RCOX models: Graphical Gaussian models with edge and vertex symmetries

Søren Højsgaard, University of Aarhus, Denmark

Joint work with

Steffen Lauritzen, University of Oxford

and

Sren Hjsgaard, Sören Höjsgaard,

Særen Hæjsgaard and S???ren H???jsgaard

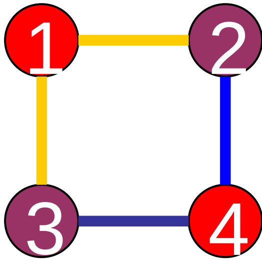
Take-home message and Outline

- **New types of graphical Gaussian models**
  - With colours; as in the logo
  - Attribute specific meanings to the colours

- **An R-package (gRc) for inference in these models**

- **Motivation**
- **RCO'X' models: RCON, RCOR, RCOP**
- **Estimation algorithms**
- **Software**

Idea…

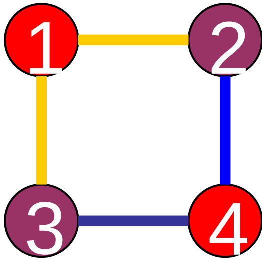


- **Apply graphical (Gaussian) models to large matrices, e.g. in gene expression**
- **Problem: d >>n, many genes few replicates**

Graphical model for $Y \sim N_d(0,\Sigma)$

- **When d >n, MLE of Σ does not exist (in saturated model)**
- **Impose restrictions on $K=\Sigma^{-1}$ to achieve more parsimonious model;**
- **In addition to conditional independence, restrict parameters to being identical**

Concentration and derived quantities



- Model for $y \sim N_d(0, \Sigma)$, $K = (k^{ij}) = \Sigma^{-1}$.
- The partial (conditional) covariance :

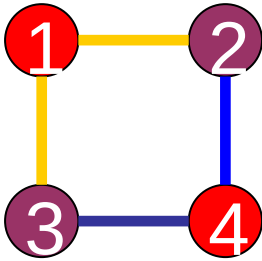
$$cov(y_i, y_j | rest) = \frac{-k^{ij}}{\det K^{ij}} \text{ where } K^{ij} = \begin{bmatrix} k^{ii} & k^{ij} \\ k^{ij} & k^{jj} \end{bmatrix}$$

- The partial (conditional) correlation:

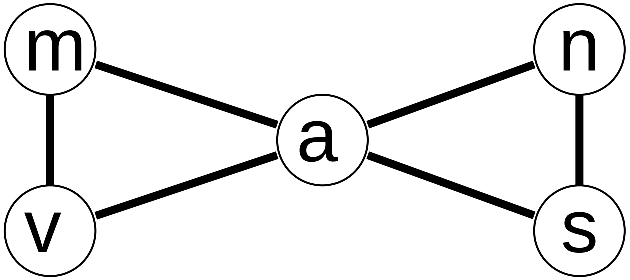
$$cor(y_i, y_j | rest) = \rho^{ij} = \frac{-k^{ij}}{\sqrt{k^{ii} k^{jj}}}$$

- Conditional independence: $y_i \perp\!\!\!\perp y_j | rest$ iff $k^{ij} = 0$

Example: Mathematics marks

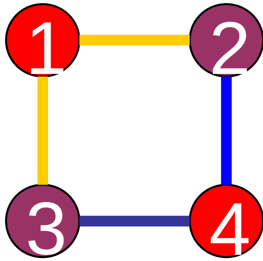


- **Mathmark data (Mardia, Kent, Bibby): 88 students marks on (a)lgebra, a(n)alysis, (m)echanics, (v)ectors and (s)tatistics**
- **Stepwise backward selection gives butterfly model**



- Convention: Black and white are neutral colours - corresponding parameters are unrestricted.

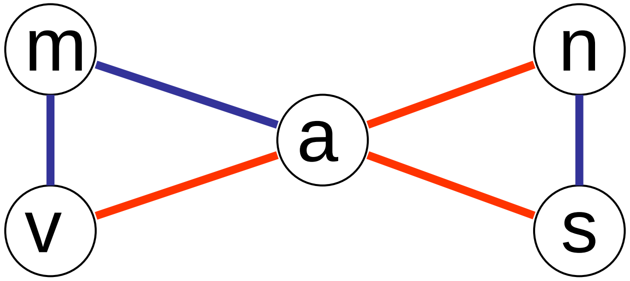
Concentrations for mathmarks



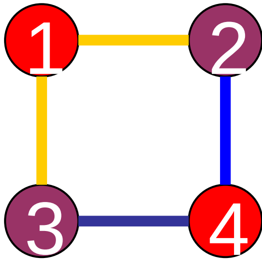
Concentrations ($\times 1000$):

	mechanics	vectors	algebra	analysis	statistics
mechanics	5.24	−2.44	−2.74	0.01	−0.14
vectors	−2.44	10.43	−4.71	−0.79	−0.17
algebra	−2.74	−4.71	26.95	−7.05	−4.70
analysis	0.01	−0.79	−7.05	9.88	−2.02
statistics	−0.14	−0.17	−4.70	−2.02	6.45

- **Some concentrations $\approx$ 0**
- **Some concentrations $\approx$ identical**



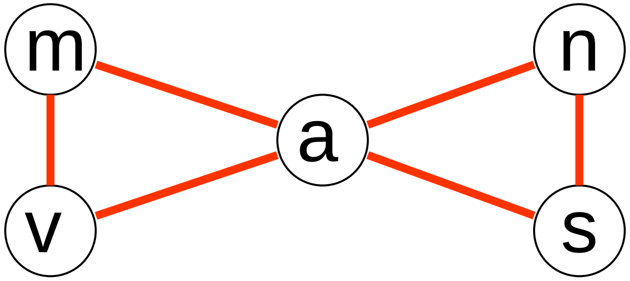
Partial correlations for mathmarks



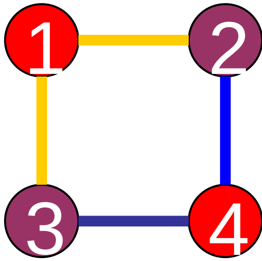
Partial correlations:

	mechanics	vectors	algebra	analysis	statistics
(m)echanics	1.00	−0.30	−0.20	0.00	0.00
(v)ectors	−0.30	1.00	−0.30	−0.10	0.00
(a)lgebra	−0.20	−0.30	1.00	−0.40	−0.40
a(n)alysis	0.00	−0.10	−0.40	1.00	−0.30
(s)tatistics	0.00	0.00	−0.40	−0.30	1.00
Partial variances	190.67	95.91	37.10	101.18	155.04

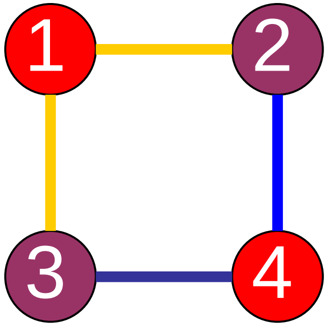
- **Some partial correlations ≈ identical**



RCON (Restricted CONcentration) models



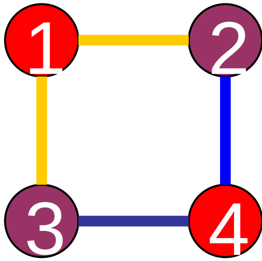
Model: $Y \sim N_d(0,\Sigma)$



$$K = \Sigma^{-1} = \begin{bmatrix} k^{11} & k^{12} & k^{13} & 0 \\ k^{12} & k^{22} & 0 & k^{24} \\ k^{13} & 0 & k^{33} & k^{34} \\ 0 & k^{24} & k^{34} & k^{44} \end{bmatrix}$$

- Markov properties as for graphical Gaussian model
- Entries of K with same colour restricted to being identical $\rightarrow$ 4 rather than 8 parameters.

# Implied restrictions (I): Equal contributions in regression



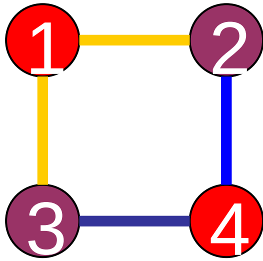
- **Regressing y_1 on y_2, y_3, y_4**

$$y_1 = a_1 - (k^{12}/k^{11})\, y_2 - (k^{13}/k^{11})\, y_3$$

- so y_2 and y_3 contribute equally because $k^{12} = k^{13}$

$$K = \Sigma^{-1} = \begin{bmatrix} k^{11} & k^{12} & k^{13} & 0 \\ k^{12} & k^{22} & 0 & k^{24} \\ k^{13} & 0 & k^{33} & k^{34} \\ 0 & k^{24} & k^{34} & k^{44} \end{bmatrix}$$

Implied restrictions (II): Parallel regressions



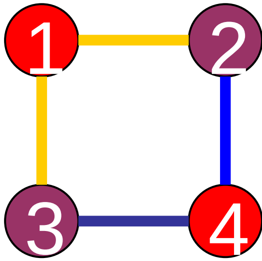
- **Regressions of y_2, y_3 on y_1, y_4 are parallel**

$$\begin{bmatrix} y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} a_2 \\ a_3 \end{bmatrix} - \begin{bmatrix} k^{22} & 0 \\ 0 & k^{33} \end{bmatrix}^{-1} \begin{bmatrix} k^{21} & k^{24} \\ k^{31} & k^{34} \end{bmatrix} \begin{bmatrix} y_1 \\ y_4 \end{bmatrix}$$

$$= \begin{bmatrix} a_2 \\ a_3 \end{bmatrix} - \begin{bmatrix} k^{21}/k^{22} & k^{24}/k^{22} \\ k^{31}/k^{33} & k^{34}/k^{33} \end{bmatrix} \begin{bmatrix} y_1 \\ y_4 \end{bmatrix}$$

- because $k^{12}=k^{13}$, $k^{24}=k^{34}$ and $k^{22}=k^{33}$

Implied restrictions (III): Partial covariances and concentrations



- This RCON model has also identical partial covariances

$$cov(y_1, y_2 | y_3, y_4) = \frac{-k^{12}}{|K^{12}|} = \frac{-k^{13}}{|K^{13}|} = cov(y_1, y_3 | y_2, y_4)$$

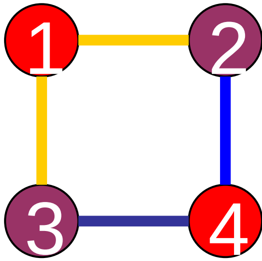
- – and identical partial correlations

$$cor(y_1, y_2 | y_3, y_4) = \frac{-k^{12}}{\sqrt{k^{11} k^{22}}} = \frac{-k^{13}}{\sqrt{k^{11} k^{33}}} = cor(y_1, y_3 | y_2, y_4)$$

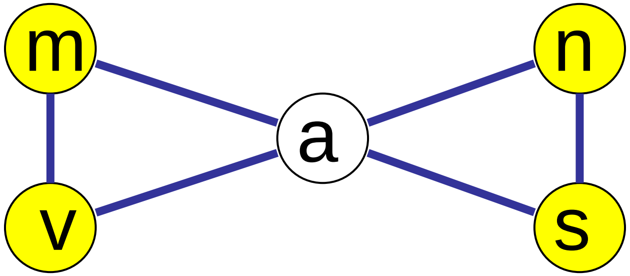
$$K = \Sigma^{-1} = \begin{bmatrix} k^{11} & k^{12} & k^{13} & 0 \\ k^{12} & k^{22} & 0 & k^{24} \\ k^{13} & 0 & k^{33} & k^{34} \\ 0 & k^{24} & k^{34} & k^{44} \end{bmatrix}$$

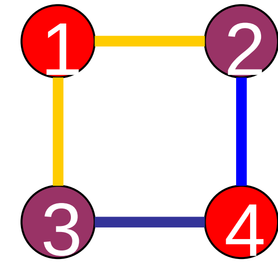
Not generally the case!!!

Example – mathematics marks



- Focus on butterfly model
  - EdgeColourClass: Edges with same colour
  - VertexColourClass: Vertices with same colour
- Note: Black and white are neutral colours; no restrictions
- Successively apply (with LR-test, 5% level)
  - JoinEdgeColourClasses()
  - JoinVertexColourClasses()

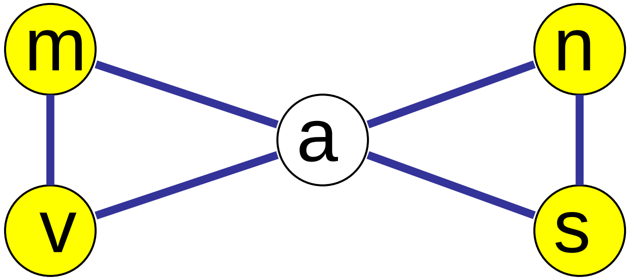




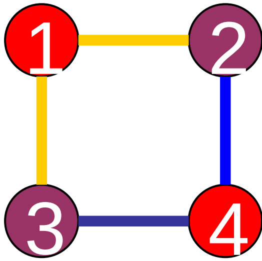
Mathematics marks...

Estimated / observed concentrations (× 1000) are

Concentrations	mechanics	vectors	algebra	analysis	statistics
mechanics	7.58	−3.49	−3.49	0.00	0.00
vectors	−3.49	7.58	−3.49	0.00	0.00
algebra	−3.49	−3.49	20.76	−3.49	−3.49
analysis	0.00	0.00	−3.49	7.58	−3.49
statistics	0.00	0.00	−3.49	−3.49	7.58
mechanics	5.24	−2.44	−2.74	0.01	−0.14
vectors	−2.44	10.43	−4.71	−0.79	−0.17
algebra	−2.74	−4.71	26.95	−7.05	−4.70
analysis	0.01	−0.79	−7.05	9.88	−2.02
statistics	−0.14	−0.17	−4.70	−2.02	6.45



RCOR (Restricted CORrelation) models

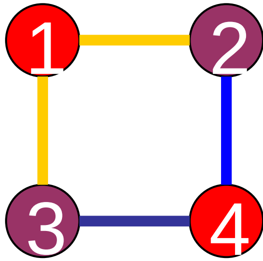


- Restricting concentrations not scale-invariant: Identical concentrations of $y \sim N(0,\Sigma)$ not generally preserved for $Ay \sim N(0,A\Sigma A)$, where A is diagonal

- Hence RCON models are only of interest in cases where the scale of measurement for different variables are comparable

- Alternatively: Focus on restricted partial correlations

RCOR (restricted correlation) models



Write –K as

$$\begin{bmatrix} \eta^{11} & 0 & 0 & 0 \\ 0 & \eta^{22} & 0 & 0 \\ 0 & 0 & \eta^{33} & 0 \\ 0 & 0 & 0 & \eta^{44} \end{bmatrix} \begin{bmatrix} 1 & \rho^{12} & \rho^{13} & 0 \\ \rho^{12} & 1 & 0 & \rho^{24} \\ \rho^{13} & 0 & 1 & \rho^{34} \\ 0 & \rho^{24} & \rho^{34} & 1 \end{bmatrix} \begin{bmatrix} \eta^{11} & 0 & 0 & 0 \\ 0 & \eta^{22} & 0 & 0 \\ 0 & 0 & \eta^{33} & 0 \\ 0 & 0 & 0 & \eta^{44} \end{bmatrix}$$

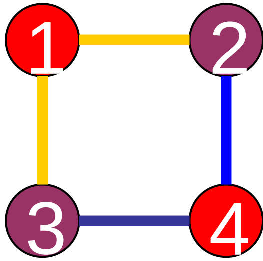
= A_η C_ρ A_η where

A_η : diagonal with positive entries (inverse partial standard deviations)

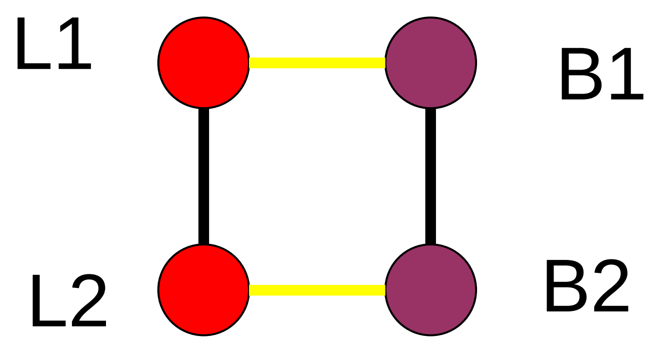
C_ρ : has 1s on diagonal (partial correlations)

- Entries of (A_η, C_ρ) restricted according to colouring
- Scale invariant - if vertices with same colour are rescaled identically

RCOP models ('P' for permution symmetry)



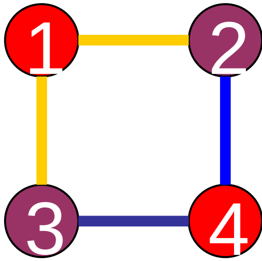
Fret's head: Length and breadth of head of first and second son:

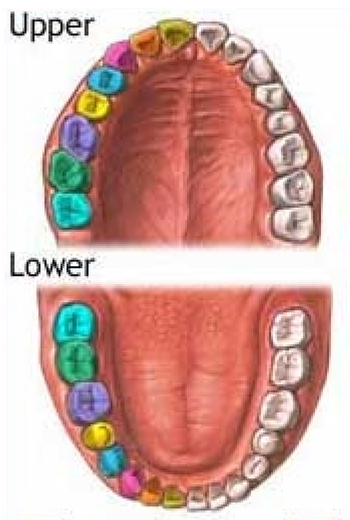


Complete symmetry between first and second son

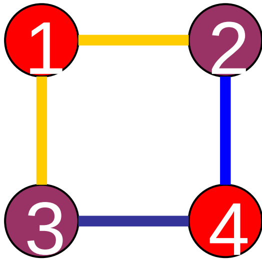
- Both RCON and RCOR
- RCOP-model (defined by permutations)
- Studied by Helene Neufeld (poster session)

Symmetry models in nature



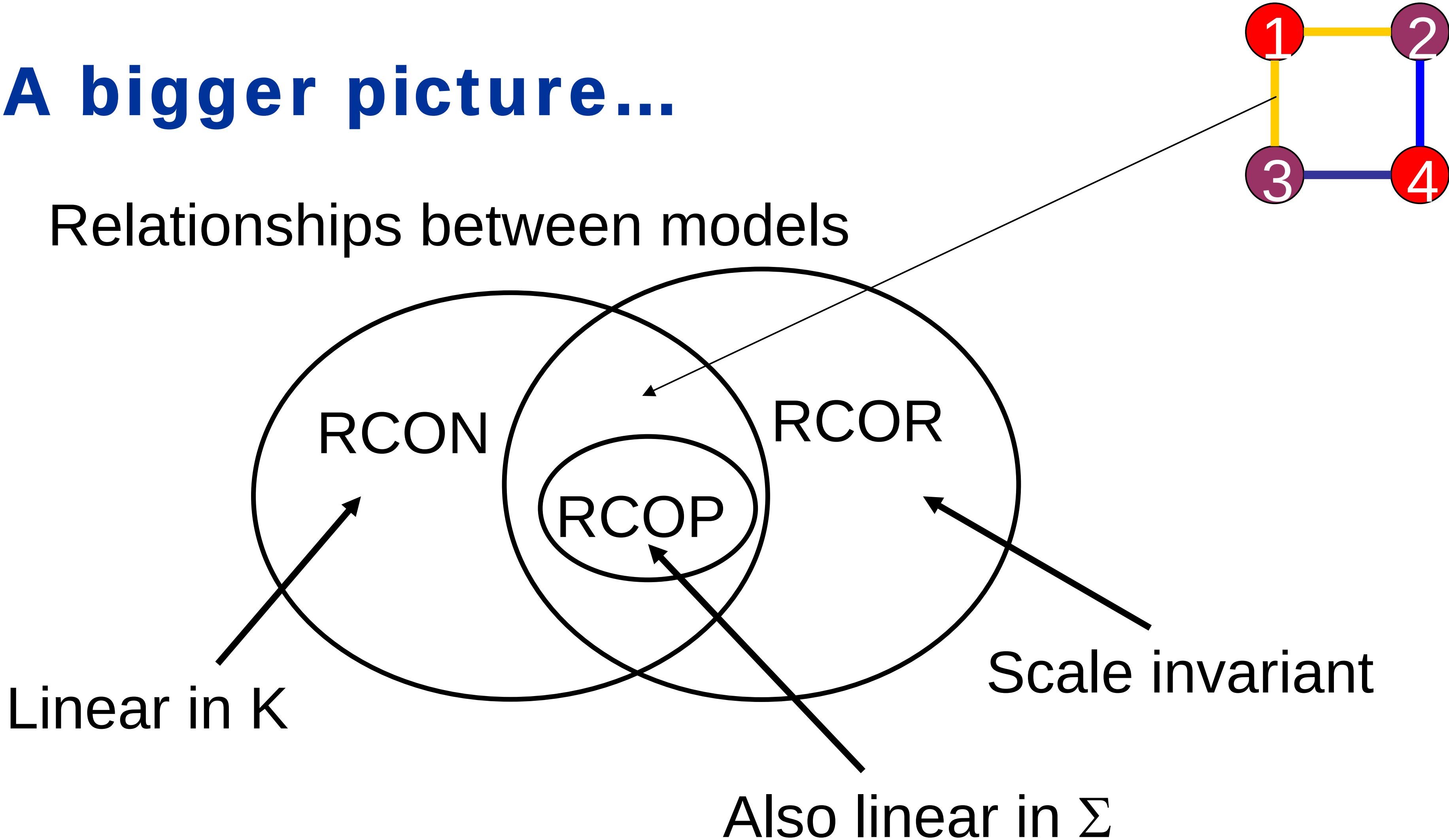


Further…

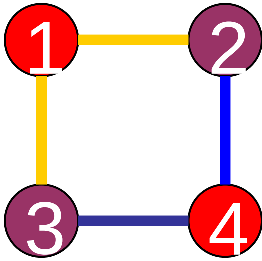


A bigger picture…

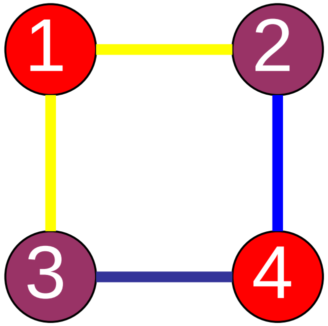
Relationships between models



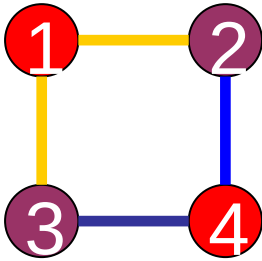
Specification of RCON / RCOR models



- **Generators of model**
  - Edge colour classes:
    EC **=** ***{{1,2}{1,3}} {{2,4}{3,4}}***
  - Vertex colour classes:
    VC **=** ***{1,4}{2,3}***
- **(EC,VC) specifies RCON / RCOR model**



RCON Estimation - Sufficient statistics



For EdgeCC s={{1,2}{1,3}}

$$T^s = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

For vertexCC s={1,4}

$$T^s = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Let W ~ Wishart(f,Σ)

RCON regular exponential family; linear in K:

Then

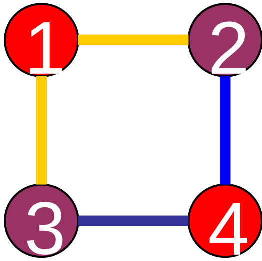
$$2 \log L = f \log |K| - tr(KW)$$
$$= f \log |K| - \Sigma_s \theta_s tr(T^s W)$$

So $\{t^s = tr(T^s W)\}$ is a set of sufficient statistics

Equated with expectations

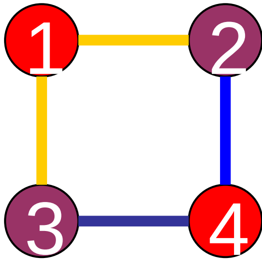
$$\{t^s\} = \{f\ tr(T^s \Sigma)\}$$

Estimation in RCON models - Modified Newton



- **Use convergent (!) algorithm of Jensen, Johansen and Lauritzen (1991):**
- **Maximize 1/L (instead of logL) cyclically over one parameter at the time by Newton iteration.**
- **In exponential families we have**
  - $\exp(t(y)\theta-\psi(\theta))$
  - $l^* = 1/L = \exp(-t(y)\theta+\psi(\theta))$

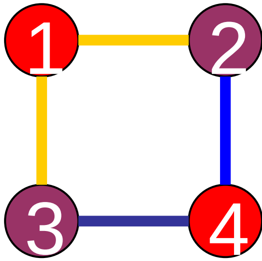
Modified Newton



- **Let $\Delta=E(t(y))-t(y)$.**
- **Then**
  - $(l^*)' = \exp(-t(y)\theta+\psi(\theta))\Delta$
  - $(l^*)'' = \exp(-t(y)\theta+\psi(\theta))[Var(t(y)) + \Delta^2]$
- **Newton step becomes:**
- $\theta \leftarrow \theta - (l^*)'/ (l^*)'' = \theta - \Delta / (Var(t(y)) + \Delta^2)$

- **For RCON models**
  - $E(T^s W)=f\ tr(T^s \Sigma)$
  - $Var(T^s W) = f\ tr(T^s \Sigma T^s \Sigma)$

Estimation – RCOR - short version



RCOR model is generally not a linear exponential family

$\log L = f/2 \log |C_\rho| + f \log |A_\eta| - \mathrm{tr}(A_\eta C_\rho A_\eta W)/2$

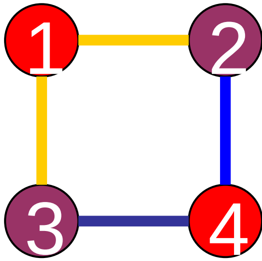
Linear exponential family for fixed η; likelihood equations quadratic for fixed ρ.

Existence/uniqueness of MLE not clear. But unique maximum in ρ for fixed η and unique maximum in η for fixed ρ.

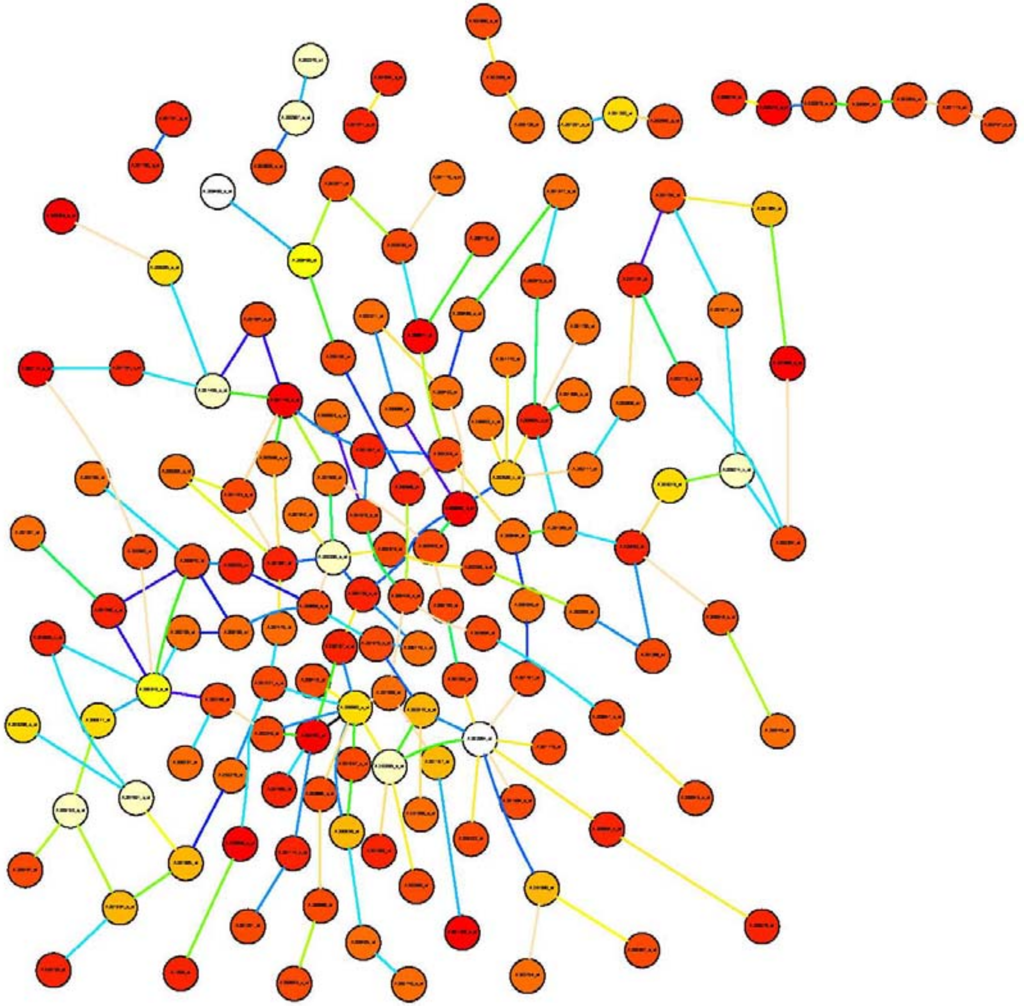
Suggests alternating algorithm

1. For given η, estimate ρ using modified Newton as for RCON
2. For given ρ, estimate η by solving system of 2nd degree equations (cyclically)

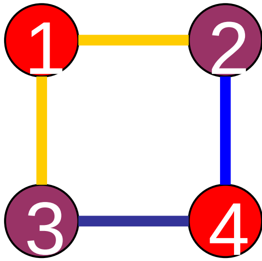
Larger model: Breast cancer genes



- **58 cases, 150 genes**
- **7 VCC, 10 ECC: 17 parameters**
- **Fitting with gRc: 0.3 sec**
- **380 parms**
- **Fitting with gRc: 1.1 sec**
- **Model selection is a big issue…**



Summing up



- SH +Lauritzen: Graphical Gaussian models with edge and vertex symmetries. *(RSSB, To appear)*
- SH + Lauritzen (2007). Inference in graphical Gaussian models with edge and vertex symmetries with the gRc package for R *(J. Stat. Soft)*
- gRc package in R (‘c’ for colour) on CRAN