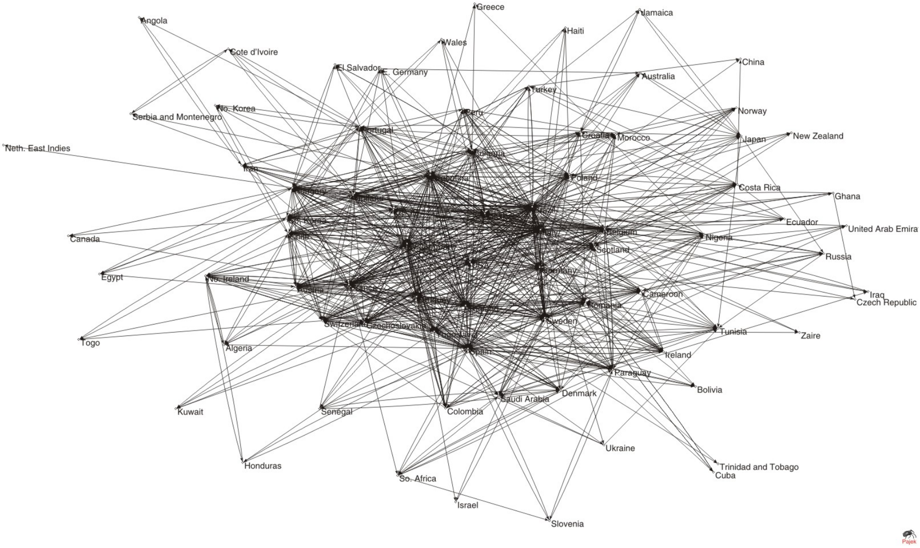
Statistical Analysis of Networks - Networks, correlation and time series

Dimitris Kugiumtzis

November 7, 2018

Contents

Introduction - Example: Games of world cup 1930 - 2006



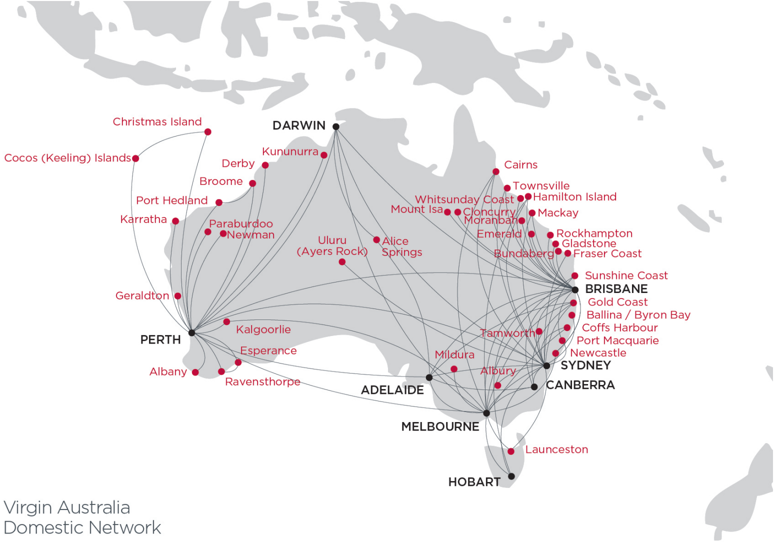
Data from: `http://gd2006.org/contest/details.php#worldcup` see [1]

Introduction - Example: Flight connections

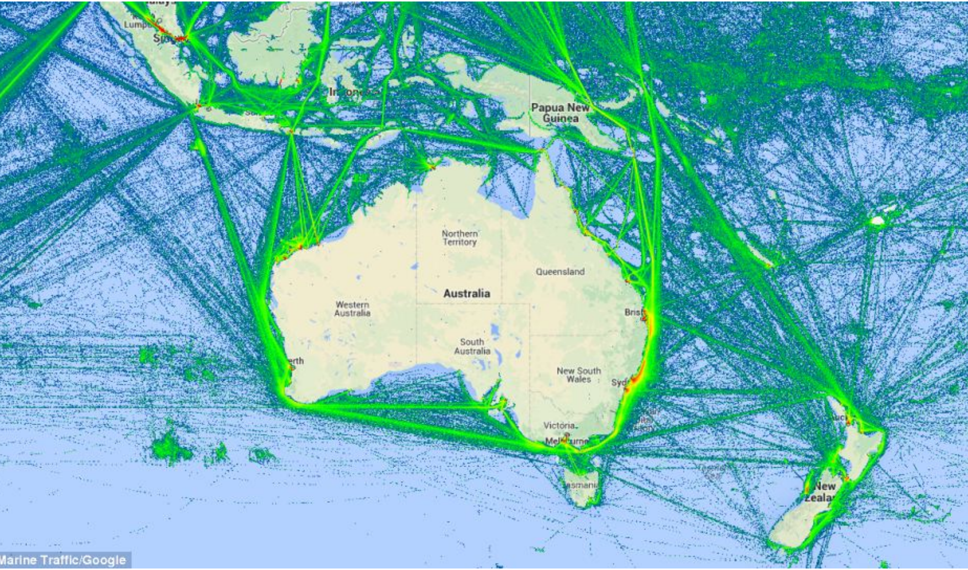
Data from: `https://au.pinterest.com/pin/488077678338752549`

Introduction - Example: Flight connections

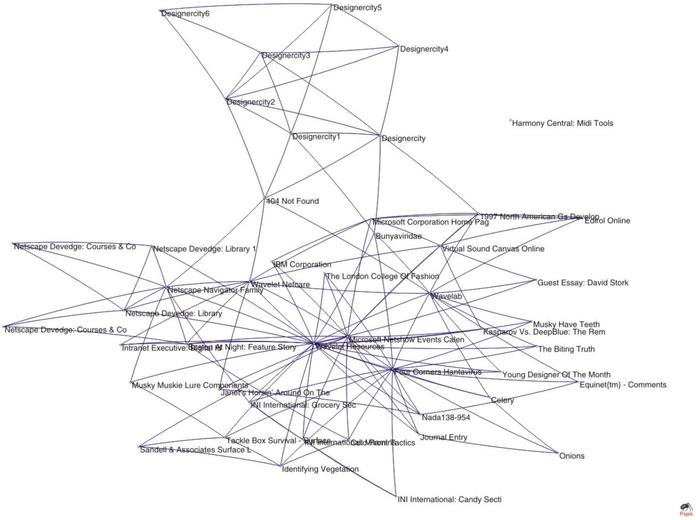
Introduction - Example: Flight connections



Introduction - Example: Ship connections



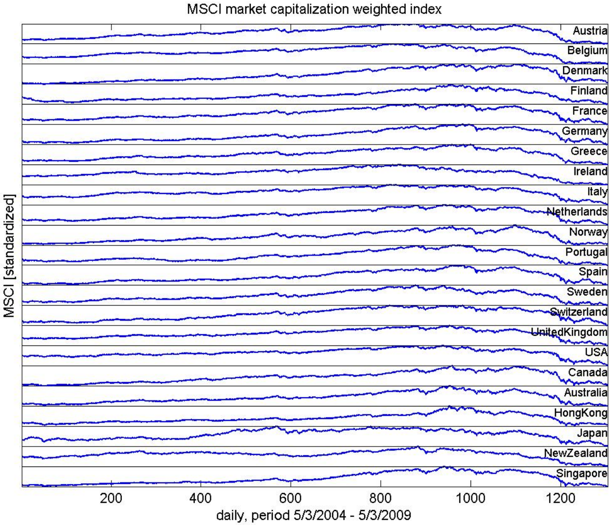
Introduction - Example: similar web-pages



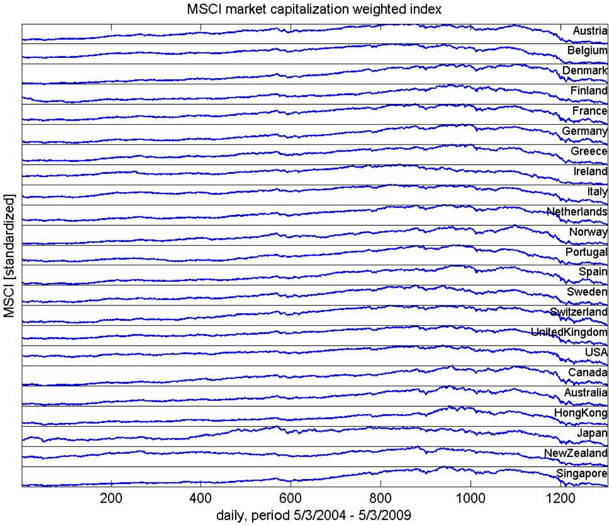
Data from: `http://vlado.fmf.uni-lj.si/pub/networks/data/GD/gd97/B97.net`

see [1]

Introduction - Example: Finance

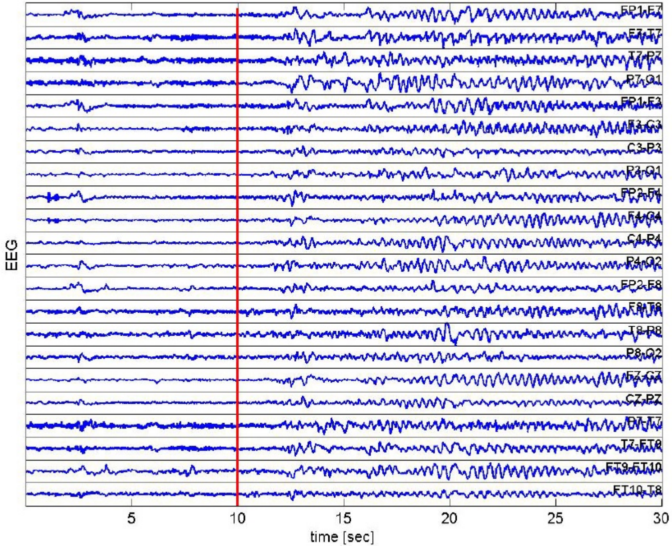


Introduction - Example: Finance

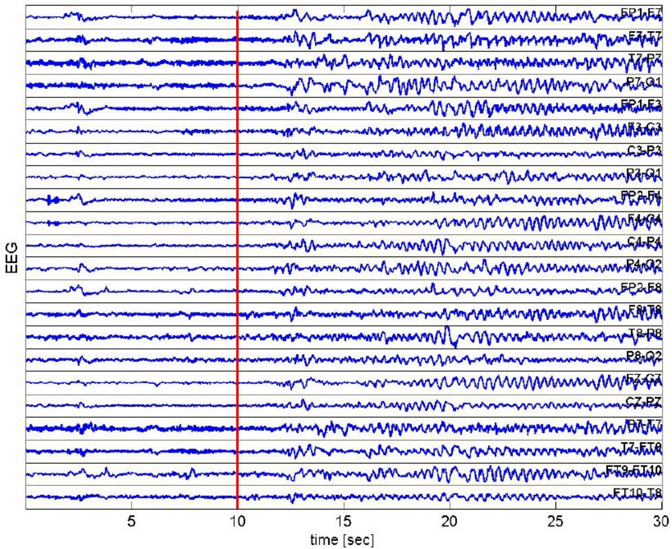


Network ?

Introduction - Example: Brain Data



Introduction - Example: Brain Data



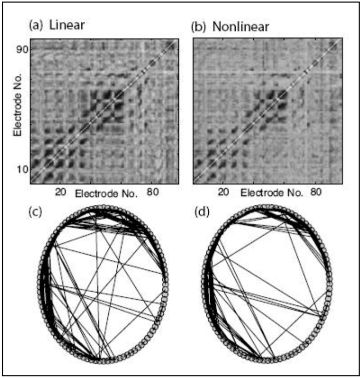
Network ?

Introduction - Example: brain network

PHYSICAL REVIEW E **79**, 061916 (2009)

Network inference with confidence from multivariate time series

Mark A. Kramer,[1,*] Uri T. Eden,[1] Sydney S. Cash,[2] and Eric D. Kolaczyk[1]
[1]*Department of Mathematics and Statistics, Boston University, Boston, Massachusetts 02215, USA*
[2]*Department of Neurology, Epilepsy Service, Harvard Medical School, ACC 835, Massachusetts General Hospital, 55 Fruit Street, Boston, Massachusetts 02114, USA*
(Received 9 March 2009; revised manuscript received 14 May 2009; published 11 June 2009)

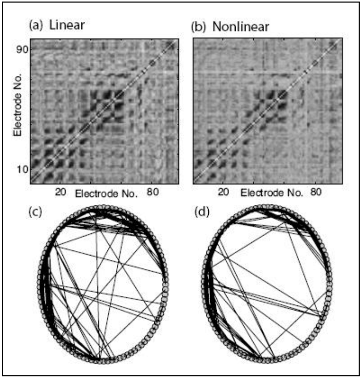


Introduction - Example: brain network

PHYSICAL REVIEW E 79, 061916 (2009)

Network inference with confidence from multivariate time series

Mark A. Kramer,[1,*] Uri T. Eden,[1] Sydney S. Cash,[2] and Eric D. Kolaczyk[1]
[1]*Department of Mathematics and Statistics, Boston University, Boston, Massachusetts 02215, USA*
[2]*Department of Neurology, Epilepsy Service, Harvard Medical School, ACC 835, Massachusetts General Hospital, 55 Fruit Street, Boston, Massachusetts 02114, USA*
(Received 9 March 2009; revised manuscript received 14 May 2009; published 11 June 2009)



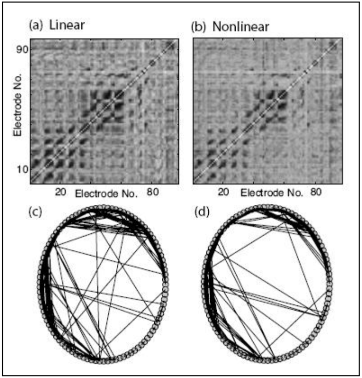
ECoG: "Linear and nonlinear association measures produce similar association matrices and networks."

Introduction - Example: brain network

PHYSICAL REVIEW E **79**, 061916 (2009)

Network inference with confidence from multivariate time series

Mark A. Kramer,[1,*] Uri T. Eden,[1] Sydney S. Cash,[2] and Eric D. Kolaczyk[1]
[1]*Department of Mathematics and Statistics, Boston University, Boston, Massachusetts 02215, USA*
[2]*Department of Neurology, Epilepsy Service, Harvard Medical School, ACC 835, Massachusetts General Hospital, 55 Fruit Street, Boston, Massachusetts 02114, USA*
(Received 9 March 2009; revised manuscript received 14 May 2009; published 11 June 2009)



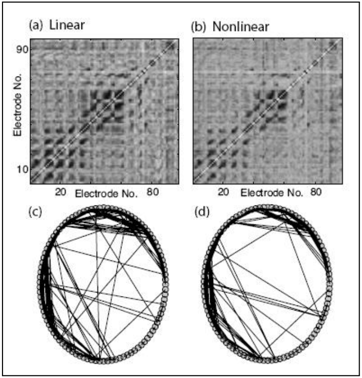
ECoG: "Linear and nonlinear association measures produce similar association matrices and networks."
?

Introduction - Example: brain network

PHYSICAL REVIEW E **79**, 061916 (2009)

Network inference with confidence from multivariate time series

Mark A. Kramer,[1,*] Uri T. Eden,[1] Sydney S. Cash,[2] and Eric D. Kolaczyk[1]
[1]*Department of Mathematics and Statistics, Boston University, Boston, Massachusetts 02215, USA*
[2]*Department of Neurology, Epilepsy Service, Harvard Medical School, ACC 835, Massachusetts General Hospital, 55 Fruit Street, Boston, Massachusetts 02114, USA*
(Received 9 March 2009; revised manuscript received 14 May 2009; published 11 June 2009)



ECoG: "Linear and nonlinear association measures produce similar association matrices and networks."

?

It is important to:

- Use appropriate measure of correlation / association / causality.

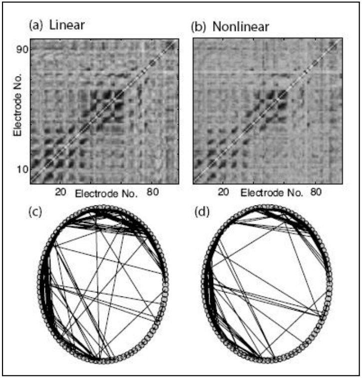
Introduction - Example: brain network

PHYSICAL REVIEW E 79, 061916 (2009)

Network inference with confidence from multivariate time series

Mark A. Kramer,[1,*] Uri T. Eden,[1] Sydney S. Cash,[2] and Eric D. Kolaczyk[1]
[1]*Department of Mathematics and Statistics, Boston University, Boston, Massachusetts 02215, USA*
[2]*Department of Neurology, Epilepsy Service, Harvard Medical School, ACC 835, Massachusetts General Hospital, 55 Fruit Street, Boston, Massachusetts 02114, USA*

(Received 9 March 2009; revised manuscript received 14 May 2009; published 11 June 2009)



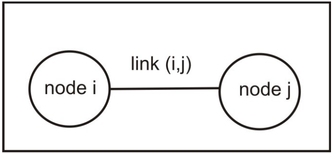
ECoG: "Linear and nonlinear association measures produce similar association matrices and networks."

?

It is important to:

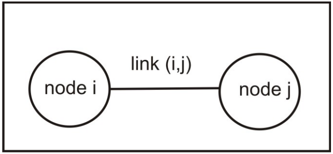
- Use appropriate measure of correlation / association / causality.
- Assess the significance of the measure.

Network



A network consists of nodes and links

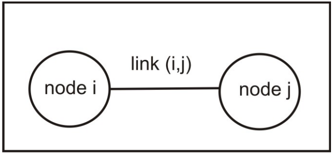
Network



A network consists of nodes and links

Node: national team, web-page, ...

Network

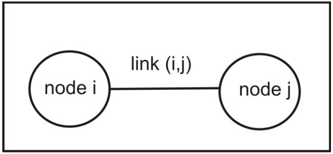


A network consists of nodes and links

Node: national team, web-page, ...

Link: match between two teams, link between two web-pages, ...

Network



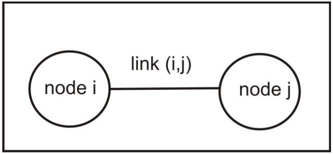
A network consists of nodes and links

Node: national team, web-page, ...

Link: match between two teams, link between two web-pages, ...

Each node is an entity and the link denotes a connection between two entities.

Network



A network consists of nodes and links

Node: national team, web-page, ...

Link: match between two teams, link between two web-pages, ...

Each node is an entity and the link denotes a connection between two entities.

Here, we will study a different (specific) type of nodes and links:

Each node type is a variable.

The link denotes some form of association or correlation between the variables.

Association Network (link: association rule) see [2], Sec 7.3

Link: association level between attributes of the two nodes.

Association is not necessarily determined by a statistical measure.

Association Network (link: association rule) see [2], Sec 7.3

Link: association level between attributes of the two nodes.

Association is not necessarily determined by a statistical measure.

Example

Association Network (link: association rule) see [2], Sec 7.3

Link: association level between attributes of the two nodes.
Association is not necessarily determined by a statistical measure.

Example

Scientometrics: Study the relationship among various scientific disciplines. see [2], Sec 3.5.1

Association Network (link: association rule) see [2], Sec 7.3

Link: association level between attributes of the two nodes.
Association is not necessarily determined by a statistical measure.

Example

Scientometrics: Study the relationship among various scientific disciplines. see [2], Sec 3.5.1
Node (unit): scientific journal

Association Network (link: association rule) see [2], Sec 7.3

Link: association level between attributes of the two nodes.
Association is not necessarily determined by a statistical measure.

Example

Scientometrics: Study the relationship among various scientific disciplines. see [2], Sec 3.5.1

Node (unit): scientific journal

Link: interaction between two journals,

Association Network (link: association rule) see [2], Sec 7.3

Link: association level between attributes of the two nodes.
Association is not necessarily determined by a statistical measure.

Example

Scientometrics: Study the relationship among various scientific disciplines. see [2], Sec 3.5.1

Node (unit): scientific journal

Link: interaction between two journals,
e.g. a link is established if journal i cites journal j at least once within a given period (directed link), and vice versa (undirected link).

Association Network (link: association rule) see [2], Sec 7.3

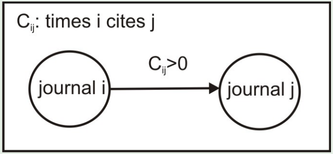
Link: association level between attributes of the two nodes.
Association is not necessarily determined by a statistical measure.

Example

Scientometrics: Study the relationship among various scientific disciplines. see [2], Sec 3.5.1

Node (unit): scientific journal

Link: interaction between two journals,
e.g. a link is established if journal *i* cites journal *j* at least once within a given period (directed link), and vice versa (undirected link).



Association Network (link: association rule) see [2], Sec 7.3

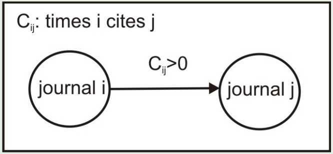
Link: association level between attributes of the two nodes.
Association is not necessarily determined by a statistical measure.

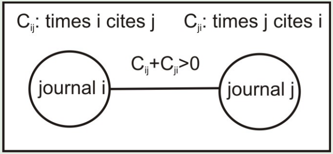
Example

Scientometrics: Study the relationship among various scientific disciplines. see [2], Sec 3.5.1

Node (unit): scientific journal

Link: interaction between two journals,
e.g. a link is established if journal *i* cites journal *j* at least once within a given period (directed link), and vice versa (undirected link).





Association Network (link: association rule) see [2], Sec 7.3

Link: association level between attributes of the two nodes.
Association is not necessarily determined by a statistical measure.

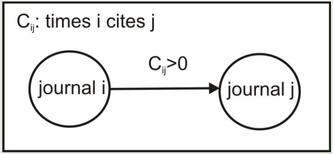
Example

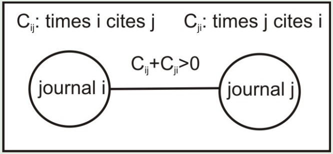
Scientometrics: Study the relationship among various scientific disciplines. see [2], Sec 3.5.1

Node (unit): scientific journal

Link: interaction between two journals,

e.g. a link is established if journal *i* cites journal *j* at least once within a given period (directed link), and vice versa (undirected link).

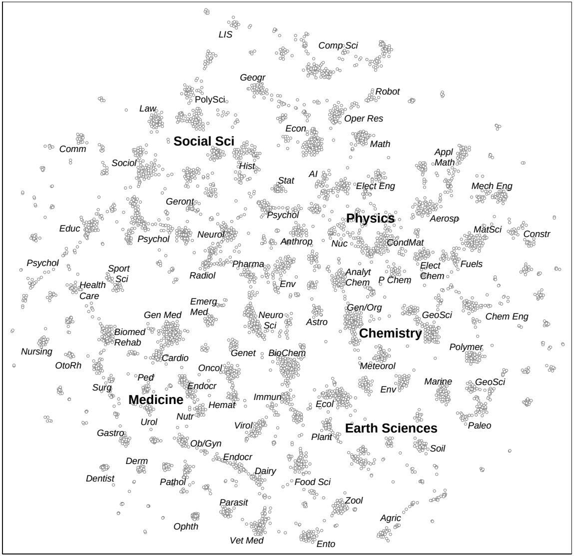




Another association rule is the Jaccard measure:

$$\mathrm{JAC}_{ij} = \mathrm{JAC}_{ji} = \frac{C_{ij}+C_{ji}}{\sum_{k\neq j} C_{ik} + \sum_{k\neq i} C_{jk}}$$

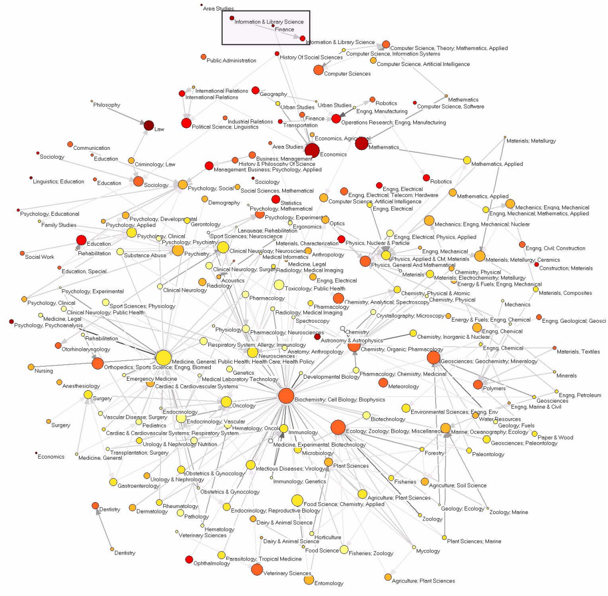
A "backbone" map of Science and Social Science: 7121 journals from year 2000



42

Source: `http://grants.nih.gov/grants/KM/OERRM/OER_KM_events/Borner.pdf`

The 212 nodes represent clusters of journals for different disciplines



44

Source: `http://grants.nih.gov/grants/KM/OERRM/OER_KM_events/Borner.pdf`

Association Network (node: vector of attributes)

Each node i is presented with a vector $\mathbf{x}_i$ of n observed attributes

$$\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]' \qquad .$$

Each node i is presented with a vector $\mathbf{x}_i$ of n observed attributes
$$\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]' \qquad .$$

A similarity measure $\mathrm{sim}(i,j)$ quantifies the level of association between two such nodes i and j.

Association Network (node: vector of attributes)

Each node i is presented with a vector $\mathbf{x}_i$ of n observed attributes

$$\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]' \quad .$$

A similarity measure $\mathrm{sim}(i,j)$ quantifies the level of association between two such nodes i and j.

$\mathrm{sim}(i,j)$ may take numerical values.

Association Network (node: vector of attributes)

Each node i is presented with a vector $\mathbf{x}_i$ of n observed attributes

$$\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]' \qquad .$$

A similarity measure sim(i, j) quantifies the level of association between two such nodes i and j.

sim(i, j) may take numerical values.

A link is assigned if the level of sim(i, j) constitutes non-trivial association between i and j.

Each node i is presented with a vector $\mathbf{x}_i$ of n observed attributes
$$\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]' \quad .$$

A similarity measure $\mathrm{sim}(i,j)$ quantifies the level of association between two such nodes i and j.

$\mathrm{sim}(i,j)$ may take numerical values.

A link is assigned if the level of $\mathrm{sim}(i,j)$ constitutes non-trivial association between i and j.

$\mathrm{sim}(i,j)$ is not directly observable but can be inferred by the information in $\mathbf{x}_i$ and $\mathbf{x}_j$.

Exercise 1: Profile of Department web-sites

Example

Consider as network an ensemble of Departments of some sort (e.g. of the same University, discipline, country etc).

Exercise 1: Profile of Department web-sites

Example

Consider as network an ensemble of Departments of some sort (e.g. of the same University, discipline, country etc).
The interest is in studying the quality / strength / similarity of the web-profiles of the Departments.

Exercise 1: Profile of Department web-sites

Example

Consider as network an ensemble of Departments of some sort (e.g. of the same University, discipline, country etc).
The interest is in studying the quality / strength / similarity of the web-profiles of the Departments.
Node: a Dept web-site, assigned with a number of attributes:

Exercise 1: Profile of Department web-sites

Example

Consider as network an ensemble of Departments of some sort (e.g. of the same University, discipline, country etc).
The interest is in studying the quality / strength / similarity of the web-profiles of the Departments.
Node: a Dept web-site, assigned with a number of attributes:

- Staff members having their home-page linked to the Department web-pages (e.g. given as percentage).

Exercise 1: Profile of Department web-sites

Example

Consider as network an ensemble of Departments of some sort (e.g. of the same University, discipline, country etc).
The interest is in studying the quality / strength / similarity of the web-profiles of the Departments.
Node: a Dept web-site, assigned with a number of attributes:

- Staff members having their home-page linked to the Department web-pages (e.g. given as percentage).
- Number of announcements within the last month.

Example

Consider as network an ensemble of Departments of some sort (e.g. of the same University, discipline, country etc).
The interest is in studying the quality / strength / similarity of the web-profiles of the Departments.
Node: a Dept web-site, assigned with a number of attributes:

- Staff members having their home-page linked to the Department web-pages (e.g. given as percentage).
- Number of announcements within the last month.
- Other ???

Exercise 1: Profile of Department web-sites

Example

Consider as network an ensemble of Departments of some sort (e.g. of the same University, discipline, country etc).
The interest is in studying the quality / strength / similarity of the web-profiles of the Departments.
Node: a Dept web-site, assigned with a number of attributes:

- Staff members having their home-page linked to the Department web-pages (e.g. given as percentage).
- Number of announcements within the last month.
- Other ???

What is an appropriate sim(i, j) to infer links?

Exercise 1: Profile of Department web-sites

Example

Consider as network an ensemble of Departments of some sort (e.g. of the same University, discipline, country etc).
The interest is in studying the quality / strength / similarity of the web-profiles of the Departments.
Node: a Dept web-site, assigned with a number of attributes:

- Staff members having their home-page linked to the Department web-pages (e.g. given as percentage).
- Number of announcements within the last month.
- Other ???

What is an appropriate $\mathrm{sim}(i,j)$ to infer links?

... the above is the **first exercise**!

Exercise 1: Profile of Department web-sites

Example

Consider as network an ensemble of Departments of some sort (e.g. of the same University, discipline, country etc).
The interest is in studying the quality / strength / similarity of the web-profiles of the Departments.
Node: a Dept web-site, assigned with a number of attributes:

- Staff members having their home-page linked to the Department web-pages (e.g. given as percentage).
- Number of announcements within the last month.
- Other ???

What is an appropriate sim(i,j) to infer links?

... the above is the **first exercise**!
You should determine and collect data for: Departments (e.g. 5), attributes (e.g. 4-5), and determine a suitable sim(i,j).

Exercise 1: Profile of Department web-sites

Example

Consider as network an ensemble of Departments of some sort (e.g. of the same University, discipline, country etc).
The interest is in studying the quality / strength / similarity of the web-profiles of the Departments.
Node: a Dept web-site, assigned with a number of attributes:

- Staff members having their home-page linked to the Department web-pages (e.g. given as percentage).
- Number of announcements within the last month.
- Other ???

What is an appropriate sim(i, j) to infer links?

... the above is the **first exercise**!
You should determine and collect data for: Departments (e.g. 5), attributes (e.g. 4-5), and determine a suitable sim(i, j).
You may use a software (e.g. `pajek`) to draw the network.

Correlation Network see [2], Sec 7.3.1

For each node i, an attribute X_i is assigned that is considered as a continuous random variable.

Correlation Network see [2], Sec 7.3.1

For each node i, an attribute X_i is assigned that is considered as a continuous random variable.

For each X_i there are n observations: $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$.

Correlation Network see [2], Sec 7.3.1

For each node i, an attribute X_i is assigned that is considered as a continuous random variable.

For each X_i there are n observations: $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$.

A similarity measure $\mathrm{sim}(i,j)$ quantifies the level of correlation between X_i and X_j.

Correlation Network see [2], Sec 7.3.1

For each node i, an attribute X_i is assigned that is considered as a continuous random variable.

For each X_i there are n observations: $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$.

A similarity measure $\text{sim}(i,j)$ quantifies the level of correlation between X_i and X_j.

A standard similarity measure is the Pearson correlation coefficient

$$\text{Corr}(X,Y) = r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

s_{XY}: sample covariance of X and Y, s_X: sample SD of X

Correlation Network see [2], Sec 7.3.1

For each node i, an attribute X_i is assigned that is considered as a continuous random variable.

For each X_i there are n observations: $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$.

A similarity measure $\text{sim}(i, j)$ quantifies the level of correlation between X_i and X_j.

A standard similarity measure is the Pearson correlation coefficient

$$\text{Corr}(X, Y) = r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

s_{XY}: sample covariance of X and Y, s_X: sample SD of X

Example

Gene Regulation from Microarray Data: Patterns of regulatory interactions among genes can be described by networks.

Correlation Network see [2], Sec 7.3.1

For each node i, an attribute X_i is assigned that is considered as a continuous random variable.

For each X_i there are n observations: $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$.

A similarity measure $\mathrm{sim}(i, j)$ quantifies the level of correlation between X_i and X_j.

A standard similarity measure is the Pearson correlation coefficient

$$\mathrm{Corr}(X, Y) = r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

s_{XY}: sample covariance of X and Y, s_X: sample SD of X

Example

Gene Regulation from Microarray Data: Patterns of regulatory interactions among genes can be described by networks.

Node: the gene

Correlation Network see [2], Sec 7.3.1

For each node i, an attribute X_i is assigned that is considered as a continuous random variable.

For each X_i there are n observations: $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$.

A similarity measure $\text{sim}(i, j)$ quantifies the level of correlation between X_i and X_j.

A standard similarity measure is the Pearson correlation coefficient

$$\text{Corr}(X, Y) = r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

s_{XY}: sample covariance of X and Y, s_X: sample SD of X

Example

Gene Regulation from Microarray Data: Patterns of regulatory interactions among genes can be described by networks.

Node: the gene

X_i: relative level of RNA expression of the gene i in a cell.

For each node i, an attribute X_i is assigned that is considered as a continuous random variable.

For each X_i there are n observations: $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$.

A similarity measure $\text{sim}(i, j)$ quantifies the level of correlation between X_i and X_j.

A standard similarity measure is the Pearson correlation coefficient

$$\text{Corr}(X, Y) = r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

s_{XY}: sample covariance of X and Y, s_X: sample SD of X

Example

Gene Regulation from Microarray Data: Patterns of regulatory interactions among genes can be described by networks.

Node: the gene

X_i: relative level of RNA expression of the gene i in a cell.

$\mathbf{x}_i$: Microarray measurements of the RNA level at n experiments (different conditions).

Correlation Network see [2], Sec 7.3.1

For each node i, an attribute X_i is assigned that is considered as a continuous random variable.

For each X_i there are n observations: $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$.

A similarity measure $\text{sim}(i, j)$ quantifies the level of correlation between X_i and X_j.

A standard similarity measure is the Pearson correlation coefficient

$$\text{Corr}(X, Y) = r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

s_{XY}: sample covariance of X and Y, s_X: sample SD of X

Example

Gene Regulation from Microarray Data: Patterns of regulatory interactions among genes can be described by networks.
Node: the gene
X_i: relative level of RNA expression of the gene i in a cell.
$\mathbf{x}_i$: Microarray measurements of the RNA level at n experiments (different conditions).
Link: regulatory relationship, $\text{sim}(i, j) := \text{Corr}(X_i, X_j) = r_{X_i, X_j} = r_{ij}$

Example: Gene Regulation from Microarray Data

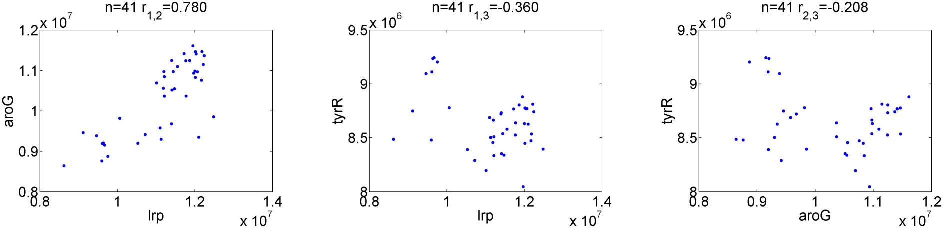
Data from: `http://m3d.bu.edu/cgi-bin/web/array/index.pl` see [2], Sec 7.3.1

Three genes: lrp, aroG, tyrR, and 41 experiments.

Example: Gene Regulation from Microarray Data

Data from: `http://m3d.bu.edu/cgi-bin/web/array/index.pl` see [2], Sec 7.3.1

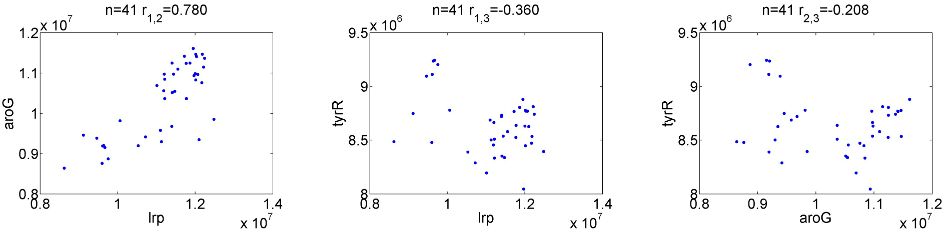
Three genes: lrp, aroG, tyrR, and 41 experiments.

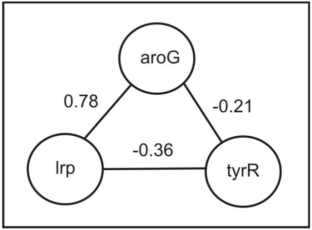


Example: Gene Regulation from Microarray Data

Data from: `http://m3d.bu.edu/cgi-bin/web/array/index.pl` see [2], Sec 7.3.1

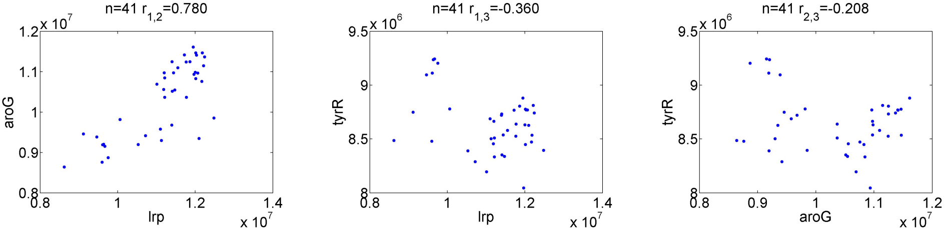
Three genes: lrp, aroG, tyrR, and 41 experiments.

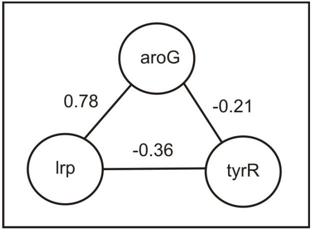




Data from: `http://m3d.bu.edu/cgi-bin/web/array/index.pl` see [2], Sec 7.3.1

Three genes: Irp, aroG, tyrR, and 41 experiments.



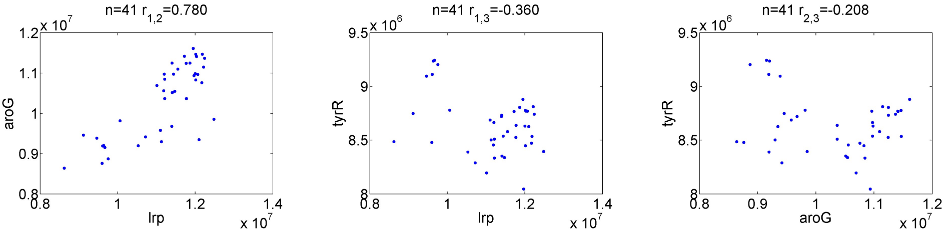


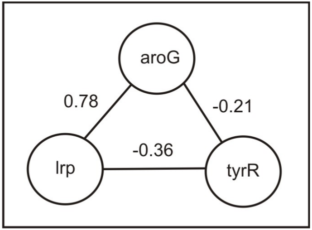
Which links are "non-trivial"?

Example: Gene Regulation from Microarray Data

Data from: `http://m3d.bu.edu/cgi-bin/web/array/index.pl` see [2], Sec 7.3.1

Three genes: lrp, aroG, tyrR, and 41 experiments.





Which links are "non-trivial"?

Significance test for correlation coefficient?

Significance of correlation (parametric test)

Let $\rho_{X_i,X_j} = \rho_{ij}$ be the true Pearson correlation coefficient of X_i and X_j.

Significance of correlation (parametric test)

Let $\rho_{X_i,X_j} = \rho_{ij}$ be the true Pearson correlation coefficient of X_i and X_j.

Hypothesis test for significance:
$\mathrm{H}_0 : \rho_{ij} = 0, \qquad \mathrm{H}_1 : \rho_{ij} \neq 0.$

Significance of correlation (parametric test)

Let $\rho_{X_i,X_j} = \rho_{ij}$ be the true Pearson correlation coefficient of X_i and X_j.

Hypothesis test for significance:
$\mathrm{H}_0 : \rho_{ij} = 0, \qquad \mathrm{H}_1 : \rho_{ij} \neq 0.$

Estimate of ρ_{ij}: $r_{ij} = \frac{s_{ij}}{s_i s_j}$.

Significance of correlation (parametric test)

Let $\rho_{X_i,X_j} = \rho_{ij}$ be the true Pearson correlation coefficient of X_i and X_j.

Hypothesis test for significance:
$\mathrm{H}_0 : \rho_{ij} = 0, \qquad \mathrm{H}_1 : \rho_{ij} \neq 0.$

Estimate of ρ_{ij}: $r_{ij} = \frac{s_{ij}}{s_i s_j}$.

Parametric testing, assuming $(X_i, X_j) \sim \mathrm{N}([\mu_i, \mu_j], [\sigma_i^2, \sigma_j^2], \rho_{ij})$:

Significance of correlation (parametric test)

Let $\rho_{X_i,X_j} = \rho_{ij}$ be the true Pearson correlation coefficient of X_i and X_j.

Hypothesis test for significance:
$\text{H}_0 : \rho_{ij} = 0, \qquad \text{H}_1 : \rho_{ij} \neq 0.$

Estimate of ρ_{ij}: $r_{ij} = \frac{s_{ij}}{s_i s_j}$.

Parametric testing, assuming $(X_i, X_j) \sim \text{N}([\mu_i, \mu_j], [\sigma_i^2, \sigma_j^2], \rho_{ij})$:

Test statistic:

- $t = \frac{r_{ij}\sqrt{n-2}}{\sqrt{1-r_{ij}^2}} \sim t_{n-2}$, or
- $z = \tanh^{-1}(r_{ij}) = \frac{1}{2}\log\left[\frac{1+r_{ij}^2}{1-r_{ij}^2}\right] \sim \text{N}(0, \frac{1}{n-3})$

Significance of correlation (parametric test)

Let $\rho_{X_i,X_j} = \rho_{ij}$ be the true Pearson correlation coefficient of X_i and X_j.

Hypothesis test for significance:
$\mathrm{H}_0 : \rho_{ij} = 0, \qquad \mathrm{H}_1 : \rho_{ij} \neq 0.$

Estimate of ρ_{ij}: $r_{ij} = \frac{s_{ij}}{s_i s_j}$.

Parametric testing, assuming $(X_i, X_j) \sim \mathrm{N}([\mu_i, \mu_j], [\sigma_i^2, \sigma_j^2], \rho_{ij})$:

Test statistic:

- $t = \frac{r_{ij}\sqrt{n-2}}{\sqrt{1-r_{ij}^2}} \sim t_{n-2}$, or
- $z = \tanh^{-1}(r_{ij}) = \frac{1}{2}\log\left[\frac{1+r_{ij}^2}{1-r_{ij}^2}\right] \sim \mathrm{N}(0, \frac{1}{n-3})$

Test all pairs at the significance level α? Multiple testing?

Example: Gene Regulation (continuing)

Parametric test for the significance of the correlation for the genes: lrp, aroG, tyrR, $n = 41$.

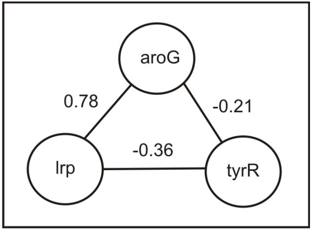
Example: Gene Regulation (continuing)

Parametric test for the significance of the correlation for the genes: lrp, aroG, tyrR, $n = 41$.

gene pair	r_{ij}	t-statistic (p-value)	z-statistic (p-value)
lrp-aroG	0.78	7.79 (0.0000)	6.48 (0.0000)
lrp-tyrR	-0.36	-2.41 (0.0208)	-2.32 (0.0202)
aroG-tyrR	-0.21	-1.36 (0.1929)	-1.30 (0.1942)

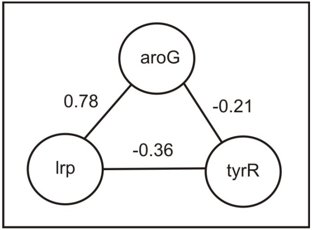
Parametric test for the significance of the correlation for the genes: lrp, aroG, tyrR, $n = 41$.

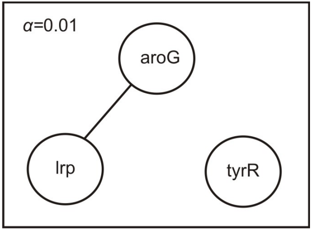
gene pair	r_{ij}	t-statistic (p-value)	z-statistic (p-value)
lrp-aroG	0.78	7.79 (0.0000)	6.48 (0.0000)
lrp-tyrR	-0.36	-2.41 (0.0208)	-2.32 (0.0202)
aroG-tyrR	-0.21	-1.36 (0.1929)	-1.30 (0.1942)



Parametric test for the significance of the correlation for the genes: lrp, aroG, tyrR, $n = 41$.

gene pair	r_{ij}	t-statistic (p-value)	z-statistic (p-value)
lrp-aroG	0.78	7.79 (0.0000)	6.48 (0.0000)
lrp-tyrR	-0.36	-2.41 (0.0208)	-2.32 (0.0202)
aroG-tyrR	-0.21	-1.36 (0.1929)	-1.30 (0.1942)

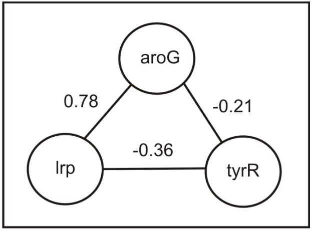


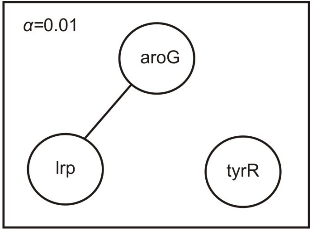


Example: Gene Regulation (continuing)

Parametric test for the significance of the correlation for the genes: lrp, aroG, tyrR, $n = 41$.

gene pair	r_{ij}	t-statistic (p-value)	z-statistic (p-value)
lrp-aroG	0.78	7.79 (0.0000)	6.48 (0.0000)
lrp-tyrR	-0.36	-2.41 (0.0208)	-2.32 (0.0202)
aroG-tyrR	-0.21	-1.36 (0.1929)	-1.30 (0.1942)



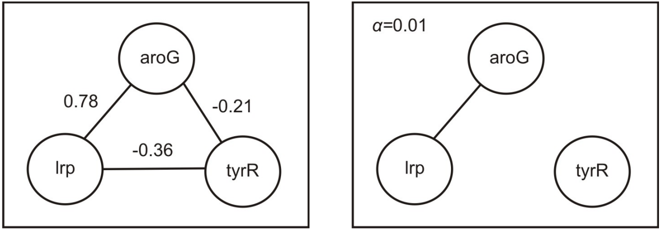


Does significance level $\alpha = 0.01$ establishes "non-trivial" links?

Example: Gene Regulation (continuing)

Parametric test for the significance of the correlation for the genes: lrp, aroG, tyrR, $n = 41$.

gene pair	r_{ij}	t-statistic (p-value)	z-statistic (p-value)
lrp-aroG	0.78	7.79 (0.0000)	6.48 (0.0000)
lrp-tyrR	-0.36	-2.41 (0.0208)	-2.32 (0.0202)
aroG-tyrR	-0.21	-1.36 (0.1929)	-1.30 (0.1942)

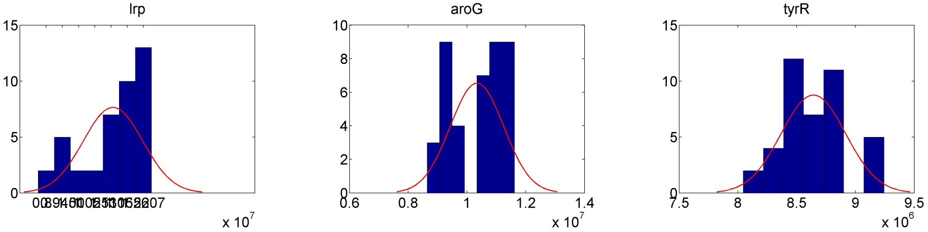


Does significance level $\alpha = 0.01$ establishes "non-trivial" links?
Does $(X_i, X_j) \sim \mathrm{N}([\mu_i, \mu_j], [\sigma_i^2, \sigma_j^2], \rho_{ij})$ hold?

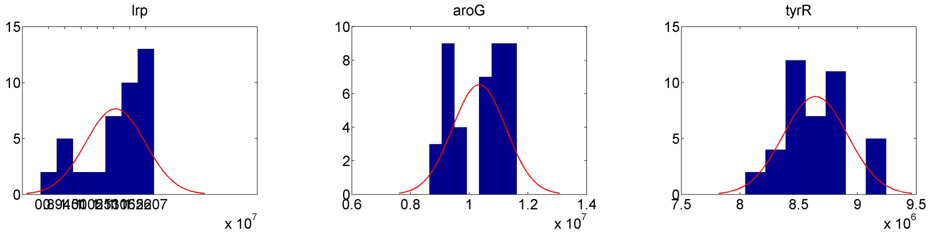
Does $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ hold?

Example: Gene Regulation (continuing)

Does $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ hold?



Does $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ hold?



The results of the parametric testing are called into question!

Significance of correlation (nonparametric test)

Nonparametric testing: draw the null distribution of r_{ij} from resampled pairs consistent to $\mathrm{H}_0 : \rho_{ij} = 0$.

Significance of correlation (nonparametric test)

Nonparametric testing: draw the null distribution of r_{ij} from resampled pairs consistent to $\mathrm{H}_0 : \rho_{ij} = 0$.

1. For an "original" pair $(\mathbf{x}_i, \mathbf{x}_j)$, generate B randomized sample pairs $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$. Generation of each b pair:

Significance of correlation (nonparametric test)

Nonparametric testing: draw the null distribution of r_{ij} from resampled pairs consistent to $\mathrm{H}_0 : \rho_{ij} = 0$.

1. For an "original" pair $(\mathbf{x}_i, \mathbf{x}_j)$, generate B randomized sample pairs $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$. Generation of each b pair:
    - Let first variable intact, $\mathbf{x}_i^{*b} = \mathbf{x}_i$.

Significance of correlation (nonparametric test)

Nonparametric testing: draw the null distribution of r_{ij} from resampled pairs consistent to $\mathrm{H}_0 : \rho_{ij} = 0$.

1. For an "original" pair $(\mathbf{x}_i, \mathbf{x}_j)$, generate B randomized sample pairs $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$. Generation of each b pair:
   - Let first variable intact, $\mathbf{x}_i^{*b} = \mathbf{x}_i$.
   - Shuffle randomly the samples of the other variable to get $\mathbf{x}_j^{*b}$.

Significance of correlation (nonparametric test)

Nonparametric testing: draw the null distribution of r_{ij} from resampled pairs consistent to $\mathrm{H}_0 : \rho_{ij} = 0$.

1. For an "original" pair $(\mathbf{x}_i, \mathbf{x}_j)$, generate B randomized sample pairs $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$. Generation of each b pair:
   - Let first variable intact, $\mathbf{x}_i^{*b} = \mathbf{x}_i$.
   - Shuffle randomly the samples of the other variable to get $\mathbf{x}_j^{*b}$.

   Each sample pair $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$ is from (X_i, X_j) under the hypothesis of independence ($\mathbf{x}_i^{*b}$ preserves the marginal distribution of $\mathbf{x}_i$, the same for j).

Significance of correlation (nonparametric test)

Nonparametric testing: draw the null distribution of r_{ij} from resampled pairs consistent to $\mathrm{H}_0 : \rho_{ij} = 0$.

1. For an "original" pair $(\mathbf{x}_i, \mathbf{x}_j)$, generate B randomized sample pairs $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$. Generation of each b pair:
   - Let first variable intact, $\mathbf{x}_i^{*b} = \mathbf{x}_i$.
   - Shuffle randomly the samples of the other variable to get $\mathbf{x}_j^{*b}$.

   Each sample pair $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$ is from (X_i, X_j) under the hypothesis of independence ($\mathbf{x}_i^{*b}$ preserves the marginal distribution of $\mathbf{x}_i$, the same for j).
2. Compute r_{ij}^{*b} on each pair $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$.

Significance of correlation (nonparametric test)

Nonparametric testing: draw the null distribution of r_{ij} from resampled pairs consistent to $\mathrm{H}_0 : \rho_{ij} = 0$.

1. For an "original" pair $(\mathbf{x}_i, \mathbf{x}_j)$, generate B randomized sample pairs $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$. Generation of each b pair:
   - Let first variable intact, $\mathbf{x}_i^{*b} = \mathbf{x}_i$.
   - Shuffle randomly the samples of the other variable to get $\mathbf{x}_j^{*b}$.

   Each sample pair $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$ is from (X_i, X_j) under the hypothesis of independence ($\mathbf{x}_i^{*b}$ preserves the marginal distribution of $\mathbf{x}_i$, the same for j).
2. Compute r_{ij}^{*b} on each pair $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$.
   The ensemble $\{r_{ij}^{*b}\}_{b=1}^{B}$ forms the empirical null distribution of r_{ij}.

Significance of correlation (nonparametric test)

Nonparametric testing: draw the null distribution of r_{ij} from resampled pairs consistent to $\mathrm{H}_0 : \rho_{ij} = 0$.

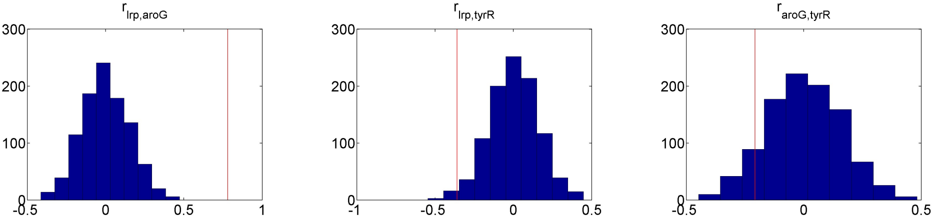
1. For an "original" pair $(\mathbf{x}_i, \mathbf{x}_j)$, generate B randomized sample pairs $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$. Generation of each b pair:
   - Let first variable intact, $\mathbf{x}_i^{*b} = \mathbf{x}_i$.
   - Shuffle randomly the samples of the other variable to get $\mathbf{x}_j^{*b}$.

   Each sample pair $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$ is from (X_i, X_j) under the hypothesis of independence ($\mathbf{x}_i^{*b}$ preserves the marginal distribution of $\mathbf{x}_i$, the same for j).
2. Compute r_{ij}^{*b} on each pair $(\mathbf{x}_i^{*b}, \mathbf{x}_j^{*b})$, $b = 1, \ldots, B$.
   The ensemble $\{r_{ij}^{*b}\}_{b=1}^{B}$ forms the empirical null distribution of r_{ij}.
3. Reject H_0 if sample r_{ij} is not in the distribution of $\{r_{ij}^{*b}\}_{b=1}^{B}$ (using rank ordering).

Example: Gene Regulation (continuing)

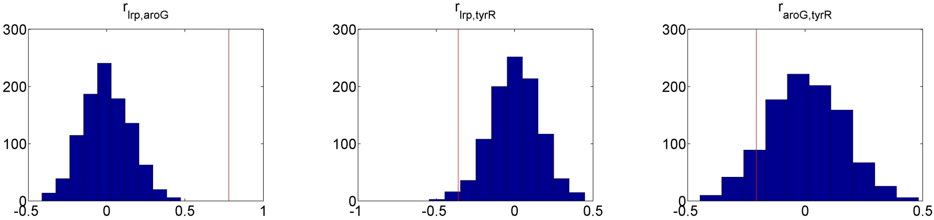
Nonparametric testing, 1000 randomized samples.

Nonparametric testing, 1000 randomized samples.



Example: Gene Regulation (continuing)

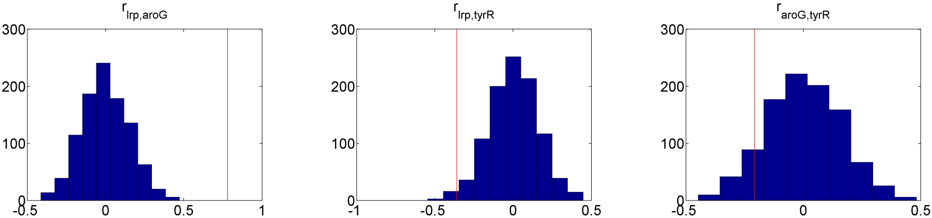
Nonparametric testing, 1000 randomized samples.



gene pair	r_{ij}	t-statistic (p-value)	rank (p-value)
lrp-aroG	0.78	6.48 (0.0000)	1001 (0.0013)
lrp-tyrR	-0.36	-2.32 (0.0202)	17 (0.0333)
aroG-tyrR	-0.21	-1.30 (0.1942)	99 (0.1971)

Example: Gene Regulation (continuing)

Nonparametric testing, 1000 randomized samples.



gene pair	r_{ij}	t-statistic (p-value)	rank (p-value)
lrp-aroG	0.78	6.48 (0.0000)	1001 (0.0013)
lrp-tyrR	-0.36	-2.32 (0.0202)	17 (0.0333)
aroG-tyrR	-0.21	-1.30 (0.1942)	99 (0.1971)

Correlation coefficient and correlation matrix

The correlation coefficients r_{ij}, $i, j = 1, \ldots, N$ form a correlation matrix (positive semidefinite)

Correlation coefficient and correlation matrix

The correlation coefficients r_{ij}, $i, j = 1, \ldots, N$ form a correlation matrix (positive semidefinite)

Establishing the statistically significant r_{ij}, $i, j = 1, \ldots, N$, the correlation matrix is converted to the adjacency matrix.

Correlation coefficient and correlation matrix

The correlation coefficients r_{ij}, $i,j = 1,\ldots,N$ form a correlation matrix (positive semidefinite)

Establishing the statistically significant r_{ij}, $i,j = 1,\ldots,N$, the correlation matrix is converted to the adjacency matrix.

Example

Correlation for the genes: lrp, aroG, tyrR

gene	r_{ij}
lrp-aroG	0.78
lrp-tyrR	-0.36
aroG-tyrR	-0.21

$$\longrightarrow \quad R = \begin{bmatrix} 1 & 0.78 & -0.36 \\ 0.78 & 1 & -0.21 \\ -0.36 & -0.21 & 1 \end{bmatrix}$$

Correlation coefficient and correlation matrix

The correlation coefficients r_{ij}, $i,j = 1,\ldots,N$ form a correlation matrix (positive semidefinite)

Establishing the statistically significant r_{ij}, $i,j = 1,\ldots,N$, the correlation matrix is converted to the adjacency matrix.

Example

Correlation for the genes: lrp, aroG, tyrR

gene	r_{ij}
lrp-aroG	0.78
lrp-tyrR	-0.36
aroG-tyrR	-0.21

$$\longrightarrow \quad R = \begin{bmatrix} 1 & 0.78 & -0.36 \\ 0.78 & 1 & -0.21 \\ -0.36 & -0.21 & 1 \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & 0.78 & -0.36 \\ 0.78 & 1 & -0.21 \\ -0.36 & -0.21 & 1 \end{bmatrix} \longrightarrow R = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Are the link(s) found statistically significant also "non-trivial"?

Exercise 2: Micro-array Data

Use any subset (3 or more) of the genes in file `Ecoliv4Build6ex1`
(in ascii or excel format, see course web-page).
Using the 41 experiments for each gene, form the correlation network for the selected genes.

Exercise 2: Micro-array Data

Use any subset (3 or more) of the genes in file `Ecoliv4Build6ex1`
(in ascii or excel format, see course web-page).
Using the 41 experiments for each gene, form the correlation network for the selected genes.
You should:

Exercise 2: Micro-array Data

Use any subset (3 or more) of the genes in file `Ecoliv4Build6ex1` (in ascii or excel format, see course web-page).

Using the 41 experiments for each gene, form the correlation network for the selected genes.

You should:

1. Use the correlation coefficient r_{ij} as similarity measure of two genes i and j.

Exercise 2: Micro-array Data

Use any subset (3 or more) of the genes in file `Ecoliv4Build6ex1` (in ascii or excel format, see course web-page).

Using the 41 experiments for each gene, form the correlation network for the selected genes.

You should:

1. Use the correlation coefficient r_{ij} as similarity measure of two genes i and j.
2. Use parametric and nonparametric test of significance for each correlation coefficient r_{ij}.

Exercise 2: Micro-array Data

Use any subset (3 or more) of the genes in file `Ecoliv4Build6ex1` (in ascii or excel format, see course web-page).

Using the 41 experiments for each gene, form the correlation network for the selected genes.

You should:

1. Use the correlation coefficient r_{ij} as similarity measure of two genes i and j.
2. Use parametric and nonparametric test of significance for each correlation coefficient r_{ij}.
3. Identify whether the significant links are the same with parametric and nonparametric testing.

Exercise 2: Micro-array Data

Use any subset (3 or more) of the genes in file `Ecoliv4Build6ex1` (in ascii or excel format, see course web-page).

Using the 41 experiments for each gene, form the correlation network for the selected genes.

You should:

1. Use the correlation coefficient r_{ij} as similarity measure of two genes i and j.
2. Use parametric and nonparametric test of significance for each correlation coefficient r_{ij}.
3. Identify whether the significant links are the same with parametric and nonparametric testing.
4. Form the networks from significant links from each test.

Exercise 2: Micro-array Data

Use any subset (3 or more) of the genes in file `Ecoliv4Build6ex1` (in ascii or excel format, see course web-page).
Using the 41 experiments for each gene, form the correlation network for the selected genes.
You should:

1. Use the correlation coefficient r_{ij} as similarity measure of two genes i and j.
2. Use parametric and nonparametric test of significance for each correlation coefficient r_{ij}.
3. Identify whether the significant links are the same with parametric and nonparametric testing.
4. Form the networks from significant links from each test.

matlab:

- for the Pearson correlation coefficient you may use the function `corrcoef`
- for random shuffling you may use the function `randperm`

Partial Correlation Network see [2], Sec 7.3.1

If X_i and X_j are found to have a large r_{ij}:

1. There is direct dependence of X_i on X_j, or of X_j on X_i, or both.

Partial Correlation Network see [2], Sec 7.3.1

If X_i and X_j are found to have a large r_{ij}:

1. There is direct dependence of X_i on X_j, or of X_j on X_i, or both.
2. Both X_i and X_j are dependent on an other variable (node) X_k or on m other variables (nodes) $\mathbf{X}_K = \{X_{k_1}, \ldots, X_{k_m}\}$, where $K = \{k_1, \ldots, k_m\}$.

If X_i and X_j are found to have a large r_{ij}:

1. There is direct dependence of X_i on X_j, or of X_j on X_i, or both.
2. Both X_i and X_j are dependent on an other variable (node) X_k or on m other variables (nodes) $\mathbf{X}_K = \{X_{k_1}, \ldots, X_{k_m}\}$, where $K = \{k_1, \ldots, k_m\}$.

Case 2 may be considered as 'trivial' correlation and may not suggest a link (i, j).

If X_i and X_j are found to have a large r_{ij}:

1. There is direct dependence of X_i on X_j, or of X_j on X_i, or both.
2. Both X_i and X_j are dependent on an other variable (node) X_k or on m other variables (nodes) $\mathbf{X}_K = \{X_{k_1}, \ldots, X_{k_m}\}$, where $K = \{k_1, \ldots, k_m\}$.

Case 2 may be considered as 'trivial' correlation and may not suggest a link (i, j).

To maintain links of only direct dependence, the appropriate similarity measure is the partial correlation

$$\rho_{ij|K} = \frac{\sigma_{ij|K}}{\sigma_{ii|K}\sigma_{jj|K}}$$

Partial Correlation Network see [2], Sec 7.3.1

If X_i and X_j are found to have a large r_{ij}:

1. There is direct dependence of X_i on X_j, or of X_j on X_i, or both.
2. Both X_i and X_j are dependent on an other variable (node) X_k or on m other variables (nodes) $\mathbf{X}_K = \{X_{k_1}, \ldots, X_{k_m}\}$, where $K = \{k_1, \ldots, k_m\}$.

Case 2 may be considered as 'trivial' correlation and may not suggest a link (i, j).

To maintain links of only direct dependence, the appropriate similarity measure is the partial correlation

$$\rho_{ij|K} = \frac{\sigma_{ij|K}}{\sigma_{ii|K}\sigma_{jj|K}}$$

$\rho_{ij|K} = 0$ if X_i and X_j are independent, conditional to $\mathbf{X}_K$.

Partial variance / covariance

$\sigma_{ij|K}$, $\sigma_{ii|K}$ and $\sigma_{jj|K}$ are components of the 2×2 partial covariance matrix

$$\Sigma_{11|2} = \begin{bmatrix} s^2_{ii|K} & s_{ij|K} \\ s_{ij|K} & s^2_{jj|K} \end{bmatrix}$$

Partial variance / covariance

$\sigma_{ij|K}$, $\sigma_{ii|K}$ and $\sigma_{jj|K}$ are components of the 2×2 partial covariance matrix

$$\Sigma_{11|2} = \begin{bmatrix} s_{ii|K}^2 & s_{ij|K} \\ s_{ij|K} & s_{jj|K}^2 \end{bmatrix}$$

$\Sigma_{11|2}$ is defined as

$$\Sigma_{11|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$\sigma_{ij|K}$, $\sigma_{ii|K}$ and $\sigma_{jj|K}$ are components of the 2×2 partial covariance matrix

$$\Sigma_{11|2} = \left[\begin{array}{cc} s^2_{ii|K} & s_{ij|K} \\ s_{ij|K} & s^2_{jj|K} \end{array} \right]$$

$\Sigma_{11|2}$ is defined as

$$\Sigma_{11|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

The matrices Σ_{11}, Σ_{12}, Σ_{22} and Σ_{21} are components of the partitioned covariance matrix

$$\mathrm{Cov}(\mathbf{W}) = \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right]$$

of all involved variables partitioned as $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2]'$, and $\mathbf{W}_1 = [X_i, X_j]'$, $\mathbf{W}_2 = \mathbf{X}_K$.

How to select the variables (nodes), to which the correlation between X_i and X_j is to be conditioned on?

Significance of Partial Correlation

How to select the variables (nodes), to which the correlation between X_i and X_j is to be conditioned on?

- How many, that is what is m?

How to select the variables (nodes), to which the correlation between X_i and X_j is to be conditioned on?

- How many, that is what is m?
- Which m variables from a total of $N-2$ variables?

Significance of Partial Correlation

How to select the variables (nodes), to which the correlation between X_i and X_j is to be conditioned on?

- How many, that is what is m?
- Which m variables from a total of $N-2$ variables?

Let us suppose we have decided the set of variables to condition on $\mathbf{X}_K = \{X_{k_1}, \ldots, X_{k_m}\}$.

Significance of Partial Correlation

How to select the variables (nodes), to which the correlation between X_i and X_j is to be conditioned on?

- How many, that is what is m?
- Which m variables from a total of $N-2$ variables?

Let us suppose we have decided the set of variables to condition on $\mathbf{X}_K = \{X_{k_1}, \ldots, X_{k_m}\}$.
To assess for a "non-trivial" link, test for significance of $\rho_{ij|K}$:
$\mathrm{H}_0 : \rho_{ij|K} = 0, \qquad \mathrm{H}_1 : \rho_{ij|K} \neq 0.$

Significance of Partial Correlation

How to select the variables (nodes), to which the correlation between X_i and X_j is to be conditioned on?

- How many, that is what is m?
- Which m variables from a total of $N-2$ variables?

Let us suppose we have decided the set of variables to condition on $\mathbf{X}_K = \{X_{k_1}, \ldots, X_{k_m}\}$.
To assess for a "non-trivial" link, test for significance of $\rho_{ij|K}$:
$\mathrm{H}_0 : \rho_{ij|K} = 0, \qquad \mathrm{H}_1 : \rho_{ij|K} \neq 0.$

The estimate of $\rho_{ij|K}$ is the sample partial correlation $r_{ij|K}$.

Significance of Partial Correlation

How to select the variables (nodes), to which the correlation between X_i and X_j is to be conditioned on?

- How many, that is what is m?
- Which m variables from a total of $N-2$ variables?

Let us suppose we have decided the set of variables to condition on $\mathbf{X}_K = \{X_{k_1}, \ldots, X_{k_m}\}$.
To assess for a "non-trivial" link, test for significance of $\rho_{ij|K}$:
$\mathrm{H}_0 : \rho_{ij|K} = 0, \qquad \mathrm{H}_1 : \rho_{ij|K} \neq 0.$

The estimate of $\rho_{ij|K}$ is the sample partial correlation $r_{ij|K}$.
Given n observations $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$ for each variable X_i, $r_{ij|K}$ is computationally derived in these steps:

Significance of Partial Correlation

How to select the variables (nodes), to which the correlation between X_i and X_j is to be conditioned on?

- How many, that is what is m?
- Which m variables from a total of $N-2$ variables?

Let us suppose we have decided the set of variables to condition on $\mathbf{X}_K = \{X_{k_1}, \ldots, X_{k_m}\}$.
To assess for a "non-trivial" link, test for significance of $\rho_{ij|K}$:
$\mathrm{H}_0 : \rho_{ij|K} = 0, \qquad \mathrm{H}_1 : \rho_{ij|K} \neq 0.$

The estimate of $\rho_{ij|K}$ is the sample partial correlation $r_{ij|K}$.
Given n observations $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$ for each variable X_i, $r_{ij|K}$ is computationally derived in these steps:

1. Compute the residuals $\mathbf{e}_i$ of multiple linear regression of X_i on $\mathbf{X}_K$.

Significance of Partial Correlation

How to select the variables (nodes), to which the correlation between X_i and X_j is to be conditioned on?

- How many, that is what is m?
- Which m variables from a total of $N-2$ variables?

Let us suppose we have decided the set of variables to condition on $\mathbf{X}_K = \{X_{k_1}, \ldots, X_{k_m}\}$.
To assess for a "non-trivial" link, test for significance of $\rho_{ij|K}$:
$\mathrm{H}_0 : \rho_{ij|K} = 0, \qquad \mathrm{H}_1 : \rho_{ij|K} \neq 0.$

The estimate of $\rho_{ij|K}$ is the sample partial correlation $r_{ij|K}$.
Given n observations $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$ for each variable X_i, $r_{ij|K}$ is computationally derived in these steps:

1. Compute the residuals $\mathbf{e}_i$ of multiple linear regression of X_i on $\mathbf{X}_K$.
2. Similarly, compute the residuals $\mathbf{e}_j$ of X_j on $\mathbf{X}_K$.

How to select the variables (nodes), to which the correlation between X_i and X_j is to be conditioned on?

- How many, that is what is m?
- Which m variables from a total of $N-2$ variables?

Let us suppose we have decided the set of variables to condition on $\mathbf{X}_K = \{X_{k_1}, \ldots, X_{k_m}\}$.
To assess for a "non-trivial" link, test for significance of $\rho_{ij|K}$:
$\mathrm{H}_0 : \rho_{ij|K} = 0, \qquad \mathrm{H}_1 : \rho_{ij|K} \neq 0.$

The estimate of $\rho_{ij|K}$ is the sample partial correlation $r_{ij|K}$.
Given n observations $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$ for each variable X_i, $r_{ij|K}$ is computationally derived in these steps:

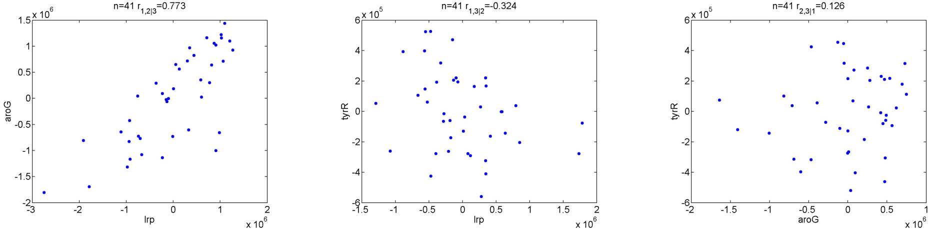
1. Compute the residuals $\mathbf{e}_i$ of multiple linear regression of X_i on $\mathbf{X}_K$.
2. Similarly, compute the residuals $\mathbf{e}_j$ of X_j on $\mathbf{X}_K$.
3. $r_{ij|K} = r_{\mathbf{e}_i,\mathbf{e}_j}$, the correlation coefficient of $\mathbf{e}_i$ and $\mathbf{e}_j$.

Example: Gene Regulation (continuing)

Partial correlation for the three genes: lrp, aroG, tyrR, and 41 experiments.

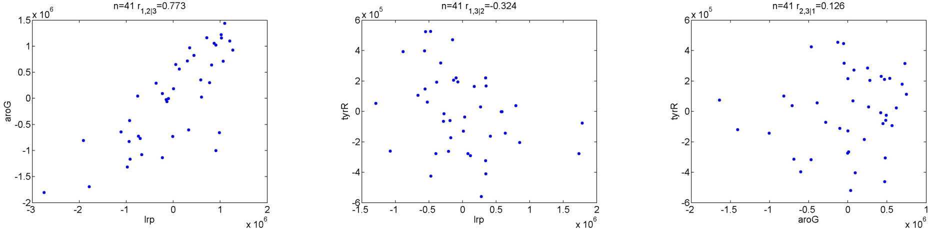
Example: Gene Regulation (continuing)

Partial correlation for the three genes: lrp, aroG, tyrR, and 41 experiments.



Example: Gene Regulation (continuing)

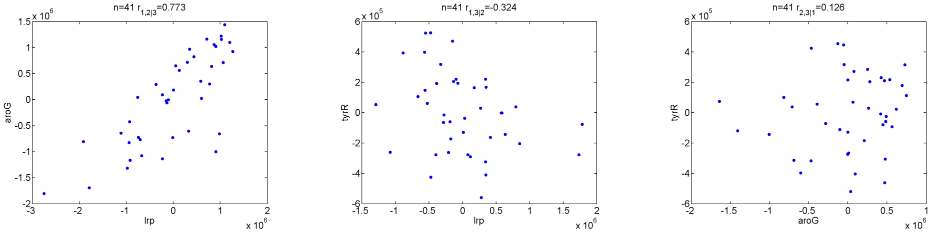
Partial correlation for the three genes: lrp, aroG, tyrR, and 41 experiments.



gene pair	r_{ij}	$r_{ij\|k}$	z-statistic (p-value)	rank (p-value)
lrp-aroG	0.78	0.77	6.25 (0.0000)	1001 (0.0013)
lrp-tyrR	-0.36	-0.32	-2.04 (0.0411)	23 (0.0453)
aroG-tyrR	-0.21	0.13	0.77 (0.4421)	765 (0.4727)

Example: Gene Regulation (continuing)

Partial correlation for the three genes: lrp, aroG, tyrR, and 41 experiments.



gene pair	r_{ij}	$r_{ij\|k}$	z-statistic (p-value)	rank (p-value)
lrp-aroG	0.78	0.77	6.25 (0.0000)	1001 (0.0013)
lrp-tyrR	-0.36	-0.32	-2.04 (0.0411)	23 (0.0453)
aroG-tyrR	-0.21	0.13	0.77 (0.4421)	765 (0.4727)

Only the partial correlation of aroG-tyrR is substantially different from the correlation coefficient.

Exercise 3: Micro-array Data

Do the same as in Exercise 2 but using the partial correlation as similarity matrix.

Exercise 3: Micro-array Data

Do the same as in Exercise 2 but using the partial correlation as similarity matrix.

matlab: for the partial correlation you may use the function `parcorr` (in the Econometrics toolbox)

Correlation Network and Time Series

So far, the n measurements of attribute X_i are independent.

Correlation Network and Time Series

So far, the n measurements of attribute X_i are independent.

Further, we suppose that the n measurements may be ordered, typically being time dependent.

So far, the n measurements of attribute X_i are independent.

Further, we suppose that the n measurements may be ordered, typically being time dependent.

The vector $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$ denotes a **time series** of X_i.

So far, the n measurements of attribute X_i are independent.

Further, we suppose that the n measurements may be ordered, typically being time dependent.

The vector $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$ denotes a **time series** of X_i.

A similarity measure $\text{sim}(i, j)$ quantifies the level of

- correlation or coupling between X_i and X_j (undirected link)

Correlation Network and Time Series

So far, the n measurements of attribute X_i are independent.

Further, we suppose that the n measurements may be ordered, typically being time dependent.

The vector $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$ denotes a **time series** of X_i.

A similarity measure $\text{sim}(i,j)$ quantifies the level of

- correlation or coupling between X_i and X_j (undirected link)
- causality from X_i and X_j, and vice versa (directed link).

Correlation Network and Time Series

So far, the n measurements of attribute X_i are independent.

Further, we suppose that the n measurements may be ordered, typically being time dependent.

The vector $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$ denotes a **time series** of X_i.

A similarity measure $\mathrm{sim}(i,j)$ quantifies the level of

- correlation or coupling between X_i and X_j (undirected link)
- causality from X_i and X_j, and vice versa (directed link).

A standard similarity measure is again $\mathrm{Corr}(X_i, X_j) = r_{X_i,Y_j}$.

Correlation Network and Time Series

So far, the n measurements of attribute X_i are independent.

Further, we suppose that the n measurements may be ordered, typically being time dependent.

The vector $\mathbf{x}_i = [x_{i1}, \ldots, x_{in}]'$ denotes a **time series** of X_i.

A similarity measure $\text{sim}(i, j)$ quantifies the level of

- correlation or coupling between X_i and X_j (undirected link)
- causality from X_i and X_j, and vice versa (directed link).

A standard similarity measure is again $\text{Corr}(X_i, X_j) = r_{X_i, Y_j}$.

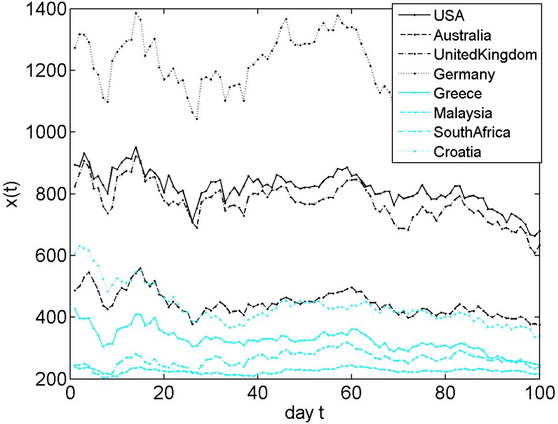
Others ???

Example: World financial markets

$N = 8$ world stock markets, daily indices, $n = 100$ days.

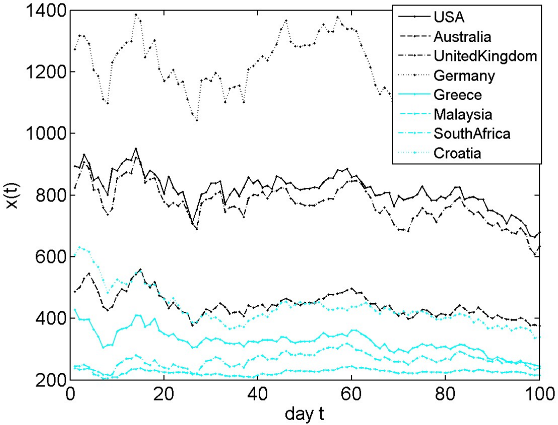
Example: World financial markets

$N = 8$ world stock markets, daily indices, $n = 100$ days.



Example: World financial markets

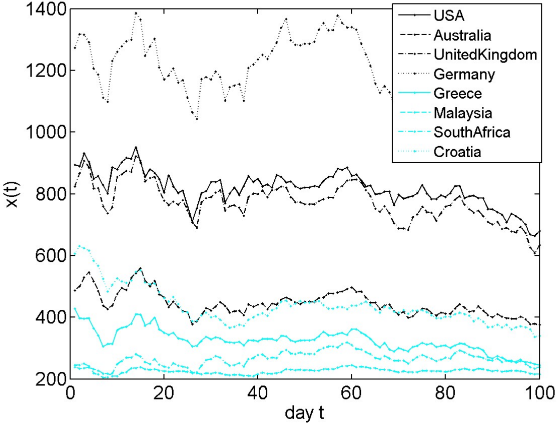
$N = 8$ world stock markets, daily indices, $n = 100$ days.



Similar indices, links among world stock markets?

Example: World financial markets

$N = 8$ world stock markets, daily indices, $n = 100$ days.



Similar indices, links among world stock markets?

Can we use the same similarity measure as for time-independent observations?

Example: World financial markets, correlation coefficient

Upper triangular: sample correlation coefficient r_{ij}.
Lower triangular: p-value for significance test for ρ_{ij} (z-statistic)

	USA	AUS	UK	GER	GRE	MAL	SAF	CRO
USA		0.86	0.92	0.88	0.89	0.33	0.27	0.75
AUS	0		0.91	0.82	0.90	0.56	0.27	0.83
UK	0	0		0.88	0.92	0.40	0.31	0.74
GER	0	0	0		0.84	0.44	0.53	0.61
GRE	0	0	0	0		0.40	0.16	0.82
MAL	0.0008	0	0	0	0		0.54	0.38
SAF	0.0057	0.0065	0.0017	0	0.1154	0		-0.15
CRO	0	0	0	0	0	0.0001	0.1408	

Example: World financial markets, correlation coefficient

Upper triangular: sample correlation coefficient r_{ij}.
Lower triangular: p-value for significance test for ρ_{ij} (z-statistic)

	USA	AUS	UK	GER	GRE	MAL	SAF	CRO
USA		0.86	0.92	0.88	0.89	0.33	0.27	0.75
AUS	0		0.91	0.82	0.90	0.56	0.27	0.83
UK	0	0		0.88	0.92	0.40	0.31	0.74
GER	0	0	0		0.84	0.44	0.53	0.61
GRE	0	0	0	0		0.40	0.16	0.82
MAL	0.0008	0	0	0	0		0.54	0.38
SAF	0.0057	0.0065	0.0017	0	0.1154	0		-0.15
CRO	0	0	0	0	0	0.0001	0.1408	

Almost all indices are strongly correlated.

Example: World financial markets, correlation network

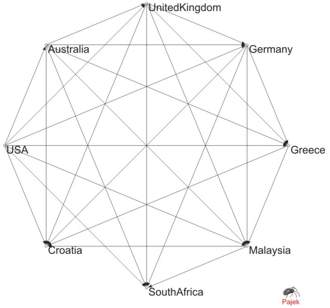
Adjacency matrix, threshold at $\alpha = 0.01$ (multiple testing?)

	USA	AUS	UK	GER	GRE	MAL	SAF	CRO
USA	0	1	1	1	1	1	1	1
AUS	1	0	1	1	1	1	1	1
UK	1	1	0	1	1	1	1	1
GER	1	1	1	0	1	1	1	1
GRE	1	1	1	1	0	1	0	1
MAL	1	1	1	1	1	0	1	1
SAF	1	1	1	1	0	1	0	0
CRO	1	1	1	1	1	1	0	0

Example: World financial markets, correlation network

Adjacency matrix, threshold at $\alpha = 0.01$ (multiple testing?)

	USA	AUS	UK	GER	GRE	MAL	SAF	CRO
USA	0	1	1	1	1	1	1	1
AUS	1	0	1	1	1	1	1	1
UK	1	1	0	1	1	1	1	1
GER	1	1	1	0	1	1	1	1
GRE	1	1	1	1	0	1	0	1
MAL	1	1	1	1	1	0	1	1
SAF	1	1	1	1	0	1	0	0
CRO	1	1	1	1	1	1	0	0



Example: World financial markets, partial correlation

Upper triangular: partial correlation $r_{ij|K}$, conditioned on all $|K| = 6$ rest variables.
Lower triangular: p-value for significance test for $\rho_{ij|K}$ (z-statistic)

	USA	AUS	UK	GER	GRE	MAL	SAF	CRO
USA		0.01	0.37	0.27	0.07	-0.27	0.11	0.27
AUS	0.9378		0.42	-0.02	0.15	0.30	0.10	0.38
UK	0.0002	0		0.08	0.36	-0.16	0.08	-0.11
GER	0.0081	0.8469	0.4693		0.38	-0.31	0.66	0.26
GRE	0.4946	0.1392	0.0003	0.0001		0.19	-0.36	0.01
MAL	0.0083	0.0033	0.1232	0.0026	0.0710		0.68	0.46
SAF	0.2908	0.3554	0.4321	0	0.0003	0		-0.70
CRO	0.0079	0.0002	0.3149	0.0099	0.9083	0	0	

Upper triangular: partial correlation $r_{ij|K}$, conditioned on all $|K| = 6$ rest variables.
Lower triangular: p-value for significance test for $\rho_{ij|K}$ (z-statistic)

	USA	AUS	UK	GER	GRE	MAL	SAF	CRO
USA		0.01	0.37	0.27	0.07	-0.27	0.11	0.27
AUS	0.9378		0.42	-0.02	0.15	0.30	0.10	0.38
UK	0.0002	0		0.08	0.36	-0.16	0.08	-0.11
GER	0.0081	0.8469	0.4693		0.38	-0.31	0.66	0.26
GRE	0.4946	0.1392	0.0003	0.0001		0.19	-0.36	0.01
MAL	0.0083	0.0033	0.1232	0.0026	0.0710		0.68	0.46
SAF	0.2908	0.3554	0.4321	0	0.0003	0		-0.70
CRO	0.0079	0.0002	0.3149	0.0099	0.9083	0	0	

Correlation between any two indices decreased when conditioned on all others.

Example: Financial markets, partial correlation network

Adjacency matrix, threshold at $\alpha = 0.01$

	USA	AUS	UK	GER	GRE	MAL	SAF	CRO
USA	0	0	1	1	0	1	0	1
AUS	0	0	1	0	0	1	0	1
UK	1	1	0	0	1	0	0	0
GER	1	0	1	0	1	1	1	1
GRE	0	0	1	1	0	0	1	0
MAL	1	1	1	1	0	0	1	1
SAF	0	0	1	1	1	1	0	1
CRO	1	1	1	1	0	1	1	0

Example: Financial markets, partial correlation network

Adjacency matrix, threshold at $\alpha = 0.01$

	USA	AUS	UK	GER	GRE	MAL	SAF	CRO
USA	0	0	1	1	0	1	0	1
AUS	0	0	1	0	0	1	0	1
UK	1	1	0	0	1	0	0	0
GER	1	0	1	0	1	1	1	1
GRE	0	0	1	1	0	0	1	0
MAL	1	1	1	1	0	0	1	1
SAF	0	0	1	1	1	1	0	1
CRO	1	1	1	1	0	1	1	0

Literature

[1] Data sets for `pajek` software,
http://vlado.fmf.uni-lj.si/pub/networks/data.

[2] Kolaczyk ED (2009) Statistical Analysis of Network Data, Springer.