Graphical Models

-Statistical Machine Learning-

Lecturer: Darren Homrighausen, PhD

Conditional independence

The core idea encoded in graphical models are conditional independence relations

A priori independent causes of an event can become not independent with a new measurement

Conditional independence

Example: Suppose you live in Los Angeles and are on a business trip to New York.

Your phone rings to notify you that your home security system has been activated

Simultaneously, you notice a news report that there has been an earthquake in LA

Given that you know from prior experience that earthquakes sometimes cause false alarms, you feel that an actual burglary is less likely.

Graphs

The expression of conditional independence relations can be expressed with a graph

A graph is a pair $G = \{V, E\}$, where

- V is a set of vertices
- E is a set of edges

  (Really, E is a set of (possibly ordered) pairs from V)

For our purposes, each vertex corresponds to a random variable

Each edge represents some aspect of their joint distribution

(For our purposes, we will only consider undirected graphs and hence the ordering doesn't matter)

Graphs

Let $x = (X_1, \ldots, X_p)^\top \sim \mathbb{P}$

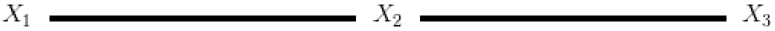
A graph G for $\mathbb{P}$ has p vertices (aka nodes)

The crucial aspect is that the absence of an edge encodes conditional independence

$$\{j,k\} \notin E \Rightarrow X_j \perp X_k | \mathrm{rest}$$

(This a Markov property that is a bit technical in full generality. See `http://www.stat.cmu.edu/ larry/=sml/GraphicalModels.pdf` for an indepth discussion)

Example:



$$X_1 \perp X_3 | X_2$$

Statistical graphical models

The Markov part is tricky as there isn't, in general, a 1-1 map between $\mathbb{P}$ and G

For our purposes, let's somewhat reductively define the following

- $I(G) = \text{all independence statements implied by } G$
- $I(\mathbb{P}) = \text{all independence statements implied by } \mathbb{P}$
- $\mathcal{P}(G) = \{\mathbb{P} : I(G) \subseteq I(\mathbb{P})\}$
- If $\mathbb{P} \in \mathcal{P}(G)$ then we say that $\mathbb{P}$ is Markov to G
- In this case, G represents a class of distributions

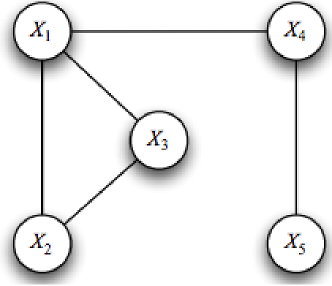
Example: The graph $X_1 \cdots X_2$ has $I(G) = \emptyset$. All bivariate distributions are in $\mathcal{P}(G)$, including $p(X_1, X_2) = p(X_1)p(X_2)$

Nonparametric statistical graphical models

Undirected (Markov) graphical models allow a decomposition into clique potentials

A clique is a fully connected subgraph

A maximal clique is such that it is not contained in any larger clique



Nonparametric statistical graphical models

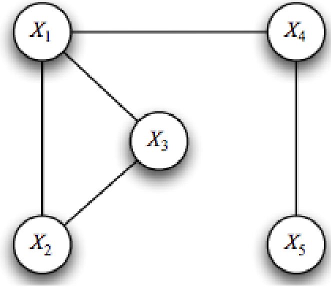
Let $\mathcal{C}$ be the set of all maximal cliques in a graph

Hammersley and Clifford: A Markov and "nice" measure $\mathbb{P}$ can be factored multiplicatively as

$$p(X_1, \ldots, X_p) = \prod_{C \in \mathcal{C}} \psi_C(X_C)$$

The ψ_C are known as clique potentials

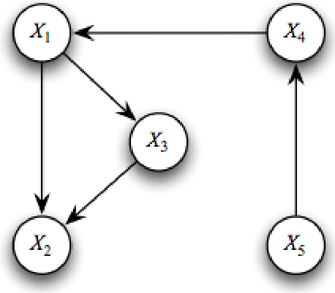
Nonparametric statistical graphical models



The family of distributions represented by this graph can be factored as

$$p(X_1, \ldots, X_5) = \psi_{1,2,3}(X_1, X_2, X_3)\psi_{1,4}(X_1, X_4)\psi_{4,5}(X_4, X_5)$$

Directed graphical models



The family of distributions represented by this graph can be factored as

$$p(X_1, \ldots, X_5) = p(X_5)p(X_4|X_5)p(X_1|X_4)p(X_3|X_1)p(X_2|X_1, X_3)$$

Parametric statistical graphical models

Goal: Given a sample $x_1, \dots, x_n \sim \mathbb{P}$, we wish to estimate (or less ambitiously, constrain) the graph G

Using properties of Gaussian distributions, we know that

$$X_j \perp X_k | \mathrm{rest} \Leftrightarrow \Omega_{jk} = 0$$

where $\Omega = \Sigma^{-1}$ is the precision matrix

Parametric statistical graphical models

Suppose we are in low dimensions

(That is, $n >> p$)

We can use the usual MLE to find $\hat{\Omega}$

$$\begin{aligned}\log p(x_1, \ldots, x_n|\Omega) &\propto \log\left(|\Omega|^{n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\hat{\mu})^\top \Omega (x_i-\hat{\mu})}\right) \\ &\propto \frac{1}{2}\left(\log|\Omega| - n\mathrm{trace}(\Omega S)\right)\end{aligned}$$

where S is the sample covariance (Here, I've maximized over μ)

This gives $S^{-1} = \hat{\Omega}$ and we can test where $\Omega_{jk} = 0$

Parametric statistical graphical models

As usual, use the MLE at your own risk, especially if n isn't extremely large relative to p

Example: Suppose we collect S& P 500 data from January 1, 2003 to January 1, 2008

(This will only be 452 stocks, as we'll only take the intersection of the listing over time)

This gives us $X_j \in \mathbb{R}^{1258}$ for $j = 1, \ldots, 452$

(X_{jt} is the price of j^{th} stock on t^{th} day)

Of course, these are autocorrelated, hence we report[1]

$$X_{jt} = \log(X_{jt}/X_{j,t-1})$$

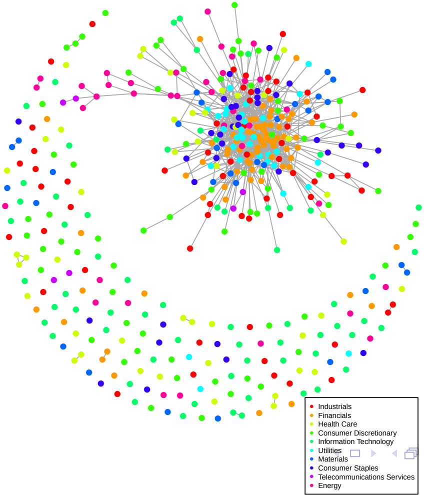
[1] Plus some outlier truncation

Parametric statistical graphical models

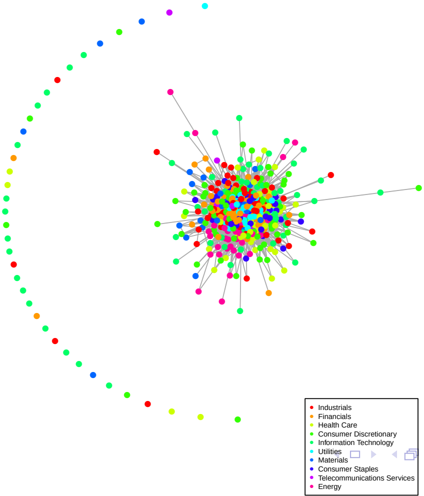
After estimating Ω, we can plot the resulting graph for various thresholds on the size of the entry in Ω

This gives us an idea of some strength of conditional independence relations

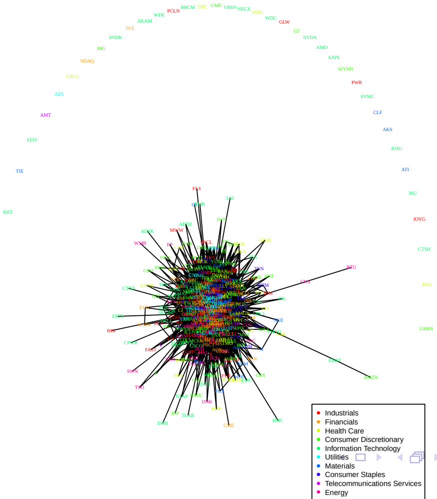
Parametric statistical graphical models



Parametric statistical graphical models



Parametric statistical graphical models



Parametric statistical graphical models

There is an R package for doing this a bit more formally: SIN

The etymology is from terminology rampant in the field of graphical models

- Faithfulness: This occurs when $I(\mathbb{P}) = I(G)$
- Moral graph: The undirected version of a DAG that has the 'same' independence relations

Hence terms related to morality persist in the field

Parametric statistical graphical models

SIN is a pseudo-acronym for partitioning the vertices into a(n)

- significant set S
- indeterminate set I
- non-significant N

This can be thought of as a way for controlling the overall error rate for incorrect edge inclusion

SIN output two graphs:

- A graph whose edges are in $S \cup I$
- A graph whose edges are in S

SIN

SIN is comprised of testing the partial correlation of each pair of covariates, given the others

In previous work, this testing was done in a backwards stepwise fashion

The largest p-value is determined and the edge is removed from the graph

(The null-hypothesis is that the correlation coefficient is 0)

This approach has some obvious flaws

(A clear mis-use of p-values, no control of familiy-wise error rate)

SIN

In the SINful approach, they do the following (details omitted)

1. Identify that the sample covariance approximately follows a Wishart distribution
2. Using the delta method + a z-transformation, we get an asymptotic normal for the sample partial correlations
3. Use a Gaussian concentration result to get family-wise p-values
4. These p-values get partitioned into S, I, and N

   (Perhaps using $S = (0, 0.05]$, $I = (0.05, .25]$, and $N = (.25, 1]$. Most common is to visualize the p-values and subjectively bin them)

(See Drton, Perlman (2004) for details)

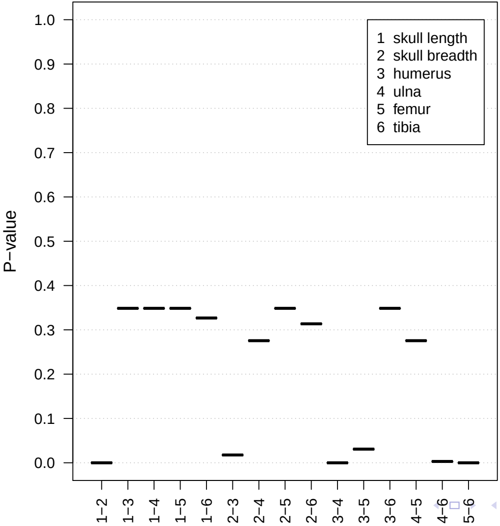
SIN

(The stock data doesn't work well with SIN (too high dimesional). Example from help file instead)

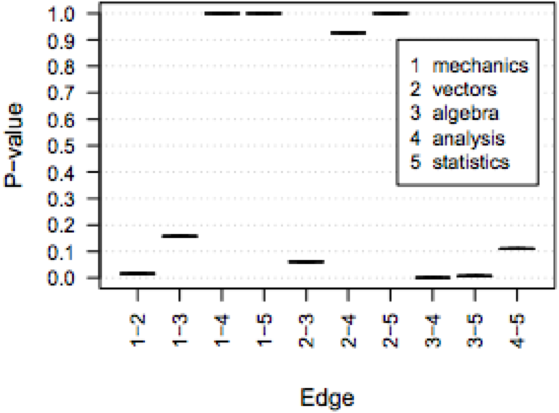
```
data(fowlbones)
pvals <- holm(sinUG(fowlbones$corr,fowlbones$n))
plotUGpvalues(pvals)
```

Note: the holm function implements a technique from a paper in 1979 that 'improves p-values while still allowing valid simultaneous testing'

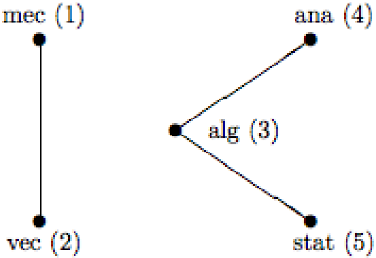
Parametric statistical graphical models



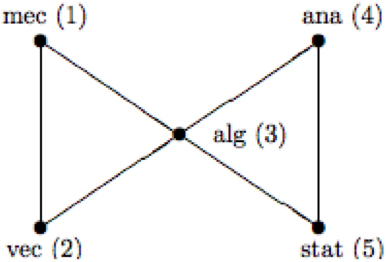
Parametric statistical graphical models



Parametric statistical graphical models



S



S, I

Regularized statistical graphical models

There are two common methods for estimation when $p > n$

- parallel lasso

  (Meinshausen and Buhlmann (2006))
- Graphical lasso

  (Banerjee et al.(2008) or Hastie et al.)

Parallel lasso

This is conceptually quite simple

1. For each $j = 1, \ldots, p$, regress X_i on all other variables using lasso
2. Put an edge between X_i and X_j if each appears in the active set of the other variable

Graphical lasso

This approach takes the usual likelihood

$$\begin{aligned}\log p(x_1,\ldots,x_n|\Omega) &\propto \log\left(|\Omega|^{n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\hat{\mu})^\top\Omega(x_i-\hat{\mu})}\right)\\ &\propto \frac{1}{2}\left(\log|\Omega| - n\mathrm{trace}(\Omega S)\right)\end{aligned}$$

and penalizes it

$$\min -\frac{1}{2}\left(\log|\Omega| - n\mathrm{trace}(\Omega S)\right) + \lambda\left|\left|\Omega\right|\right|_1$$

($||\cdot||_1$ is the matrix functional given by the sum of the absolute values of the entries)

Regularized statistical graphical models in R

Both can be accomplished with the huge package

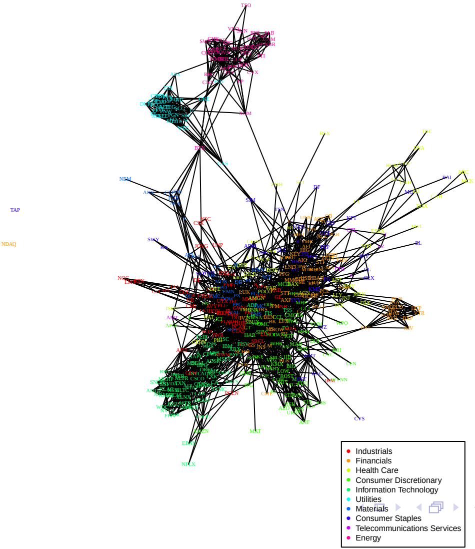
- parallel lasso

  ```
  out.parallel  = huge(cov.hat,method = "mb")
  ```

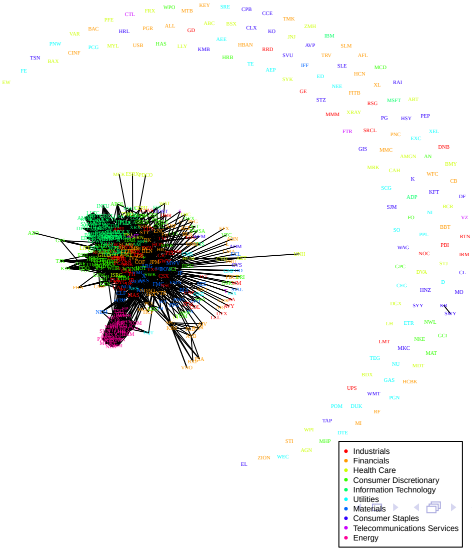
- Graphical lasso

  ```
  out.glasso    = huge(cov.hat,method = "glasso")
  ```

Parallel lasso



Graphical lasso



R code

```
g = graph.adjacency(adj.mat,mode="undirected",diag=F)
plot(g, layout=layout.auto,edge.color=’black’,
     vertex.size=.1,vertex.label=labels,
     vertex.frame.color=NA,vertex.label.cex=.3,
     vertex.label.color=col.vec)
legend(x=x,y=y,legend=legend,col=col.palette,pch=16)
```

There is a nice book called Graphical models with R, along with an online document with the same title