Analyzing (complex) systems with Structural Equation Modelling

Mathieu Emily

16 mars 2021
NetBio

Examples of the use of SEM

- Economics, Social Science, Psychology
  - Structural equation models and the **quantification of behavior** (Bollen *et al.*, 2011)

Examples of the use of SEM

- Economics, Social Science, Psychology
  - Structural equation models and the **quantification of behavior** (Bollen *et al.*, 2011)
- Ecology
  - Structural Equation Modeling and Natural Systems (Grace, 2009)
  - Applications of structural equation modeling in ecological studies (Fan, 2016)

Examples of the use of SEM

- Economics, Social Science, Psychology
  - Structural equation models and the **quantification of behavior** (Bollen *et al.*, 2011)
- Ecology
  - Structural Equation Modeling and Natural Systems (Grace, 2009)
  - Applications of structural equation modeling in ecological studies (Fan, 2016)
- Medicine and Genomics
  - Structural equation models for **pathway identification** (Xiong, 2001)
  - Application of Structural Equation Models to **GWAS** (Kim *et al.*, 2010)
  - The mediating effects of **public genomic knowledge** in precision medicine implementation: A structural equation model approach (Mogaka and Chimbari, 2020)
  - Bayesian structural equation modeling in **multiple omics data** (Maity, 2020)

Examples of the use of SEM

- Economics, Social Science, Psychology
  - Structural equation models and the **quantification of behavior** (Bollen *et al.*, 2011)
- Ecology
  - Structural Equation Modeling and Natural Systems (Grace, 2009)
  - Applications of structural equation modeling in ecological studies (Fan, 2016)
- Medicine and Genomics
  - Structural equation models for **pathway identification** (Xiong, 2001)
  - Application of Structural Equation Models to **GWAS** (Kim *et al.*, 2010)
  - The mediating effects of **public genomic knowledge** in precision medicine implementation: A structural equation model approach (Mogaka and Chimbari, 2020)
  - Bayesian structural equation modeling in **multiple omics data** (Maity, 2020)
  - A comparison of methods for **inferring causal relationships between genotype and phenotype** using additional biological measurements (Ainsworth *et al.*, 2017)

Examples of the use of SEM

- Economics, Social Science, Psychology
  - Structural equation models and the **quantification of behavior** (Bollen *et al.*, 2011)
- Ecology
  - Structural Equation Modeling and Natural Systems (Grace, 2009)
  - Applications of structural equation modeling in ecological studies (Fan, 2016)
- Medicine and Genomics
  - Structural equation models for **pathway identification** (Xiong, 2001)
  - Application of Structural Equation Models to **GWAS** (Kim *et al.*, 2010)
  - The mediating effects of **public genomic knowledge** in precision medicine implementation: A structural equation model approach (Mogaka and Chimbari, 2020)
  - Bayesian structural equation modeling in **multiple omics data** (Maity, 2020)
  - A comparison of methods for **inferring causal relationships between genotype and phenotype** using additional biological measurements (Ainsworth *et al.*, 2017)

B. Shipley, *Cause and correlation in Biology*, 2016

- SEM is a tool for modeling a **global system**
- SEM is one of the most **popular tool for investigating causality**

Outline

1. From Linear model to Path model
2. Latent variables
3. Model
4. SEM and Explanatory Factor Analysis
5. Ending words

Outline

1. From Linear model to Path model
2. Latent variables
3. Model
4. SEM and Explanatory Factor Analysis
5. Ending words

Introductive example : Electroencephalography for Alzheimer's patients

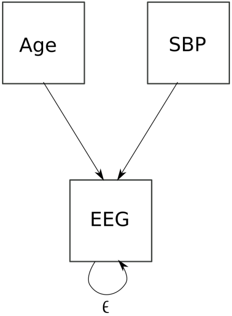
Multiple linear regression

- Three variables: z-scores for brain rate in the frontal region (=*EEG*), *Age* and Systolic Blood Pressure (*SBP*)
- Linear regression
  - $EEG = \beta_0 + \beta_1 Age + \beta_2 SBP + \varepsilon$
  - Coefficients (β_0, β_1 and β_2) are estimated by minimizing the residual variance $\sum(EEG - EEG_{Mod})^2$
- From a **system** point-of-view
  - *Age* and *SBP* values are determined outside the model and are imposed on the model (=**Exogeneous** variables)
  - *EEG* values are determined by the model (**=Endogeneous** variable)

Introductive example : Electroencephalography for Alzheimer's patients DAG visualisation

- Visualisation using a **Directed Acyclic Graph (DAG)**

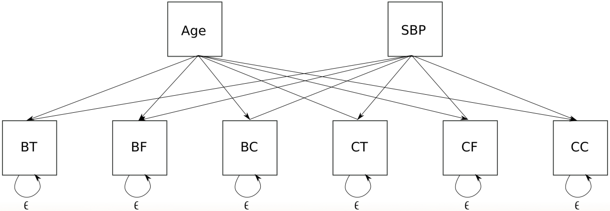
$$EEG = \beta_0 + \beta_1 Age + \beta_2 SBP + \varepsilon$$



Introductive example : Electroencephalography for Alzheimer's patients

Multivariate regression

- **6 measures for EEG**: 3 regions (frontal, temporal, central) and 2 features (brain rate, complexity)
- **Multivariate** regression ($\sim$ Manova)
  - Basics for the estimation: minimizing the distance between the observed covariance for "response" variables and the model covariance
- DAG for a multivariate regression model

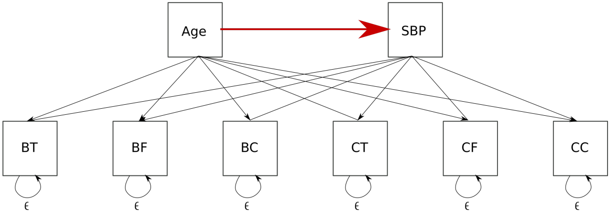


Introductive example : Electroencephalography for Alzheimer's patients Path modeling (1)

- *"An increase in (systolic)* **blood pressure** *has always been taken as an inevitable consequence of* **ageing**" (Pinto, 2007)
- How can we modify the modeling of the system?

Introductive example : Electroencephalography for Alzheimer's patients Path modeling (1)

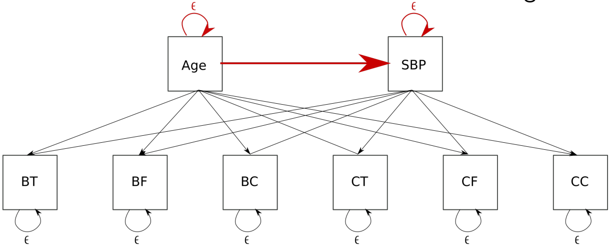
- *"An increase in (systolic)* **blood pressure** *has always been taken as an inevitable consequence of* **ageing**" (Pinto, 2007)
- How can we modify the modeling of the system?



- SBP is now an endogeneous variable

Introductive example : Electroencephalography for Alzheimer's patients Path modeling (2)

- **Measurement error** is also accounted for SBP and Age



Paradigm shift

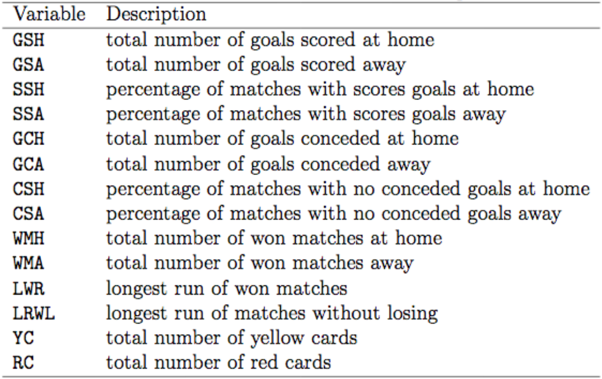
- In **path modeling**, all observed variables in the system are considered in the estimation of the model
- The aim is to model the covariance matrix

Outline

1. From Linear model to Path model
2. Latent variables
3. Model
4. SEM and Explanatory Factor Analysis
5. Ending words

Football example

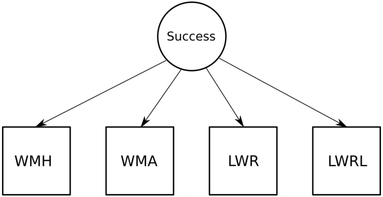
- How to define a strategy of **success**?
- Data obtained from all teams in an entire season.



Variable	Description
GSH	total number of goals scored at home
GSA	total number of goals scored away
SSH	percentage of matches with scores goals at home
SSA	percentage of matches with scores goals away
GCH	total number of goals conceded at home
GCA	total number of goals conceded away
CSH	percentage of matches with no conceded goals at home
CSA	percentage of matches with no conceded goals away
WMH	total number of won matches at home
WMA	total number of won matches away
LWR	longest run of won matches
LRWL	longest run of matches without losing
YC	total number of yellow cards
RC	total number of red cards

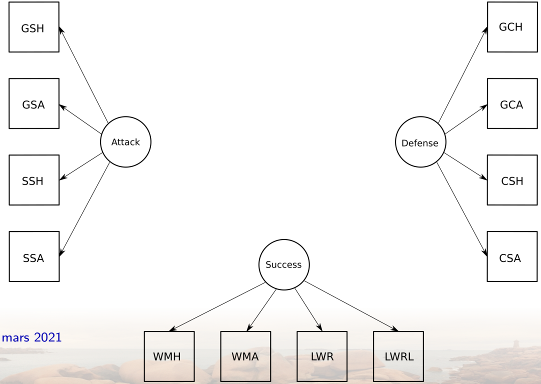
Football example
The concept of Success

- Success is easy to observe/measure but understanding how to achieve success is more complicated
  - Attack strategy
  - Defense strategy
  - Adapt to the opponent
- 4 variables are related to **concept the success**: WMH, WMA, LWR and LRWL



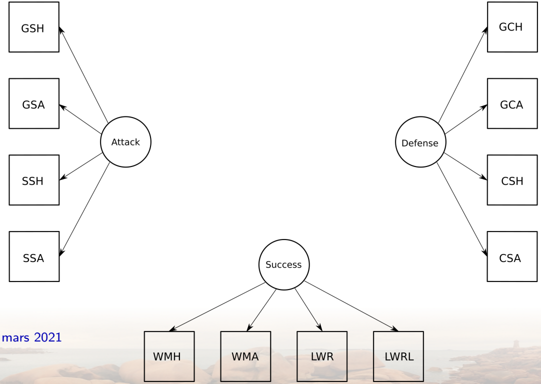
Football example
Latent modeling

- Similarly, the **concepts of Attack** and **Defense** can be modeled as:
  - Attack: GSH, GSA, SSH and SSA
  - Defense: GCH, GCA, CSH and CSA



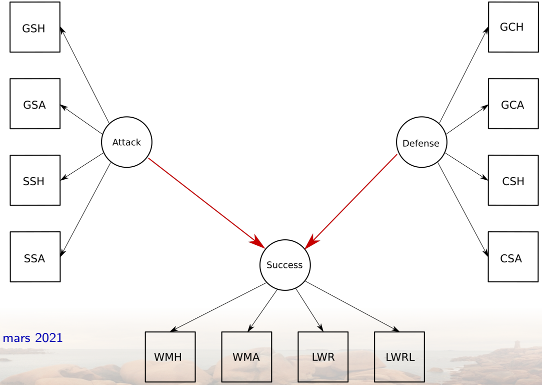
Football example
Latent modeling

- Similarly, the **concepts of Attack** and **Defense** can be modeled as:
  - Attack: GSH, GSA, SSH and SSA
  - Defense: GCH, GCA, CSH and CSA
- How to **link observed** and/or **latent** variables?



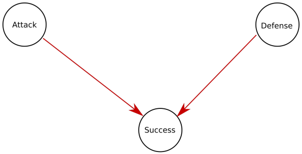
Football example
Latent modeling

- Similarly, the **concepts of Attack** and **Defense** can be modeled as:
  - Attack: GSH, GSA, SSH and SSA
  - Defense: GCH, GCA, CSH and CSA
- How to **link observed** and/or **latent** variables?

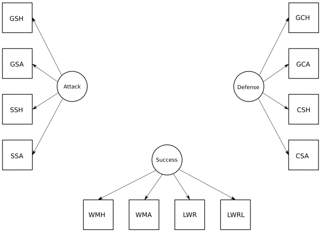


Structural model

- A structural model is made by 2 models:



Latent model



Measurement model

- Each arrow is a **linear** link between variables:
  - $Success = f(Attack, Defense) = \beta_1 Attack + \beta_2 Defense + \varepsilon$
  - $GSH = f(Attack) = \gamma_1 Attack + \varepsilon$
  - ...
- Remark: Success is an endogeneous latent variable while Attack and Defense are two exogeneous latent variables.

Outline

1. From Linear model to Path model
2. Latent variables
3. Model
4. SEM and Explanatory Factor Analysis
5. Ending words

Outline

3 Model

General definition

Identification rules

Estimation and tests

Latent model

- Let consider a model with m endogeneous latent variables and n exogeneous variables

$$\eta = \mathbf{B}\eta + \mathbf{\Gamma}\xi + \zeta$$

  - $\mathbf{B}$ is a $m \times m$ matrix of coefficients for latent endogeneous variables
  - $\mathbf{\Gamma}$ is a $m \times n$ matrix of coefficients for latent exogeneous variables
  - $\Phi = \mathbb{E}[\xi\xi']$ is a $n \times n$ covariance matrix for ξ
  - $\Psi = \mathbb{E}[\zeta\zeta']$ is a $m \times m$ covariance matrix for ζ
- Assumptions:
  - $\mathbb{E}[\eta] = 0$
  - $\mathbb{E}[\xi] = 0$
  - $\mathbb{E}[\zeta] = 0$
  - $Cov(\zeta, \xi) = 0$
  - $(I - B)$ nonsingular

Measurement model

- Let consider a model with p endogeneous observed variables and q exogeneous observed variables

$$\mathbf{x} = \mathbf{\Lambda_x}\xi + \delta$$
$$\mathbf{y} = \mathbf{\Lambda_y}\eta + \varepsilon$$

  - $\mathbf{\Lambda_x}$ is a $q \times n$ matrix of coefficients relating x to ξ
  - $\mathbf{\Lambda_y}$ is a $p \times m$ matrix of coefficients relating y to η
  - $\Theta_\delta = \mathbb{E}[\delta\delta']$ is a $q \times q$ covariance matrix for δ
  - $\Theta_\varepsilon = \mathbb{E}[\varepsilon\varepsilon']$ is a $p \times p$ covariance matrix for ε
- Assumptions:
  - $\mathbb{E}[\delta] = 0$
  - $\mathbb{E}[\varepsilon] = 0$
  - $Cov(\delta, \varepsilon) = 0$
  - $Cov(\delta, \zeta) = 0$ and $Cov(\delta, \xi) = 0$
  - $Cov(\varepsilon, \zeta) = 0$ and $Cov(\varepsilon, \xi) = 0$

Toy example of prostate cancer

Observed variables:

- Gleason score from biopsy
- PSA test from a blood sample
- HPC1 (hereditary prostate cancer 1) expression
- PcaP (predisposing for prostate cancer) expression
- PG1 (prostate cancer susceptibility gene 1) expression
- BMI
- Exposure to pollution
- Age

Toy example of prostate cancer

Observed variables:

- Gleason score from biopsy
- PSA test from a blood sample

Cancer measures

- HPC1 expression
- PcaP expression
- PG1 expression

Genetic measures

- BMI
- Exposure to pollution
- Age

Environnemental measures

Toy example of prostate cancer

$$\mathbf{B} = \begin{bmatrix} 0 \end{bmatrix}$$

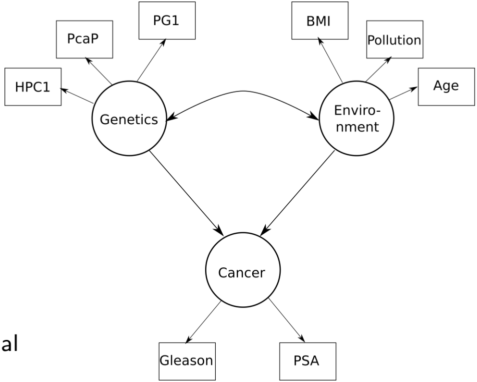
$$\mathbf{\Gamma} = \begin{bmatrix} \beta_{11} \\ \beta_{21} \end{bmatrix}$$

$$\mathbf{\Lambda_x} = \begin{bmatrix} \lambda^x_{11} & 0 \\ \lambda^x_{21} & 0 \\ \lambda^x_{31} & 0 \\ 0 & \lambda^x_{12} \\ 0 & \lambda^x_{22} \\ 0 & \lambda^x_{32} \end{bmatrix}$$

$$\mathbf{\Lambda_y} = \begin{bmatrix} \lambda^y_{11} \\ \lambda^y_{21} \end{bmatrix}$$

$$\mathbf{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}$$

Ψ, Θ_δ and Θ_ε are diagonal



Covariance implied by the model

- Examples

$$\begin{aligned}
Cov(HPC1, PSA) &= Cov(\lambda^x_{11} Genetics + \delta_{11}, \lambda^y_{21} Cancer + \varepsilon_2) \\
&= \lambda^x_{11}\lambda^y_{21} Cov(Genetics, Cancer) \\
&= \lambda^x_{11}\lambda^y_{21} Cov(Genetics, \beta_{11} Genetics + \beta_{21} Environ. + \zeta_1) \\
&= \lambda^x_{11}\lambda^y_{21}\beta_{11}\phi_{11} + \lambda^x_{11}\lambda^y_{21}\beta_{21}\phi_{12}
\end{aligned}$$

$$\begin{aligned}
Cov(HPC1, PG1) &= Cov(\lambda^x_{11} Genetics + \delta_{11}, \lambda^x_{31} Genetics + \delta_{31}) \\
&= \lambda^x_{11}\lambda^x_{31}\phi_{11}
\end{aligned}$$

Covariance implied by the model

- Examples

$$\begin{aligned}
Cov(HPC1, PSA) &= Cov(\lambda^x_{11} Genetics + \delta_{11}, \lambda^y_{21} Cancer + \varepsilon_2) \\
&= \lambda^x_{11}\lambda^y_{21} Cov(Genetics, Cancer) \\
&= \lambda^x_{11}\lambda^y_{21} Cov(Genetics, \beta_{11} Genetics + \beta_{21} Environ. + \zeta_1) \\
&= \lambda^x_{11}\lambda^y_{21}\beta_{11}\phi_{11} + \lambda^x_{11}\lambda^y_{21}\beta_{21}\phi_{12}
\end{aligned}$$

$$\begin{aligned}
Cov(HPC1, PG1) &= Cov(\lambda^x_{11} Genetics + \delta_{11}, \lambda^x_{31} Genetics + \delta_{31}) \\
&= \lambda^x_{11}\lambda^x_{31}\phi_{11}
\end{aligned}$$

- Similarly, all covariances can be obtained thus leading to the **implied covariance** $\mathbf{\Sigma}(\theta)$ where θ is the set of unknown parameters of the model

Estimation principle

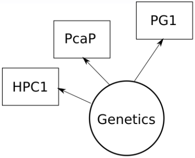
- Choosing θ for $\mathbf{\Sigma}(\theta)$ to be as close to S as possible

Outline

3 Model

- General definition
- Identification rules
- Estimation and tests

Issue with identification



- θ is **identified** if $\nexists\ \theta_1$ and θ_2 such as $\mathbf{\Sigma}(\theta_1) = \mathbf{\Sigma}(\theta_2)$
- Example:

	HPC1	PcaP	PG1
HPC1	$(\lambda^x_{11})^2\phi_{11} + \Theta^\delta_{11}$		
PcaP	$\lambda^x_{11}\lambda^x_{21}\phi_{11}$	$(\lambda^x_{21})^2\phi_{11} + \Theta^\delta_{22}$	
PG1	$\lambda^x_{11}\lambda^x_{31}\phi_{11}$	$\lambda^x_{21}\lambda^x_{31}\phi_{11}$	$(\lambda^x_{31})^2\phi_{11} + \Theta^\delta_{33}$

- 7 parameters for only 6 observations: a **need for constraint**
  - Set the variance of the latent variable to 1 ($\phi_{11} = 1$)
  - Set $\lambda^x_{11} = 1$ to scale the *Genetics* to $HPC1$
  - Set $\lambda^x_{11} = \lambda^x_{21} = \lambda^x_{31}$ to balance the amount of variance/covariance in the latent space ($\tau-$equivalence)

Conditions for identification (Bollen, 1989)

- **The** *$t-rule$*

$$t \leq \frac{(p+q)(p+q+1)}{2}$$

  where t is the number of free parameters in θ

  - A necessary but not sufficient condition ($t = 19$ in the general prostate model with $p+q=8$ observed variables)

- **Two-Step rules**
  - Step 1 : Consider y and η as exogeneous variables (CFA)
    - Three-indicator rule
    - **Two-indicator rule**
  - Step 2 : Consider the identification as the latent model (as a measurement model)
  - A sufficient condition

- **MIMIC rule** (for Multiple Indicators and MultIple Causes model)

Outline

3 Model

General definition

Identification rules

Estimation and tests

Estimation

The **closeness** of $\boldsymbol{\Sigma}(\theta)$ to S is measured by fitting functions $F(S, \boldsymbol{\Sigma}(\theta))$ (with $F \geq 0$ and $F = 0$ iif $\boldsymbol{\Sigma}(\theta) = S$)

Estimation

The **closeness** of $\mathbf{\Sigma}(\theta)$ to S is measured by fitting functions $F(S, \mathbf{\Sigma}(\theta))$ (with $F \geq 0$ and $F = 0$ iif $\mathbf{\Sigma}(\theta) = S$)

- **ML (Maximum Likelihood)**

$$F_{ML} = log|\mathbf{\Sigma}(\theta)| + tr(S\mathbf{\Sigma}^{-1}(\theta)) - \log|S| - (p+q)$$

  - Asymptotically unbiased
  - Consistent
  - Asymptotically efficient
  - Scale freeness
  - Availibity of a Confidence Interval

Estimation

The **closeness** of $\mathbf{\Sigma}(\theta)$ to S is measured by fitting functions $F(S, \mathbf{\Sigma}(\theta))$ (with $F \geq 0$ and $F = 0$ iif $\mathbf{\Sigma}(\theta) = S$)

- **ML (Maximum Likelihood)**

$$F_{ML} = log|\mathbf{\Sigma}(\theta)| + tr(S\mathbf{\Sigma}^{-1}(\theta)) - \log|S| - (p+q)$$

  - Asymptotically unbiased
  - Consistent
  - Asymptotically efficient
  - Scale freeness
  - Availibity of a Confidence Interval

- **ULS (Unweighted Least Squares)**

$$F_{ULS} = \frac{1}{2} tr\left([S - \mathbf{\Sigma}(\theta)]^2\right)$$

- **GLS (Generalized Least Squares)**

$$F_{GLS} = \frac{1}{2} tr\left(\left[I - \mathbf{\Sigma}(\theta)S^{-1}\right]^2\right)$$

lavaan R package - syntax and estimation

- Package loading

```
> library(lavaan)
```

- Model specification

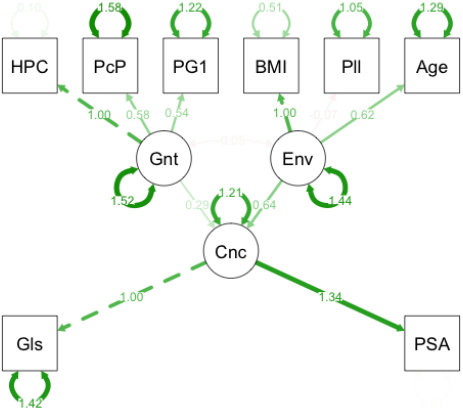
```
> FitModel <- ’
    Genetics =∼ HPC1+PcaP+PG1
    Environment =∼ BMI+Pollution+Age
    Cancer =∼ Gleason+PSA
    Cancer ∼ Genetics+Environment
    Genetics ∼∼ Environment
’
```

- Model estimation

```
> EstimModel <- sem(FitModel, myData)
```

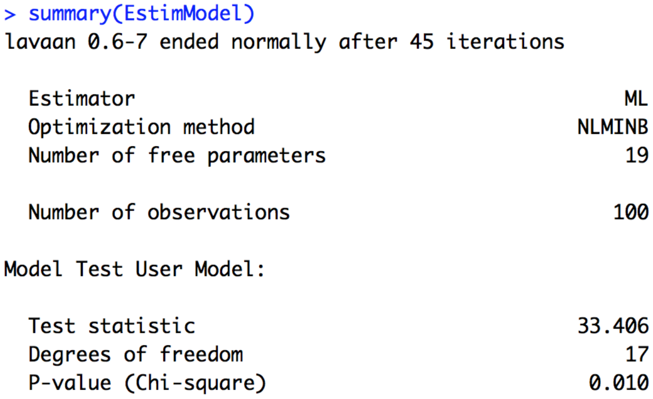
semPlot R package - visualisation

```
> library(semPlot)
> semPaths(EstimModel,what="est",sizeLat=10,edge.label.cex = 1,sizeMan=10)
```



Global summary

```
> summary(EstimModel)
```



```
> summary(EstimModel)
lavaan 0.6-7 ended normally after 45 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         19

  Number of observations                           100

Model Test User Model:

  Test statistic                                33.406
  Degrees of freedom                                17
  P-value (Chi-square)                           0.010
```

Global Fit Measures

- Principle: **comparaison with the saturated model**
  - $\mathcal{M}_s$: Saturated model: no latent variable and one parameter for each variance/covariance for manifest variables
  - $\mathcal{D} = -2(\ell(\mathcal{M}) - \ell(\mathcal{M}_s)) \sim_{\mathcal{H}_0} \chi^2(df)$
  - $p = 0.010$: the model is rejected

- **Other measures** are proposed but *"their purpose is to determine the degree to which the rejected model is approximately correct"* (Shipley, 2016):
  - RMSEA (Root Mean Square Error of Approximation)
  - CFI (Bentler's comparative fit index)

Sample size: N

- Determining the sample size: a **challenge** faced by investigators, peer reviewers, and grant writers
- In the early 80's (Boomsma, 1985)
  - Reasonable results could be obtained with **N of the order of 100**
- In the late 1980's: Bollen consider the **N:q ratio** (where q is the number of free parameters)
  - $N : q = 5$ seems to be enough for normally distributed variables
  - $N : q = 10$ seems to be enough for other distribution
- More recent simulation-based results show the **complex interplay** between (Wolf *et al.*, 2013, Deng *et al.*, 2018)
  - Effect of number of factors
  - Effect of number of indicators
  - Effect of magnitude of factor loadings and regression paths

Interpretation

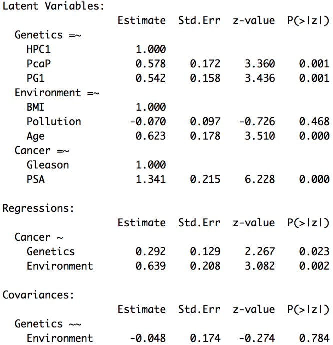
The proposed model is rejected: game over?

- Yes in Confirmatory Factor Analysis (**CFA**)
  - ▸ The model is not confirmed by observed data
- No in Explanatory Factor Analysis (**EFA**)
  - ▸ How can we propose a more likely model?

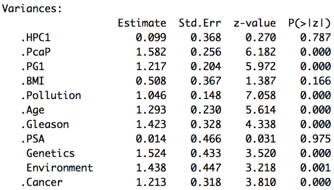
Outline

1. From Linear model to Path model
2. Latent variables
3. Model
4. SEM and Explanatory Factor Analysis
5. Ending words

Caution with coefficients summary



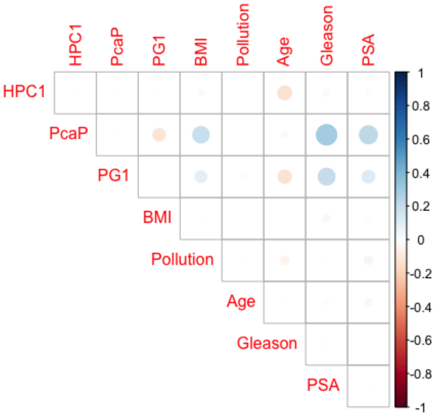
```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Genetics =~
    HPC1              1.000
    PcaP              0.578    0.172    3.360    0.001
    PG1               0.542    0.158    3.436    0.001
  Environment =~
    BMI               1.000
    Pollution        -0.070    0.097   -0.726    0.468
    Age               0.623    0.178    3.510    0.000
  Cancer =~
    Gleason           1.000
    PSA               1.341    0.215    6.228    0.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Cancer ~
    Genetics          0.292    0.129    2.267    0.023
    Environment       0.639    0.208    3.082    0.002

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
  Genetics ~~
    Environment      -0.048    0.174   -0.274    0.784
```



```
Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .HPC1              0.099    0.368    0.270    0.787
   .PcaP              1.582    0.256    6.182    0.000
   .PG1               1.217    0.204    5.972    0.000
   .BMI               0.508    0.367    1.387    0.166
   .Pollution         1.046    0.148    7.058    0.000
   .Age               1.293    0.230    5.614    0.000
   .Gleason           1.423    0.328    4.338    0.000
   .PSA               0.014    0.466    0.031    0.975
    Genetics          1.524    0.433    3.520    0.000
    Environment       1.438    0.447    3.218    0.001
   .Cancer            1.213    0.318    3.810    0.000
```

- By default, latent variables are of the **scale** of “its” first manifest variable
  - Interpretation depends on the constraint
  - Changing the constraint on the latent variable does not modify the global fit

Residuals



- *PcaP* and *PG*1 are badly fitted

Outline

4 SEM and Explanatory Factor Analysis

Model modification

Variable selection using R-square

Outline

4 SEM and Explanatory Factor Analysis

Model modification

Constraints relaxation

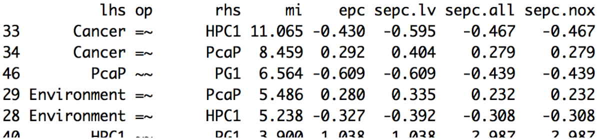
Adding constraint

Model comparison

Modification Indices

- A model can be modified by **relaxing fixed coefficients**
- **Modification index** is based on Lagrangian multiplier (LM)

```
> modindices(EstimModel)
```



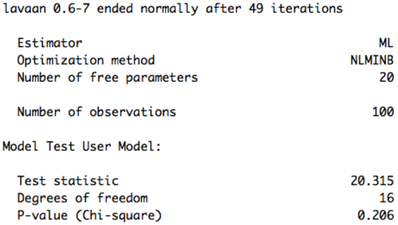
```
               lhs op    rhs     mi    epc sepc.lv sepc.all sepc.nox
33          Cancer =~   HPC1 11.065 -0.430  -0.595   -0.467   -0.467
34          Cancer =~   PcaP  8.459  0.292   0.404    0.279    0.279
46            PcaP ~~    PG1  6.564 -0.609  -0.609   -0.439   -0.439
29     Environment =~   PcaP  5.486  0.280   0.335    0.232    0.232
28     Environment =~   HPC1  5.238 -0.327  -0.392   -0.308   -0.308
40            HPC1 ~~    PG1  3.900  1.038   1.038    2.987    2.987
```

Stepwise approach using modification indices

- Freeing $Cancer =\sim HPC1$ and $Cancer =\sim PcaP$ is nonsense
- We try to **add a covariance** between $PcaP$ and $PG1$

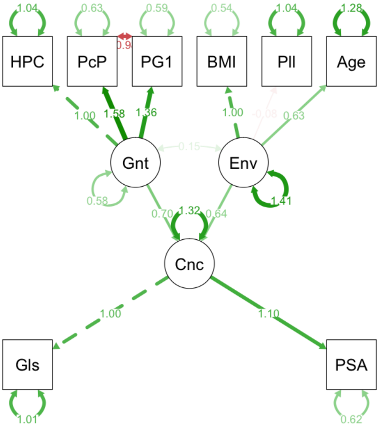
```
> FitModel.2 <- '
    Genetics =~ HPC1+PcaP+PG1
    Environment =~ BMI+Pollution+Age
    Cancer =~ Gleason+PSA
    Cancer ~ Genetics+Environment
    Genetics ~~ Environment
    PcaP ~~ PG1
'
```

- Global fit measure



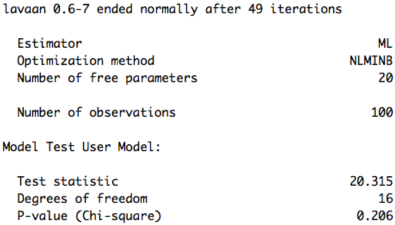
```
lavaan 0.6-7 ended normally after 49 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         20

  Number of observations                           100

Model Test User Model:

  Test statistic                                20.315
  Degrees of freedom                                16
  P-value (Chi-square)                           0.206
```

Updated DAG

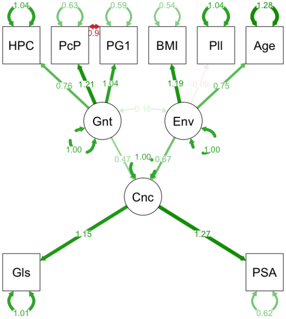

1.04
0.63
0.59
0.54
1.04
1.28
0.9
HPC
PcP
PG1
BMI
PII
Age
1.00
1.58
1.36
1.00
-0.08
0.63
0.15
Gnt
Env
0.58
1.41
0.70
1.32
0.64
Cnc
1.00
1.10
Gls
PSA
1.01
0.62

Constraint modification with lavaan

- Freeing latent coefficient: `Genetics =~ NA*HPC1+PcaP+PG1`
- Fixing latent variance: `Genetics ~~ 1*Genetics`



```
lavaan 0.6-7 ended normally after 49 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         20

  Number of observations                           100

Model Test User Model:

  Test statistic                                20.315
  Degrees of freedom                                16
  P-value (Chi-square)                           0.206
```



- **Global fit remains unchanged**

Outline

4 SEM and Explanatory Factor Analysis

Model modification

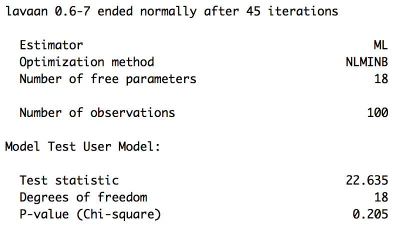
Constraints relaxation

Adding constraint

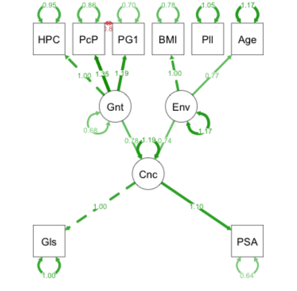
Model comparison

Modification of the models based on coefficient testing

- Latent model
  - The estimated covariance between **Genetics and Environnement** is not significant
- Measurment model
  - The loading between **Pollution and Environnement** is not significant



```
lavaan 0.6-7 ended normally after 45 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of free parameters                         18

  Number of observations                           100

Model Test User Model:

  Test statistic                                22.635
  Degrees of freedom                                18
  P-value (Chi-square)                           0.205
```



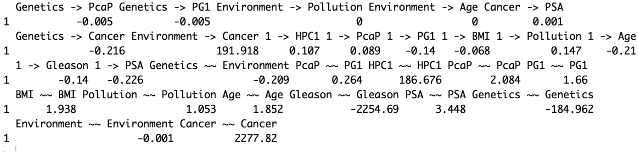
Regularized SEM

- Jacobucci (2019) has proposed a **regularized version** of SEM:

$$F_{ML}^{Reg} = log|\Sigma(\theta)| + tr(S\Sigma^{-1}(\theta)) - \log|S| - (p+q) + \lambda P(.)$$

  - where $P(.)$ is a penalized function (for ex. Lasso, Ridge, ...)

```
> fitRegSem <- regsem(EstimModelRegSem, lambda=1,
  type="lasso", pars_pen=c("regressions","loadings"))
> fitRegSem$coefficients
```



```
  Genetics -> PcaP Genetics -> PG1 Environment -> Pollution Environment -> Age Cancer -> PSA
1           -0.005          -0.005                        0                  0         0.001
  Genetics -> Cancer Environment -> Cancer 1 -> HPC1 1 -> PcaP 1 -> PG1 1 -> BMI 1 -> Pollution 1 -> Age
1             -0.216               191.918     0.107     0.089    -0.14   -0.068          0.147    -0.21
  1 -> Gleason 1 -> PSA Genetics ~~ Environment PcaP ~~ PG1 HPC1 ~~ HPC1 PcaP ~~ PcaP PG1 ~~ PG1
1        -0.14   -0.226                  -0.209       0.264      186.676        2.084       1.66
  BMI ~~ BMI Pollution ~~ Pollution Age ~~ Age Gleason ~~ Gleason PSA ~~ PSA Genetics ~~ Genetics
1      1.938                  1.053      1.852           -2254.69      3.448             -184.962
  Environment ~~ Environment Cancer ~~ Cancer
1                     -0.001          2277.82
```

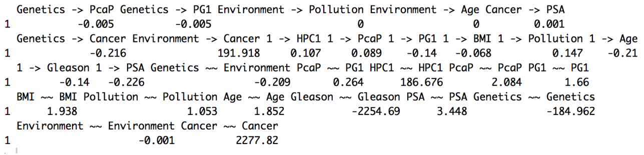
Regularized SEM

- Jacobucci (2019) has proposed a **regularized version** of SEM:

$$F_{ML}^{Reg} = log|\mathbf{\Sigma}(\theta)| + tr(S\mathbf{\Sigma}^{-1}(\theta)) - \log|S| - (p+q) + \lambda P(.)$$

  - where $P(.)$ is a penalized function (for ex. Lasso, Ridge, ...)

```
> fitRegSem <- regsem(EstimModelRegSem, lambda=1,
  type="lasso", pars_pen=c("regressions","loadings"))
> fitRegSem$coefficients
```



```
  Genetics -> PcaP Genetics -> PG1 Environment -> Pollution Environment -> Age Cancer -> PSA
1           -0.005          -0.005                        0                  0         0.001
  Genetics -> Cancer Environment -> Cancer 1 -> HPC1 1 -> PcaP 1 -> PG1 1 -> BMI 1 -> Pollution 1 -> Age
1             -0.216               191.918     0.107     0.089    -0.14   -0.068          0.147    -0.21
  1 -> Gleason 1 -> PSA Genetics ~~ Environment PcaP ~~ PG1 HPC1 ~~ HPC1 PcaP ~~ PcaP PG1 ~~ PG1
1        -0.14   -0.226                  -0.209       0.264      186.676        2.084       1.66
  BMI ~~ BMI Pollution ~~ Pollution Age ~~ Age Gleason ~~ Gleason PSA ~~ PSA Genetics ~~ Genetics
1      1.938                  1.053      1.852           -2254.69      3.448             -184.962
  Environment ~~ Environment Cancer ~~ Cancer
1                     -0.001          2277.82
```

- **Choosing λ** is still an issue

Outline

4 SEM and Explanatory Factor Analysis

Model modification

Constraints relaxation

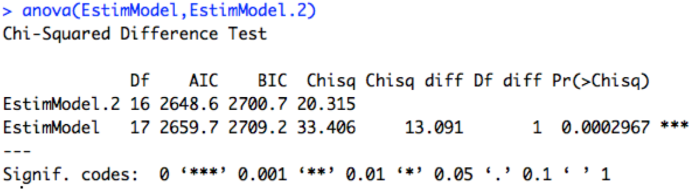
Adding constraint

Model comparison

Model comparison

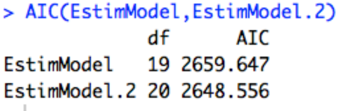
Usual model comparison tools are available

- **Nested model**



```
> anova(EstimModel,EstimModel.2)
Chi-Squared Difference Test

             Df    AIC    BIC  Chisq Chisq diff Df diff Pr(>Chisq)
EstimModel.2 16 2648.6 2700.7 20.315
EstimModel   17 2659.7 2709.2 33.406     13.091       1  0.0002967 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Non-nested model**



```
> AIC(EstimModel,EstimModel.2)
             df      AIC
EstimModel   19 2659.647
EstimModel.2 20 2648.556
```

- ...

Outline

4 SEM and Explanatory Factor Analysis

Model modification

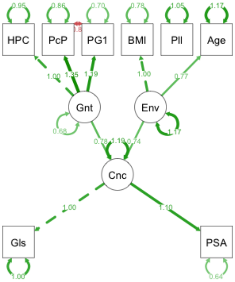
Variable selection using R-square

R-square

- What is the variance for Pollution **explained by the model**?
  - $R^2_{Pollution} = \frac{\lambda^2_{Pollution} \times \mathbb{V}[Env]}{\lambda^2_{Pollution} \times \mathbb{V}[Env] + \mathbb{V}[Pollution]}$
    $R^2_{Pollution} = 0.007824397$
- Interpretation?
  - Pollution seems not to be correlated with the other manifest variables

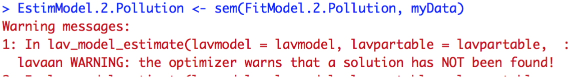
R-square

- What is the variance for Pollution **explained by the model**?
  - $R^2_{Pollution} = \frac{\lambda^2_{Pollution} \times \mathbb{V}[Env]}{\lambda^2_{Pollution} \times \mathbb{V}[Env] + \mathbb{V}[Pollution]}$
    $R^2_{Pollution} = 0.007824397$
- Interpretation?
  - Pollution seems not to be correlated with the other manifest variables

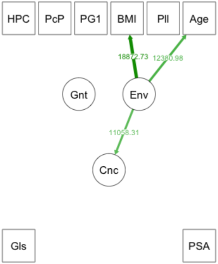


Remark on the importance of the constraint

- Loading constraint should be carefully done



```
> EstimModel.2.Pollution <- sem(FitModel.2.Pollution, myData)
Warning messages:
1: In lav_model_estimate(lavmodel = lavmodel, lavpartable = lavpartable,  :
  lavaan WARNING: the optimizer warns that a solution has NOT been found!
```



Outline

1. From Linear model to Path model

2. Latent variables

3. Model

4. SEM and Explanatory Factor Analysis

5. Ending words

Outline

5 Ending words

Remarks on causality

Conclusion

Eight myths about causality and SEM (Bollen and Pearl, 2013)

- Although SEM aims at incorporating causal assumptions, their ability to infer causality is still a matter of debate

Eight myths about causality and SEM (Bollen and Pearl, 2013)

- Although SEM aims at incorporating causal assumptions, their ability to infer causality is still a matter of debate
- Here 8 myths :
    1. SEMs aim to establish causal relations from associations alone
    2. SEMs and regression are essentially equivalent
    3. No causation without manipulation
    4. SEMs are not equipped to handle nonlinear causal relationships
    5. A potential outcome framework is more principled than SEMs
    6. SEMs are not applicable to experiments with randomized treatments
    7. Mediation analysis in SEMs is inherently non causal
    8. SEMs do not test any major part of the theory against the data.

Myth ♯1: SEMs aim to establish causal relations from associations alone

- Inputs of SEM:
  - Qualitative causal assumptions
  - Empirical data
- Outputs of SEM
  - Failure to fit the data
    - Doubt on causal assumptions (*e.g.* zero coefficients or zero covariance)
    - Guides to repair structural misspecifications
  - Fitting the data
    - Not a proof of causal assumptions...but it makes more plausible

"Positive results need to be replicated and to withstand the criticisms of researchers who suggest other models for the same data"

Tools for testing causality

- **D-separation** in graph theroy
  - Are two nodes independent given a set of others nodes?
  - Hardly applicable for SEM with **latent variables**
- Isolation and **pseudo-isolation**
- **Temporal** component of causality
  - Temporal priority should determining the direction of influence
  - An unsolvable issue for **experimental** design?

Outline

5 Ending words

Remarks on causality

Conclusion

Take-home messages

- SEM is a tool for **modeling (complex) systems via causal assumptions**
- Design of models should not be performed with a pure statistical point-of-view
- SEM can used for **CFA** and **EFA**
- SEM are **easy to use in R**

- Modeling specification and estimation can lead to **unusable models**
  - Convergence issues
  - Constraint sensitivity
  - Negative variance
  - ...
- SEM does **not solve causal inference**

Extensions

- Multilevel SEM modeling
- Meta-Analysis in SEM
  - testing the consistency of the estimates and effect sizes in different studies
  - estimation of a polled effect size
  - identification of potential moderators that influence the model's structure
- Multi-group SEM
- Latent growth curve modeling (LGCM)
- Non-linear SEM
  - Package `piecewiseSEM`

Thank you for your attention!