Graphical Models

-Applied Multivariate Analysis & Statistical Learning-

MMA chapter 7.2 and these notes

Lecturer: Darren Homrighausen, PhD

Preamble:

- Define graphical models
- Discuss estimating covariance graphs
- Discuss estimating correlation graphs
- Discuss estimating partial correlation graphs

Graphical Model

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

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

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

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

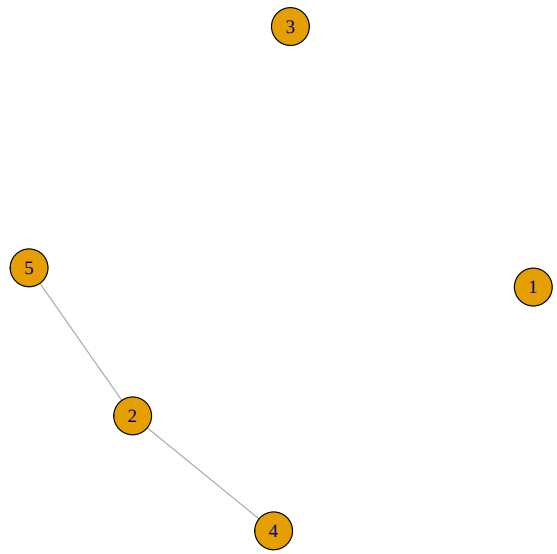
For our purposes, each vertex corresponds to a feature

Each edge represents some aspect of the relationship between features

(For our purposes, we will only consider undirected graphs)

Example graphical model

Suppose we have 5 features x_j, $j = 1, \ldots, 5$. A possible undirected graphical model would look like:



There is a 'relationship' between x_2 and x_5, for instance

The details of this relationship depends on what we are estimating

Another example: Spam

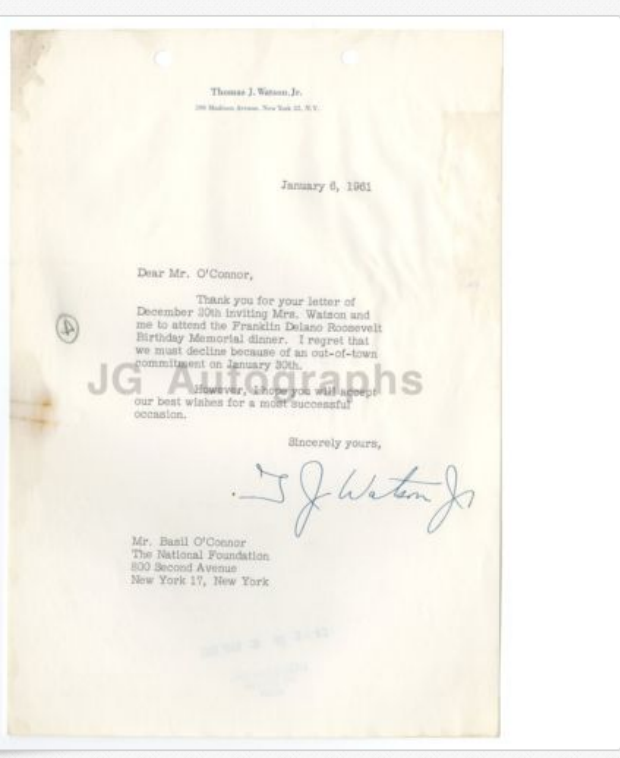
Suppose we are investigating whether an email is 'spam' or 'not spam'

We can text process a corpus of emails into a bag of words

The idea is that the words used in an email should provide information about the content, regardless of the semantic structure

I'll just give a brief idea what those concepts are here

Brief overview of text processing



Thomas Watson, Jr. - IBM Chairman - Authentic Autographed Letter (TLS)

Item condition: --

Ended: May 27, 2014 16:59:11 PDT

Winning bid: **US $11.61** [6 bids]

Shipping: **$3.99** Standard Shipping | See details
Item location: **United States**
Ships to: **Worldwide**

Delivery: Estimated within 3-6 business days

Payments: PayPal | See details

Returns: 14 days money back, buyer pays return shipping | See details

Guarantee: ebay MONEY BACK GUARANTEE | See details
Get the item you ordered or get your money back.
Covers your purchase price and original shipping.

Seller information
jgautographs (64927 ★)
100% Positive feedback

Follow this seller
See other items
Visit store: JG Autographs

Brief overview of text processing

Buyer:

Seller:

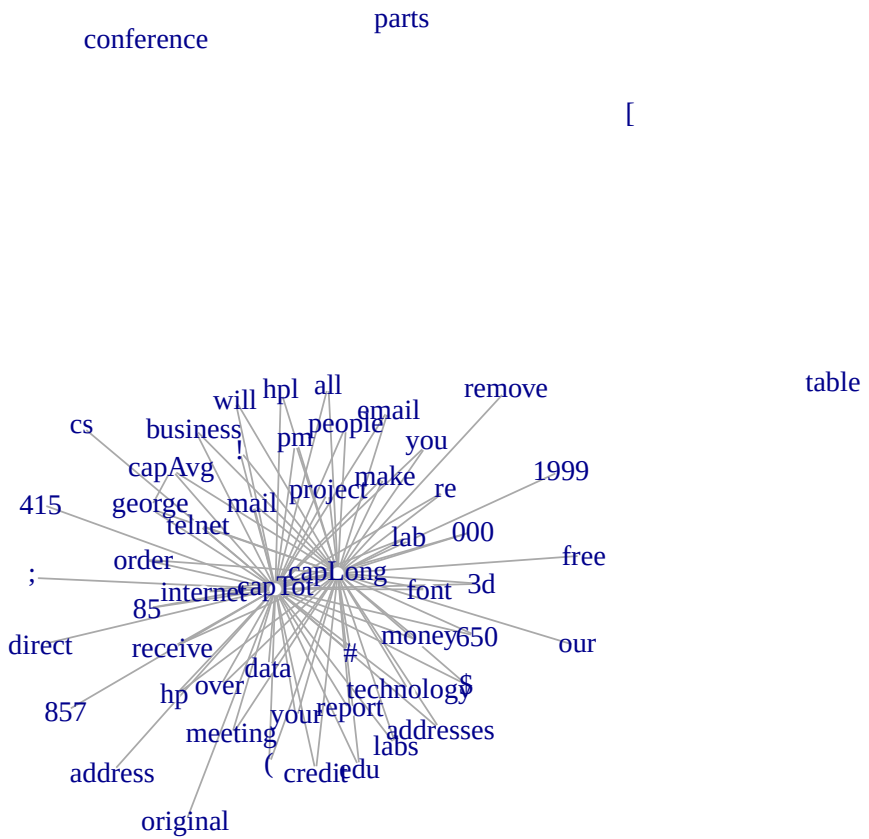
The feature matrix can then be written as $\mathbb{X} = \begin{bmatrix} X_1^\top \\ X_2^\top \\ \vdots \end{bmatrix}$ where...

	always	pleas	smooth	transact	great	commun	busin
$X_1^\top =$	[1	2	1	1	0	0	0]
$X_2^\top =$	[0	1	0	0	1	1	1]

Another example: Spam

Now, the features are the counts of the words in each email

(There is some normalization to account for the emails having different lengths)



Covariance graphs

Covariance graphs

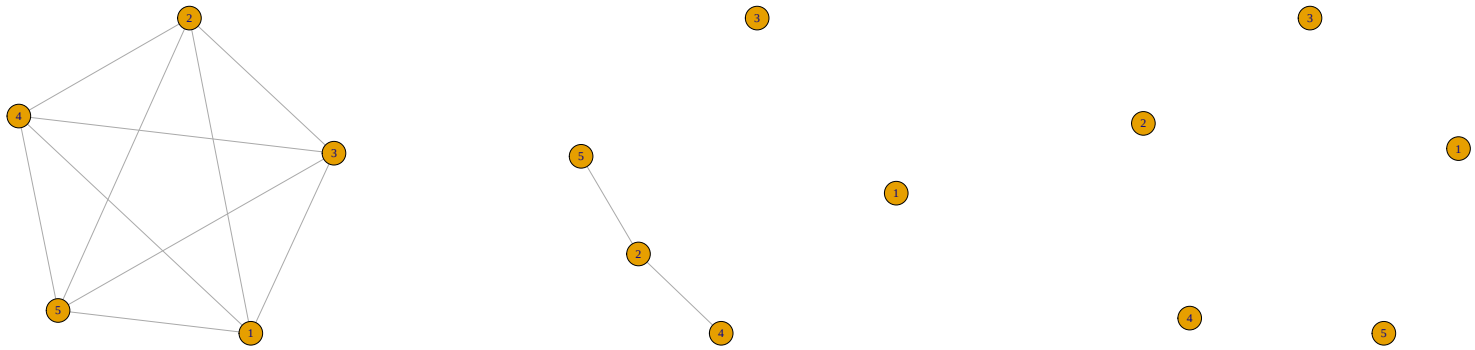
The first type of relationship we will talk about is covariance

Using the estimator S, we can put can edge between two features x_j and x_k if $|S_{jk}| > \epsilon$

Looking at the example data set again, we can plot the graphical model for various thresholds

Covariance graphs: Interpretation

$|S_{jk}| > 0$ $|S_{jk}| > 0.05$ $|S_{jk}| > 0.2$



Interpretation: For $|S_{ij}| > 0.05$, which features do we estimate are related?

We need to address two issues:

- We would like to strengthen the interpretation to independence
- We should use statistical theory to make a statement about Σ

Covariance graphs: Independence

To make an independence statement, we need to check normality

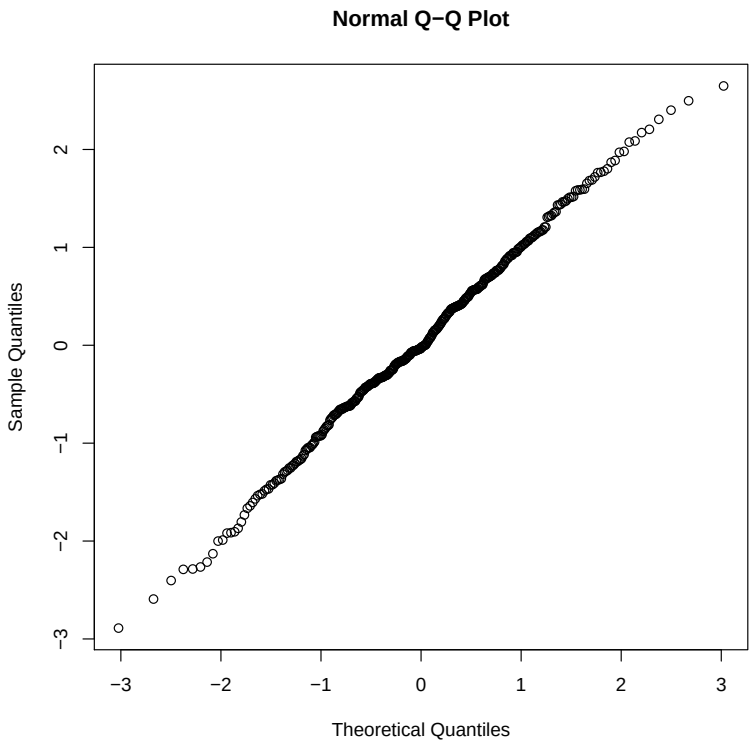
We discussed two ways of doing this:

- Check the marginal distributions
- Look at the sample Mahalanobis distances

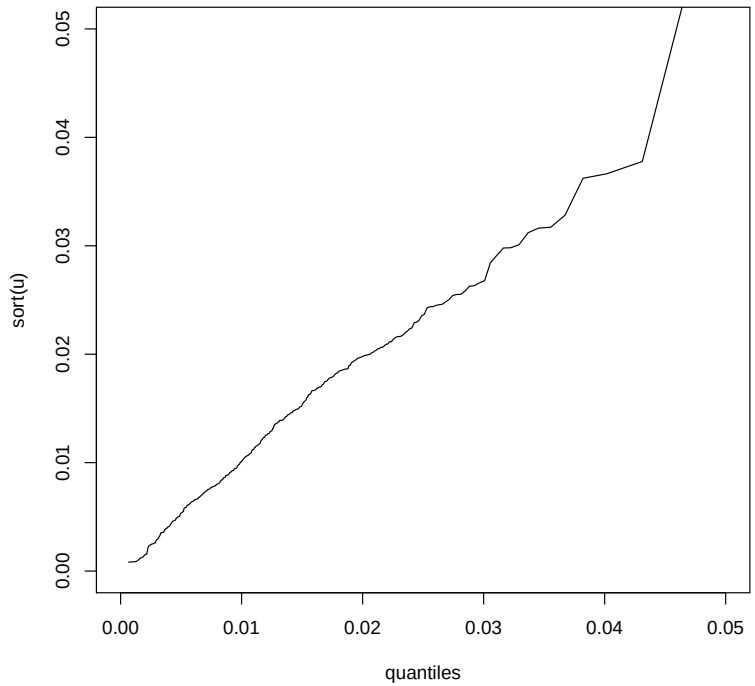
Both of these checks can be done with QQ plots

Covariance graphs: Independence

One of the marginal QQ plots



QQplot of u vs. beta



Covariance graphs: Testing

Our first take on formalizing the inference about the covariance is a test of sphericity

$$H_0 : \Sigma = \sigma^2 I$$

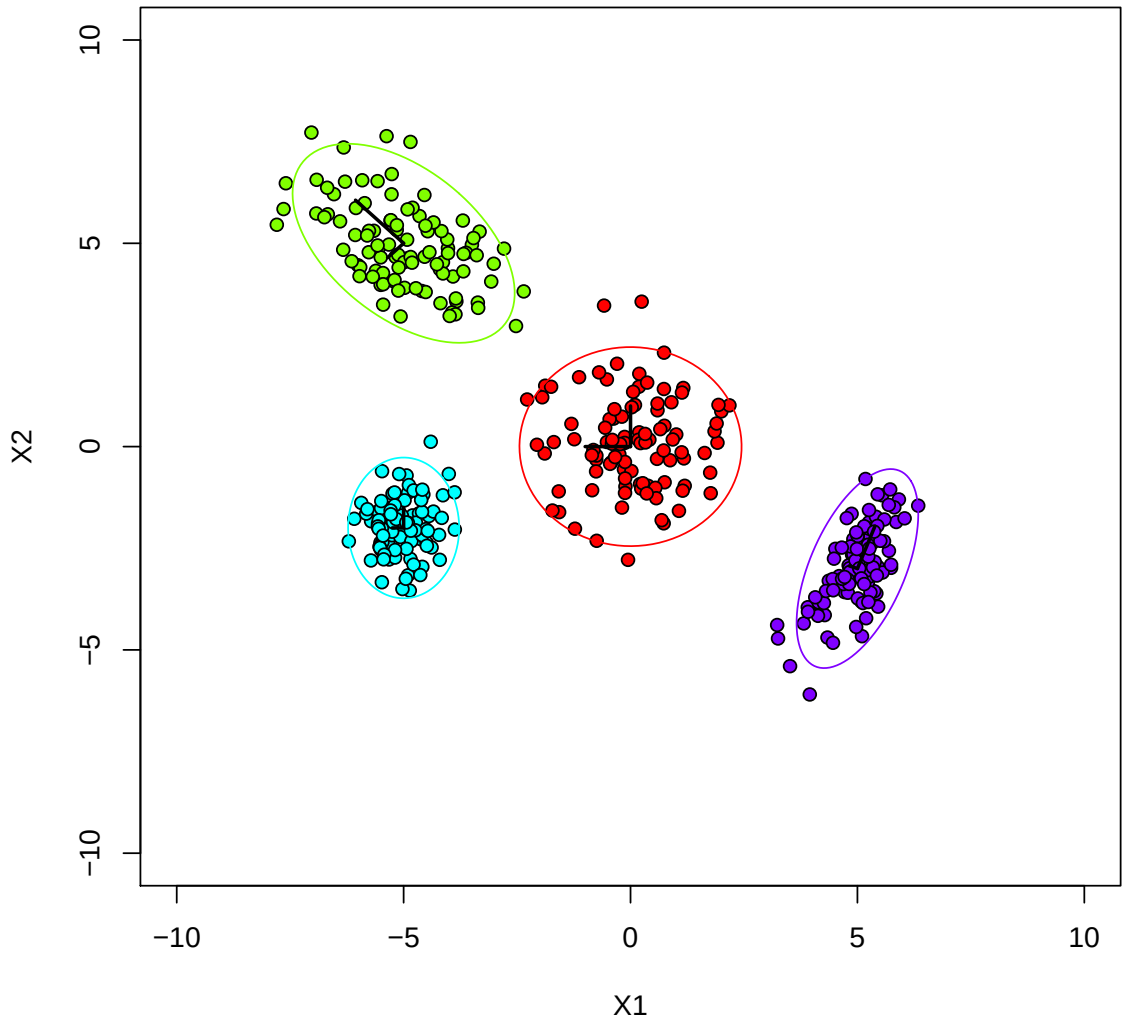
We know that $(X - \mu)\Sigma^{-1}(X - \mu) = c^2$ forms an ellipse

So, if $\Sigma = \sigma^2 I \Rightarrow \Sigma^{-1} = \sigma^{-2} I$ and

$$(X - \mu)^\top \Sigma^{-1}(X - \mu) = c^2 \Rightarrow (X - \mu)^\top (X - \mu) = \sigma^2 c^2,$$

which is the equation of a sphere

MMA chapter 7.2.2


10
5
0
−5
−10
X2
−10
−5
0
5
10
X1

Covariance graphs: Testing

To test H_0, we can use a likelihood ratio test (LRT)

Likelihood ratio test: A general method for testing two hypotheses is to form the ratio of the likelihoods (LR) maximized over each hypotheses:

$$LR = \frac{\max_{\theta \in H_0} L(\theta)}{\max_{\theta \in H_A} L(\theta)}$$

(θ is some parameter (Σ in our case) and L is a likelihood)

Then, the LRT says:

$$-2\log(LR) \to \chi^2_\nu$$

ν is the (# parameters under H_A) - (# parameters under H_0)

(In this case, $\nu =$?)

Covariance graphs: Testing

For testing for sphericity, the LRT looks like

$$-2\log(LR) = -n\log\left(\frac{|S|}{\text{trace}(S/p)^p}\right)$$

This can be computed efficiently via the SVD of $S = UD^2U^\top$ as

- $|S| = \prod_{j=1}^p d_j^2$
- $\text{trace}(S) = \sum_{j=1}^p d_j^2$

MMA chapters 2.107 & 2.108

Covariance graphs: Testing

A slight modification of $-2\log(LR)$ replaces $-n$ with $-\left(n-1-\frac{2p^2+p+2}{6p}\right)$, which provides better small sample behavior

Hence, we can reject H_0 if

$$-\left(n-1-\frac{2p^2+p+2}{6p}\right)\log\left(\frac{|S|}{\text{trace}(S/p)^p}\right) > \chi^2_{p(p+1)/2-1}(1-\alpha)$$

MMA chapters 7.2.2

Covariance graphs: Testing

Back to our example, using testStat as the (modified) LRT

```
testStat
[1] 14.07331
> qchisq(.95,p*(p+1)/2 - 1)
[1] 23.68479
```

What's the interpretation?

Covariance graphs: Testing

Let's modify the simulation to add covariance

Let Σ be $\rho = .1$ on the off diagonals and $\sigma^2 = 1$ on the diagonal

Then:

```
testStat
[1] 67.98457
> qchisq(.95,p*(p+1)/2 - 1)
[1] 23.68479
```

So, we can reject H_0 and conclude that there exists dependency in these features

Note that we can instead test blocks of Σ in the same way if we are interested in particular subsets of the features

In order to produce a sensible graphical model, we should seek to make confidence intervals...

Correlation graphs

Correlation graphs

Correlation graphs are similar to covariance graphs in that we put an edge between two nodes (features) if $|\rho_{jk}| > \epsilon$

(Often, $\epsilon = 0$ is chosen)

Here, ρ is some measure of association

A common choice is to use Pearson's correlation:

$$\rho_{jk} = \frac{\Sigma_{jk}}{\sqrt{\Sigma_{jj}\Sigma_{kk}}}$$

Correlation graphs: Other measures

Note that there are others valid choices for ρ such as

- Spearman's
- Kendall's τ

In fact, there are some such that no correlation implies independence regardless of normality!

- 'distance correlation'
- 'Bergsma-Dassios τ correlation'

The other correlations drop the implicit 'linear' notation of Pearsons and instead rely on an idea called concordance and discordance

Correlation graphs: Testing

Using Pearson's correlation, we can test

$$H_0 : \rho_{jk} = 0 \text{ vs. } H_A : \rho_{jk} \neq 0$$

via

- an asymptotic test. This uses a transformation of the sample correlation + normal approximation
- permutation test. This recomputes the correlations after permuting the features

Correlation graphs: Asymptotic test

Define the following transformation:

$$Z_{jk} = \frac{1}{2} \log \left(\frac{1 + r_{jk}}{1 - r_{jk}} \right)$$

This random variable has an approximate distribution:

$$Z_{jk} \to N \left(\theta_{jk}, \frac{1}{n-3} \right) \text{ under } H_0$$

where

$$\theta_{jk} = \frac{1}{2} \log \left(\frac{1 + \rho_{jk}}{1 - \rho_{jk}} \right)$$

We can reject H_0 if $|Z_{jk}| \sqrt{n-3} > Z(1-\alpha/2)$

(Here, we indicate that $Z(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile of the standard normal)

Correlation graphs: Asymptotic test

This is what the test would look like:

```
fisherStat = 1/2 * log((1+cor(X))/(1-cor(X)))
diag(fisherStat) = 0
abs(fisherStat) > qnorm(.975)/sqrt(n-3)
```

Uncorrelated example:

```
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] FALSE FALSE FALSE FALSE FALSE
[2,] FALSE FALSE FALSE  TRUE FALSE
[3,] FALSE FALSE FALSE FALSE FALSE
[4,] FALSE  TRUE FALSE FALSE FALSE
[5,] FALSE FALSE FALSE FALSE FALSE
```

Correlated example ($\rho_{jk} = .1$):

```
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] FALSE  TRUE  TRUE  TRUE  TRUE
[2,]  TRUE FALSE  TRUE  TRUE  TRUE
[3,]  TRUE  TRUE FALSE  TRUE FALSE
[4,]  TRUE  TRUE  TRUE FALSE  TRUE
[5,]  TRUE  TRUE FALSE  TRUE FALSE
```

Correlation graphs: Asymptotic test

Let's compute confidence intervals for the correlated simulation

The CI for θ_{jk} is straight-forward

$$CI = [\text{lower}, \text{upper}] = \left[Z_{jk} - \frac{Z(1-\alpha/2)}{\sqrt{n-3}}, Z_{jk} - \frac{Z(1-\alpha/2)}{\sqrt{n-3}} \right]$$

Transforming back to ρ_{jk} involves inversion:

$$\theta = \frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right)$$

```
untransformedLower = fisherStat - qnorm(.975)/sqrt(n-3)
untransformedUpper = fisherStat + qnorm(.975)/sqrt(n-3)
transformedLower   = (exp(2*untransformedLower)-1)/
                          (exp(2*untransformedLower)+1)
transformedUpper   = (exp(2*untransformedUpper)-1)/
                          (exp(2*untransformedUpper)+1)
```

Correlation graphs: Asymptotic test

```
transformedLower[upper.tri(transformedLower, diag=TRUE)] = NA
transformedUpper[upper.tri(transformedUpper, diag=TRUE)] = NA
```

```
> round(transformedLower,3)
      [,1]  [,2]   [,3]  [,4] [,5]
[1,]    NA    NA     NA    NA   NA
[2,] 0.021    NA     NA    NA   NA
[3,] 0.002 0.013     NA    NA   NA
[4,] 0.019 0.108  0.047    NA   NA
[5,] 0.006 0.098 -0.006 0.029   NA
> round(transformedUpper,3)
      [,1]  [,2]  [,3]  [,4] [,5]
[1,]    NA    NA    NA    NA   NA
[2,] 0.214    NA    NA    NA   NA
[3,] 0.196 0.207    NA    NA   NA
[4,] 0.212 0.296 0.239    NA   NA
[5,] 0.200 0.286 0.189 0.222   NA
```

Correlation graphs: Asymptotic test

We really should be controlling for multiple testing

(We are after all testing $\binom{p}{2}$ parameters (10 in this case))

An easy and theoretically sound choice in this case is the Bonferroni correction

The idea behind the Bonferroni correction is called the union bound

$$\mathbb{P}(A \text{ or } B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

Correlation graphs: Asymptotic test

For M tests, we can control the overall Type 1 error by:

$$\mathbb{P}(\text{Type I error on test 1 or ... Type I error on test } M)$$
$$\leq \sum_{m=1}^{M} \mathbb{P}(\text{Type I error on test m}) = M\alpha$$

We need to test at α/M to keep the overall $\mathbb{P}(\text{type I error}) = \alpha$

```
> qnorm(.975)/sqrt(n-3)
[1] 0.09836777
> qnorm(1-.025/10)/sqrt(n-3)
[1] 0.140881
```

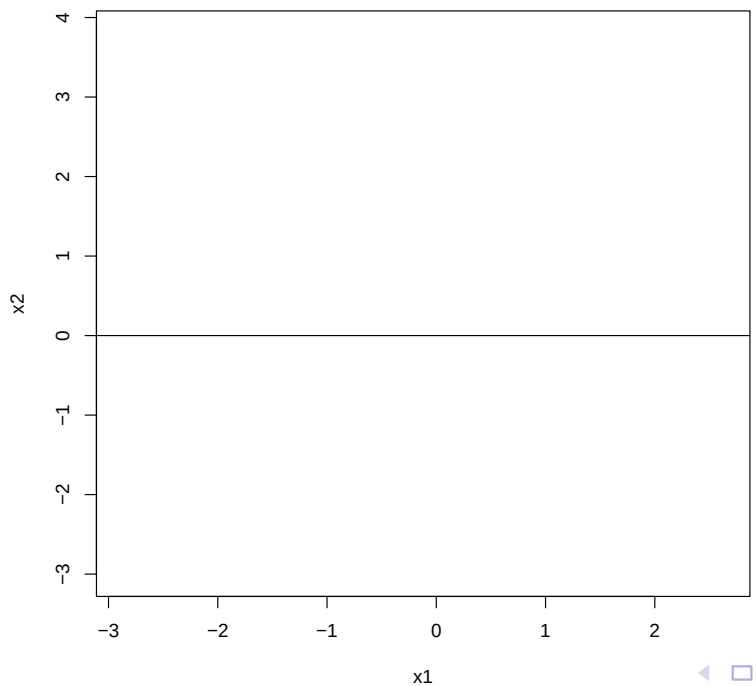
After controlling for multiple testing, we can reject H_0 if $|Z_{jk}|\sqrt{n-3} > Z(1-\alpha/(2M))$

Correlation graphs: Permutation test

A general tool in statistics is to apply random permutations in order to test hypotheses

The idea is that if H_0 is true, then the features are uncorrelated.

If indeed the features are uncorrelated, then permuting a feature shouldn't affect the sample correlation very much



Correlation graphs: Permutation test

The idea is as follows:

- Choose a number of permutations B (something like 10,000)
- Then, for each pair of features j, k, pick one of them and permute it B times
- For each permutation b, recompute the sample correlation: r_{jk}^b
- We can compute a p-value via

$$\text{p-value}_{jk} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}(|r_{jk}^b| \geq |r_{jk}|)$$

- Reject H_0 if $\text{p-value}_{jk} \leq \alpha / M$

Note that we are using $\mathbf{1}$ as an indicator function:

$$\mathbf{1}(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false} \end{cases}$$

Partial correlation graphs

Partial correlation

Another measure of 'relationship' between two features is partial correlation

The partial correlation between two features x_1 and x_2 is the amount they are associated after removing the effect of another feature x_3

Partial correlation

Let R be the matrix of (population) partial correlations of each feature, after removing the effect of all the other $p-2$ features

It turns out

$$R_{jk} = -\frac{\Omega_{jk}}{\sqrt{\Omega_{jj}\Omega_{kk}}}$$

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

We can estimate R by plugging in S^{-1} for Ω

Partial correlation: Gaussians

Using properties of multivariate Gaussian distributions, we know that x_j and x_k are independent of everything else if and only if $\Omega_{jk} = 0$

(Again, Ω is called the precision matrix)

So, we can apply the same ideas as correlation graphs, but to the matrix R and its estimate $\hat{R}$

Partial correlation: Bootstrap

For partial correlations, let's make confidence intervals via the bootstrap

We will just briefly discuss the bootstrap now

We will return to it later in the semester when we discuss (decision) trees

Partial correlation: Bootstrap

Suppose we are estimating a parameter θ with an estimator $\hat{\theta}$

There are typically two ways that we can create confidence intervals about θ

- By knowing the sampling distribution of $\hat{\theta}$
  (Example: $\overline{X} \sim N(\mu, \Sigma/n)$ if $X \sim N(\mu, \Sigma)$)
- By using an approximation to the distribution of $\hat{\theta}$
  (Example: $\overline{X} \to N(\mu, \Sigma/n)$ if X isn't Gaussian via CLT)

However, if n isn't large or an approximation isn't desirable/available there is a third option: the bootstrap

Partial correlation: Bootstrap confidence intervals

The idea is that we can sample with replacement from our data to get an idea of the sampling distribution of $\hat{\theta}$

Specifically, for $b = 1, \ldots, B$

1. Draw n observations from the sample with replacement
2. Compute $\hat{\theta}^b$

Now, find the $\alpha/2$ and $1 - \alpha/2$ quantiles of the values $\hat{\theta}^1, \ldots, \hat{\theta}^B$.

It turns out this interval has good theoretical properties and doesn't require distributional assumptions

(Small point: technically, this is called the 'percentile interval'. Another variation based on the 'empirical cumulative distribution function' has the good theoretical properties. Generally, these two approaches are quite similar in practice)

Partial correlation: Bootstrap confidence intervals

For our problem, we want to form bootstrap confidence intervals for the entries in $R \in \mathbb{R}^{p \times p}$

Specifically, for $b = 1, \ldots, B$

1. Draw n observations from $\mathbb{X}$ with replacement
2. Compute $\hat{R}^b$

Now, find the $\alpha/2$ and $1 - \alpha/2$ quantiles of the values $\hat{R}^1_{jk}, \ldots, \hat{R}^B_{jk}$.

Let's look at an example... partialCorrelation.html

Postamble:

- Define graphical models

  (A discrete math object that consists of nodes (or vertices) and edges between those nodes)
- Discuss estimating covariance graphs

  (We can test for particular covariance structures using the sample covariance matrix)
- Discuss estimating correlation graphs

  (We can either use an asymptotic test or use a permutation test)
- Discuss estimating partial correlation graphs

  (We discussed how the bootstrap could be used to get confidence intervals)