Score-Based Tests of Measurement Invariance with Respect to Continuous and Ordinal Variables

Achim Zeileis, Edgar C. Merkle, Ting Wang

http://eeecon.uibk.ac.at/~zeileis/

Overview

- Motivation
- Framework
- Score-based tests
  - Continuous variables
  - Ordinal variables
  - Categorial variables
- Illustration

Motivation

Psychometric models: Typically measure latent scales based on certain manifest variables, e.g., item response theory (IRT) models or confirmatory factor analysis (CFA, today's focus).

Crucial assumption: Measurement invariance (MI). Otherwise observed differences in scales cannot be reliably attributed to the latent variable that the model purports to measure.

Parameter stability: In parametric models, the MI assumption corresponds to stability of parameters across all possible subgroups.

Inference: The typical approach for assessing MI is

- to split the data into reference and focal groups,
- assess the stability of selected parameters (all or only a subset) across these groups
- by means of standard tests: likelihood ratio (LR), Wald, or Lagrange multiplier (LM or score) tests.

Motivation

Problems:

- Subgroups have to be formed in advance.
- Continuous variables are often categorized into groups in an ad hoc way (e.g., splitting at the median).
- In ordinal variables the ordering of the categories is often not exploited (assessing only if at least one group differs from the others).
- When likelihood ratio or Wald tests are employed, the model has to be fitted to each subgroup which can become numerically challenging and computationally intensive.

Motivation

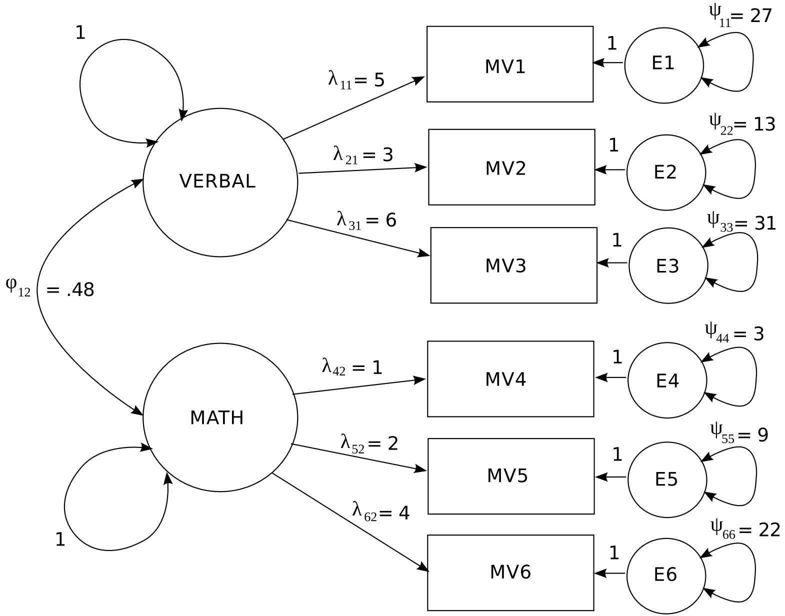
Idea:

- Generalize the LM test.
- Thus, the model only has to be fitted once under the MI assumption to the full data set.
- Catpure model deviations along a variable that is suspected to cause MI violations.
- Exploit ordering to assess if there is (at least) one split so that the model parameters before and after the split differ.
- The split does *not* have to be known or guessed in advance.

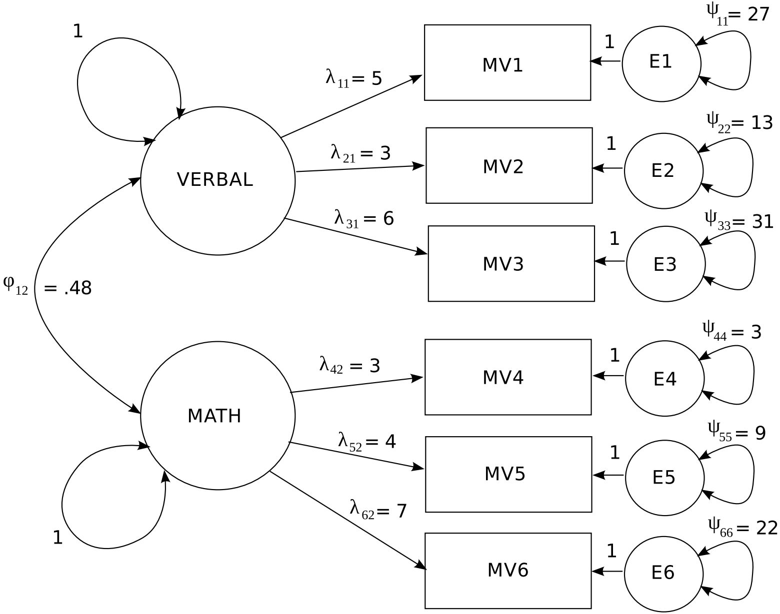
Illustration: CFA for artificial data.

- Model with two latent scales (verbal and math).
- Three manifest variables for each scale.
- Violation of MI for the math loadings along the age of the subjects.

Motivation: CFA for age ≤ 16



Motivation: CFA for age > 16



Framework

Model: Based on log-likelihood $\ell(\cdot)$ for p-dimensional observations $\boldsymbol{x}_i$ ($i = 1, \dots, n$) based on k-dimensional parameter $\boldsymbol{\theta}$.

Estimation: Maximum likelihood.

$$\hat{\boldsymbol{\theta}} \;=\; \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{x}_i).$$

Equivalently: Solve first order conditions

$$\sum_{i=1}^{n} \boldsymbol{s}(\hat{\boldsymbol{\theta}}; \boldsymbol{x}_i) \;=\; 0,$$

where the score function is the partial derivative of the casewise likelihood contributions w.r.t. the parameters $\boldsymbol{\theta}$.

$$\boldsymbol{s}(\boldsymbol{\theta}; \boldsymbol{x}_i) \;=\; \left(\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{x}_i)}{\partial \theta_1}, \dots, \frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{x}_i)}{\partial \theta_k} \right)^{\top}.$$

Framework

Assumption: Distribution/likelihood of $\boldsymbol{x}_i$ depends only on the latent scales (through the parameters $\boldsymbol{\theta}$) – but not on any other variable v_i.

Alternative view: Parameters $\boldsymbol{\theta}$ do not depend any such variable v_i. Hence assess for $i = 1, \dots, n$

$$\begin{aligned} H_0 : \boldsymbol{\theta}_i &= \boldsymbol{\theta}_0, \\ H_1 : \boldsymbol{\theta}_i &= \boldsymbol{\theta}(v_i). \end{aligned}$$

Special case: Two subgroups resulting from one split point ν.

$$H_1^* : \boldsymbol{\theta}_i = \begin{cases} \boldsymbol{\theta}^{(A)} & \text{if } v_i \leq \nu \\ \boldsymbol{\theta}^{(B)} & \text{if } v_i > \nu \end{cases}$$

Tests: LR/Wald/LM tests can be easily employed if pattern $\boldsymbol{\theta}(v_i)$ is known, specifically for H_1^* with fixed split point ν.

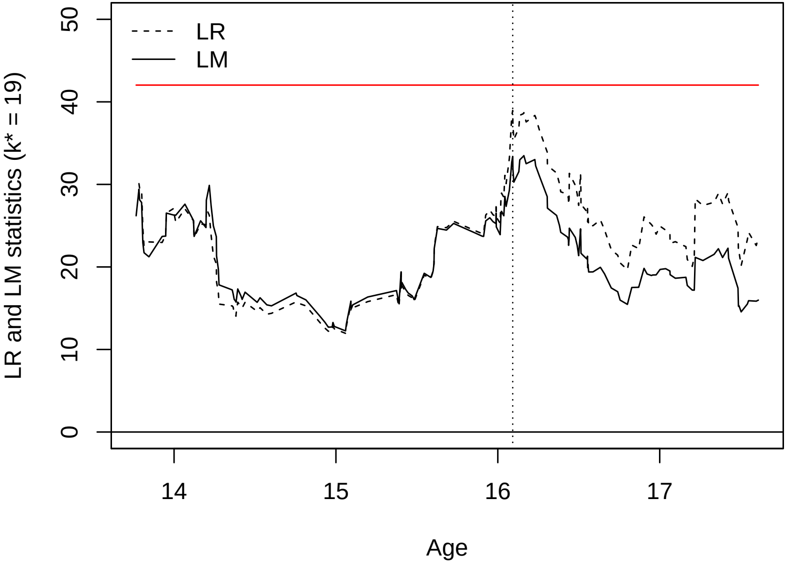
Framework

For unknown split points: Compute LR/Wald/LM tests for each possible split point $v_1 \leq v_2 \leq \dots \leq v_n$ and reject if the maximum statistic is large.

Caution: By maximally selecting the test statistic different critical values are required (not from a χ^2 distribution)!

Illustration: Assess all $k^* = 19$ model parameters from the artificial CFA example along the continuous variable age (v_i).

Framework



Framework

Note: For the maxLM test the parameters $\hat{\theta}$ only have to be estimated once. Only the model scores $\boldsymbol{s}(\hat{\theta}; \boldsymbol{x}_i)$ have to be aggregated differently for each split point.

More generally: Consider a class of tests that assesses whether the model "deviations" $\boldsymbol{s}(\hat{\theta}; \boldsymbol{x}_i)$ depend on v_i. This can consider only a subset $k*$ of all k parameters/scores or try to capture other patterns than H_1^*.

Score-based tests

Fluctuation process: Capture fluctuations in the cumulative sum of the scores ordered by the variable v.

$$\boldsymbol{B}(t; \hat{\theta}) \; = \; \hat{\boldsymbol{I}}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor n \cdot t \rfloor} \boldsymbol{s}(\hat{\theta}; \boldsymbol{x}_{(i)}) \qquad (0 \le t \le 1).$$

- $\hat{\boldsymbol{I}}$ – estimate of the information matrix.
- t – proportion of data ordered by v.
- $\lfloor n \cdot t \rfloor$ – integer part of $n \cdot t$.
- $x_{(i)}$ – observation with the i-th smallest value of the variable v.

Functional central limit theorem: Under H_0 convergence to a (continuous) Brownian bridge process $\boldsymbol{B}(\cdot; \hat{\theta}) \stackrel{d}{\to} \boldsymbol{B}^0(\cdot)$, from which critical values can be obtained – either analytically or by simulation.

Score-based tests: Continuous variables

Test statistics: The empirical process can be viewed as a matrix $\boldsymbol{B}(\hat{\theta})_{ij}$ with rows $i = 1, \dots, n$ (observations) and columns $j = 1, \dots, k$ (parameters). This can be aggregated to scalar test statistics along continuous the variable v.

$$
\begin{aligned}
DM &= \max_{i=1,\dots,n} \max_{j=1,\dots,k} |\boldsymbol{B}(\hat{\theta})_{ij}| \\
CvM &= n^{-1} \sum_{i=1,\dots,n} \sum_{j=1,\dots,k} \boldsymbol{B}(\hat{\theta})_{ij}^2, \\
\max LM &= \max_{i=\underline{i},\dots,\bar{\imath}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1,\dots,k} \boldsymbol{B}(\hat{\theta})_{ij}^2.
\end{aligned}
$$

Critical values: Analytically for DM. Otherwise by direct simulation or further refined simulation techniques.

Score-based tests: Ordinal variables

Test statistics: Aggregation along ordinal variables v with m levels.

$$
\begin{aligned}
WDM_o &= \max_{i \in \{i_1, \dots, i_{m-1}\}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1/2} \max_{j = 1, \dots, k} |\boldsymbol{B}(\hat{\theta})_{ij}|, \\
\max LM_o &= \max_{i \in \{i_1, \dots, i_{m-1}\}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j = 1, \dots, k} \boldsymbol{B}(\hat{\theta})_{ij}^2,
\end{aligned}
$$

where $i_1, \dots, i_{m-1}$ are the numbers of observations in each category.

Critical values: For WDM_o directly from a multivariate normal distribution. For $\max LM_o$ via simulation.

Score-based tests: Categorical variables

Test statistic: Aggregation within the *m* (unordered) categories of *v*.

$$LM_{uo} = \sum_{\ell=1,\dots,m} \sum_{j=1,\dots,k} \left(\boldsymbol{B}(\hat{\theta})_{i_\ell j} - \boldsymbol{B}(\hat{\theta})_{i_{\ell-1} j} \right)^2,$$

Critical values: From a χ^2 distribution (as usual).

Asymptotically equivalent: LR test.

Illustration

Software: In R system for statistical computing.

- *strucchange* implements this general framwork for parameter instability tests.
- Object-oriented implementation that can be applied to many model classes, including *lavaan* objects for CFA models.

Data:

- Application of adult gratitude scale to $n = 1401$ youth aged 10–19 years.
- GQ-6 scale has five Likert scale items with seven points each.
- Assess the factor loadings of a one-factor model.
- Question: Measurement invariance across six age groups?

Illustration

Packages:

```
R> library("lavaan")
R> library("strucchange")
```

Data: Omitting incomplete cases.

```
R> data("YouthGratitude", package = "psychotools")
R> compcases <- apply(YouthGratitude[, 4:28], 1,
+    function(x) all(x %in% 1:9))
R> yg <- YouthGratitude[compcases, ]
```

Estimation: One-factor CFA with loadings restricted to be equal across age groups.

```
R> gq6cfa <- cfa("f1 =~ gq6_1 + gq6_2 + gq6_3 + gq6_4 + gq6_5",
+    data = yg, group = "agegroup", meanstructure = TRUE,
+    group.equal = "loadings")
```

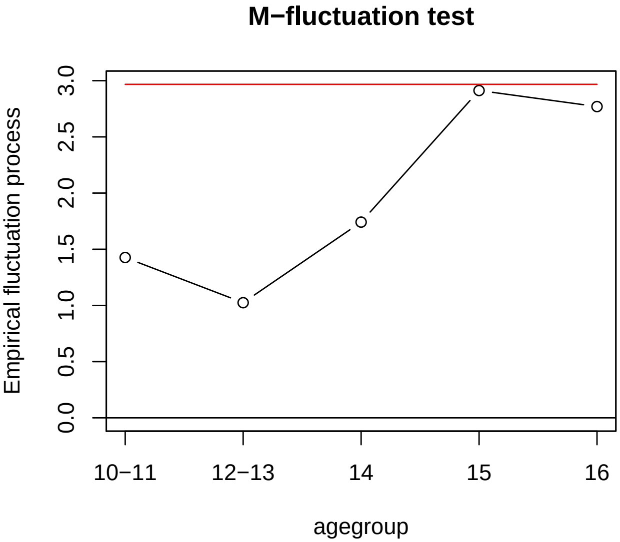
Illustration

Measurement invariance tests:

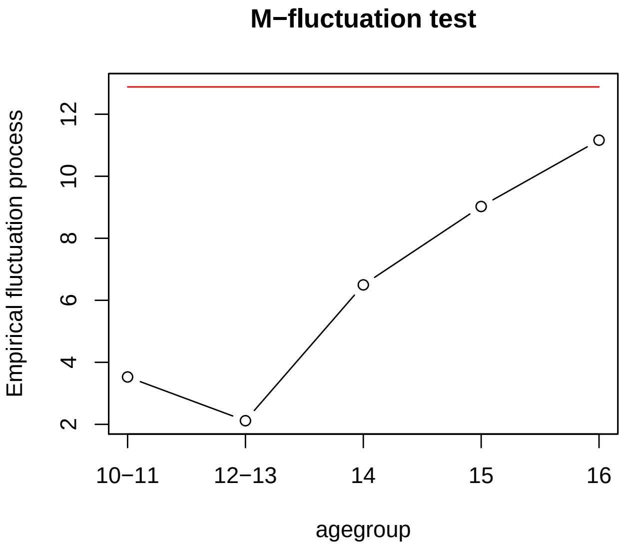
```
R> sctest(gq6cfa, order.by = yg$agegroup, parm = 1:4,
+    vcov = "info", functional = "WDMo", plot = TRUE)

M-fluctuation test
data:  gq6cfa
f(efp) = 2.9129, p-value = 0.05874

R> sctest(gq6cfa, order.by = yg$agegroup, parm = 1:4,
+    vcov = "info", functional = "maxLMo", plot = TRUE)

M-fluctuation test
data:  gq6cfa
f(efp) = 11.163, p-value = 0.09765
```

Both tests reflect only moderate parameter instability across age groups and do not show significant violations of measurement invariance at 5% level.

Illustration



Illustration



Summary

- General score-based test framework for assessing measurement invariance in parametric psychometric models.
- Assessment is along some variable v which can be continuous, ordinal, or categorical.
- Tests can be seen as generalizations of the Lagrange multiplier test.
- Computation of critical values might require simulation from certain stochastic processes (Brownian bridges).
- Easy-to-use implementation available in R package *strucchange*.
- Can be re-used in model-based recursive partitioning.

Acknowledgments: This work was supported by National Science Foundation grant SES-1061334.

References

Merkle EC, Zeileis A (2013). "Tests of Measurement Invariance without Subgroups: A Generalization of Classical Methods." *Psychometrika*, **78**(1), 59–82. `doi:10.1007/s11336-012-9302-4`

Merkle EC, Fan J, Zeileis A (2014). "Testing for Measurement Invariance with Respect to an Ordinal Variable." *Psychometrika*, **79**(4), 569–584. `doi:10.1007/s11336-013-9376-7`

Wang T, Merkle EC, Zeileis A (2014). "Score-Based Tests of Measurement Invariance: Use in Practice." *Frontiers in Psychology*, **5**(438). `doi:10.3389/fpsyg.2014.00438`