STA 437/2005: Methods for Multivariate Data

Week 10: Factor Analysis

Piotr Zwiernik

University of Toronto

Random vector and independence

Let (X, Y) be a vector of two random variables.

Joint distribution

Density function $f_{XY}(x, y)$ if continuous.

Probability mass function $f_{XY}(x, y) = \mathbb{P}(X = x, Y = y)$ if discrete.

Marginal distribution

continuous: $f_X(x) = \int_{\mathbb{R}} f_{XY}(x, y)\mathrm{d}y$.

discrete: $f_X(x) = \sum_y f_{XY}(x, y) = \mathbb{P}(X = x)$.

This can be generalized to random vectors.

Independence

If $f_{XY}(x, y)$ is the joint density (or PMF) of (X, Y) then X and Y are independent if and only if

$$f_{XY}(x, y) = f_X(x) f_Y(y) \qquad \text{for all } x, y.$$

We write $X \perp\!\!\!\perp Y$.

Recall:

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \quad \text{and} \quad \mathrm{var}(X) = \text{cov}(X, X).$$

The correlation $\rho_{X,Y}$ between X, Y is:

$$\rho_{X,Y} \; = \; \frac{\text{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} \; \in \; [-1, 1].$$

If $X \perp\!\!\!\perp Y$ then $\rho_{X,Y} = 0$.

(but in general not the other way around, see slide 17)

Conditional distribution

Conditional distribution

In the discrete case the conditional probability mass function is defined as

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X=x, Y=y)}{\mathbb{P}(Y=y)}$$

for all x, y such that $\mathbb{P}(Y = y) > 0$ and so

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} \text{ for all } x, y \text{ s.t. } f_Y(y) > 0.$$

In the continuous case we use the same definition.

Important reformulation of independence

$X \perp\!\!\!\perp Y$ if and only if $f_{X|Y}(x|y) = f_X(x)$.
(knowing Y brings no extra information about X)

A cautionary note

Note: $f_{X|Y}(x|y) \neq f_{Y|X}(y|x)$.

Example: A medical test for a disease D has outcomes $+$ and - with probabilities

	D	D^c
$+$	.009	.099
-	.001	.891

As needed $\mathbb{P}(+|D) = 0.9$ and $\mathbb{P}(-|D^c) = 0.9$. However, $\mathbb{P}(D|+) \approx 0.08$ (!)

Appendix: Conditional expectation

Let X, Y have joint distribution $f_{XY}(x,y)$ and conditional $f_{X|Y}(x|y)$. Then the conditional expectation $\mathbb{E}[X|Y]$ is the expectation of X with respect to the conditional distribution $X|Y=y$.

$$\mathbb{E}[X|Y=y] \;=\; \begin{cases} \sum_x x\, f_{X|Y}(x|y) \\ \int_{\mathbb{R}} x\, f_{X|Y}(x|y)\mathrm{d}x \end{cases} .$$

It is clear that $\mathbb{E}[g(Y)X|Y] = g(Y)\mathbb{E}[X|Y]$.

Note that $\mathbb{E}[X|Y]$ is a function of Y and so a random variable!

A powerful result states that

$$\mathbb{E}\big[\mathbb{E}(X|Y)\big] \;=\; \mathbb{E}(X).$$

Example: two binary variables

Suppose $f_{XY}(0,0) = 0.4$, $f_{XY}(0,1) = 0.2$, $f_{XY}(1,0) = 0.1$, $f_{XY}(1,1) = 0.3$. Then...

Conditional independence

X, Y, Z random variables.

X is independent of Y given Z (write $X \perp\!\!\!\perp Y|Z$) if

$$f_{XY|Z}(x, y|z) \;=\; f_{X|Z}(x|z) f_{Y|Z}(y|z) \qquad \text{for every } z.$$

We can define conditional expectation $\mathbb{E}(X|Z)$ and conditional correlation :

$$\rho_{X,Y|Z} \;:=\; \frac{\mathrm{cov}(X,Y|Z)}{\sqrt{\mathrm{var}(X|Z)\mathrm{var}(Y|Z)}}$$

Note that these are functions of Z!

If $X \perp\!\!\!\perp Y|Z$ then $\rho_{X,Y|Z} \equiv 0$
(but not the other way around)

Partial correlation

In practice we often work with the partial correlation:

$$\rho_{X,Y\cdot Z} \;:=\; \frac{\rho_{X,Y}-\rho_{X,Z}\rho_{X,Y}}{\sqrt{(1-\rho_{X,Z}^2)(1-\rho_{Y,Z}^2)}}.$$

We have $\rho_{X,Y\cdot Z} = 0$ if and only if in the **linear** regression of X on Y, Z the coefficient of Y is zero.

We have $\rho_{X,Y\cdot Z} = \rho_{X,Y|Z}$ in the case of Gaussian, elliptical, multinomial and Dirichlet distributions. But not in general.

Important: If $\boldsymbol{X} = (X_1, \dots, X_m)$ is a random vector with covariance matrix Σ, denote $K = \Sigma^{-1}$, then for each $i, j \in \{1, \dots, m\}$

$$\rho_{X_i,X_j\cdot X_{\{1,\dots,m\}\setminus\{i,j\}}} \;=\; -\frac{K_{ij}}{\sqrt{K_{ii}K_{jj}}}.$$

So normalizing the **concentration matrix** gives the partial correlations.

- $\rho_{X_i,X_j\cdot X_{\{1,\dots,m\}\setminus\{i,j\}}} = 0 \quad \Longleftrightarrow \quad K_{ij} = 0$
- $\rho_{X_i,X_j\cdot X_{\{1,\dots,m\}\setminus\{i,j\}}} \geq 0 \quad \Longleftrightarrow \quad K_{ij} \leq 0$

1.2 Testing independence

Recall: A statistical test

Given a statistical hypothesis $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$, a statistical test consists of a test statistics $T(X^{(1)}, \dots, X^{(n)})$ and a rejection region, typically of the form

$$R \;=\; \{T(X^{(1)}, \dots, X^{(n)}) > t\}.$$

If the null hypothesis is true T is unlikely to take large values.

Type I error: $\mathbb{P}(T \in R | H_0)$

although H_0 is true, it is rejected

Type II error: $\mathbb{P}(T \notin R | H_1)$

although H_0 is false, it is retained

A good test should minimize probabilities of both types of errors.

The idea is that T has some known distribution under H_0 so that we can compute the probability of $T \in R$ easily.

Testing independence

Data: $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} P_{X,Y}$.

Goal: Decide whether $X \perp\!\!\!\perp Y$.

Statistical test: $\qquad H_0 : X \perp\!\!\!\perp Y, \quad H_A : X \not\perp\!\!\!\perp Y$

There are many tests of independence.

We discuss some examples.

Test for vanishing correlation

Fisher's z-transform test for Gaussian data

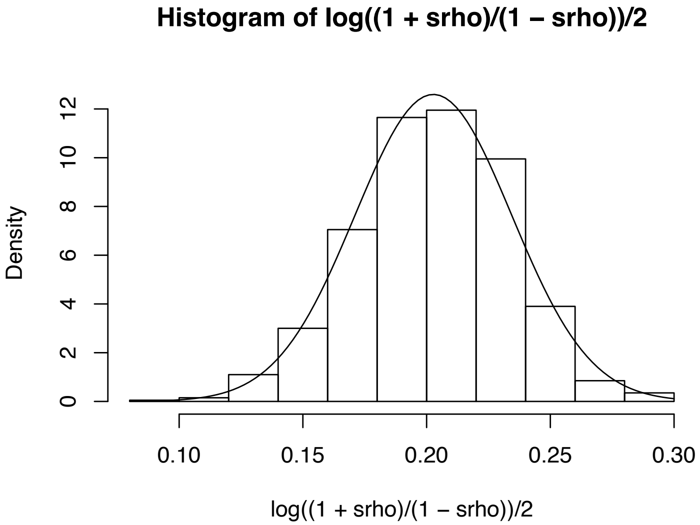
Let r_n is the sample correlation coefficient from an *iid* sample $(X^{(i)}, Y^{(i)})$.

Define $Z_n = \frac{1}{2}\log\left(\frac{1+r_n}{1-r_n}\right)$.

If (X, Y) is bivariate normal with correlation ρ then Z_n has asymptotically normal distribution with mean $\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$ and variance $\frac{1}{n-3}$.

```
# illustrate the normal approximation
library(MCMCpack); n <- 1000; iter <- 1000; rho <- 0.2
srho <- rep(0,iter)
for (i in 1:iter) {srho[i] <- cor(mvrnorm(n,c(0,0),matrix(c(1,rho
hist(log((1+srho)/(1-srho))/2,prob=TRUE,ylim=c(0,13))
curve(dnorm(x, mean=log((1+rho)/(1-rho))/2, sd=1/sqrt(n-3)), add=
```

Compare the sample distribution of Z_n with its theoretical asymptotic distribution.



Fisher's z-transform test is implemented in R as `cor.test`.

```
> set.seed(1); n <- 100; rho <- 0.2
> x <- mvrnorm(n,c(0,0),matrix(c(1,rho,rho,1),2,2))
> cor.test(x[,1], x[,2], method = "pearson")

^^IPearson's product-moment correlation

data:  x[, 1] and x[, 2]
t = 6.2913, df = 998, p-value = 4.704e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1349527 0.2542267
sample estimates:
      cor
0.1953118
```

Try the same with $\rho = 0$.

Non-gaussianity may invalidate the test and affect its power.

Basic nonparametric test

Kendall's tau test for non-Gaussian data

Suppose a bivariate sample (x_i, y_i) for $i = 1, \ldots, n$ is given.

Pair (x_i, y_i) , (x_j, y_j) is concordant if $(x_i, y_i) < (x_j, y_j)$ or $(x_i, y_i) > (x_j, y_j)$. Otherwise discordant.

Define $\tau_{XY} = \frac{(\#\text{concordant})-(\#\text{discordant})}{\binom{n}{2}} \in [-1, 1]$.

Test based on Kendell's τ statistic is implemented in R as `cor.test`.

```
> set.seed(1); n <- 200; rho <- 0.2; Z <- runif(n);
> X <- runif(n)^2+sqrt(rho)*Z; Y <- runif(n)+sqrt(rho)*Z
> cor.test(X, Y, method = "pearson")$p.value
[1] 0.03417231
> cor.test(X, Y, method = "kendall")$p.value
[1] 0.01100592
```

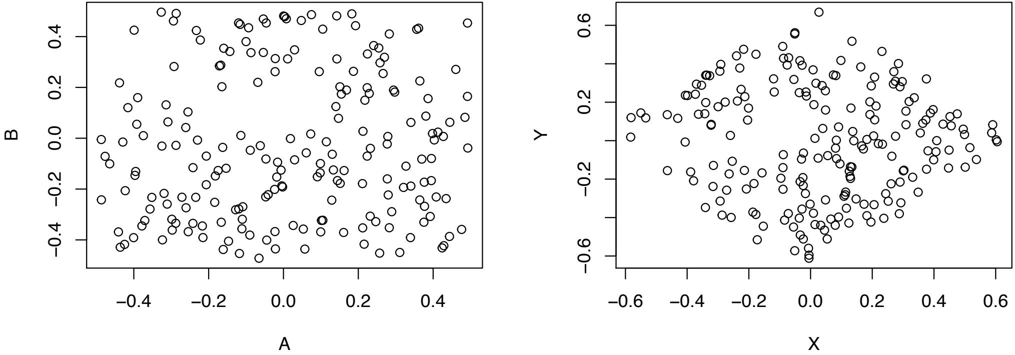
Non-Gaussianity issue

Vanishing covariance does not imply independence!

```
# generate sample from two uncorrelated but dependent random vari
> set.seed(1); n <- 200
> A <- runif(n)-1/2; B <- runif(n)-1/2
> X <- t(c(cos(pi/4),-sin(pi/4)) %*% rbind(A,B))
> Y <- t(c(sin(pi/4),cos(pi/4)) %*% rbind(A,B))
> cor.test(X,Y, method = "pearson")

 ^^IPearson's product-moment correlation

 data:  X and Y
 t = -0.84711, df = 198, p-value = 0.398
 alternative hypothesis: true correlation is not equal to 0
 95 percent confidence interval:
  -0.1971897  0.0793095
 sample estimates:
```



We see that X and Y are highly dependent.

Test based on distance correlation

Distance correlation $\mathcal{R}(X, Y)$ provides a test which applies when X, Y are two random **vectors** of any dimensions.

$\mathcal{R}(X, Y) = 0$ if and only if X and Y are independent.

The sample version of $\mathcal{R}(X, Y)$ gives a nonparametric test of independence.

This is also implemented in R in package `energy`.

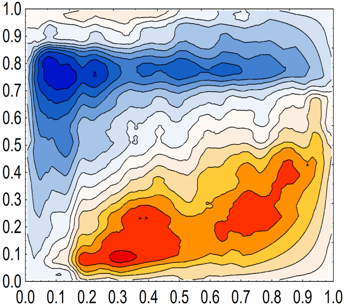
```
> library(energy); set.seed(1); n <- 200
> A <- runif(n)-1/2; B <- runif(n)-1/2
> X <- t(c(cos(pi/4),-sin(pi/4)) %*% rbind(A,B))
> Y <- t(c(sin(pi/4),cos(pi/4)) %*% rbind(A,B))
> dcor.test(X,Y,R=1000)

^^IdCor independence test (permutation test)

data:  index 1, replicates 1000
dCor = 0.21161, p-value = 0.004995
```

Another cautionary example

Bowman& Azzalini (1997) analyse aircraft wing span and speed data.



```
> library(sm); set.seed(1);
> X <- aircraft$Span
> Y <- aircraft$Speed
> cor.test(X,Y)$p.value
 [1] 0.7816014
> dcor.test(X,Y,R=1000)$p.value
```

Tests for discrete data

χ^2-test for discrete data

```
> M <- as.table(rbind(c(762, 327, 468), c(484, 239, 477)))
> dimnames(M) <- list(gender = c("F", "M"),
+                     party = c("Democrat","Independent", "Republ

> (Xsq <- chisq.test(M))  # Prints test summary

^^IPearsons Chi-squared test

data:  M
X-squared = 30.07, df = 2, p-value = 2.954e-07

> Xsq$expected   # expected counts under the null
      party
gender Democrat Independent Republican
```

Testing conditional independence

Testing conditional independence is hard in general.

For discrete data we have an asymptotic χ^2-test.

Some parametric tests are implemented in the library `bnlearn`.
Many non-parametric methods have been implemented in `CondIndTest`

```
> library(CondIndTests); library(bnlearn); set.seed(1); n <- 100
> Z <- rnorm(n); X <- 4 + 2 * Z + rnorm(n); Y <- 3 * X^2 + Z + rn
> CondIndTest(X,Y,Z, method = "KCI")$pvalue
[1] 2.419926e-10
> bnlearn::ci.test(X,Y,Z)$p.value
[1] 1.15458e-25
```

See Section 3 in: C. Heinze-Deml, J. Peters, N. Meinshausen, Invariant Causal Prediction for Nonlinear Models, Journal of Causal Inference, 2018.

See: `http://www.bnlearn.com/documentation/man/conditional.independence.tests.html`

Simpson’s paradox: UC Berkeley admissions example

The admission figures of the grad school at UC Berkeley in 1973: 8442 (44%) men, 4321 (35%) women admitted.

The same data conditioned on the department are:

2*Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

“Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.”

(Bickel et al, *Sex Bias in Graduate Admissions: Data From Berkeley*, Science, 1975)

In R:

```
> library(gRim); data(UCBAdmissions)
> bnlearn::ci.test(x = "Gender" , y = "Admit", z = "Dept", test="

^^IPearson's X^2


data:  Gender ~ Admit | Dept
x2 = 0, df = 6, p-value = 1
alternative hypothesis: true value is greater than 0

# gRim gives a slightly more refined output
> gRim::ciTest(as.data.frame(UCBAdmissions),set=~Gender+Admit+Dep

set: [1] "Gender" "Admit"  "Dept"
Testing Gender _|_ Admit | Dept
Statistic (DEV):    0.000 df: 6 p-value: 1.0000 method: CHISQ

Slice information:
```

Florida murderers

Sentences in 4863 murder cases in Florida over the six years 1973-78

	Sentence	
Murderer	Death	Other
Black	59	2547
White	72	2185

The table shows a greater proportion of white murderers receiving death sentence than black (3.2% vs. 2.3%).

```
> flor <- matrix(c(59,72,2547,2185),2,2)
> dimnames(flor) <- list(Murderer=c("Black","White"),Sentence=c("
> chisq.test(flor)

^^IPearson's Chi-squared test with Yates' continuity correction

data:  flor
X-squared = 3.6117, df = 1, p-value = 0.05737
# in the alternative we observe whites to be more often sentenced
```

Controlling for colour of victim

		Sentence	
Victim	Murderer	Death	Other
Black	Black	11	2309
	White	0	111
White	Black	48	238
	White	72	2074

Now the table for given colour of victim shows a very different picture.

In particular, note that 111 white murderers killed black victims and none were sentenced to death.

```
> flor <- c(11,48,0,72,2309,238,111,2074); dim(flor) <- c(2,2,2)
> dimnames(flor) <- list(Victim=c("Black","White"),Murderer=c("Bl
> ciTest_table(flor,set=~Sentence+Murderer+Victim)
 Testing Sentence _|_ Murderer | Victim
 Statistic (DEV):   67.980 df: 2 p-value: 0.0000 method: CHISQ
```

1.3 Multivariate normal distribution

Recall: The density function

Multivariate normal distribution, $X = (X_1, \dots, X_m)$

Let $\mu \in \mathbb{R}^m$ and Σ symmetric positive definite $m \times m$ matrix. We write $X \sim N_m(\mu, \Sigma)$ if the density of the vector X is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{m/2}} (\det \Sigma)^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu)^T \Sigma^{-1} (\boldsymbol{x} - \mu)\right).$$

If $Z \sim N_m(\mathbf{0}_m, \mathbb{I}_p)$ and $X = \mu + \Sigma^{1/2} Z$ then $X \sim N_m(\mu, \Sigma)$.
If $X \sim N_m(\mu, \Sigma)$ then $\Sigma^{-1/2}(X - \mu) \sim N_m(\mathbf{0}_m, \mathbb{I}_m)$.

To sample from $N_m(\mu, \Sigma)$ you can use `mvrnorm()` in package `MCMCpack` or

```
library(expm)
m <- 3; mu <- c(0,0,0); Sigma <- matrix(c(1,1/3,1/3,1/3,1,1/3,1/3
n <- 1000; Z <- matrix(rnorm(m*n),m,n)
X <- mu + sqrtm(Sigma)%*%Z # sqrtm() not sqrt()!
# run cov(t(X)) to check the result
```

Recall: Marginal and conditional distributions

Split X into two blocks $X = (X_A, X_B)$. Denote

$$\mu = (\mu_A, \mu_B) \qquad \text{and} \qquad \Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}.$$

Marginal distribution

$X_A \sim N_{|A|}(\mu_A, \Sigma_{AA})$

Conditional distribution

$X_A | X_B = x_B \sim N_{|A|} \left(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} \right)$

- Note that the conditional covariance is constant.

Some other properties

Linear transformations

If $A \in \mathbb{R}^{m \times p}$ for $m \leq p$ and $X \sim N_p(\mu, \Sigma)$ then $AX \sim N_m(A\mu, A\Sigma A^T)$.

Moments, $X \sim N_m(\mu, \Sigma)$, $\mathbb{E}X = \mu$, $\mathrm{var}(X) = \Sigma$

Other moments are determined by the expectation and the covariance matrix.

Independence and conditional independence

$X_i \perp\!\!\!\perp X_j$ if and only if $\Sigma_{ij} = 0$.

$X_i \perp\!\!\!\perp X_j | X_C$ if and only if $\Sigma_{ij} - \Sigma_{i,C}\Sigma_{C,C}^{-1}\Sigma_{C,j} = 0$

Let $R = V \setminus \{i, j\}$. The following are equivalent:

- $X_i \perp\!\!\!\perp X_j | X_R$
- $\Sigma_{ij} - \Sigma_{i,R}\Sigma_{R,R}^{-1}\Sigma_{R,j} = 0$
- $(\Sigma^{-1})_{ij} = 0$

Modelling with Gaussian distributions

The m-dimensional Gaussian distribution has $m + \binom{m+1}{2}$ free parameters, which is prohibitive in high dimensions, and not very interesting from the modelling point of view.

Structured covariance

Toeplitz, graphical models, linear covariance/concentration models, etc.

Some further limitations

In practice distributions with heavier tails may be more suitable in certain applications.

The Gaussian distribution is unimodal and very symmetric (around its mean). This is sometimes too limiting.