Undirected Graphical Models (at a glance!)

Pierpaolo Brutti

Statistical Methods for Data Science

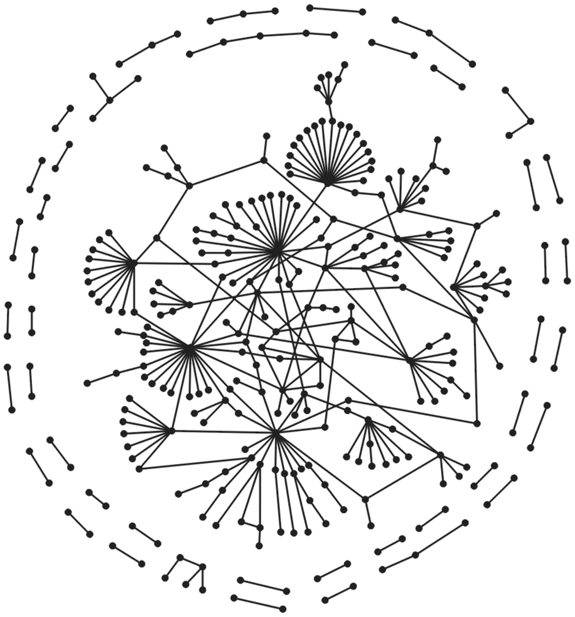
Introduction

Graphical models are a way of representing the relationships between features (variables). There are two main brands: **directed** and **undirected**. Here we shall focus on the latter.

Undirected graphis with attached a probabilistic semantic (i.e. undirected graphical models) come in different flavors, such as:

1. Marginal Correlation Graphs.
2. Partial Correlation Graphs.
3. Conditional Independence Graphs.

In each case, there are parametric and nonparametric versions but for the sake of these notes, we just talk about the parametric case. The figure shows an example of an undirected graph associated to protein networks from Maslov and Sneppen (2002).



Let $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ be random vector IID from some distribution $P(\cdot)$ where $\mathbf{X}_i = \left[X_i(1), \ldots, X_i(\mathrm{D})\right]^{\mathsf{T}} \in \mathbb{R}^{\mathrm{D}}$.

The **vertices** (nodes) of the graph refer to the D features: each node corresponds to a single feature. **Edges** instead represent *relationships* between the features.

The graph is represented by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{V_1, ..., V_{\mathrm{D}}\}$ is the vertex–set and $\mathcal{E}$ the edge–set. We can regard the edge–set $\mathcal{E}$ as a $(\mathrm{D} \times \mathrm{D})$ matrix $\mathbf{E}$ where $\mathbf{E}(j,k) = 1$ if there is an edge between feature j and feature k and 0 otherwise. Alternatively, you can regard $\mathcal{E}$ as a list of *unordered* pairs where $\{j,k\} \in \mathcal{E}$ if there is an edge between j and k.

As always, we write $X \perp Y$ to mean that X and Y are *independent* $\Leftrightarrow p(x,y) = p(x) \cdot p(y)$, whereas we write $(X \perp Y \,|\, Z)$ to mean that X and Y are independent given $Z \Leftrightarrow p(x,y \,|\, z) = p(x \,|\, z)\, p(y \,|\, z)$.

Marginal Correlation Graphs

In a **marginal correlation graph** – or *association graph* – we put an edge between V_j and V_k if

$$\left| \, \rho(j,k) \, \right| \geqslant \epsilon,$$

where $\rho(j,k)$ is *some* measure of *association*. Often we set $\epsilon = 0$ in which case there is an edge if and only if $\rho(j,k) \neq 0$. We also write $\rho_{j,k}$ or $\rho(X_j, X_k)$ to mean the same as $\rho(j,k)$.

The parameter $\rho(j,k)$ is required to have the following property:

$$X \perp Y \quad \Rightarrow \quad \rho(X,Y) = 0.$$

In general, the reverse may **not** be true. We will say that ρ is *strong* if

$$X \perp Y \quad \Leftrightarrow \quad \rho(X,Y) = 0.$$

In this section, we will not dig into these last type of association measures – see the Appendix for details. Let's just mention the recent work by Bergsma and Dassios (2014) who extended the well–known Kendall's τ into a *strong* correlation, and also the now well–known distance covariance defined in Szekely et al. (2007).

In general, we would like ρ to have several properties:

1. easy to compute;
2. robust to outliers;
3. there must be some way to calculate a confidence interval for the parameter.

Pearson Correlation. In spite to all its weaknesses and lack of robustness to outliers – it's just an expected value after all – a common choice of ρ is the **Pearson correlation coefficient**:

$$\rho(X,Y) = \frac{\mathbb{C}\mathrm{ov}(X,Y)}{\sigma_X \cdot \sigma_Y},$$

with the associated well–known plug–in estimator given by

$$\widehat{\rho}(X,Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{S_X \cdot S_Y}.$$

Remember that, since we are dealing with D features, to simplify notation we will write $\rho_{j,k} = \rho(j,k) \equiv \rho\big(X(j), X(k)\big)$, and the corresponding sample correlation will be denoted by $\widehat{\rho}_{j,k} = \widehat{\rho}(j,k)$.

Statistically speaking, checking if an edge $\{j,k\}$ is in the vertex–set $\mathcal{E}$ or not, is equivalent to test the null–hypothesis $\mathrm{H}_0 : \rho(j,k) = 0$ versus the alternative $\mathrm{H}_1 : \rho(j,k) \neq 0$. To this end we can use an *asymptotic* test or an *exact* test.

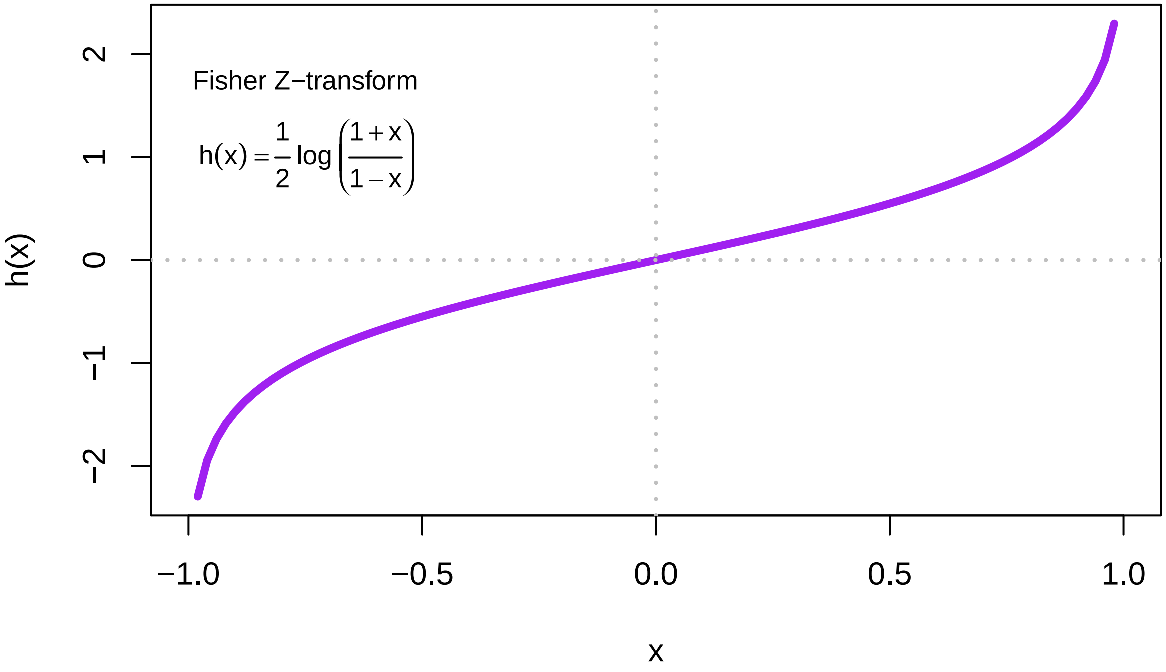
For the *asymptotic* test we'd like to invoke some sort of large sample normality of our estimator $\widehat{\rho}_{j,k}$, but by definition it is constrained to live between -1 and $+1$ $\Rightarrow$ we need to find a suitable transformation that maps smoothly $[-1,+1]$ into $\mathbb{R}$, the support of a Normal. As done before – see notes and code – let's define,

$$h(x) = \frac{1}{2}\log\left(\frac{1+x}{1-x}\right) = \mathrm{arctanh}(x) \quad \rightsquigarrow \quad Z_{j,k} = h(\widehat{\rho}_{j,k}) = \frac{1}{2}\log\left(\frac{1+\widehat{\rho}_{j,k}}{1-\widehat{\rho}_{j,k}}\right).$$

For this particular variance–stabilizing transformation, Fisher proved that

$$Z_{j,k} \mathrel{\dot\sim} \mathrm{N}_1\left(\theta_{j,k}, \frac{1}{n-3}\right) \text{ where } \theta_{j,k} = h(\rho_{j,k}) = \frac{1}{2}\log\left(\frac{1+\rho_{j,k}}{1-\rho_{j,k}}\right) \rightsquigarrow \rho_{j,k} = h^{-1}(\theta_{j,k}) = \tanh(\theta_{j,k}) = \frac{\mathrm{e}^{2\theta_{j,k}}-1}{\mathrm{e}^{2\theta_{j,k}}+1}.$$

Since $h(x) = 0 \Leftrightarrow x = 0$, following the usual path, we may reject the null if $|Z_{j,k}| > z_{\frac{\alpha}{2}} \cdot \frac{1}{\sqrt{n-3}}$, where $z_{\frac{\alpha}{2}} = -\mathrm{qnorm}\big(\frac{\alpha}{2}\big) \overset{\mathrm{sym}}{=} \mathrm{qnorm}\big(1-\frac{\alpha}{2}\big)$. So, for example, if $\alpha = 0.05$ then $z_{\frac{0.05}{2}} = -\mathrm{qnorm}(0.025) \approx 1.96$.



The result we just obtained is of course perfectly fine in case we have only a small number of edges to check. In general though, the graphs we have to deal with are characterized by a large number of nodes and, consequently, a huge amount of edges to test: $m = \binom{\mathrm{D}}{2}$ to be precise.

Take Home Message: we necessarily need to control for multiplicity in order to avoid a ridiculous overflow of false discoveries! The easiest to implement – although quite conservative recipe – here is the so called Bonferroni correction that simply asks for testing each single hypothesis at a level equal to α/m if α is the desired overall level and m the number of hypotheses. Hence, in our case, we just reject the null if $|Z_{j,k}| > z_{\frac{\alpha}{2m}} \cdot \frac{1}{\sqrt{n-3}}$ where $m = \binom{\mathrm{D}}{2}$.

Out of the asymptotic normality, we can also get confidence sets $\mathrm{C}_{n,\alpha} = [L, U]$ for each *single* $\rho_{i,j}$, where

- $L = h^{-1}\big(Z_{j,k} - z_{\frac{\alpha}{2}}/\sqrt{n-3}\big)$,
- $U = h^{-1}\big(Z_{j,k} + z_{\frac{\alpha}{2}}/\sqrt{n-3}\big)$.

A **simultaneous** confidence set for **all** the correlations can be obtained using the **bootstrap** (Wasserman et al., 2013). This is especially important if we want to put an edge when $\big|\rho(j,k)\big| \geqslant \epsilon$. If we have a confidence interval $\mathrm{C}_{n,\alpha}$ then we can put an edge whenever $[-\epsilon, +\epsilon] \cap \mathrm{C}_{n,\alpha} = \emptyset$.

1. Let $\mathbf{R}$ be the $(\mathrm{D} \times \mathrm{D})$ matrix of **true** correlations and let $\widehat{\mathbf{R}}$ be the $(\mathrm{D} \times \mathrm{D})$ matrix of sample correlations.
2. Now let $\{\mathbf{X}_1^\star, \ldots, \mathbf{X}_n^\star\}$ denote a **bootstrap** sample and let $\widehat{\mathbf{R}}^\star$ be the $(\mathrm{D} \times \mathrm{D})$ matrices of correlations from the bootstrap sample.
3. After taking B bootstrap samples we have $\{\widehat{\mathbf{R}}_1^\star, \ldots, \widehat{\mathbf{R}}_B^\star\}$
4. Now, for each bootstrap sample $b \in \{1, \ldots, B\}$, define the following bootstrapped replicate of a **simultaneous** test statistics
$$\Delta_b = \sqrt{n} \max_{j,k} \left| \widehat{\mathbf{R}}_b^\star[j,k] - \widehat{\mathbf{R}}[j,k] \right|,$$
and its associated boostrapped ECDF:
$$\widehat{F}_n(t) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\big(\Delta_b \leqslant t\big).$$
Within the usual bootstrap analogy, for large n and B, $\widehat{F}_n(t)$ should be a good approximation to
$$F_n(t) = \mathbb{P}\left(\sqrt{n} \max_{j,k} \left| \widehat{\mathbf{R}}[j,k] - \mathbf{R}[j,k] \right| \leqslant t\right).$$
5. Finally, to build our simultaneous confidence set, consider the sample quantile at level $1-\alpha$ of the bootstrapped distribution $\widehat{F}_n(t)$, say $t_\alpha = \widehat{F}_n^{-1}(1-\alpha)$, and set
$$\mathrm{C}_{j,k}(\alpha) = \left[\widehat{\mathbf{R}}[j,k] - \frac{t_\alpha}{\sqrt{n}}, \widehat{\mathbf{R}}[j,k] + \frac{t_\alpha}{\sqrt{n}}\right].$$

Theorem. *If* $\mathrm{D} = o\big(\exp(n^{1/6})\big)$, *then*

$$\mathbb{P}\big(\mathbf{R}[j,k] \in \mathrm{C}_{j,k}(\alpha) \text{ for all } (j,k)\big) \overset{n}{\to} 1-\alpha$$

Partial Correlation Graphs

Let X, Y be random variables and $\mathbf{Z}$ be a random vector. The partial correlation between X and Y, given $\mathbf{Z}$, is a measure of association between X and Y after removing the effect of $\mathbf{Z}$. Specifically, $\rho(X, Y \,|\, \mathbf{Z})$ is the correlation between the *residuals* ϵ_X and ϵ_Y where

$$\epsilon_X = X - \Pi_{\mathbf{Z}}(X) \quad \text{and} \quad \epsilon_Y = Y - \Pi_{\mathbf{Z}}(Y),$$

Here, $\Pi_{\mathbf{Z}}(X)$ is the *projection* of X onto the linear space spanned by $\mathbf{Z}$. That is $\Pi_{\mathbf{Z}}(X) = \widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{Z}$ where

$$\widehat{\boldsymbol{\beta}} = \underset{\text{All } \boldsymbol{\beta}}{\operatorname{argmin}}\, \mathbb{E}\big[Y - \boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}\big].$$

In other words, $\Pi_{\mathbf{Z}}(X)$ is the *linear regression* of X on $\mathbf{Z}$. Similarly, for $\Pi_{\mathbf{Z}}(Y)$. We will give an explicit formula for the partial correlation shortly.

Now let's go back to graphs. Let $\mathbf{X} = \big[X(1), \ldots, X(\mathrm{D})\big]^{\mathsf{T}}$ and let $\rho_{j,k}^{(p)}$ denote this time the partial correlation between $X(j)$ and $X(k)$ **given all** the other variables. Let $\mathbf{R}_{(p)} = \big[\,\rho_{j,k}^{(p)}\,\big]_{j,k}$ be the $(\mathrm{D} \times \mathrm{D})$ matrix of partial correlations. Then we have:

Theorem. *The matrix* $\mathbf{R}_{(p)}$ *is given by*

$$\mathbf{R}_{(p)}[j,k] = -\frac{\Lambda_{j,k}}{\sqrt{\Lambda_{j,j} \cdot \Lambda_{k,k}}},$$

where $\Lambda = \Sigma^{-1}$ *is the precision matrix associated to the covariance matrix* Σ.

The partical correlation graph $\mathcal{G}$ has an edge between j and k when $\rho_{j,k}^{(p)} \neq 0$.

In the **low–dimensional** setting ($\mathrm{D} << n$), we can estimate $\mathbf{R}_{(p)}$ as follows. Let $\widehat{\Sigma}_n$ be the sample covariance, and $\widehat{\Lambda} = \widehat{\Sigma}_n^{-1}$ its inverse. Then define

$$\widehat{\mathbf{R}}_{(p)}[j,k] = \widehat{\rho}_{j,k}^{(p)} = -\frac{\widehat{\Lambda}_{j,k}}{\sqrt{\widehat{\Lambda}_{j,j} \cdot \widehat{\Lambda}_{k,k}}}.$$

The easiest and reliable way to construct the graph is then to get **simultaneous** confidence intervals $\mathrm{C}_{j,k}(\alpha)$ using **boostrap** (again), and put in an edge $\{j,k\}$ if $0 \notin \mathrm{C}_{j,k}(\alpha)$.

There is also a *Normal approximation* similar to correlations. Define once again

$$Z_{j,k} = h\big(\widehat{\rho}_{j,k}^{(p)}\big) = \frac{1}{2}\log\left(\frac{1+\widehat{\rho}_{j,k}^{(p)}}{1-\widehat{\rho}_{j,k}^{(p)}}\right) \quad \text{then} \quad Z_{j,k} \mathrel{\dot\sim} \mathrm{N}_1\left(\theta_{j,k}, \frac{1}{n-g-3}\right)$$

where $\theta_{j,k} = h\big(\rho_{j,k}^{(p)}\big)$ and $g = (\mathrm{D}-2)$.

As before, via a Bonferroni correction we reject $\mathrm{H}_0 : \rho_{j,k}^{(p)} = 0$ if $|Z_{j,k}| > z_{\frac{\alpha}{2m}} \cdot \frac{1}{\sqrt{n-g-3}}$ where $m = \binom{\mathrm{D}}{2}$.

An implementation of this method is provided by the function `sinUG()` in the package `R` package `SIN`. The only difference is that, instead of Bonferroni, it adopts a finer correction for multiplicity.

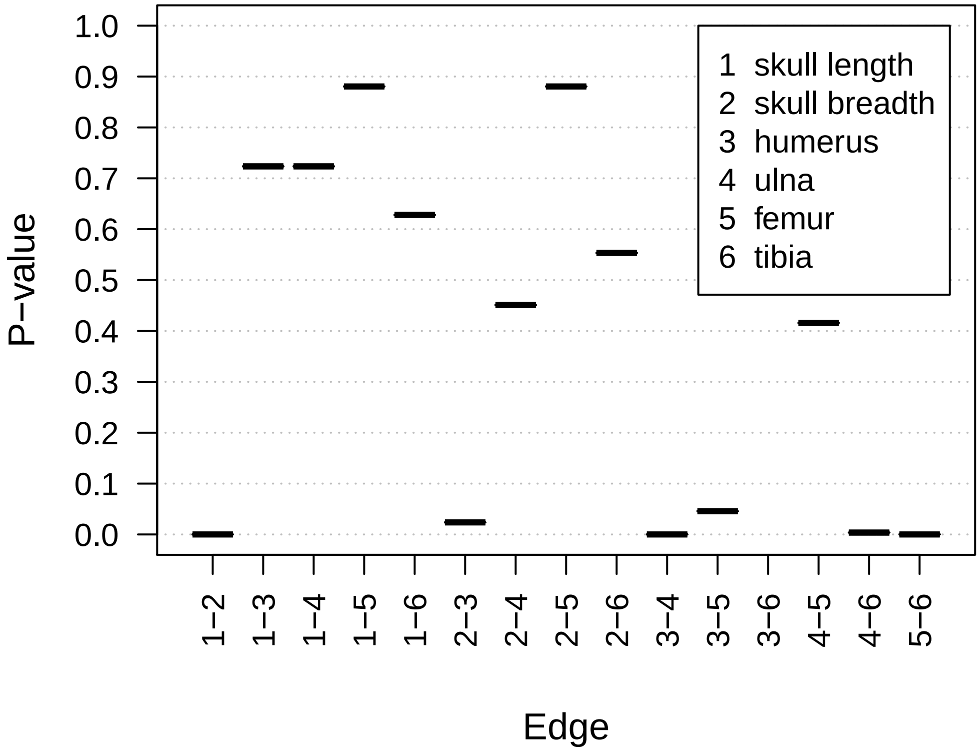
```r
# Package
library(SIN); library(help = SIN); ?sinUG

# Some data: are there dependencies between biometric measurements?
data(fowlbones)
?fowlbones

# Run & take a look
out <- sinUG(fowlbones$corr,fowlbones$n)
round(out,3)          # estimated partial correlation matrix
```

```
##               skull length skull breadth humerus  ulna femur tibia
## skull length            NA         0.000   0.723 0.723 0.880 0.628
## skull breadth        0.000            NA   0.024 0.451 0.880 0.553
## humerus              0.723         0.024      NA 0.000 0.046 0.723
## ulna                 0.723         0.451   0.000    NA 0.416 0.004
## femur                0.880         0.880   0.046 0.416    NA 0.000
## tibia                0.628         0.553   0.723 0.004 0.000    NA
```

```
plotUGpvalues(out)    # take a look at the p-values
```



```
# Get the graph associated to a specific confidence level
library(igraph)

par(mfrow = c(1,2))
alpha = 0.05
E.SIN = getgraph(out, alpha)
print(E.SIN)
```

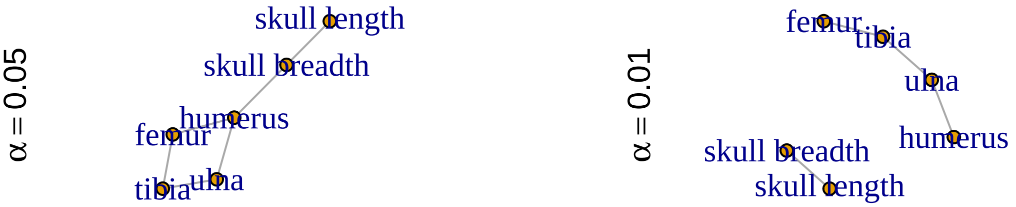
```
##               skull length skull breadth humerus ulna femur tibia
## skull length             0             1       0    0     0     0
## skull breadth            1             0       1    0     0     0
## humerus                  0             1       0    1     1     0
## ulna                     0             0       1    0     0     1
## femur                    0             0       1    0     0     1
## tibia                    0             0       0    1     1     0
```

```
G.SIN = graph.adjacency(E.SIN, mode = "undirected")
plot(G.SIN, ylab = expression(alpha == 0.05))

alpha = 0.01
E.SIN = getgraph(out, alpha)
print(E.SIN)
```

```
##               skull length skull breadth humerus ulna femur tibia
## skull length             0             1       0    0     0     0
## skull breadth            1             0       0    0     0     0
## humerus                  0             0       0    1     0     0
## ulna                     0             0       1    0     0     1
## femur                    0             0       0    0     0     1
## tibia                    0             0       0    1     1     0
```

```
G.SIN = graph.adjacency(E.SIN, mode = "undirected")
plot(G.SIN, ylab = expression(alpha == 0.01))
```



In high–dimensions, this protocol will **not** work since $\widehat{\Sigma}_n$ is **not** invertible. In fact one can show that

$$\mathbb{V}\mathrm{ar}\big(\,\widehat{\rho}^{(p)}_{j,k}\,\big) \approx \frac{1}{n - \mathrm{D}},$$

and this blows up when D is close to, or even larger than, n.

In such a bad situation, we have at least three exit strategies:

1. Compute a **correlation graph** instead. This is easy, works well, and often reveals similar dependence structures.
2. Apply some **shrinkage** (statistical lingo) or **regularization** (more math inclined) scheme to “fix” the ill–conditioned matrix $\widehat{\Sigma}_n$ and make it non–singular + bootstrap on the entries of the resulting matrix. See, for example, Schager and Strimmer (2005) and Ledoit and Wolf (2004).
3. Use the *graphical lasso* (`glasso` package). **Warning!** The reliability of the graphical lasso depends on lots of non-trivial, uncheckable assumptions.

Conditional Independence Graphs

The strongest type of undirected graph is a **conditional independence graph**. In this case, we omit the edge between j and k if $X(j)$ is independent of $X(k)$ given the *rest* of the variables. We write this as

$$\big(X(j) \perp X(k)\,\big|\,\mathrm{rest}\big)$$

Conditional independence graphs are the most informative undirected graphs but they are also the hardest to estimate! This final type of undirected graphical models is treated in some details in *Ch.18* and *Ch.19* of the purple book, so we are not going much further.

Here we just mention an interesting fact related to the so called Gaussian Random Fields (i.e. undirected graphical models over Normal random vectors). In fact, in the special case of Normality, conditional independence graphs are equivalent to partial correlation graphs.

More specifically we have the following result.

Theorem: *Suppose that*

$$\mathbf{X} = \big[X(1),\ldots,X(\mathrm{D})\big]^{\mathsf{T}} \sim \mathrm{N}_{\mathrm{D}}(\boldsymbol{\mu},\Sigma).$$

Let $\Lambda = \Sigma^{-1}$ be the precision matrix. Then $X(j)$ is independent of $X(k)$ given the rest, if and only if $\Lambda_{j,k} = 0$. So, in the Normal case, we are back to partial correlations.

Appendix: Other measures of association

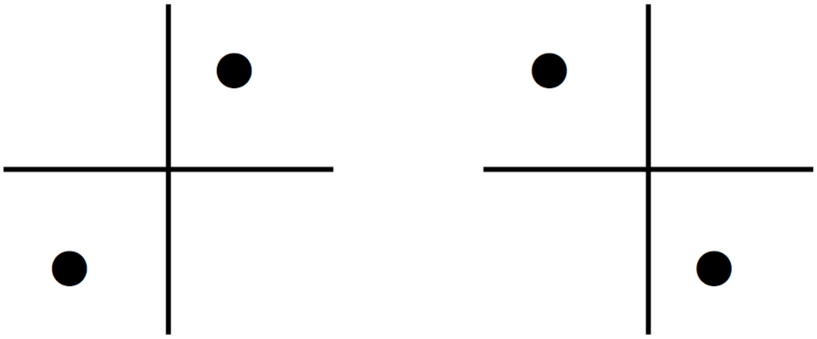
Kendall's τ. The Pearson linear correlation is **not** very robust to outliers. Talking about copula models, we introduced Spearman's ρ as Pearson correlation applied to ranks. Another robust measure of association is Kendall's τ.

So consider a random vector (X, Y) with joint distribution $F_{X,Y}(\cdot)$, and denote by (X', Y') an *independent* copy of (X, Y) – in other words (X, Y) and (X', Y') are independent and identically distributed.

We say that (X, Y) and (X', Y') are **concordant** if

$$(X - X')(Y - Y') \geqslant 0$$

and **discordant** otherwise.



Concordant Pair

Discordant Pair

Kendall's τ is defined as:

$$\tau(X, Y) = \mathbb{P}\big(\text{concordant}\big) - \mathbb{P}\big(\text{discordant}\big) = \mathbb{P}\big((X - X')(Y - Y') \geqslant 0\big) - \mathbb{P}\big((X - X')(Y - Y') < 0\big).$$

After some algebra it can be shown that

$$\tau(X, Y) = \mathbb{E}\Big[\text{sign}\big((X - X')(Y - Y')\big)\Big],$$

so that an obvious estimator is given by

$$\widehat{\tau}(X, Y) = \frac{1}{\binom{n}{2}} \sum_{j,k} \Big[\text{sign}\big((X_j - X_k)(Y_j - Y_k)\big)\Big].$$

A statistics of this form is called a U–statistics, and there's a corpus of dedicated literature that gives us all the relevant information about its (asymptotic) sampling distribution.

In particular, going back to checking the presence of a $\{j, k\}$ edge in our marginal association graph, under the null hypothesis $\mathrm{H}_0 : \tau_{j,k} = 0$ we know that asymptotically

$$\widehat{\tau}_{j,k} \,\dot{\sim}\, \mathrm{N}\left(0, \tfrac{4}{9 \cdot n}\right),$$

so we reject when

$$\big|\widehat{\tau}_{j,k}\big| > z_{\frac{\alpha}{2m}} \cdot \frac{2}{3\sqrt{n}}.$$

A finite sample alternative would be a permutation test (see Section 10.5 of the purple book).

Distance Correlation. There are various other nonparametric measures of association. The most common are the **distance correlation** and the RKHS correlation (from Reproducing Kernel Hilbert Space).

Let's briefly focus on **distance correlation**. So the **squared distance covariance** between two random vectors $\mathbf{X}$ and $\mathbf{Y}$ is defined by (Szekely et al. 2007)

$$\gamma^2(\mathbf{X}, \mathbf{Y}) = \mathbb{C}\text{ov}\big(\|\mathbf{X} - \mathbf{X}'\|, \|\mathbf{Y} - \mathbf{Y}'\|\big) - 2\,\mathbb{C}\text{ov}\big(\|\mathbf{X} - \mathbf{X}'\|, \|\mathbf{Y} - \mathbf{Y}''\|\big),$$

where $(\mathbf{X}, \mathbf{Y})$, $(\mathbf{X}', \mathbf{Y}')$ and $(\mathbf{X}'', \mathbf{Y}'')$ are independent copies, and $\|\cdot\|$ denotes the Euclidean norm or any other suitable norm. Please notice that the random vector $\mathbf{X}$ and $\mathbf{Y}$ can be of different dimension.

The *squared distance correlation* is then

$$\rho^2(\mathbf{X}, \mathbf{Y}) = \frac{\gamma^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\gamma^2(\mathbf{X}, \mathbf{X})\,\gamma^2(\mathbf{Y}, \mathbf{Y})}}.$$

Another expression (Lyons 2013) for γ is

$$\gamma^2(\mathbf{X},\mathbf{Y}) = \mathbb{E}\big[\delta(\mathbf{X},\mathbf{X}') \cdot \delta(\mathbf{Y},\mathbf{Y}')\big],$$

where

$$\delta(\mathbf{X},\mathbf{X}') = d(\mathbf{X},\mathbf{X}') - 2 \cdot \int d(\mathbf{X},\mathbf{u})\,\mathrm{d}P(\mathbf{u}) + \int\int d(\mathbf{u},\mathbf{v})\,\mathrm{d}P(\mathbf{u})\,\mathrm{d}P(\mathbf{v}),$$

and $d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$. In fact, other metrics $d(\cdot,\cdot)$ can be used.

An important result is the following

Theorem. *We have that* $\rho(\mathbf{X},\mathbf{Y}) \in [0,1]$ *and* $\rho(\mathbf{X},\mathbf{Y}) = 0 \Leftrightarrow \mathbf{X} \perp \mathbf{Y}$.

Given a random sample $\big\{(\mathbf{X}_i,\mathbf{Y}_i)\big\}_{i=1}^n$, a plug-in estimator for γ can be defined as follows. Let,

- $a_{j,k} = \|\mathbf{X}_j - \mathbf{X}_k\|$,
- $a_{j,\bullet}$, $a_{\bullet,k}$, and $a_{\bullet,\bullet}$ be the row, column and grand means of the matrix $\mathbb{A} = [a_{j,k}]_{j,k}$,
- $A_{j,k} = a_{j,k} - (a_{j,\bullet} + a_{\bullet,k}) + a_{\bullet,\bullet}$,
- $b_{j,k} = \|\mathbf{Y}_j - \mathbf{Y}_k\|$,
- $b_{j,\bullet}$, $b_{\bullet,k}$, and $b_{\bullet,\bullet}$ be the row, column and grand means of the matrix $\mathbb{B} = [b_{j,k}]_{j,k}$,
- $B_{j,k} = b_{j,k} - (b_{j,\bullet} + b_{\bullet,k}) + b_{\bullet,\bullet}$.

Then,

$$\widehat{\gamma}^2(\mathbf{X},\mathbf{Y}) = \frac{1}{n^2}\sum_{j,k} A_{j,k} \cdot B_{j,k}.$$

The asymptotic distribution of $\widehat{\gamma}^2$ is complicated. But we can easily test the hypothesis $\gamma^2 = 0$ adopting a permutation test (see Section 10.5 of the purple book).

Fortunately, there's an handy `R` package called `energy` that ships with all the basic procedures already neatly implemented.