - \hat{B}_\tau) + \tilde{R}_\tau \right\} \right),$$
Fractional prior

\[ p^F(B_\tau, \Omega_\tau) \propto |\Omega_\tau| \frac{^{a_D+n_0-|p\Lambda(\tau)|-|\tau|-2}}{2} \]
\[ \cdot e^{-\frac{n_0}{2} \text{tr} \left( \Omega_\tau \left\{ (B_\tau - \hat{B}_\tau)^\top \tilde{C}_\tau (B_\tau - \hat{B}_\tau) + \tilde{R}_\tau \right\} \right)} , \]

\[ \hat{B}_\tau = (X_\tau^\top X_\tau)^{-1} X_\tau^\top Y_\tau \]
EG Chain component $\tau$

Fractional prior

$$p^F(B_\tau, \Omega_\tau) \propto |\Omega_\tau|^{a_D+n_0-|p_{\mathcal{A}}(\tau)|-|\tau|-2} \cdot e^{-\frac{n_0}{2} \text{tr} \left( \Omega_\tau \left\{ (B_\tau - \hat{B}_\tau)^\top \tilde{C}_\tau (B_\tau - \hat{B}_\tau) + \tilde{R}_\tau \right\} \right)}$$

- $\hat{B}_\tau = (X_\tau^\top X_\tau)^{-1} X_\tau^\top Y_\tau$
- $\hat{E}_\tau = (Y_\tau - X_\tau \hat{B}_\tau)$
Fractional prior

\[ p^F(B_\tau, \Omega_\tau) \propto |\Omega_\tau|^{a_D+n_0-|\text{pa}_G(\tau)|-|\tau|-2} \]
\[ \cdot e^{-\frac{n_0}{2} \text{tr}(\Omega_\tau \{ (B_\tau - \hat{B}_\tau)^\top \tilde{C}_\tau (B_\tau - \hat{B}_\tau) + \tilde{R}_\tau \})} , \]

- \( \hat{B}_\tau = (X_\tau^\top X_\tau)^{-1} X_\tau^\top Y_\tau \)
- \( \hat{E}_\tau = (Y_\tau - X_\tau \hat{B}_\tau) \)
- \( \tilde{C}_\tau = n^{-1}X_\tau^\top X_\tau \)
EG  Chain component $\tau$

Fractional prior

$$p^F(B_\tau, \Omega_\tau) \propto |\Omega_\tau|^\frac{a_D+n_0-|p_{A^G(\tau)}|-|\tau|-2}{2} \cdot e^{-\frac{n_0}{2} \text{tr}(\Omega_\tau \{(B_\tau-\hat{B}_\tau)^\top \tilde{C}_\tau (B_\tau-\hat{B}_\tau) + \tilde{R}_\tau\})},$$

- $\hat{B}_\tau = (X_\tau^\top X_\tau)^{-1} X_\tau^\top Y_\tau$
- $\hat{E}_\tau = (Y_\tau - X_\tau \hat{B}_\tau)$
- $\tilde{C}_\tau = n^{-1} X_\tau^\top X_\tau$
- $\tilde{R}_\tau = n^{-1} \hat{E}_\tau^\top \hat{E}_\tau$
Fractional prior

\[ p^F(B_\tau, \Omega_\tau) \propto |\Omega_\tau|^{\frac{aD+n_0 - |\nu_G(\tau)| - |\tau| - 2}{2}} \cdot e^{-\frac{n_0}{2} \text{tr} \left( \Omega_\tau \left\{ (B_\tau - \hat{B}_\tau)^\top \tilde{C}_\tau (B_\tau - \hat{B}_\tau) + \tilde{R}_\tau \right\} \right)}, \]

- \( \hat{B}_\tau = (X_\tau^\top X_\tau)^{-1} X_\tau^\top Y_\tau \)
- \( \hat{E}_\tau = (Y_\tau - X_\tau \hat{B}_\tau) \)
- \( \tilde{C}_\tau = n^{-1} X_\tau^\top X_\tau \)
- \( \tilde{R}_\tau = n^{-1} \hat{E}_\tau^\top \hat{E}_\tau \)

- Matrix normal Wishart conjugate to the conditional sampling model in chain component \( \tau \)
Marginal density of data submatrix in $\tau$

$$m_{\tau}(Y_{J,\tau} | X_{\tau}) = \pi^{-\frac{(n-n_0) |J|}{2}} \frac{\Gamma |J| (\frac{a_D+n-|pa_G(\tau)|-1-|\bar{J}|}{2})}{\Gamma |J| (\frac{a_D+n_0-|pa_G(\tau)|-1-|\bar{J}|}{2})} \cdot \left(\frac{n_0}{n}\right)^{\frac{|J|(a_D+n_0-|\bar{J}|)}{2}} |\hat{E}_{J,\tau} \hat{E}_{J,\tau}|^{-\frac{n-n_0}{2}}$$

- $\bar{J} = \tau \setminus J$, $|\bar{J}| = |\tau| - |J|$
Marginal density of data submatrix in $\tau$

$$m_\tau(Y_{J,\tau} \mid X_{\tau}) = \pi^{-(n-n_0|J|)/2} \frac{\Gamma{|J|}{\left(\frac{a_D+n_0-pa_G(\tau)-1-|J|}{2}\right)}}{\Gamma{|J|}{\left(\frac{a_D+n-n_0-pa_G(\tau)-1-|\bar{J}|}{2}\right)}} \cdot \left(\frac{n_0}{n}\right)^{\frac{|J|(a_D+n_0-|\bar{J}|)}{2}} \left|\hat{E}_{J,\tau}^\top \hat{E}_{J,\tau}\right|^{-\frac{n-n_0}{2}}$$

- $\bar{J} = \tau \setminus J$, $|\bar{J}| = |\tau| - |J|$
- $\hat{E}_{J,\tau} = (Y_{J,\tau} - X_{\tau} \hat{B}_{J,\tau})$
Marginal density of data submatrix in $\tau$

$$m_\tau(Y_{J,\tau} | X_\tau) = \pi \left( \frac{n - n_0}{2} \right) \frac{\Gamma|J|}{\Gamma|\bar{J}|} \left( \frac{a_D + n - |\text{pa}_G(\tau)| - 1 - |\bar{J}|}{2} \right) \left( \frac{a_D + n_0 - |\text{pa}_G(\tau)| - 1 - |\bar{J}|}{2} \right) \cdot \left( \frac{n_0}{n} \right)^{\frac{|J|(a_D + n_0 - |\bar{J}|)}{2}} |\hat{E}_{J,\tau}^\top \hat{E}_{J,\tau}|^{-\frac{n - n_0}{2}}$$

- $\bar{J} = \tau \setminus J$, $|\bar{J}| = |\tau| - |J|
- $\hat{E}_{J,\tau} = (Y_{J,\tau} - X_\tau \hat{B}_{J,\tau})$
- $\hat{B}_{J,\tau} = (X_\tau^\top X_\tau)^{-1} X_\tau^\top Y_{J,\tau}$