

Structural Equation Modeling (SEM): Concepts, Methods, and Applications in Linguistics

结构方程模型：概念、方法与语言学应用

guoshujian@tongji.edu.cn

2025年12月3日

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Outline / 提纲

- 1 Concepts & Methods (概念与方法)
- 2 Case Studies (语言学案例分析)
- 3 Common Issues (容易出现的问题)
- 4 R Programming Examples (R语言编程实例)
- 5 Research Design (研究设计)

What is SEM? / 什么是结构方程模型?

Definition (定义)

- A multivariate statistical analysis technique that is used to analyze structural relationships.
- It is the combination of **Factor Analysis** and **Multiple Regression**.

Key Features (核心特征)

- **Latent Variables (潜变量)**: Constructs not directly measured (e.g., Anxiety).
- **Observed Variables (显变量)**: Measured items (e.g., Test scores, Survey items).
- **Measurement Error (测量误差)**: Explicitly modeled.

Two Parts of SEM / 模型的两部分

1. Measurement Model (测量模型)

Relationships between latent variables and their indicators (CFA).

描述潜变量与显变量之间的关系（验证性因子分析）。

2. Structural Model (结构模型)

Relationships between latent variables (Path Analysis).

描述潜变量之间的路径关系（路径分析）。

Goal: To explain the covariance among observed variables.

目标：解释观测变量之间的协方差。

目标：解释观测变量之间的协方差。

CFA or EFA / 验证性因子分析 还是 探索性因子分析

维度	探索性因子分析 (EFA)	验证性因子分析 (CFA)
研究目的	发现结构 ：我不知道数据里有几个因子，也不知道哪个题归哪个因子，我让数据自己“告诉我”。	验证结构 ：我已经有理论假设（如：这10题分2个因子，前5题归因子A），我来看看数据支不支持我的假设。
理论基础	数据驱动 (Data-driven) ：理论较弱或未知时使用。	理论驱动 (Theory-driven) ：必须基于坚实的文献或理论框架。
因子载荷	自由估计 ：默认假设所有题目都可能与所有因子有关（会有交叉载荷）。	限制估计 ：人为限制题目只能归属于特定的因子（其他路径强制设为0）。
判断标准	看载荷大小、特征值、碎石图，解释是否合理。	看拟合指数（RMSEA, CFI, TLI, χ^2 ），看模型是否拟合数据。
常用软件	SPSS, SAS, R (psych包)	AMOS, Mplus, LISREL, R (lavaan包)
属于SEM吗	不属于（它是SEM的前置步骤）。	属于（CFA就是SEM中的“测量模型”部分）。

Fit Indices / 拟合指标

Index (指标)	Full Name (全称)	Threshold (标准)
χ^2	Chi-square	Non-significant ($p > .05$) (Sensitive to sample size N)
CFI	Comparative Fit Index	> 0.90 (Good), > 0.95 (Excellent)
TLI	Tucker-Lewis Index	> 0.90 (Good), > 0.95 (Excellent)
RMSEA	Root Mean Square Error of Approximation	< 0.08 (Acceptable), < 0.05 (Good)
SRMR	Standardized Root Mean Square Residual	< 0.08

表: Commonly Used Model Fit Indices (常用模型拟合优度指标)

Case 1: Psycholinguistics / 心理语言学案例

Topic: L2 Motivation Self System & Anxiety

Hypothesis: Ideal L2 Self reduces Anxiety, which in turn increases Proficiency.

- **Latent Variables:**

- Ideal Self (3 questionnaire items)
- L2 Anxiety (4 questionnaire items)

- **Observed Variable:** English Final Exam Score

- **Structure:** Ideal Self $\rightarrow$ (-) Anxiety $\rightarrow$ (-) Score

Analysis: Mediation Analysis (中介效应分析). The direct path from Ideal Self to Score is significant, but partially mediated by Anxiety.

Case 2: Sociolinguistics / 社会语言学案例

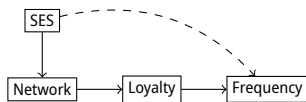
Topic: Language Attitude and Dialect Usage

Context: Urbanization and dialect loss in a specific region.

Model Structure:

- **Exogenous (外生变量):** SES (Socioeconomic Status), Network Strength.
- **Endogenous (内生变量):** Local Loyalty, Dialect Frequency.

Findings: High SES predicts weaker Network Strength, leading to lower Local Loyalty and reduced Dialect Frequency.



Case 3: Experimental Phonetics / 实验语音学案例

Topic: Articulation-Acoustics-Perception Link

Method: Using SEM to model the causal chain of speech production.

- **Latent V1 (Tongue Pos):** Measured by Ultrasound (Tongue height, retraction).
- **Latent V2 (Acoustics):** Formant frequencies (F1, F2, F3).
- **Latent V3 (Perception):** Identification accuracy by listeners.

Innovation: Unlike traditional correlation, SEM handles measurement errors in ultrasound data (noisy images) and acoustic extraction simultaneously to predict intelligibility.

Common Pitfalls / 常见陷阱

1 Sample Size (样本量不足):

- SEM requires large samples. Rule of thumb: $N > 200$, or 10-20 cases per parameter estimated.

2 Normality Violation (违反主要态分布):

- Likert scales are ordinal, not continuous. Use Robust ML (MLR) or Diagonally Weighted Least Squares (DWLS).

3 Identification (模型识别):

- More unknown parameters than data points ($df < 0$).
- Solution: Fix the scale of latent variables (set one loading to 1).

4 Post-hoc Modifications (事后修正):

- Modifying models solely based on Modification Indices (MI) without theoretical justification constitutes "fishing."

Setup: The lavaan Package

The lavaan (Latent Variable Analysis) package is the standard for SEM in R.

Installation & Loading

```
1 install.packages("lavaan")
2 library(lavaan)
3 library(semPlot) # For visualization
```

Syntax Guide / 语法指南:

- = : Defined by (Latent variable definition / 测量模型)
- : Regressed on (Structural relationship / 回归关系)
- : Covariance/Variance (Residuals / 协方差)

Example 1: CFA (Confirmatory Factor Analysis)

Scenario: *Validating a survey with 2 factors: Visual and Textual skills.*

```
1 # 1. Define the Model / 定义模型
2 cfa_model <- '
3     visual  =~ x1 + x2 + x3
4     textual =~ x4 + x5 + x6
5 '
6
7 # 2. Fit the Model / 拟合模型
8 # std.lv=TRUE fixes latent variance to 1 (standardized)
9 fit_cfa <- cfa(cfa_model, data = HolzingerSwineford1939,
10               std.lv = TRUE)
11
12 # 3. Summary / 结果摘要
13 summary(fit_cfa, fit.measures = TRUE, standardized = TRUE)
```

Example 2: Path Analysis (Mediation)

Scenario: $X \rightarrow M \rightarrow Y$. Testing indirect effects.

```
1 # Model with labeled parameters (a, b, c)
2 med_model <- '
3   # Direct associations
4   M ~ a*X
5   Y ~ b*M + c*X
6
7   # Defined parameters (Indirect and Total effects)
8   indirect := a*b
9   total    := c + (a*b)
10 '
11
12 fit_med <- sem(med_model, data = myData)
13 summary(fit_med, standardized = TRUE)
```

Note: The ':= ' operator defines new parameters for significance testing.

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Example 3: Full SEM (Structural + Measurement)

Scenario: Psycholinguistics model (Motivation → Anxiety).

```
1 full_model <- '  
2   # Measurement Model (Latent definitions)  
3   Motiv  =~ m1 + m2 + m3 + m4  
4   Anx    =~ a1 + a2 + a3  
5  
6   # Structural Model (Regressions)  
7   Anx    ~ Motiv      # Motivation predicts Anxiety  
8   Score ~ Anx         # Anxiety predicts Score  
9  
10  # Covariance (if needed)  
11  m1 ~~ m2            # Correlated errors  
12 '  
13  
14 fit_full <- sem(full_model, data = l2_data, estimator = "MLR")  
15 # MLR handles non-normality robustly  
16 summary(fit_full, fit.measures = TRUE, standardized = TRUE)
```

Designing for SEM / 如何设计符合SEM的研究

1. Theory Driven (理论驱动)

Do not use SEM for data mining. Draw the diagram based on literature before collecting data.
切勿使用SEM进行数据挖掘。在收集数据前，必须基于文献绘制假设路径图。

2. Instrumentation (工具编制)

Each latent variable needs at least 3 indicators (items).
每个潜变量至少需要3个观测指标（题目）。

3. Sample Planning (样本规划)

Conduct a Power Analysis. Simple model: $N = 200$. Complex model: $N = 500+$.
进行效能分析。简单模型200人，复杂模型建议500人以上。

Checklist for Reviewers / 审稿人检查清单

When writing your paper using SEM:

- ☐ Did you report the software and estimator used? (e.g., lavaan, MLR)
- ☐ Did you report χ^2 , df, CFI, RMSEA, and SRMR?
- ☐ Are factor loadings significant (> 0.4 or > 0.5)?
- ☐ Did you address missing data (FIML or Imputation)?
- ☐ Is the model theoretically plausible?

Thank You!

谢谢!

Questions?
guoshujian@tongji.edu.cn