Objective Bayesian model selection in general Gaussian graphical models

Mathilde Bouriga

EDF R&D Département OSIRIS - Universite Paris-Dauphine

8 décembre 2011

Application context

Market-risk assessment for high-dimensional asset portfolio.

- Portfolio variation $V_{t,t+h}$ between t and $t+h$ affected by risk factors, specifically by price returns X of portfolio products.
- A widely used risk measure : *Value-at-risk*.
  $\mathrm{VaR}_{1-\alpha}$ at a risk level α over a given time horizon h
  $=$ the α-quantile of the portfolio variation between t and $t+h$.

$$Pr(V_{t,t+h} < VaR_{1-\alpha}) = \alpha\%.$$

VaR Computation

- Method = the analytic VaR, built on 2 assumptions :
  1. portfolio variation as a linear combination of product returns, $V_{t,t+h} = a^T X_{t,t+h}$,
  2. normal distribution assumptions about returns, $X_{t,t+1}|\Sigma \sim \mathcal{N}_p(\mathbf{0}, \Sigma)$.

  $\Rightarrow \text{VaR}_{1-\alpha} = \sqrt{h}\sqrt{a^T \Sigma a}\Phi^{-1}(\alpha)$, calculated from $\hat{\Sigma}$,
  with $\Phi^{-1}(\alpha)$ the α-quantile of the standard normal distribution.

- Problem : sensitivity of VaR results to variations of $\hat{\Sigma}$ + unstable matrix estimator, as with a small sample.

  $\hookrightarrow$ Requirement : stable covariance matrix between returns.

Data

- Portfolio made of 27 energy products.
- The covariance matrix for the returns X on the products in the portfolio to estimate, *i.e* 378 elements to estimate.
- Matrix to estimate from 200 observations.

Problem formalization

$$X|\Sigma \sim \mathcal{N}_p(\mathbf{0}, \Sigma),$$

where Σ is a $p \times p$ symmetric positive-definite matrix.

Problem : **Estimation of** Σ from a sample of X, $\mathbf{X}$=$(X_1, ..., X_n)$ where p **is close to** n.

Classical estimator based on the scatter matrix $S_n = \mathbf{X}^T\mathbf{X}$: inappropriate.

- unstable estimator
- distortion of the eigenstructure
- S_n no longer positive definite if $p \geq n$.

Alternatives

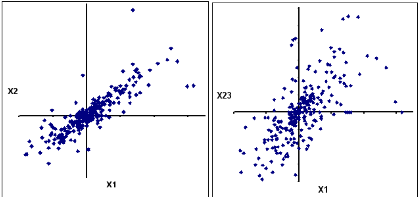
General approaches to induce stability over the unstructured classical estimator of the covariance matrix :

- by shrinking of eigenvalues,
- by shrinking this estimate toward a parsimonious, structured form of the matrix,
- by imposing various restrictions on the model and then estimating covariance matrix related to these structural assumptions.

Alternatives

General approaches to induce stability over the unstructured classical estimator of the covariance matrix :

- by shrinking of eigenvalues,
- by shrinking this estimate toward a parsimonious, structured form of the matrix,
- by imposing various restrictions on the model and then estimating covariance matrix related to these structural assumptions.



Alternatives

General approaches to induce stability over the unstructured classical estimator of the covariance matrix :

- by shrinking of eigenvalues,
- by shrinking this estimate toward a parsimonious, structured form of the matrix,
- by imposing various restrictions on the model and then estimating covariance matrix related to these structural assumptions.

Bayesian inference on covariance matrices
in Gaussian Graphical models
$\Rightarrow$ Visual aid - interpretation / Aid in parameter estimation

Graph theory

An undirected graph is a pair $G = (V, E)$ with vertex set V and edge set $E = \{(i, j)\}$ for some pairs $(i, j) \in V$.

A *clique* C of G is a set of pairwise adjacent vertices.

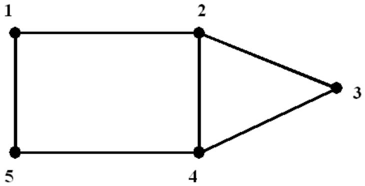


Figure: Graph G with 5 nodes and 6 edges.

Matrix theory

Let Σ be a matrix, the *G-incomplete symmetric matrix* Σ^E is defined as an incomplete symmetric matrix indexed by $V \times V$, in which the elements are those of Σ_{ij} for all $(i,j) \in E$, and with the remaining elements unspecified.

$$\Sigma^E = \begin{pmatrix} \sigma_{11} & \sigma_{12} & * & * & \sigma_{15} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{14} & * \\ * & \sigma_{32} & \sigma_{33} & \sigma_{34} & * \\ * & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} \\ \sigma_{51} & * & * & \sigma_{54} & \sigma_{55} \end{pmatrix}$$

A *completion* of an incomplete matrix is a specific choice of values for the unspecified entries.

Gaussian graphical model GGM (1)

A GGM uses a graphical structure to define a set of pairwise conditional independence relationships.

- With precision matrix $\Omega = \Sigma^{-1}$, X_i and X_j of X are conditionally independent (given the neighboring variables of each) iff $\omega_{ij} = 0$.
- If $G = (V, E)$ is an undirected graph whose vertices are associated with X, $(|V| = p)$, $\omega_{ij} = 0$ for all pairs $(i, j) \notin E$.

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} & 0 & 0 & \omega_{15} \\ \omega_{21} & \omega_{22} & \omega_{23} & \omega_{14} & 0 \\ 0 & \omega_{32} & \omega_{33} & \omega_{34} & 0 \\ 0 & \omega_{42} & \omega_{43} & \omega_{44} & \omega_{45} \\ \omega_{51} & 0 & 0 & \omega_{54} & \omega_{55} \end{pmatrix} \Leftrightarrow X_1 \perp X_3 | X_2, X_4, X_5 \; ...$$

GGM (2)

Let $G = (V, E)$ and $M^+(G)$ denote the cone of $|V| \times |V|$ positive definite matrices such that ij entry is equal to 0 whenever $(i, j) \notin E$.

A GGM with graph G is

$$\mathcal{M}_G = \left\{ \mathcal{N}(\mathbf{0}, \Sigma) \middle| \ \Omega = \Sigma^{-1} \text{ and } \Omega \in M^+(G) \right\}.$$

On the covariance space, incomplete matrices Σ^E to handle : far from simple.

Two challenging problems for covariance estimation in GGM

1. graphical model selection problem
   = problem of estimating the zero-pattern of Ω,
2. covariance matrix estimation based on the model selected.

Two challenging problems for covariance estimation in GGM

1. graphical model selection problem
   = problem of estimating the zero-pattern of Ω,
2. covariance matrix estimation based on the model selected.

in a Bayesian framework.

$$X|\Sigma \sim \mathcal{N}_p(\mathbf{0}, \Sigma),\ \Omega = \Sigma^{-1} \in M^+(G)$$

Parameters : Ω, *nuisance parameter*, and G, *parameter of interest*.

- priors to handle : $\pi(\Omega, G) = \pi(\Omega|G)\pi(G)$,
- posterior to handle : $\pi(G|X) = \int \pi(\Omega, G|X)d\Omega$,
- estimator to choose : $\hat{G}$.

With our real data

Example : focus on the 9 first variables.
Starting from the empirical covariance matrix, we seek to reduce problem complexity and find the underlying conditional-dependency structures.

$$\begin{pmatrix} 1.00 & 0.85 & 0.76 & 0.75 & 0.64 & 0.72 & 0.66 & 0.59 & 0.60 & \\ 0.85 & 1.00 & 0.84 & 0.79 & 0.72 & 0.74 & 0.69 & 0.63 & 0.62 & \\ 0.76 & 0.84 & 1.00 & 0.88 & 0.78 & 0.75 & 0.73 & 0.64 & 0.66 & \\ 0.75 & 0.79 & 0.88 & 1.00 & 0.80 & 0.80 & 0.76 & 0.67 & 0.70 & \\ 0.64 & 0.72 & 0.78 & 0.80 & 1.00 & 0.83 & 0.69 & 0.77 & 0.63 & \cdots \\ 0.72 & 0.74 & 0.75 & 0.80 & 0.83 & 1.00 & 0.71 & 0.90 & 0.64 & \\ 0.66 & 0.69 & 0.73 & 0.76 & 0.69 & 0.71 & 1.00 & 0.60 & 0.89 & \\ 0.59 & 0.63 & 0.64 & 0.67 & 0.77 & 0.90 & 0.60 & 1.00 & 0.59 & \\ 0.60 & 0.62 & 0.66 & 0.70 & 0.63 & 0.64 & 0.89 & 0.59 & 1.00 & \\ & & & & \cdots & & & & & \end{pmatrix} \Rightarrow$$

0	1	1	1	0	1	0	0	1
1	0	1	0	0	0	0	0	0
1	1	0	1	1	0	1	0	0
1	0	1	0	1	1	1	1	0
0	0	1	1	0	1	0	1	0
1	0	0	1	1	0	1	1	1
0	0	1	1	0	1	0	1	1
0	0	0	1	1	1	1	0	1
1	0	0	0	0	1	1	1	0

Figure: $\mathbf{X}^T\mathbf{X}$ and the underlying structure.

Substantial problems

- Which priors, $\pi(\Sigma|G)$ and $\pi(G)$, for efficient model search ? (explicit expression for prior in the decomposable case)
- Properness conditions for the posterior distribution ? (easier to derive in the decomposable case)
- Which tool to model comparison ? (depending on the choice of priors : proper or not)
- Which graphical model-selection procedure ? (search computationally less expensive in the decomposable case)

Decomposition

(A, B, C), subsets of V, form a *decomposition* of G if C is complete, *i.e* a set of pairwise adjacent vertices, and C is separator of A, B, *i.e* any path from A to B goes through C.

A sequence of subgraphs that cannot be decomposed further are the *prime components* of a graph ; if every prime component is clique, the graph is *decomposable*.

Any given graph can G be embedded in a decomposable graph by adding edges, the decomposable graph is called a *triangulation* of G.

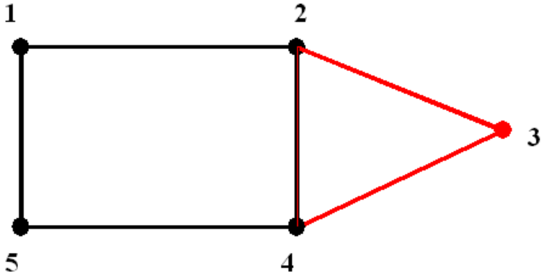


Figure: Graph decomposition.

$A = \{X_1, X_2, X_4, X_5\}$ is a prime component, $B = \{X_2, X_3, X_4\}$ is a clique and $C = \{X_2, X_4\}$ is a separator.

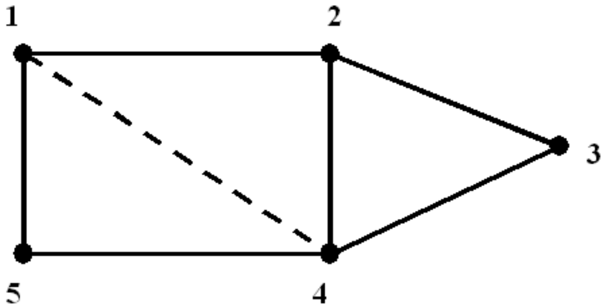


Figure: Triangulated graph.

All the prime components are cliques.

Although, in the literature, attention is often restricted to the decomposable case, only a small fraction of the total number of graphs on p nodes is decomposable.

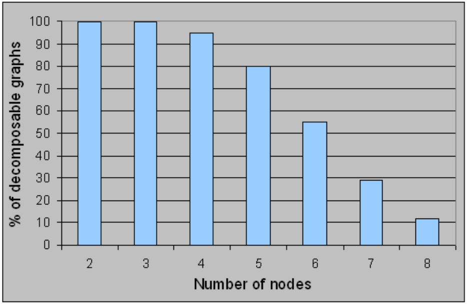


Figure: Proportion of decomposable graphs depending on the number of vertices.

$\Longrightarrow$ Graphical model selection for general graphs.

Standard prior for G

We choose to consider a Bernoulli distribution on the edge inclusion indicators with success probability β.

$$\pi(G \text{ with } k \text{ edges}|\beta) \propto \beta^k (1-\beta)^{m-k}$$

with $m = \frac{p(p-1)}{2}$, the maximum number of possible edges.

$\beta = \frac{1}{p-1}$ will encourage sparse graphs.

Standard prior for Ω in the literature (1)

As the GGM with graph $G = (V, E)$ is a regular exponential family [Lau96] with canonical parameter Ω, the standard conjugate prior for Ω in $M^+(G)$ can be written as

$$\pi_G(\Omega|\delta, D^E) = \frac{1}{Z(G, \delta, D^E)} |\Omega|^{(\delta-2)/2} \exp\left\{-\frac{1}{2} tr(\Omega D^E)\right\}$$

where δ, D^E are such that the normalizing constant $Z(G, \delta, D^E)$ is finite.

Standard prior for Ω (2)

$$\int_{M^+(G)} |\Omega|^{(\delta-2)/2} \exp\left\{-\frac{1}{2}tr(\Omega D^E)\right\} d\Omega < \infty$$

if $\delta > 2$ and the incomplete matrix D^E admits a positive completion.

In this case, it is called G-Wishart distribution with parameters (δ, D^E).

- In decomposable cases, $Z(G, \delta, D^E)$ available in a closed form,
- in non-decomposable cases, $Z(G, \delta, D^E)$ not available in a closed form.

A new objective prior for Ω (1)

We propose to consider this noninformative prior for Ω of a GGM with arbitrary graph G :

$$\pi_N(\Omega|G) \propto |\Omega|^{-1} \text{ for } \Omega \in M^+(G).$$

- Choice motivation : the involved default-procedure for GGM selection yields efficient posterior seperation of models.
- A particular case : this distribution corresponds to the prior proposed by [CS07] for model selection when considering only the decomposable graphs.

Proposition : The posterior density of G

$$\pi(G|\mathbf{X}) \propto \frac{\frac{1}{p-1}^{m}(p-2)^{m-k}}{\sqrt{2\pi}^{np}} Z(G, n, \mathbf{X}^T\mathbf{X}) \text{ with } m = \frac{p(p-1)}{2}$$

is proper iff

$$Z(G, n, (\mathbf{X}^T\mathbf{X})^E) = \int_{M^+(G)} |\Omega|^{\frac{n-2}{2}} \exp\left\{-1/2tr(\Omega(\mathbf{X}^T\mathbf{X})^E)\right\} d\Omega$$

is finite.

Sufficient conditions :

- $n > 2$
- $(\mathbf{X}^T\mathbf{X})^E$ has a positive completion : condition hard to find for general graphs.

Proposition : Let $G^+ = (V, E^+)$ be a minimal triangulation of G - a decomposable graph where $E^+ \supset E$, with the property that removal of any edge in G^+ which is not an edge in G will not be decomposable.

Let $\mathcal{C}^+$ denote the set of cliques of G^+.

$$n > \max_{C \in \mathcal{C}^+} |C^+| \Rightarrow (\mathbf{X}^T\mathbf{X})^E \text{ has a positive completion.}$$

Particular case : for the full graph, well-known condition.

Conclusion :

- $\pi(G|\mathbf{X})$ proper for all the graphs, when $n > p$.
- If $n \leq p$, restriction on the graphs under consideration. $\pi(G|\mathbf{X})$ proper for any graph in $\mathcal{S}_G = \left\{G|\ Z(G, n, (\mathbf{X}^T\mathbf{X})^E) < \infty\right\}$.

Structural learning in Gaussian graphical models usually involves assessing the posterior probability of the graphs to evaluate

$$\frac{\pi(G_1|\mathbf{X})}{\pi(G_2|\mathbf{X})} = \frac{\pi(G_1)}{\pi(G_2)} BF_{12}(\mathbf{X}),$$

where

$$BF_{12}(\mathbf{X}) = \frac{f(X|G_1)}{f(X|G_2)},$$

where $f(X|G_i) = \int_{M^+(G)} f(X|\Omega_i, G_i)\pi_i(\Omega|G_i)d\Omega_i$ is the marginal likelihood of G_i.

Bayesian model comparison is usually based on Bayes factors.

Definition

Using improper priors for parameters in alternative models $\Rightarrow$ Bayes factors not well defined :

$$BF_{12}(\mathbf{X}) = \frac{c_1}{c_2}\frac{f(X|G_1)}{f(X|G_2)}\text{ , with }\frac{c_1}{c_2}\text{ unknown.}$$

Alternative key : Fractional Bayes factors (FBF) introduced by

[O'H95] among Partial Bayes factors (PBF) [Per05].

$$FBF_{12}(\mathbf{X}) = \frac{q(X|G_1,g)}{q(X|G_2,g)},$$

with $q(X|G,g) = \int_{M^+(G)} f(X|\Omega)^{1-g}\pi_g(\Omega|G,\mathbf{X},g)d\Omega$, the *fractional marginal likelihood* of G.

Graph score based on Laplace approximations

$$q(X|G,g) = \frac{1}{\sqrt{2\pi}^{np}} \frac{Z(G, n, \mathbf{X}^T\mathbf{X})}{Z(G, gn, g\mathbf{X}^T\mathbf{X})} \text{ for } ng > 2.$$

We use the diagonal Laplace approximation proposed by [LD10] to estimate $Z(G,.,.)$ for any graph G.

A challenging issue

- p nodes in a graph $\Rightarrow m = \frac{p(p-1)}{2}$ possible edges
  $\Rightarrow 2^m$ possible graphs.
  Beyond $p = 7$, enumeration becomes a practical impossibility.
- Need to scalable search methodologies that are capable of finding good models, or at least distinguishing the important edges from the irrelevant ones.
- Main classes of graphical model-selection procedures : compositional methods and direct search.

$\Longrightarrow$ Framework proposed by [BMM09] which is a direct search method initialized with a set of graphs issued from a compositional method.

An heuristic search technique

An iterative algorithm which tries to identify the most likely graphs, inspired by [SC08].

At time t, starting with

- $t-1$ distinct explored graphs $(G_1, ..., G_{t-1})$,
- $t-1$ scores $q(X|G_1, g), ..., q(X|G_{t-1}, g)$
- estimated edge-inclusion probabilites $Pr(\omega_{ij} \neq 0|G_1, ..., G_{t-1})$, $i, j = 1, ..., p$,

3 steps :

1. perform a stochastic local update to the graph based on edge-inclusion probabilities $\Rightarrow$ new graph G_t,
2. score the graph $\Rightarrow q(X|G_t, g)$,
3. update the edge-inclusion probabilities $\Rightarrow Pr(\omega_{ij} \neq 0|G_1, ..., G_t)$.

Local update *via* 2 kinds of moves :

- local moves : choose randomly to add or delete an edge.
  If add, do so in proportion to their estimated edge-inclusion probabilites. If delete, do so in inverse proportion to them,
- resampling step : revisite one of $(G_1, ..., G_{t-1})$ in proportion to their score and make local move from the resampled graph.

Before SLS, good to initialize the search with a set of promising graphs for resampling.

Initialization strategy using Neighborhood Fusion

- Neighborhood Fusion (NF) to quickly produce large sets of high quality GGM structures.
- In the space of conditional regressions,
  1. it exploits the sparse linear regression method LASSO through LARS algorithm [Tib96] to compute a set of candidate neighborhood structures for each variable,
  2. it specifies a mechanism for sampling and
  3. a mechanism for combining these neighborhoods to form undirected graphs.

Model choice

We consider

- the graph with the highest score among those explored,
- the *median probability model* G_{med} :

$$G_{med} = (V, E_{med}),$$
$$\text{where } E_{med} = \{(i,j) : Pr\left(\omega_{ij} \neq 0 | G_1, ..., G_T\right) \geq 0.5\},$$

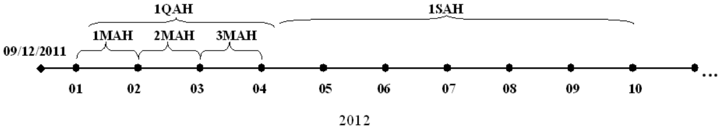
Choose it if its score is bigger and if it was not explored.

Portfolio : 27 energy products, called futures contracts, on the UK energy market.

Futures : contracts between two parties to exchange a specified commodity of standardized quantity and quality for a price agreed today with delivery occurring at a specified future date, the delivery date.
Here 27 futures :

- 9 of different delivery periods on the Electricity market (1Month AHead-2MAH-3MAH-1Quarter AHead-2QAH-1Season AHead-2SAH-3SAH-4SAH),
- 18 of different delivery periods on the Gaz market (1MAH-2MAH-3MAH-4MAH-5MAH-6MAH-7MAH-8MAH-9MAH-10MAH-11MAH-12MAH-13MAH-14MAH-15MAH-16MAH-17MAH-18MAH).

To understand futures... :



We apply our proposed model-selection procedure from 200 price returns in dimension 27. All the graphs are considered.

Result : matrix where element ij =1 if (i,j) is an edge of the selected graph.

	1MAH	2MAH	3MAH	1QAH	2QAH	1SAH	2SAH	3SAH	4SAH	1MAH	2MAH	3MAH	4MAH	5MAH	6MAH	7MAH	8MAH	9MAH	10MAH	11MAH	12MAH	13MAH	14MAH	15MAH	16MAH	17MAH	18MAH
1MAH	0	1	1	1	0	1	0	0	1	1	1	0	1	0	1	0	1	1	0	1	1	0	0	0	0	0	0
2MAH	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1	1	1	0	1	0	0	1	1	1	1
3MAH	1	1	0	1	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1
1QAH	1	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	1
2QAH	0	0	1	1	0	1	0	1	0	0	0	0	0	0	1	1	1	0	1	1	0	0	0	1	1	1	0
1SAH	1	0	0	1	1	0	1	1	1	0	0	1	0	1	0	0	1	0	0	1	1	0	0	0	0	0	1
2SAH	0	0	1	1	0	1	0	1	1	0	1	1	1	0	1	0	0	1	1	0	0	0	1	0	0	0	0
3SAH	0	0	0	1	1	1	1	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	1	1	1	1
4SAH	1	0	0	0	0	1	1	1	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	1
1MAH	1	1	1	0	0	0	0	0	1	0	1	0	1	1	0	0	0	1	1	1	1	1	0	0	1	0	1
2MAH	1	1	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	0
3MAH	0	0	1	0	0	1	1	0	0	0	1	0	1	0	1	1	0	1	1	1	1	1	0	0	0	1	1
4MAH	1	1	0	0	0	0	1	0	1	1	1	1	0	1	1	1	0	0	0	0	1	1	0	1	1	1	0
5MAH	0	1	0	0	0	1	0	1	0	1	1	0	1	0	1	0	1	1	0	1	1	1	1	0	1	1	1
6MAH	1	1	1	0	1	0	1	1	0	0	1	1	1	1	0	1	0	1	0	1	0	1	0	0	1	1	1
7MAH	0	0	0	0	1	0	0	0	0	0	1	1	1	0	1	0	1	1	1	1	1	1	0	0	1	1	1
8MAH	1	1	0	0	1	1	0	0	1	0	0	0	0	1	0	1	0	1	1	1	1	0	0	1	0	0	0
9MAH	1	1	0	1	0	0	1	0	0	1	1	1	0	1	1	1	1	0	1	1	0	1	1	0	0	1	1
10MAH	0	1	0	1	1	0	1	1	0	1	1	1	0	0	0	1	1	1	0	1	1	1	0	1	0	0	1
11MAH	1	0	1	0	1	1	0	0	1	1	1	1	0	1	1	1	1	1	1	0	1	1	0	1	0	1	1
12MAH	1	1	0	0	0	1	0	0	0	1	0	1	1	1	0	1	1	0	1	1	0	0	1	0	1	0	1
13MAH	0	0	1	0	0	0	0	0	0	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1
14MAH	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	1	1	0	1	1	1	1
15MAH	0	1	0	0	1	0	0	1	1	0	1	0	1	0	0	0	1	0	1	1	0	1	1	0	1	0	1
16MAH	0	1	0	1	1	0	0	1	0	1	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	1	1
17MAH	0	1	0	0	1	0	0	1	0	0	1	1	1	1	1	1	0	1	0	1	0	1	1	0	1	0	1
18MAH	0	1	1	1	0	1	0	1	1	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0

Figure: An idea of conditional-independence relationships between asset returns

- GGM : tractable model for covariance matrices in many dimensions and/or small samples + knowledge discovery.
- Study problem : estimation of the graph structure associated to a GGM.
- Main contributions : complete methodology to perform objective Bayesian model selection in general GGM - new objective matrix prior, properness condition for posterior, tools for model comparison and exploration of large model space.
- Perspective : estimation of the associated covariance matrix.

M.E Kahn B. Moghaddam, B. Merlin and K.P. Murphy.
Accelerating bayesian structural inference for non-decomposable gaussian graphical models.
NIPS, 2009.

C.M. Carvalho and J.G. Scott.
Objective bayesian model selection in gaussian graphical models.
Biometrika, 96(3) :497–512, 2007.

S.L. Lauritzen.
Graphical Models.
Oxford University Press, 1996.

A. Lenkoski and A. Dobra.
Bayesian structural learning and estimation in gaussian graphical models.
(545) :1–18, 2010.

A. O'Hagan.

Fractional bayes factors for model comparison.

57(1) :99–138, 1995.

L.R. Perrichi.

Model selection and hypothesis testing based on objective probabilities and bayes factors.

25 :*Bayesian Thinking Modeling and Computation*(D. K. Dey and C. R. Rao, eds.) 115Ű149. North–Holland, New York, 2005.

J.G. Scott and C.M. Carvalho.

Feature-inclusion stochastic search for gaussian graphical models.

Journal of computational and graphical statistics, 17(4) :790–808, 2008.

R. Tibshirani.

Regression shrinkage and selection via the lasso.

58(1) :267–288, 1996.