Modeling change: A gentle introduction to cross-lagged and latent growth curve approach: course materials

Štulhofer, Aleksandar; Ružojčić, Mitja

SEM primer

Modeling change: A gentle introduction to cross-lagged and latent growth curve approach

Mitja Ružojčić

September 26-27, 2023

Zagreb, Croatia

Structural equation modeling (SEM)

- Structural models are **systems of regression equations**.
- Unlike regression analysis, SEM enables us to do much more:
  - Constrain the parameters to a fixed value.
  - Use more outcome variables.
  - Analyses with latent factors, controlling for less than perfect measurement of constructs.
  - Comparison of the fit of two or more different models to the data.
- In SEM, we use the principle of **parsimony** - we want to find the simplest possible model that describes the observed relationships well - this way we get more useful and understandable theories.

Structural equation modeling (SEM)

- All SEM models are based on at least one of the following equations:

- $X_i = \mu_i + \lambda_i*\xi + 1*\varepsilon_i$ or simply $X_i = I_i + f_i*Factor + 1*e_i$ – equation for the measurement part of the structural model, describing how latent factor explains the observed score on item X_i

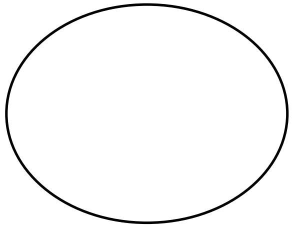
- $Y_i = \beta_0 + \beta_1*X + 1*\zeta_i$ or simply $Y_i = I_0 + b_1*X + 1*e_i$ – typical regression model equation

SEM diagrams

- Notation

Observed variable

Directional relationship (e.g. regression coefficient, factor loading)



Latent variable

Non directional relationship (e.g. covariance, variance)

Diagram of confirmatory factor analysis (CFA)

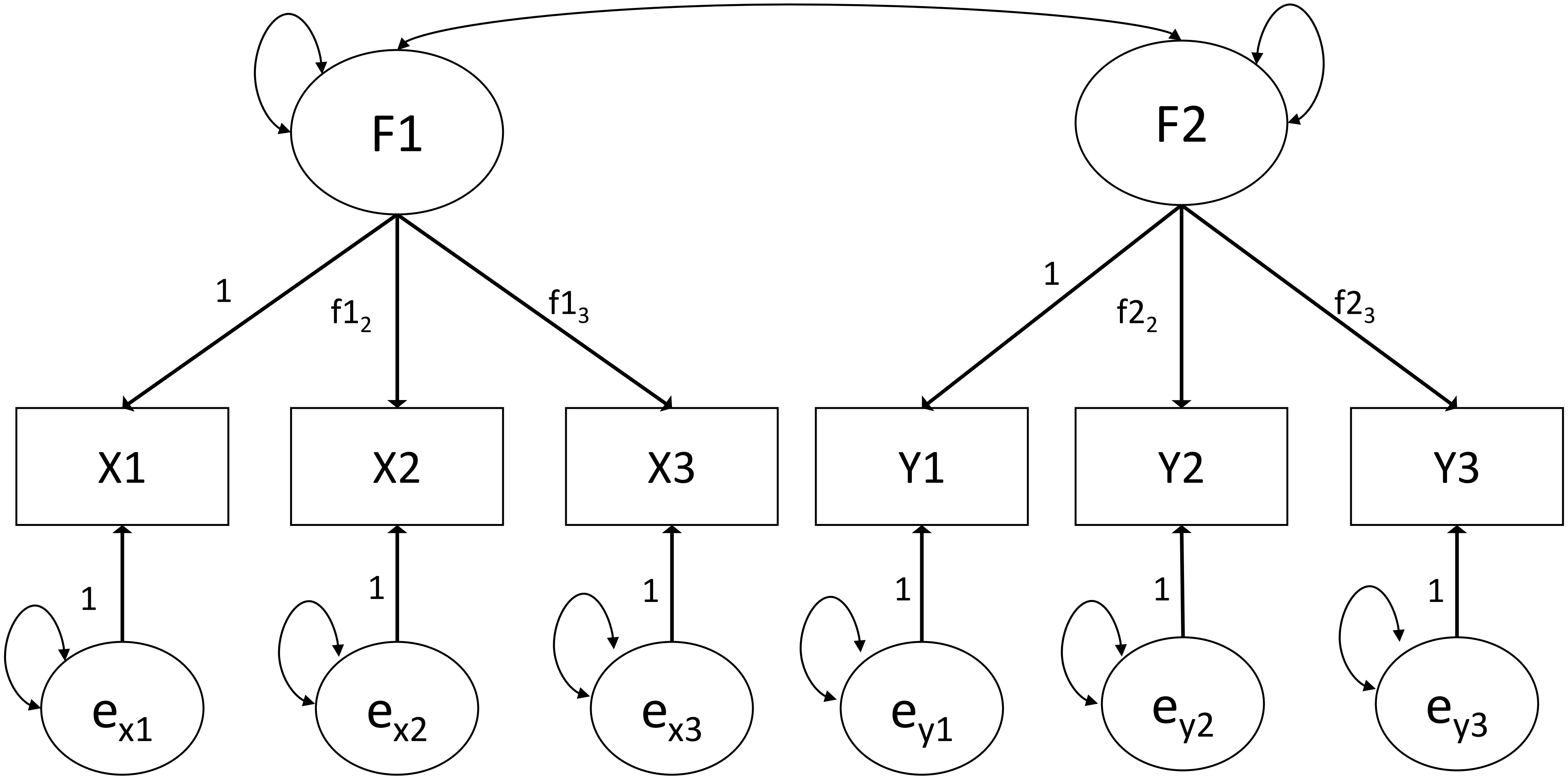


Diagram of confirmatory factor analysis (CFA)

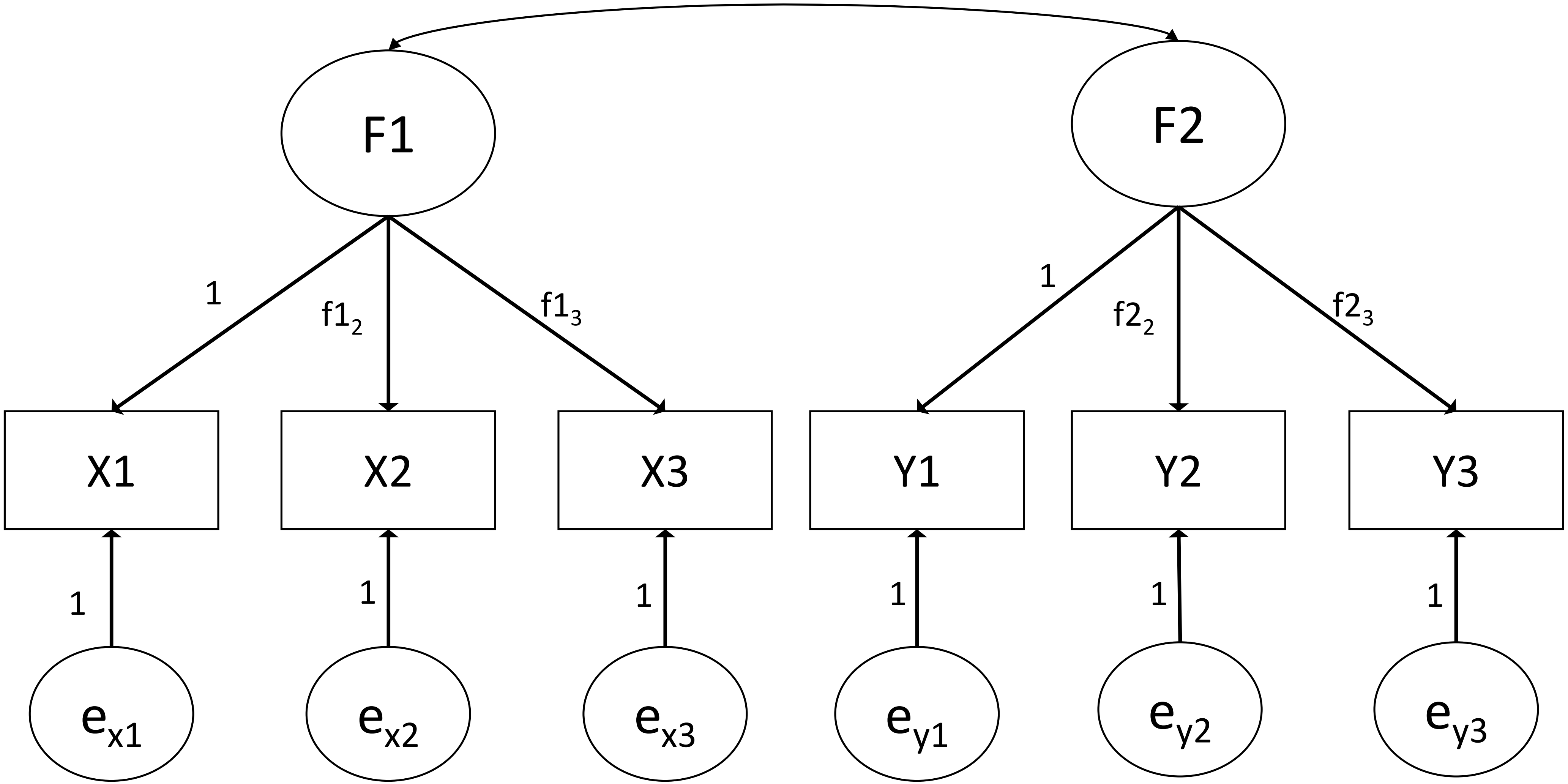


Diagram of structural model (SEM)

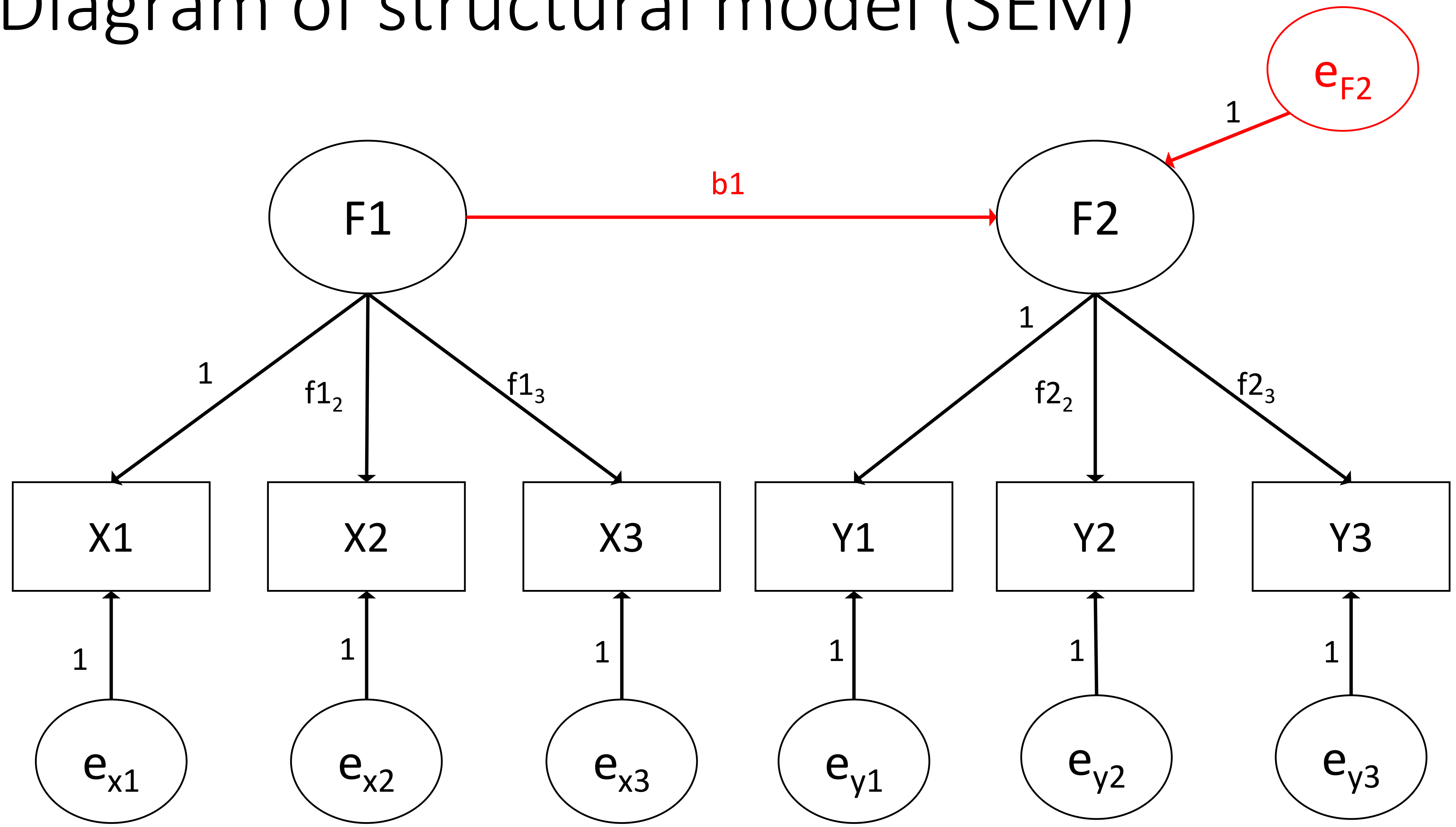
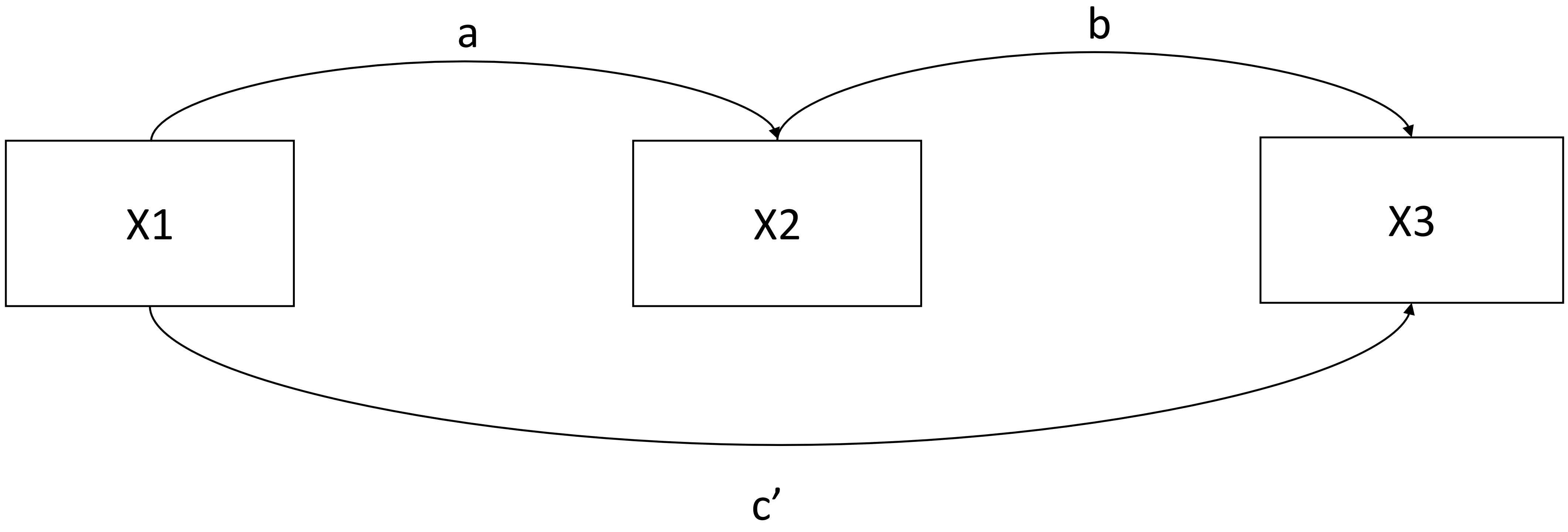


Diagram of path analysis



Model specification

- The process by which we describe what our model looks like, that is, which parameters we want to calculate, and which parameters we will constrain.
- We can specify our model:
  - By writing equations.
  - By drawing a diagram.
  - In the program in which we perform the analysis.

More important SEM terminology

- **Exogenous** variables -> independent variables, i.e. variables whose variance is not explained by any other variable in the model.
- **Endogenous** variables -> dependent variables, other variables in the model explain their variance.
- **Indicators**
  - variables with which we "define" the latent variable, they reflect the latent variable, and the latent variable explains their variance.
  - In psychology, these are usually the items of a questionnaire which measures a certain psychological construct.
- **Constraining a parameter** -> assigning a certain value to a parameter (usually 0 or 1) or equating its value to the value of some other parameter.

Identification

- A precondition for model estimation.

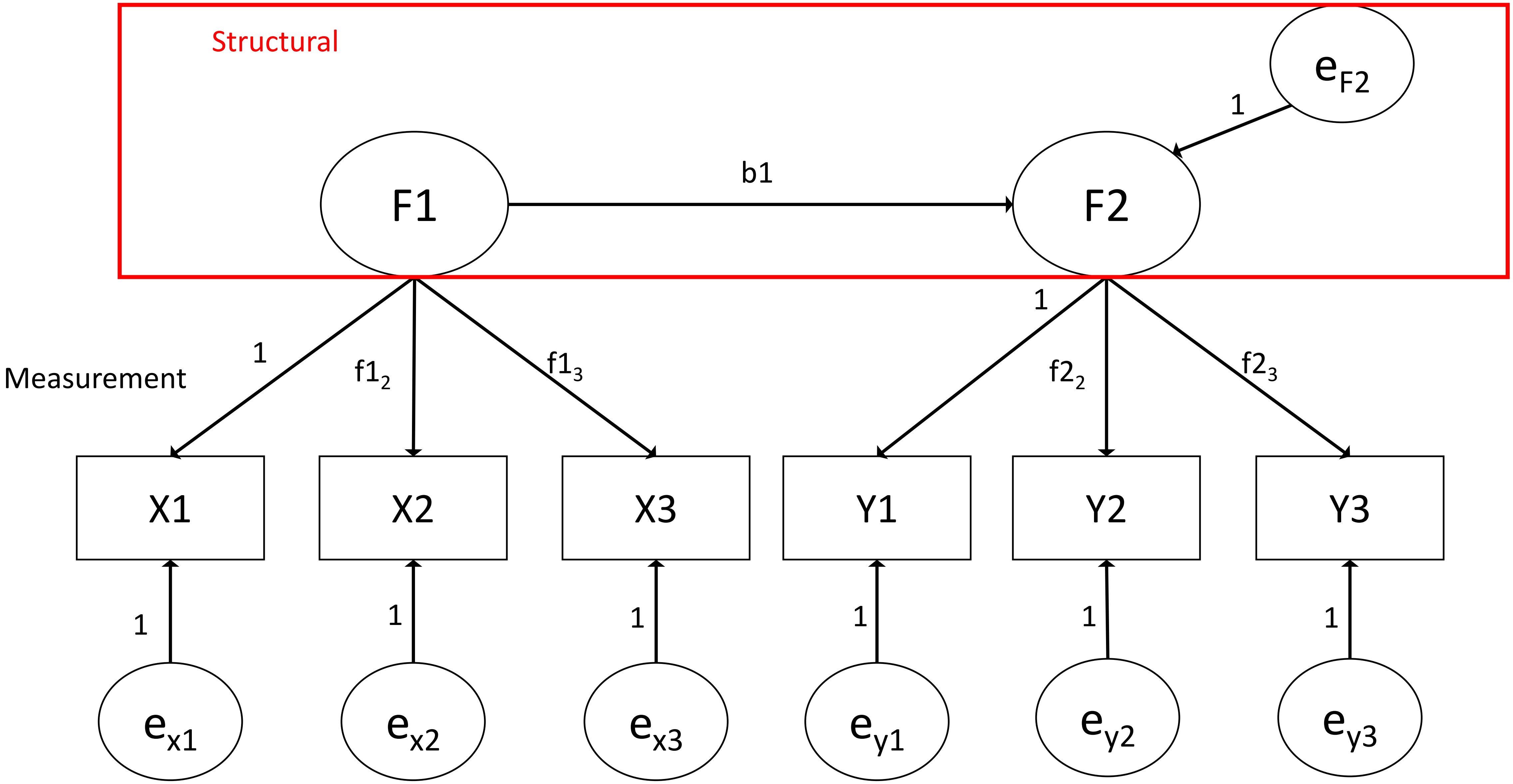
1. We cannot estimate more parameters than we have data with which we enter the analysis - the data with which we enter the analysis are the variances and covariances between the observed variables.
   - We calculate the number of variances and covariances as **{k*(k+1)}/2**, where k = the number of observed variables.
   - If we include mean structure/intercepts, the formula above is **{k*(k+1)}/2 + k**

- The degree of identification of the model is expressed through the degrees of freedom of the model (df) -> df = [{k*(k+1)}/2] – t (t is the number of parameters we estimate in the model)
  - If **df < 0**, the model is **unidentified** and we will not be able to calculate the parameters.
  - If **df = 0** the model is **just identified** – we can calculate the parameters, but it fits the data perfectly and is not parsimonious.
  - If **df > 0** the model is **identified** – we can calculate the parameters and indicators of the global fit of the model.

Identification

2. If our model includes latent variables, we must make sure that
   a) Each latent variable has at least two indicators (but recommended at least 3)
   b) Errors of indicators of the same latent variable are not correlated
   c) We assigned a measurement scale to all latent variables
      - We do this by constraining the loading of one of the items to 1
3. Even if the overall model satisfies rules 1 and 2, each of the model components must be identified on its own
   - If we have a measurement (CFA) and a structural component of the model (regression), each of these components must be identified.
   - If we have higher order factors, the measurement model at each level should be identified.

Measurement and structural components



Sample size

- Relatively large samples are necessary.
- No clear cutoff, according to some authors 100-200 is the minimum.
- 20:1 rule (Jackson, 2003) – a minimum of 20 subjects for each estimated parameter.

Estimators

- **Maximum likelihood (ML)** method is the most common estimation method in SEM.
- ML requires multivariate normality.
- If this assumption is not met – **robust ML** (basically ML with robust standard errors).
- If you do not have continuous variables – **diagonally weighted least squares (DWLS)** -> no distributional assumptions.
  - Be careful, DWLS tends to artificially inflate model fit

Model fit indices

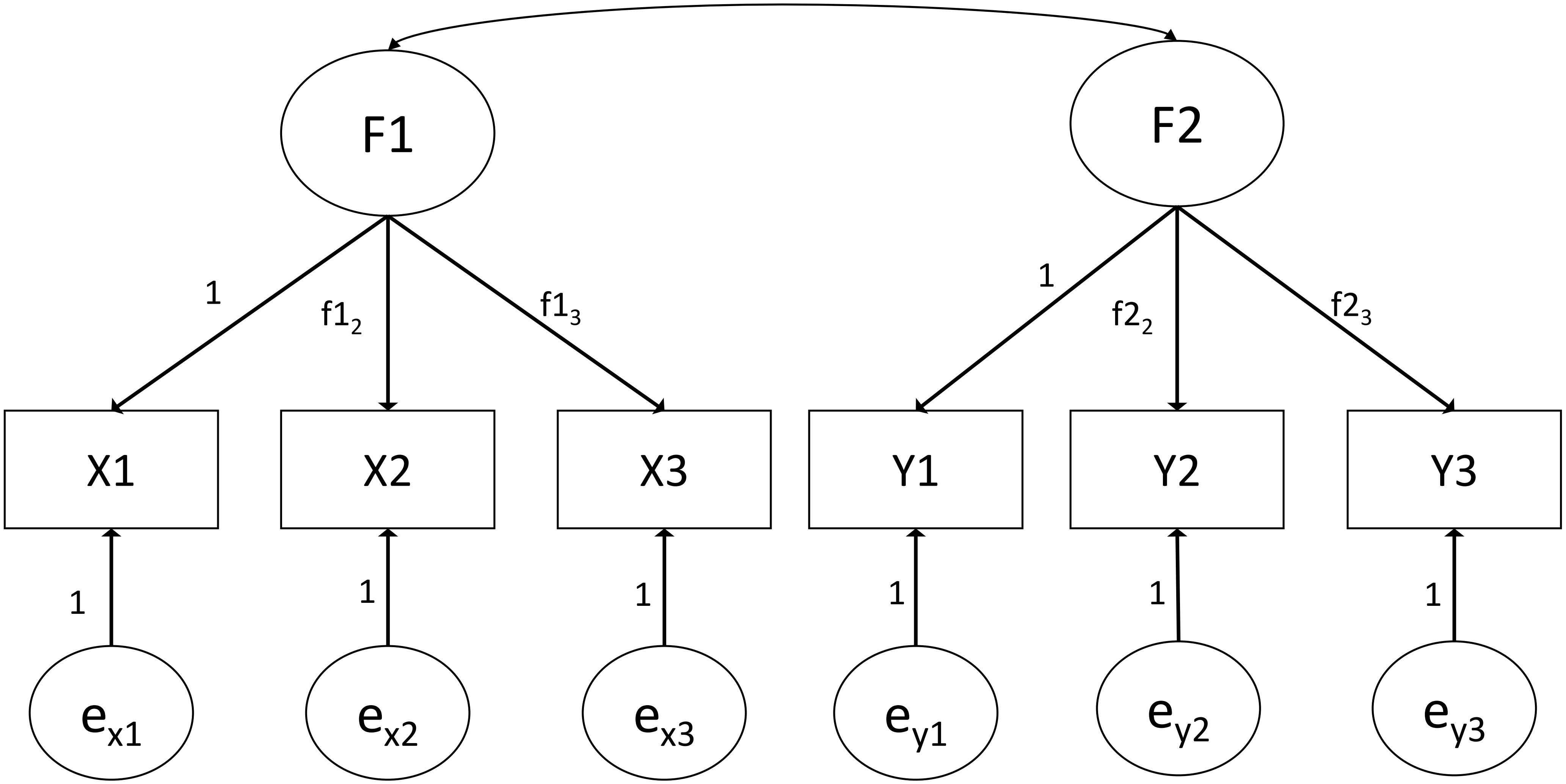
- Chi square (χ2)
  - χ^2 is the only fit index that provides information about the statistical significance of the model fit.
  - Our goal is the opposite of what we are used to - we want χ^2 to be non-significant, we want to confirm the null hypothesis.
  - Problem –very sensitive to sample size.
- Comparative fit index (**CFI**) and Tucker-Lewis index (**TLI**)
  - Goodnes-of-fit indices – higher values mean better fit.
  - Range from 0 to 1.
  - Rule of thumb – values from 0.90 (Awang, 2012) to 0.95 (Hu i Bentler, 1999) are considered an acceptable fit.

Model fit indicators

- The Root Mean Square Error of Approximation (**RMSEA**)
  - Badness of fit index – lower values mean better fit.
  - Ranges from 0 to 1.
  - Rule of thumb – values lower than 0.05 (Hu & Bentler, 1999) or 0.08 (Awang, 2012) are considered as an acceptable fit.
  - It favours simpler, parsimonious models.
- Standardized Root Mean Squared Residual (**SRMR**)
  - Badness of fit index – lower values mean better fit.
  - Ranges from 0 to 1.
  - Rule of thumb – values lower than 0.05 (Hu & Bentler, 1999) are considered as an acceptable fit.

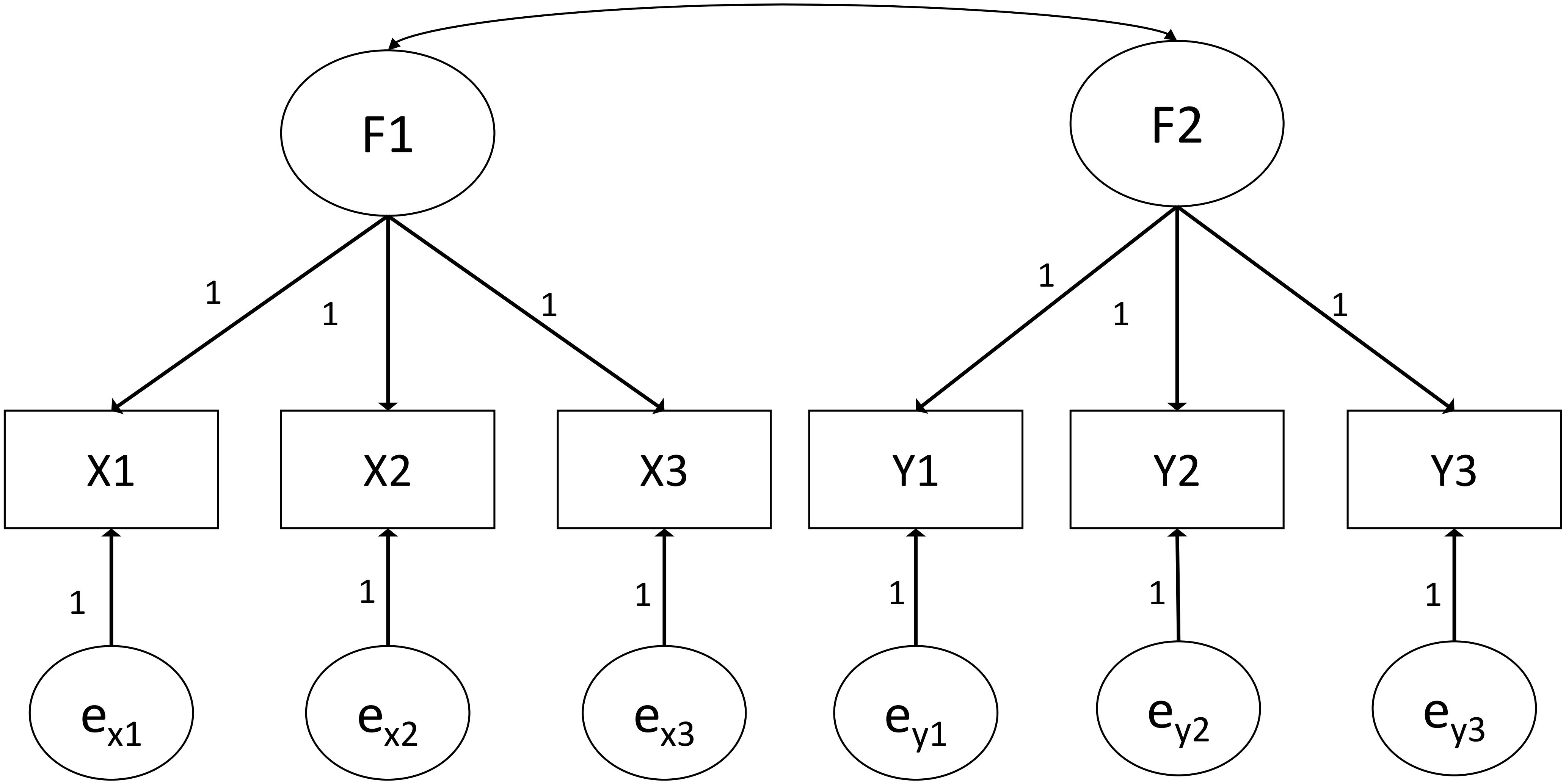
How to know if our model is fit?

- Consider multiple fit indices.
- Compare your model with alternative theorethically plausible models.
  - χ^2 difference test but also compare other fit indices – in larger samples χ^2 tends to identify significant differences between models even if the difference in fit is quite small.
  - Models need to be **nested to be comparable** – simpler model can be specified by constraining the parameters of a more complex model.

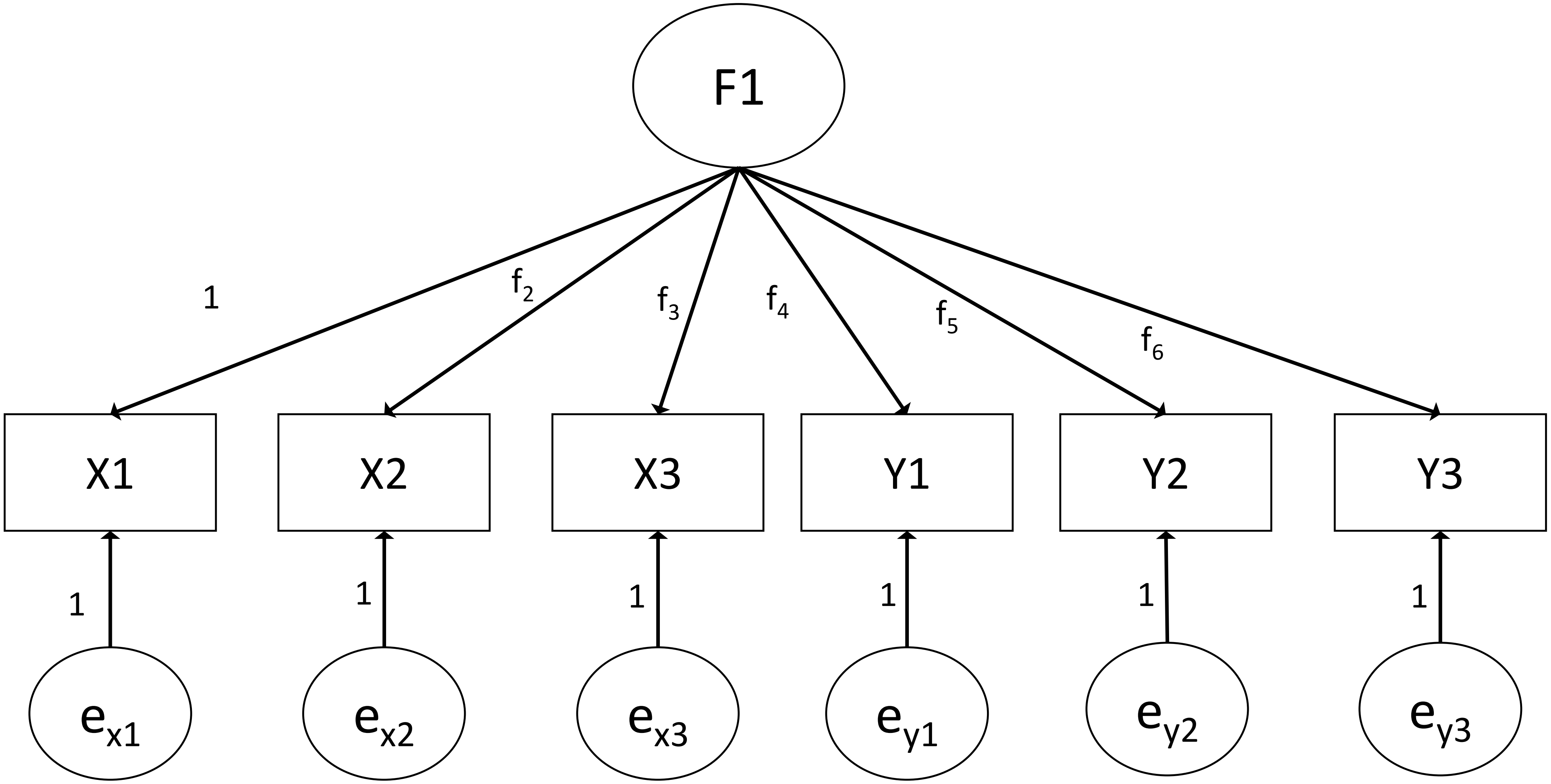


F1
F2
1
f1₂
f1₃
1
f2₂
f2₃
X1
X2
X3
Y1
Y2
Y3
1
1
1
1
1
1
e_x1
e_x2
e_x3
e_y1
e_y2
e_y3

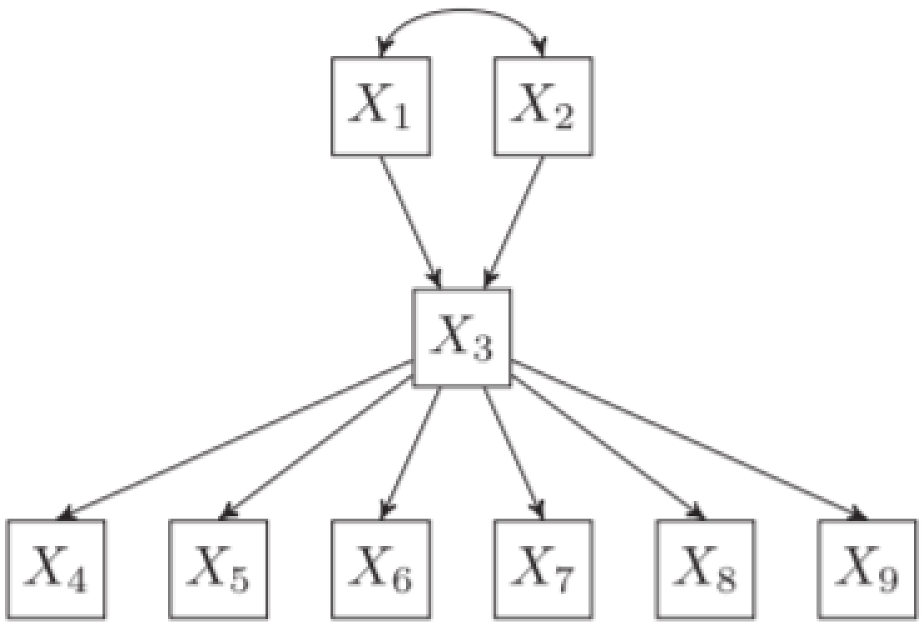
This model is nested in the model above



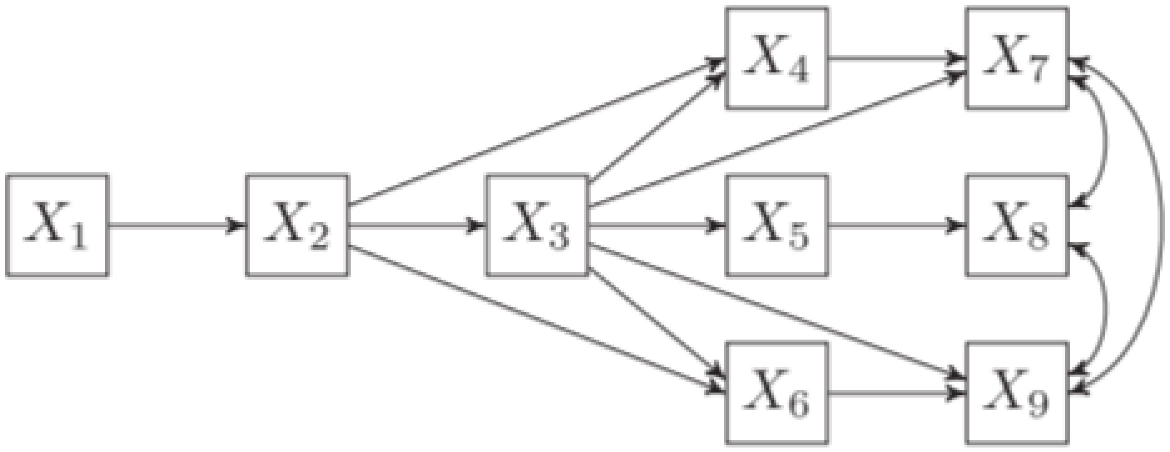
Is this model nested in the first one?



These models are not nested



Model A ($df = 27$)



Model B ($df = 21$)

How to compare non-nested models?

- Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC).
- They estimate the probability that a specified model is replicated in a sample of the same size drawn from the same population.
- The lower the value, the better.
- They do not have a cutoff values and are mostly used for comparing two models.