Nonlinear structural equation mixture modeling

Holger Brandt

University of Kansas

KU Summer Stats Camp

Structure

Introduction

Linear models: Limitations

- Linear models such as regression models or SEM often assume linear relationships between the variables.
- This assumption is often either wrong or a very crude approximation of the actual functional relationship.
- From an applied perspective, interaction effects are important because they indicate if the relationship between two variables is moderated by a third variable.
- In this workshop, we will discuss a variety of options to analyze deviations from linearity in the latent variable framework.

Why latent variables?

- Latent variable models, or more specifically here: Structural equation models (SEM) allow to consider a variety of aspects in data analysis:
  - They take measurement error into account by using multiple indicators and by specifying relationships between latent variables that are theoretically measurement error free
  - More complex relationships can be analyzed
  - Model fit can be investigated
- Particularly for nonlinear effects, the measurement error aspect is important because a typical effect size in social sciences is around 2.5%.

What is nonlinear?

- Nonlinearity can involve very different aspects:
  - Interaction effects between latent variables
  - Curvilinear relationships between latent variables
  - Nonlinear measurement models, for example for dichotomous or count data
  - Nonlinear parameters (such as λ^2)
- In this workshop, we focus on the first two aspects.

Interaction effects

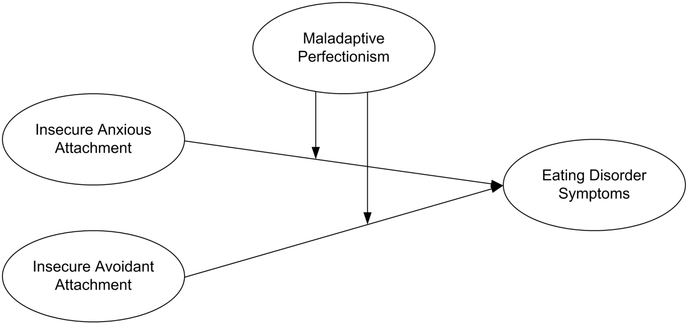


Figure: Taken from Dakanalis et al. (2014)

Quadratic and other polynomial effects

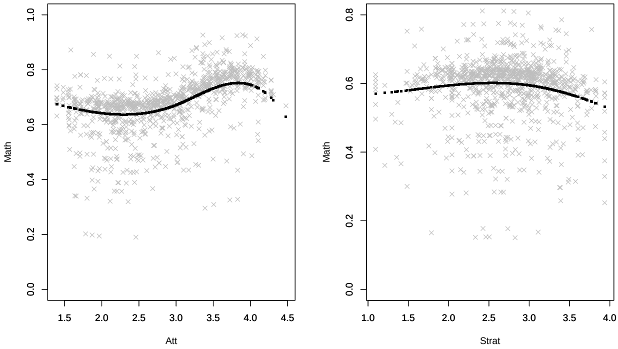


Figure: Taken from Kelava and Brandt (2015). Relationship between between pupils' math skills (Math) and their attitude toward reading (Att; left), and the reported teaching strategies (Strat; right).

Overview workshop

1. Interaction effects: Traditional product indicator approaches
2. Interaction and quadratic effects: Other approaches
3. Mixture models for curvilinear relationships
4. Mixture models for interaction and quadratic effects
5. Multilevel models with interaction effects

Overview sem-3

1. Interaction effects: Traditional product indicator (PI) approaches
   - Constrained approach
   - GAPI approach
   - Unconstrained approach
   - Main package: lavaan
2. Interaction and quadratic effects: Other approaches
   - Moment-based approaches
   - "Distribution-analytic" maximum likelihood approaches (LMS)
   - Standardization and illustration of results
   - Main package: nlsem and experimental syntax for 2SMM

Overview sem-4

1. Structural equation mixture models (SEMM)
    - Direct and indirect applications
    - Growth curve mixture models (direct application)
    - Optional: Alternative models for heterogeneous growth
    - Example for curvilinear relationship (indirect application)
2. Background: EM algorithm
3. Nonlinear structural equation mixture models (NSEMM)
    - Standardization of effects in NSEMM
4. Main software: Mplus, nlsem and plotsemm

Overview sem-5

1. Multilevel models with interaction effects
    - Bayesian modeling (priors, logic of Bayesian modeling)
    - Multilevel models with and without interaction effects
    - Multilevel SEM with and without interaction effects
    - Main packages: rstan and stan

Limitations of *"parametric"* nonlinear effects

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \beta_4 x^2 + \epsilon \quad (1)$$

- Interaction or quadratic effects are modeled with product terms
- Functional forms need to be specified a priori
- Advantage: Straightforward interpretation and standardization (Brandt, Umbach, & Kelava, 2015)
- Limitations:
  - Functional form can be misspecified (e.g., ceiling effect in development of math skill)
  - Identification of such misspecification in latent models hardly detectable (e.g., by model fit)

Mixture Modeling

- Extraction of latent classes with class-specific parameters, for example:

$$\eta_{ic} = \alpha_c + \gamma_{1c}\xi_{ic} + \zeta_{ic} \tag{2}$$

  for persons $i = 1 \dots N$ in latent classes $C = c$
- With $\xi_c \sim N(\kappa_c, \phi_c)$ and $\zeta_c \sim N(0, \psi_c)$
- Here: Linear model with normally distributed variables within each class
- Across classes: Modeling of nonnormal distributions and nonlinear (curvilinear) relationships (Bauer, 2005; Jedidi, Jagpal, & DeSarbo, 1997)

Illustration of approximation of nonnormal distributions

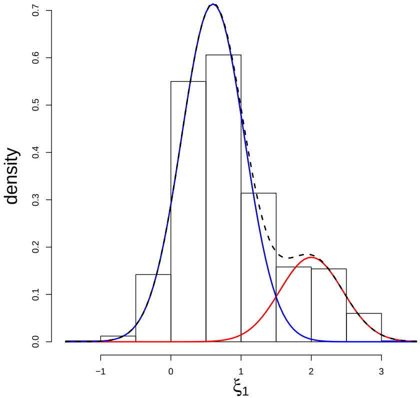


Figure: The mixture model allows to approximate the nonnormal distribution of variables by a (weighted) sum of normal distributions.

Illustration of approximation of a curvilinear relationship

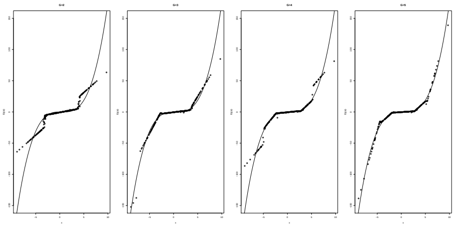


Figure: True (line) and approximated relationship. The mixture model allows to approximate the nonlinear relationships by a (weighted) sum of linear relationships; with more classes the approximation is more precise.

Interpretations of mixture models

- *Direct* application
  - Interpretation of parameters within each class
  - Often: Interpretation of classes as referring to substantive subpopulations
  - Interpretation should be based on more information (e.g., theoretical foundation; covariates that predict classes)
- *Indirect* application
  - Interpretation of parameters across classes
  - No interpretation of classes as referring to substantive subpopulations
  - Statistical tool for nonlinear modeling
- In both cases: *Restrict measurement model* such that latent factors have same interpretation in all classes

Averaging distributions across classes

- Nonnormal distributions are approximated by a mixture of normal distributions with class specific parameters.

→ Mean (κ) and variance (ϕ) of a mixture variable:

$$\kappa = \sum_{c=1}^{C} \pi_c \kappa_c \text{ and } \phi = \sum_{c=1}^{C} \pi_c(\phi_c + \kappa_c^2) - \kappa^2$$

with $\pi_c = P(C = c)$ is the class probability.

- For example: $\kappa_1 = 1, \kappa_2 = 3$, $\pi_1 = \pi_2 = .5$ and $\phi_1 = \phi_2 = 1$

→ $\kappa = 2$, $\phi = 4$

Averaging relationships across classes

- Nonnormal relationships are approximated by a mixture of linear relationships with class specific parameters.
- Two parts
  - Conditional class probability given a specific value in the predictor variable

$$P(C=c|X=x)=\frac{P(C=c)\varphi(X=x|\kappa_c,\phi_c)}{\sum_{c=1}^{C}[P(C=c)\varphi(X=x|\kappa_c,\phi_c)]} \tag{3}$$

    where φ is the pdf with class-specific parameters
  - Weighted average of the class-specific linear effects

$$E(y|X=x)=\sum_{c=1}^{C}[P(C=c|X=x)E(y|X=x,C=c)] \tag{4}$$

Application of mixture models: Heterogeneous growth processes

A model for individual growth

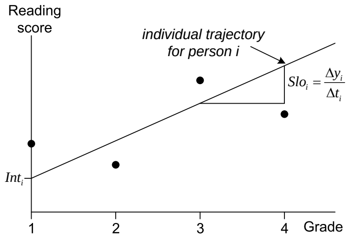


Figure: *Left:* Individual trajectory modeled by an individual intercept (Int_i) and slope (Slo_i). *Right:* Several trajectories that can be described by their means $(E(Int), E(Slo))$ and variances $(V(Int), V(Slo))$.

The Latent Growth Curve Model (LGM)

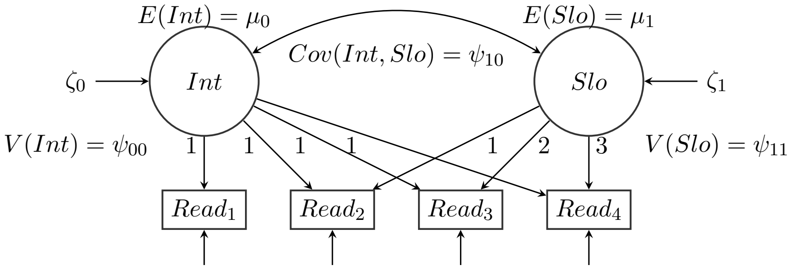
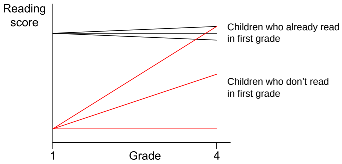


Figure: Path diagram for a simple latent growth curve model with intercept factor (Int) and linear slope factor (Slo).

Research questions relevant in longitudinal data

1. What is the functional form of the growth trajectories?
2. "Correlates of change" (Rogosa & Willett, 1985)
   - Is there a relationship between initial status and growth trajectories?
   - Can baseline covariates explain/predict change?
3. Homogeneity of growth patterns
   - Do different growth patterns exist for subgroups (*latent classes*)?
   - Does the dispersion of the trajectories change across individual starting conditions (*heteroscedasticity*)?



Heterogeneous growth processes

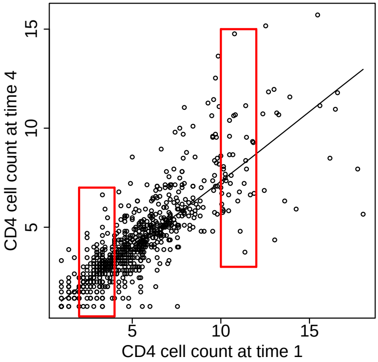


Figure: CD4 cell counts at 2 measurement occasions. The (conditional) variance of the CD4 cell counts at time 4 changes across different initial CD4 cell counts at time 1.

→ Prediction for subjects with low initial CD4 cell counts should be more precise.

Different models to represent growth processes

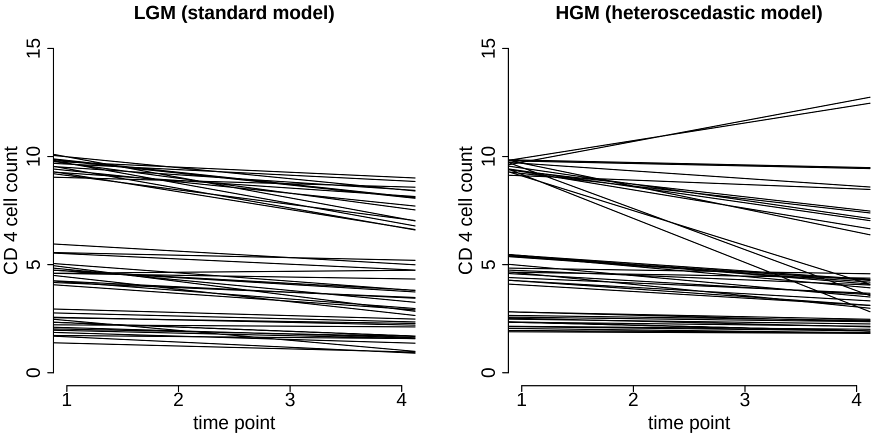
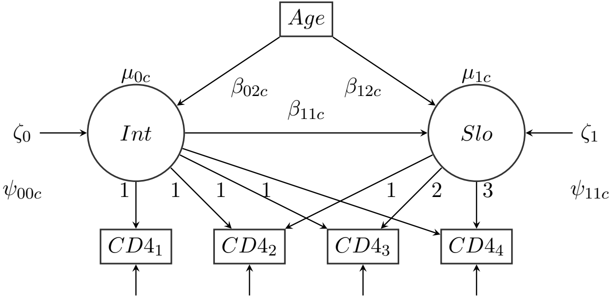


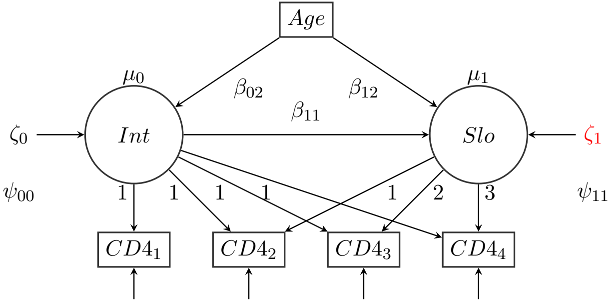
Figure: Estimated growth trajectories using an LGM or a heterogeneous growth curve model (HGM).

Growth curve mixture model (GMM)



Heteroskedasticity is modeled with class-specific paramaters (indicated with c

Heterogeneous growth curve model (HGM) (Brandt & Klein, 2015; Klein & Muthén, 2006)



Heteroskedastic variance function for the slope factor:

$$\zeta_1 = (\gamma_0 + \gamma_1 Int + \gamma_2 Age)\zeta_2 + \zeta_3 \quad (5)$$

$$\rightarrow Var(Slo|Int, Age) = \underbrace{(\gamma_0 + \gamma_1 Int + \gamma_2 Age)^2}_{\text{heteroscedastic}} + \underbrace{\psi_{33}}_{\text{homoscedastic}} \quad (6)$$

Property of the HGM

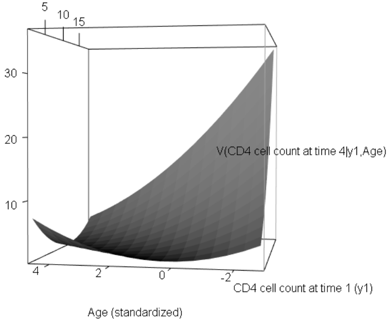


Figure: Heteroscedastic variance function for the CD4 cell counts at time 4 given those at time 1 (y_1) and patients' age (Brandt & Klein, 2015).

Exercise I

In this exercise, a growth curve mixture model is used to investigate the development of CD4 cells in HIV patients. We specify a model with different number of classes and try to extract and interpret subgroups of the patients.
Further, we compare the model to a latent growth curve model with heteroskedastic residuals (HGM).

Exercise II

In this exercise, a semiparametric mixture model is investigated in order to illustrate the functional form of the relationship between the dependent variable Reading skills and the two predictor variables Reading Attitude and Online Activities.

For the model it is assumed that the measurement models are equal across classes. The structural model, however, is class-specific with linear relationships between the variables. For this example, we restrict the analysis to two latent classes.

Exercise II

1. Conceptualize the model: Write down the equations and count the parameters. Which parameters are class-specific?
2. Complete Mplus input file.
   - Specify the measurement and structural models
   - For class #2, specify the measurement model, the structural model and latent means and variances, if necessary. Use parameter labels to constrain parameters across classes or to produce class-specific estimates.
3. How many classes do you need?
4. Interpret the parameter estimates.
5. Visualize the results using the R syntax.

Background: The EM algorithm

- The expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is general method to produce maximum likelihood estimates in situation with missing information.
- Missing information can occur in many scenarios, for example
  - Missing data (e.g., EM imputation)
  - Truncated data
  - Missing class-membership information (such as in mixture modeling)
- Furthermore, the EM algorithm is the basis of LMS (see yesterday).

Basic principles of the EM algorithm

- The EM algorithm consists of two steps that it cycles until convergence:
  1. E-step: missing data is imputed. A data set is generated as it is **expected** given the parameters (initial or from previous iteration) and data
  2. M-step: given the completed data set the parameters are estimated using a maximization function (which can look very different depending on the application)
- For SEMM, for example, the M-step involves a standard ML-fitting function for SEM.

The principle of the EM algorithm in LMS and SEMM

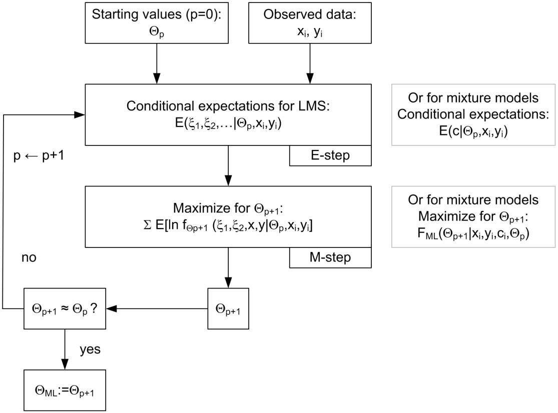


Figure: Based on Klein et al. (1997).

Selection of starting values

- The EM algorithm needs an input of starting values Θ_0
- For SEMM, final results might depend on these starting values and provide a local instead of a global optimum.
- As a consequence, several sets of (random) starting values should be chosen, and the model estimated repeatedly, to ensure that the final solution is global. [Note: this procedure is not fool-proof: a local optima can be found repeatedly, too. But there is no general way to find global optima for mixture models.]

Characteristics of Structural Equation Mixture Modeling

- Semi-parametric estimation of curvilinear relationships with mixture models allows one to identify curvilinear effects
- Disadvantages:
  - Effects remain descriptive
  - No significance test of nonlinearity because nonnormality and nonnormality are confounded (information criterion refer to overall model, and do not allow significance test)
  - Type of curvilinearity is unspecified
  - No standardization for effects possible

Nonlinear Structural Equation Mixture Modeling (NSEMM)

- Combination of mixture models and parametric nonlinear effects

$$\eta_{ic} = \alpha + \gamma_1\xi_{1ic} + \gamma_2\xi_{2ic} + \gamma_3\xi_{1ic}\xi_{2ic} + \gamma_4\xi_{1ic}^2 + \gamma_5\xi_{2ic}^2 + \zeta_{ic} \quad (7)$$

  for persons $i = 1 \ldots N$ in latent classes $C = c$
- With $\xi_c \sim N(\kappa_c, \phi_c)$ and $\zeta_c \sim N(0, \psi)$
- Predictor variables can be nonnormally distributed
- Regression coefficients are constrained across latent classe
  - Interpretation straightforward
  - Significance test
  - Type of curvilinearity is specified
  - Standardization for effects possible

NSEMM: Estimation

- For estimation, the overall likelihood is derived as a mixture of densities

$$L = \prod_i \left(\sum_c \pi_c f_c \left((\mathbf{x}', \mathbf{y}')_i' \right) \right) \tag{8}$$

  (see Kelava, Nagengast, & Brandt, 2014) with mixture component weights $\pi_c = P(C = c)$.
- f_c is a complicated density function (see LMS in sem-3) that can be approximated by ML (e.g., Klein & Moosbrugger, 2000) or quasi ML (e.g., Klein & Muthén, 2007) within each mixture.
- The overall likelihood can be approximated by applying numerical methods such as the EM algorithm.
- If ML is, NSEMM is estimated using two nested EM cycles (which makes it very time consuming).

Exercise III

In the last part of the exercise, a semiparametric nonlinear mixture model is used in order to approximate a potentially nonnormal distribution of the latent predictor variables. Here, we introduce an interaction effect and two quadratic effects into the model.

For the model it is again assumed that the measurement models are equal across classes. The structural model is restricted to be the same across classes such that the interpretation of the regression coefficients is straightforward across classes. The means and variances of the latent predictors finally have class-specific estimates.

Exercise III

1. How many parameters are class-specific?
2. How many classes are necessary?
    - For class #2, specify the measurement model, the structural model and latent means and variances, if necessary. Use parameter labels to constrain parameters across classes or to produce class-specific estimates.
3. Interpret the parameter estimates. Compare the results to the first model (Part II). Can you give evidence if the mixture model is necessary?
4. Use the R syntax.
    - Use the standardization routine. How much of the reading skills' variance can the nonlinear model explain?
    - Illustrate the results using simple slopes or the 3d graph.
    - Illustrate the nonnormal distribution of the predictor variables using a 3d graph.

Conclusions

- Nonlinear SEM include a variety of possibilities.
- Decision upon model based on data features (e.g., distribution) and scope of analysis (identify curvilinearity vs. specific effects).
- Standardization and estimation of effects needs special attention to nonnormality (and means) of the latent predictor variables.
- Even more flexibility can be achieved by using Bayesian Modeling and/or splines.

Model fit in nonlinear SEM

- Model fit evaluation in nonlinear SEM is complicated
- There is no global model fit index as in linear SEM
- There are some specific tests to enhance model fit:
  - Model comparisons to explicitly defined models
  - Tests to assess homoskedasticity that might indicate omitted nonlinear effects
  - Modification indices from misspecified linear models that may (or may not) indicate misspecification in the measurement model

KU

Model fit in linear SEM

- The chi^2 is based on a comparison of the sampling distribution of the model-implied covariance matrix of the target model vs. the one of a saturated model.
- This sampling distribution is called Wishart distribution (see Hayduk, 1989, pp. 136-138):

$$W(S, \Sigma, n) = \exp(-n/2tr(S\Sigma^{-1})) \cdot det(\Sigma)^{-n/2} \cdot C \tag{9}$$

  where S and Σ are the sample and population covariance matrices, $n = N - 1$ refers to the sample size, and C is a constant. tr and det are the trace and determinant of a matrix.

Model fit in linear SEM

- For a saturated model, we plug in the actual sample covariance matrix S as the population covariance matrix.
- For the target model, we use the model-implied covariance matrix $\Sigma(\theta)$
- A comparison of both is then called likelihood ratio and is given by

$$LR = \frac{W(S, \Sigma(\theta), n)}{W(S, S, n)} \tag{10}$$

$$= \frac{\exp(-\frac{n}{2}tr(S\Sigma(\theta)^{-1})) \cdot det(\Sigma(\theta))^{-\frac{n}{2}} \cdot C}{\exp(-\frac{n}{2}tr(SS^{-1})) \cdot det(S)^{-\frac{n}{2}} \cdot C} \tag{11}$$

and simplifies to

$$LR = \exp(-\frac{n}{2}tr(S\Sigma(\theta)^{-1}))det(\Sigma(\theta))^{-\frac{n}{2}} \exp(-\frac{n}{2}tr(SS^{-1}))det(S)^{-\frac{n}{2}} \tag{12}$$

Model fit in linear SEM

- Taking the natural logarithm of this equation then yields :

$$\log LR = -\frac{n}{2} tr(S\Sigma(\theta)^{-1}) \\ - \frac{n}{2}\log(det(\Sigma(\theta))) - \frac{n}{2} tr(SS^{-1}) - \frac{n}{2}\log(det(S)) \tag{13}$$

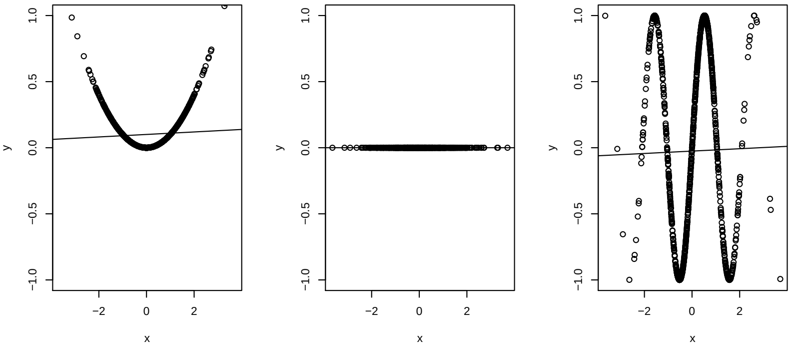
which can be maximized by maximizing

$$F = tr(S\Sigma(\theta)^{-1}) + \log(det(\Sigma(\theta))) + p + \log(det(S)) \tag{14}$$

which is the well know fitting function for SEM.

Model fit in nonlinear SEM

- The problem for nonlinear SEM now comes with the fact that the sampling covariance is not a sufficient statistic. It averages across all persons and any nonlinearity does not show up in the covariances.
- Three scenarios with $cov(x, y) \approx 0$:



Model fit in nonlinear SEM

- As a consequence the χ^2 test is not appropriate as a global test statistic
- Corrections for the χ^2 test (e.g., Satorra-Bentler) do not attenuate the problem because although they correct for nonnormality they still depend on the covariance matrices for the likelihood ratio
- All other tests based on the χ^2 test do not indicate model misfit either (e.g., RMSEA, CFI, TLI).

Model fit in nonlinear SEM: Alternatives

- **Specific omitted nonlinear effects** can be tested by comparing nested models using χ^2 difference tests (or their robust alternatives)
- For example: Model including a linear and cubic effect vs. a model with a linear effect only.
- This procedure depends on the specified comparison models. There is no possibility to guarantee that the tested models are the best models (e.g., a curvilinear relationship can be described by a quadratic effect but might actually follow an exponential effect)
- For testing interaction, Gerhard et al. (2015) propose use a model including quadratic and interaction effects as a comparison model instead of a linear model to reduce spurious results. The artifacts can occur because quadratic and interaction effects can be correlated.

Model fit in nonlinear SEM: Alternatives

- **Unspecific omitted nonlinear effects** can be tested by using tests for heteroskedasticity (e.g., Gerhard et al., 2017; Klein et al., 2016; Nestler, 2015).
- All these test work similarly and test if the squared residuals of the dependent variable (e.g., of η) follow a homoskedastic or heteroskedastic distribution.
- Significance of the test can imply that some kind of nonlinear effect was not included in the model
- Initial simulations show that the test has good power and type I error rates in situations with stronger effects or normal data.

Model fit in nonlinear SEM: Alternatives

- **Misspecified measurement models** can be investigated using misspecified linear models.
- As long as the misspecification in the measurement model is linear, for example cross-loadings or residual correlations, they show up in – for example – modification indices.
- The logic behind modification indices is similar to nested comparison models above. However, nonlinear misspecifications or completely misspecified models (e.g., wrong number of factors) do not show up in modification indices.

Model fit in mixture models

- Assessing model fit in mixture models is even more complicated.
- **Decision on the number of classes** is typically based on information criteria:
  - BIC: is more conservative and is often proposed for direct applications
  - AIC: is more liberal and is often proposed for indirect applications (it might overestimate the number of classes)
- In general, models with different numbers of classes cannot be assumed to be nested within each other (McLachlin & Peel, 2001). Hence, model difference test cannot be applied.

Model fit in mixture models

- Up to the present, the SEMM approach (or any other semi-/non-parametric approach, to my knowledge) does not allow to identify the specific form of a curvilinear relationship (e.g., quadratic or cubic).
- For example, even bootstrapped confidence interval bands are difficult to use in order to test if an effect is curvilinear.
- There is no significance test for nonlinearity in SEMM. Tthis is why NSEMM is an alternative to test specific nonlinear tests.

Summary

- Mixture models can be used to approximate
  - Nonlinear relationships
  - Nonnormal distributions
- Specific care needs to be taken for
  - How many classes? Which parameters are class-specific?
  - Indirect vs. direct interpretation
  - Standardization of nonlinear effects
- Model fit investigation for nonlinear models is complicated
- Tomorrow: Bayesian modeling.

Thank you for your attention.

Bauer, D. J. (2005). A semiparametric approach to modeling nonlinear relations among latent variables. *Structural Equation Modeling*, *12*, 513–535.

Brandt, H., & Klein, A. G. (2015). A heterogeneous growth curve model for non-normal data. *Multivariate Behavioral Research*, *50*, 416–435.

Brandt, H., Umbach, N., & Kelava, A. (2015). The standardization of linear and nonlinear effects in direct and indirect applications of structural equation mixture models for normal and nonnormal data. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.01813

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *Ser. B, 39*, 1–38.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response based segmentation and unobserved heterogeneity. *Marketing Science*, *16*, 39–59.

Kelava, A., Nagengast, B., & Brandt, H. (2014). A nonlinear structural equation mixture modeling approach for non-normally distributed latent predictor variables. *Structural Equation Modeling*, *21*, 468–481.

Klein, A. G., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*, 457–474.

Klein, A. G., & Muthén, B. O. (2006). Modeling heterogeneity of latent growth depending on initial status. *Journal of Educational and Behavioral Statistics*, *31*, 357–375.

Klein, A. G., & Muthén, B. O. (2007). Quasi maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research*, *42*, 647–674.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, *50*, 203–228.