

# Matrix algebras and fit indexes

Sören Möller  
moeller@health.sdu.dk

Epidemiology, Biostatistics and Biodemography, Institute of Public Health,  
University of Southern Denmark, Denmark

June 2014

# What is a matrix?

## Matrices

From model  
to matrix

Fit indices

- A rectangular array of numbers
- Usually enclosed in some sort of brackets
- With specified number of rows and columns (**dimensions**)

$$A = {}_2A_2 = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

$$B = {}_3B_2 = \begin{pmatrix} 1 & 2 & -5 & -5 \\ 3 & 4 & 4 & 2 \\ 1.2 & 1.1 & -5 & -5 \end{pmatrix}$$

$$C = {}_1C_3 = (1 \quad 2 \quad -5)$$

## ■ Data matrices

- Rows are subjects/observations, columns are measures/variables

$${}_nA_m = (a_{ij})_{i=1,\dots,n,j=1,\dots,m}$$

- $n$  number of subjects/observations
- $m$  number of measures/variables

## ■ Variance-covariance matrices

- Describe variance/covariance between variables of a model
- **Correlation matrices** are normalized by variances

## ■ Model parameter matrices

- Entries are parameters of a model
- Useful for concise representation of models
- Can make model calculations very efficient (for computers)

# Example: Data matrix

## Matrices

From model  
to matrix

Fit indices

| ##    | pair   | Sex  | Zyg | bmi      | age      | twin |
|-------|--------|------|-----|----------|----------|------|
| ## 1  | 100001 | male | DZ  | 26.33289 | 57.57974 | 1    |
| ## 2  | 100001 | male | DZ  | 25.46939 | 57.57974 | 2    |
| ## 3  | 100002 | male | MZ  | 28.65014 | 57.04860 | 1    |
| ## 5  | 100003 | male | DZ  | 28.40909 | 57.67830 | 1    |
| ## 7  | 100004 | male | DZ  | 27.25089 | 53.51677 | 1    |
| ## 8  | 100004 | male | DZ  | 28.07504 | 53.51677 | 2    |
| ## 9  | 100005 | male | DZ  | 27.77778 | 52.57495 | 1    |
| ## 11 | 100006 | male | DZ  | 28.04282 | 52.57221 | 1    |
| ## 12 | 100006 | male | DZ  | 22.30936 | 52.57221 | 2    |
| ## 13 | 100007 | male | DZ  | 28.06642 | 52.49007 | 1    |

# Example: Covariance and correlation matrix

## Matrices

From model  
to matrix

Fit indices

```
##          bmi1      age1      bmi2      age2
## bmi1 13.140030  7.723399  6.155991  7.723399
## age1  7.723399 60.949171  7.589034 60.949171
## bmi2  6.155991  7.589034 12.662559  7.589034
## age2  7.723399 60.949171  7.589034 60.949171
```

```
##          bmi1      age1      bmi2      age2
## bmi1 1.0000000 0.2729144 0.4772424 0.2729144
## age1 0.2729144 1.0000000 0.2731756 1.0000000
## bmi2 0.4772424 0.2731756 1.0000000 0.2731756
## age2 0.2729144 1.0000000 0.2731756 1.0000000
```

# Example: Model parameter matrices

## Matrices

From model  
to matrix

Fit indices

For a twin model we could assume the covariance matrices to be

$$\text{Cov}(\epsilon_1, \epsilon_2)_{MZ} = \begin{pmatrix} \sigma_1^2 & \rho_{MZ}\sigma_1\sigma_2 \\ \rho_{MZ}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$$\text{Cov}(\epsilon_1, \epsilon_2)_{DZ} = \begin{pmatrix} \sigma_1^2 & \rho_{DZ}\sigma_1\sigma_2 \\ \rho_{DZ}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

with parameters  $\sigma_1, \sigma_2, \rho_{MZ}$  and  $\rho_{DZ}$ .

# Matrix operations

## Matrices

From model  
to matrix

Fit indices

- Matrices can be added, subtracted and multiplied by numbers entrywise, if they have the same dimensions

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 3 & 1 \\ 5 & 2 \end{pmatrix} = \begin{pmatrix} 1+3 & 2+1 \\ 3+5 & 4+2 \end{pmatrix} = \begin{pmatrix} 4 & 3 \\ 8 & 6 \end{pmatrix}$$

- Matrices can be multiplied, if the left matrix has the same number of columns as the right one has rows

$$({}_m A_n \cdot {}_n B_p)_{i,j} = \sum_k^n a_{ik} \cdot b_{kj}$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} 3 & 1 \\ 5 & 2 \end{pmatrix} = \begin{pmatrix} 1 \cdot 3 + 2 \cdot 5 & 1 \cdot 1 + 2 \cdot 2 \\ 3 \cdot 3 + 4 \cdot 5 & 3 \cdot 1 + 4 \cdot 2 \end{pmatrix} \\ = \begin{pmatrix} 13 & 5 \\ 29 & 11 \end{pmatrix}$$

## Matrices

From model  
to matrix

Fit indices

- Identity matrix (1 on diagonal, 0 everywhere else)

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- Zero matrix (0 everywhere)

$$O = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

# Exercise

Assume our model is

$$Y_i = \mu_i + \epsilon_i, \quad i = 1, 2$$

$$\text{Cov}(\epsilon_1, \epsilon_2)_{MZ} = \begin{pmatrix} \sigma_1^2 & \rho_{MZ}\sigma_1\sigma_2 \\ \rho_{MZ}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$$\text{Cov}(\epsilon_1, \epsilon_2)_{DZ} = \begin{pmatrix} \sigma_1^2 & \rho_{DZ}\sigma_1\sigma_2 \\ \rho_{DZ}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

and we observe

$$\text{Cov}(\epsilon_1, \epsilon_2)_{MZ} = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}$$

$$\text{Cov}(\epsilon_1, \epsilon_2)_{DZ} = \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$$

Determine the values of  $\rho_{MZ}$  and  $\rho_{DZ}$ .

# Solution

## Matrices

From model  
to matrix

Fit indices

- From the diagonals we must have  $\sigma_1^2 = \sigma_2^2 = 4$  for both MZ and DZ
- Hence  $\sigma_1 = \sigma_2 = 2$  for both MZ and DZ
- So  $\rho_{MZ} = \frac{3}{2 \cdot 2} = \frac{3}{4}$
- So  $\rho_{DZ} = \frac{1}{2 \cdot 2} = \frac{1}{4}$
- This indicates some genetic effect as  $\rho_{MZ} > \rho_{DZ}$
- In reality these numbers would not be so nice, and we would have to find statistical optimal solutions, instead of exact solutions.

# From model to matrix

Matrices

From model  
to matrix

Fit indices

- We often specify models by path diagrams

- Rectangular boxes are **observed** variables

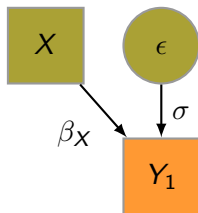


- Circles are unobserved **latent** variables



- Twoheaded arrows indicate correlations, numbers on arrows specify the coefficient of linear correlation
- Oneheaded arrows indicate assumed direction of causality

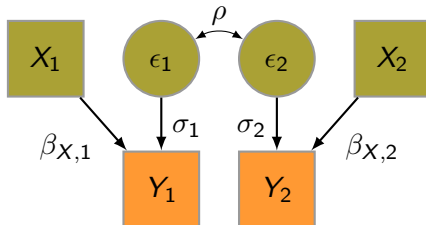
# Example: Univariate linear model



$$Y = \beta_X \cdot X + \epsilon$$

$$\text{Cov}(\epsilon, \epsilon) = \sigma^2$$

# Example: Simple twin model



$$Y_i = \beta_{X,i} \cdot X_i + \epsilon_i$$

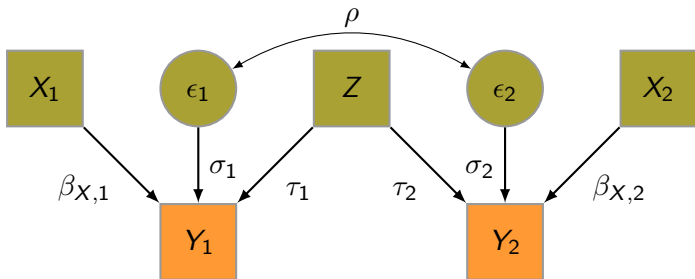
$$\text{Cov}(\epsilon_1, \epsilon_2) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

# Exercise

Matrices

From model  
to matrix

Fit indices



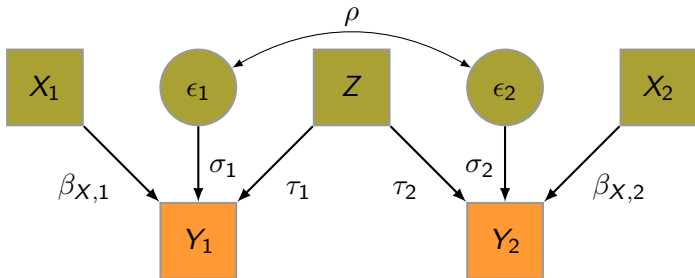
Specify this model as matrix equations

# Solution

Matrices

From model  
to matrix

Fit indices



$$Y_i = \beta_{X,i} \cdot X_i + \tau_i Z + \epsilon_i$$

$$\text{Cov}(\epsilon_1, \epsilon_2) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

# What is a fit index?

Matrices

From model  
to matrix

Fit indices

- A fit index is a measure for how good a model fits to the data
- We use it to compare different models fits to **the same** data
- We can normally not compare fit to different data sets
- In some cases we can not only determine, which model fits better, but also if it fits **significantly** better
- Fit indices can be utilized to choose the appropriate model
- Finding the right balance between being simple and describing the observed data well

# (Log-)likelihood

Matrices

From model  
to matrix

Fit indices

The likelihood of model parameters  $\theta$  if we observed the data  $x$  is given by

$$L(\theta | x) = P(x | \theta),$$

where  $P(x | \theta)$  is the probability of observing  $x$  if the true parameters are  $\theta$ . A higher likelihood indicates a better fit. Many statistical algorithms try to maximize the likelihood. For practical reasons we often consider the log-likelihood

$$\log L(\theta | x) = \log(L(\theta | x))$$

instead. A higher log-likelihood indicates a better fit. The (log)-likelihood is only useful for comparing nested models.

# Example: Log-likelihood

Matrices

From model  
to matrix

Fit indices

```
fitAge=twinlm(bmi~age,data=d,id="pair",zyg="Zyg")  
fitAgeSex=twinlm(bmi~age+Sex,data=d,id="pair",zyg="Zyg")
```

```
logLik(fitAge)
```

```
## 'log Lik.' -22138.46 (df=5)
```

```
logLik(fitAgeSex)
```

```
## 'log Lik.' -22019.54 (df=6)
```

The higher likelihood of `fitAgeSex` indicates that this model fits the data better.

# Testing for model differences

- We can test if one model fits the data **significantly** better than another by using log-likelihood

$$-2\log L_1 - (-2\log L_2) \sim \chi_j^2$$

where  $j$  is the number of additional parameters in model 1

- This only works for nested models!

```
lrtest(fitAgeSex,fitAge)

## Likelihood ratio test
##
## Model 1: bmi ~ age + Sex
## Model 2: bmi ~ age
##      #Df LogLik Df  Chisq Pr(>Chisq)
## 1      6 -22020
## 2      5 -22138 -1  237.83 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# AIC (Akaike information criterion)

Matrices

From model  
to matrix

Fit indices

- AIC is defined by

$$AIC = 2 \cdot k - 2 \cdot \log L$$

where  $k$  is the number of free parameters in the model

- Smaller AIC indicates a better fitting model
- AIC penalizes overfitting data with too many parameters
- AIC can be used to compare non-nested models
- Hence very useful when comparing different path models
- But we **can not test** for significance of the difference

# Example: AIC

```
fitSex=twinlm(bmi~Sex,data=d,id="pair",zyg="Zyg")
```

Matrices

From model  
to matrix

Fit indices

```
AIC(fitAge)
```

```
## [1] 44286.91
```

```
AIC(fitSex)
```

```
## [1] 44477.94
```

```
AIC(fitAgeSex)
```

```
## [1] 44051.08
```

The lower AIC of `fitAgeSex` indicates that this model fits the data best.

# BIC (Bayesian information criterion)

- BIC is defined by

$$BIC = k \cdot \log(n) - 2 \cdot \log L$$

where  $k$  is the number of free parameters in the model and  $n$  is the number of observations.

- Penalizes heavier for overfitting than AIC
- Motivated by good predictive properties of models
- Useful in high-dimensional (i.e. genetic) data
- We mainly use AIC in twin studies

```
BIC(fitAge)
```

```
## [1] 44318.71
```

```
BIC(fitSex)
```

```
## [1] 44509.74
```