

# Score-Based Approaches to Heterogeneity in Psychological Models

## DISSERTATION

zur Erlangung des akademischen Grades  
Doctor rerum naturalium (Dr. rer. nat.)

eingereicht an der  
Lebenswissenschaftlichen Fakultät der  
Humboldt-Universität zu Berlin

von  
Manuel Arnold (M.Sc. Psychologie, M.Sc. Statistik)

Präsidentin  
der Humboldt-Universität zu Berlin  
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät  
der Humboldt-Universität zu Berlin  
Prof. Dr. Dr. Christian Ulrichs

Gutachter:

1. Prof. Dr. Manuel Völkle
2. Prof. Dr. Ulman Lindenberger
3. Prof. Dr. Ross Jacobucci

Tag der mündlichen Prüfung: 14.12.2021

## **Selbständigkeitserklärung**

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben.

Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad.

Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde.

Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin vom 5. März 2015.

Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsbearbeiterinnen/Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

Berlin, 30.08.2021

## Summary

Statistical models of human cognition and behavior often rely on data that have been aggregated across study participants and fail to consider heterogeneity, that is, differences across individuals or groups. If overlooked, heterogeneity can bias parameter estimates and may lead to false-positive or false-negative findings. Often, heterogeneity can be detected and predicted with the help of covariates, such as demographic variables, personality traits, or biomarkers. However, identifying predictors of heterogeneity may prove to be a challenging task, especially in situations with a large number of candidate covariates and vague hypotheses about the sources of heterogeneity. To solve this issue, I propose two novel approaches for detecting and predicting individual and group differences with covariates in a wide range of models used in contemporary psychological research.

The theoretical foundation of the work presented in this dissertation is the case-wise partial derivative of the log-likelihood function with respect to the model parameters. This derivative is also referred to as the score function. Analyses of the score function have a long history in hypothesis testing and model modification. One important advantage of many score-based approaches is their computational efficiency.

This cumulative dissertation is composed of three projects. Project 1 advances the individual parameter contribution (IPC) regression framework. IPC regression allows studying heterogeneity in structural equation model (SEM) parameters by regressing them on a set of covariates. By means of a Monte Carlo simulation study, I evaluate the use of IPC regression for dynamic panel models and show that IPC regression is a promising method for detecting differences in the stability and interrelationships of processes. Moreover, I demonstrate that the estimates provided by IPC regression can be biased and, as a remedy, develop a bias correction procedure. As a contribution on a theoretical level, I derive IPCs for general maximum likelihood estimators, which opens the framework for other model classes beyond SEMs. Project 2 illustrates how IPC regression can be used in practice. To this end, I provide a step-by-step introduction to the IPC regression implementation in the **ipcr** package for the R system for statistical computing. Finally, Project 3 progresses the SEM tree framework. SEM trees are a model-based recursive partitioning method for finding covariates that predict group differences in SEM parameters. A SEM tree is a data-driven method that divides a data set recursively into homogeneous subsets. The original SEM tree implementation uses a likelihood-ratio criterion to search for predictive covariates, which is computationally demanding. As a solution to this problem, I combine SEM trees with a family of score-based tests that have been recently popularized in psychometrics. The resulting score-guided SEM trees compute quickly, solving the runtime issues of the likelihood-ratio-guided SEM trees, and show favorable statistical properties in a Monte Carlo simulation.

## Zusammenfassung

Statistische Modelle menschlicher Kognition und Verhaltens stützen sich häufig auf aggregierte Daten der Studienteilnehmenden und vernachlässigen dadurch oft Heterogenität, das heißt Unterschiede zwischen Personen oder Gruppen. Die Nichtberücksichtigung vorliegender Heterogenität kann zu verzerrten Parameterschätzungen und zu falsch positiven oder falsch negativen Tests führen. Häufig kann Heterogenität mithilfe von Kovariaten wie demografischen Variablen, Persönlichkeitsmerkmalen oder Biomarkern erkannt und vorhergesagt werden. Allerdings erweist sich die Identifizierung von Prädiktoren von Heterogenität oft als schwierige Aufgabe, insbesondere wenn viele potenzielle Prädiktoren vorliegen und Hypothesen über die Ursache der Heterogenität vage sind. Zur Lösung dieses Problems schlage ich zwei neue Ansätze vor, die individuelle und gruppenspezifische Unterschiede mithilfe von Kovariaten vorhersagen. Beide Ansätze können auf eine Vielzahl von Modellen angewendet werden, die in der aktuellen psychologischen Forschung gebräuchlich sind.

Die theoretische Grundlage dieser Dissertation stellt die fallweise partielle Ableitung der Log-Likelihood-Funktion hinsichtlich der Modellparameter dar. Diese Ableitung wird auch als Score-Funktion bezeichnet. Die Score-Funktion wird seit langem zur Testung von statistischen Hypothesen und zur Modifizierung von Modellen eingesetzt. Eine wichtige Eigenschaft vieler Score-basierter Ansätze ist ihre geringe Laufzeit.

Die vorliegende kumulative Dissertation setzt sich aus drei Projekten zusammen. Projekt 1 widmet sich dem Verfahren IPC-Regression (Individual Parameter Contribution). IPC-Regression ermöglicht die Exploration von Parameterheterogenität in Strukturgleichungsmodellen (SEM), indem Modellparameter auf Kovariaten regrediert werden. Mittels einer Monte-Carlo-Simulationsstudie evaluiere ich den Nutzen von IPC-Regression für die Exploration von Heterogenität in dynamische Panelmodelle und zeige, dass IPC-Regression eine vielversprechende Methode zur Schätzung von Unterschieden in autoregressiven und kreuzverzögerten Parametern ist. Weiterhin demonstriere ich, dass die IPC-Regressionschätzer unter bestimmten Bedingungen verzerrt sein können und entwickle dafür ein Korrekturverfahren. Als theoretischen Beitrag leite ich IPCs für allgemeine Maximum-Likelihood-Schätzer her, wodurch der Einsatz der Methode auf weitere Modellklassen ermöglicht wird. Projekt 2 veranschaulicht, wie IPC-Regression in der Praxis eingesetzt werden kann. Dazu führe ich schrittweise in die Implementierung von IPC-Regression im **ipcr**-Paket für die statistische Programmiersprache R ein. Schließlich werden in Projekt 3 SEM-Trees weiterentwickelt. SEM-Trees sind eine modellbasierte rekursive Partitionierungsmethode und finden datengeleitete Kovariaten, die Gruppenunterschiede in Parameterwerten eines SEM vorhersagen. Dabei unterteilt der SEM-Tree den Datensatz in homogene Gruppen. Bisher verwenden SEM-Trees ein Likelihood-Ratio als Kriterium, um

nach prädiktiven Kovariaten zu suchen. Dieses Vorgehen ist jedoch sehr rechenaufwändig. In Projekt 3 kombiniere ich SEM-Trees mit unterschiedlichen Score-basierten Tests, die in den letzten Jahren in der Psychometrie Verbreitung gefunden haben. Die daraus resultierenden Score-Guided-SEM-Trees lassen sich deutlich schneller als SEM-Trees berechnen, die das Likelihood-Ratio-Kriterium verwenden. Zudem zeigen Score-Guided-SEM-Trees bessere statistische Eigenschaften als die herkömmlichen SEM-Trees.

## Acknowledgments

Many people have contributed to this dissertation in various ways. I am especially grateful to ...

- my advisors and co-authors Manuel Voelkle and Andreas Brandmaier,
- my other co-author Daniel Oberski,
- my colleagues at the chair of Psychological Research Methods of Humboldt University of Berlin and the International Max-Planck Research School on Computational Methods in Psychiatry and Ageing Research,
- and my family.

## Abbreviations

CFI	comparative fit index
EPC	expected parameter change
IPC	individual parameter contribution
LASSO	least absolute shrinkage and selection operator
MGSEM	multigroup structural equation model
MI	modification index
RMSEA	root mean square error of approximation
SEM	structural equation model
SRMR	standardized root mean square residual

## Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Heterogeneity and Structural Equation Models</b>	<b>4</b>
2.1	Structural Equation Modeling . . . . .	4
2.2	Model Fit Evaluation . . . . .	6
2.3	Effects of Heterogeneity . . . . .	8
2.4	Methods for Addressing Heterogeneity . . . . .	9
2.5	Individual Parameter Contribution Regression and Score-Guided Structural Equation Model Trees . . . . .	12
<b>3</b>	<b>Score-Based Approaches to Model Evaluation</b>	<b>13</b>
3.1	The Score Function . . . . .	13
3.2	Score-Based Model Checking . . . . .	16
3.3	Individual Scores and Parameter Heterogeneity . . . . .	19
<b>4</b>	<b>Identifying Heterogeneity in Dynamic Panel Models with Individual Parameter Contribution Regression</b>	<b>21</b>
<b>5</b>	<b>Predicting Differences in Model Parameters with Individual Parameter Contribution Regression Using the R Package ipcr</b>	<b>23</b>
<b>6</b>	<b>Score-Guided Structural Equation Model Trees</b>	<b>24</b>
<b>7</b>	<b>Discussion</b>	<b>26</b>
7.1	Comparison of Methods . . . . .	26
7.1.1	Individual Parameter Contribution Regression and Score-Guided Structural Equation Model Trees . . . . .	26
7.1.2	Established Methods . . . . .	28
7.2	Applications and Caveats . . . . .	29
7.2.1	Model Modification . . . . .	29
7.2.2	Causal Interpretation . . . . .	32
7.2.3	Predicting Interindividual Differences in Intraindividual Variation . . . . .	33
7.2.4	Latent Covariates . . . . .	34
7.3	Areas of Applications . . . . .	34
7.4	Conclusion and Outlook . . . . .	35
<b>8</b>	<b>References</b>	<b>36</b>





## 1 Introduction

Heterogeneity is ubiquitous in psychological and social research. Heterogeneity can be defined as differences across individuals or groups that influences the outcome and may correlate with characteristics of primary scientific interest. While heterogeneity may have various causes and manifestations, it always poses the same problem: if overlooked or ignored, heterogeneity can bias and invalidate statistical analyses, leading to incorrect conclusions derived from the sample. To avoid this problem, researchers face the challenging task of determining whether there is a substantially relevant amount of heterogeneity in the sample and, if so, must account for this variability in the statistical model.

Heterogeneity can be observed or unobserved (e.g., Lubke & Muthén, 2005). If the sources of heterogeneity are observed, differences across individuals or groups can be addressed by incorporating covariates that act as moderator variables into the model. For instance, it may be well known that the effectiveness of a type of psychotherapy differs between female and male patients. A researcher interested in the effectiveness of the therapy for a certain psychological disorder may then account for this source of heterogeneity by using gender as a grouping variable that predicts the differences between females and males. Conversely, unobserved heterogeneity occurs when the sources of heterogeneity in the data are not known beforehand. However, some of the unobserved heterogeneity can often be explained with the help of covariates.

Explaining heterogeneity with observed covariates often appears to be a promising endeavor since many data sets studied in psychological research are vast and contain numerous variables. The question remains how researchers can decide between variables stemming from potentially multiple questionnaires, genetic data, and biomarkers in situations when hypotheses about the potential causes of heterogeneity are vague. Which of these covariates could possibly reduce heterogeneity and improve the predictive performance of the model? Not only is testing a large number of covariates labor-intensive, but it also heightens the risk of overfitting the model by including too many covariates (e.g., Yarkoni & Westfall, 2017). Such an overfitted model may perform well in the sample at hand but generalizes poorly to new data.

The complexity of the model poses another challenge for the identification of heterogeneity with covariates. A widely used statistical technique to fit complex networks is structural equation modeling. Structural equation models (SEMs; see Bollen, 1989) generalize diverse statistical approaches such as the  $t$ -test, analysis of variance, linear regression, and factor analysis models. Typically, SEMs consist of many different parameters. Some of these parameters may describe how a latent construct is derived from directly observable variables, whereas others indicate relationships among distinct constructs. Thus, it appears highly improbable that a source of heterogeneity affects all parameters of a

## 1. Introduction

SEM uniformly, rather than just certain parts. Therefore, researchers need to determine in which parts of their model heterogeneity can be accounted for with a covariate and which parts are not influenced. Like overfitting a model by including too many covariates, complex models entail the risk of overfitting a model due to specification errors, that is, specifying superfluous ways a covariate influences parts of the model.

The work summarized in this dissertation aims to improve the treatment of heterogeneity in contemporary psychological models, particularly in SEMs. In the first line of work (Project 1 and 2), I advance a recently proposed method called individual parameter contribution (IPC) regression to model heterogeneity in parameters as a function of covariates. IPC regression stands out from other methods addressing heterogeneity in SEMs due to its flexibility and computational efficiency. IPC regression allows testing and estimating if and how a parameter changes with respect to a discrete or a continuous covariate. The method encompasses every type of SEM parameter and can also be used to study linear regression models, generalized models, and mixed models. IPC regression was first introduced by Oberski (2013). Among other things, I contribute to the method by providing a software implementation, investigate its finite-sample properties for different models with simulation studies, and propose two alternative versions of IPC regression that either exhibit less bias or minimize the risk of overfitting. In the second line of work (Project 3), I advance the SEM tree framework by developing score-guided SEM trees. SEM trees are a model-based recursive partitioning method (Zeileis et al., 2008) that finds covariates to predict group differences in the parameters of a SEM by growing tree structures that divide a data set recursively into homogeneous subsets. The original SEM trees put forward by Brandmaier et al. (2013) use a likelihood ratio criterium to select splits. However, this likelihood-ratio-based split search procedure requires the estimation of many SEMs, making it computationally demanding. Therefore, I propose an alternative split search based on the score function, which is computationally more efficient. In addition to the runtime improvement, I also compare the performance of different likelihood-ratio-guided and score-guided SEM trees in a simulation study.

IPC regression and score-guided SEM trees are based on a common statistical foundation. Both methods make use of the so-called score function to assess heterogeneity. The score function is defined as the partial derivative or gradient of the log-likelihood function with respect to the model parameters. Analysis of the score function has a long history in the assessment of parameter invariance (Zeileis & Hornik, 2007). In econometrics, the score function is often studied to detect change points in time series. More recently, a family of score-based tests has been proposed to uncover heterogeneity in factor analysis models (Merkle et al., 2014; Merkle & Zeileis, 2013). An important advantage of most score-based approaches is their computational efficiency.

The outline of this dissertation is as follows. Chapter 2 gives an overview of SEMs and highlights the consequences of heterogeneity on SEMs. Chapter 3 lays the theoretical

## 1. Introduction

foundation for IPC regression and score-guided SEM trees by introducing the score function. Chapter 4, 5, and 6 summarize the individual projects encompassing this cumulative dissertation, which are reprinted in the Appendix. Finally, Chapter 7 concludes this dissertation with a general discussion, featuring a detailed comparison of IPC regression and score-guided SEM trees and covers their strengths, limitations, areas of applications, and potential future developments.

## 2 Heterogeneity and Structural Equation Models

IPC regression and score-guided SEM trees aim primarily at uncovering and addressing heterogeneity in SEMs. The following chapter provides an introduction to SEMs, which complements the rather concise treatments of SEMs in the individual projects. I will first give an overview of SEMs. Then, I highlight that the predominant methods to evaluate SEMs are not suited to detect heterogeneity. Afterwards, I discuss the consequences of heterogeneity for SEMs and summarize popular approaches to address heterogeneity in SEMs.

### 2.1 Structural Equation Modeling

Structural equation modeling is a widely applied statistical method in psychology and the social sciences. Diverse methods such as linear regression models, confirmatory factor analysis models, errors-in-variables models, simultaneous equation models, growth curve models, and dynamic panel models can be specified within the SEM framework. SEMs allow the joint analysis of observed and latent variables and their interrelations and give ways to account for measurement errors. Moreover, SEMs provide a unified and comprehensive approach to test hypotheses on these interrelations and permit complex inferences on multivariate, correlational data. A large body of work has been published about SEMs. The classic textbook of Bollen (1989) remains the conventional reference. Hoyle (2012) covers more recent developments. Kline (2016) provides a non-technical introduction to SEMs. Yuan and Bentler (2006) give a more concise treatment of SEMs well suited for mathematically inclined readers.

SEMs are most conveniently described in terms of mean and covariance structures. Let  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  denote the mean vector and covariance matrix of an unknown population. A SEM implies a certain structure, where the mean vector and covariance matrix are expressed as functions of a vector of unknown model parameters  $\boldsymbol{\theta}$ . The model is estimated by minimizing a fitting function  $F$  that measures the discrepancy between the model-implied mean vector  $\boldsymbol{\mu}(\boldsymbol{\theta})$  and the model-implied covariance matrix  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  and the corresponding sample equivalents, that is the observed sample mean  $\bar{\mathbf{y}}$  and the observed covariance matrix  $\mathbf{S}$ . In many SEM applications, the mean structure is not of particular interest. In these situations, the sample mean is simply ignored and SEMs are also called covariance structure models.

A SEM is specified by formulating a system of linear equations between a set of variables. Conventionally, these equations are grouped into a measurement part defined in Equation 2.1 and a structural part shown in Equation 2.2. Given a sample of size  $N$ , a

## 2. Heterogeneity and Structural Equation Models

SEM can be defined as follows:

$$\mathbf{y}_i = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \quad (2.1)$$

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i, \quad i = 1, \dots, N \quad (2.2)$$

The measurement part relates the  $p$ -variate vector of observed variables  $\mathbf{y}_i$  from individual  $i$  to the  $m$ -variate vector of latent variables  $\boldsymbol{\eta}_i$  weighted by the factor loading matrix  $\boldsymbol{\Lambda}$  of dimension  $p \times m$ . The  $p$ -variate vector  $\boldsymbol{\tau}$  contains the intercepts of the measurement part. The  $p$ -variate random vector  $\boldsymbol{\varepsilon}_i$  represents measurement errors or the unexplained, unique part of the observed variables and is assumed to be independent and identically normally distributed with mean zero and covariance matrix  $\boldsymbol{\Psi}$ , that is,  $\boldsymbol{\varepsilon}_i \stackrel{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ . In the structural part, the latent variables  $\boldsymbol{\eta}_i$  are related to each other via the  $m \times m$  matrix  $\mathbf{B}$  of regression coefficients. The  $m$ -variate vector  $\boldsymbol{\alpha}$  contains the intercepts of the structural part. The  $m$ -variate random vector  $\boldsymbol{\zeta}_i$  contains the error terms of the latent variables and is assumed to be independent and identically normally distributed with mean zero and covariance matrix  $\boldsymbol{\Phi}$ , that is,  $\boldsymbol{\zeta}_i \stackrel{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi})$ . It is further assumed that  $\boldsymbol{\varepsilon}_i$  and  $\boldsymbol{\zeta}_i$  are uncorrelated.

The vectors  $\boldsymbol{\tau}$  and  $\boldsymbol{\alpha}$  and the matrices  $\boldsymbol{\Lambda}$ ,  $\boldsymbol{\Psi}$ ,  $\mathbf{A}$ , and  $\boldsymbol{\Phi}$  contain the parameters of the model. Any entry of these objects can either be a free parameter or may be set to a fixed value. Furthermore, some of the free parameters may be constrained in certain ways, for instance, to have the same values. The free parameters are stored in the  $q$ -variate parameter vector  $\boldsymbol{\theta}$ . Unlike other statistical methods, such as linear regression, SEMs are not automatically identified.<sup>1</sup> An identified SEM provides sufficient data points in the observed mean vector  $\bar{\mathbf{y}}$  and the observed covariance matrix  $\mathbf{S}$  to establish a unique solution to infer each free parameter from the data. Identification can be a challenging task. Bollen (1989) provides a list of necessary or sufficient conditions for model identification.

In order to derive the model-implied mean and covariance structures, Equation 2.2 has to be rearranged so that the latent variables  $\boldsymbol{\eta}_i$  only appear on one side of the equality sign. Given an identity matrix  $\mathbf{I}_{m \times m}$  of order  $m$  and assuming that  $\mathbf{I}_{m \times m} - \mathbf{B}$  is non-singular, the so-called reduced form equation can be derived as follows:

$$\begin{aligned} \boldsymbol{\eta}_i &= \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i \\ \boldsymbol{\eta}_i - \mathbf{B}\boldsymbol{\eta}_i &= \boldsymbol{\alpha} + \boldsymbol{\zeta}_i \\ (\mathbf{I}_{m \times m} - \mathbf{B})\boldsymbol{\eta}_i &= \boldsymbol{\alpha} + \boldsymbol{\zeta}_i \\ \boldsymbol{\eta}_i &= (\mathbf{I}_{m \times m} - \mathbf{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\zeta}_i) \end{aligned} \quad (2.3)$$

---

<sup>1</sup>However, one can argue that linear regression models are not identified, when the number of predictors is larger than the number of data points.

## 2. Heterogeneity and Structural Equation Models

Next, we substitute the reduced form into the measurement part in Equation 2.1:

$$\mathbf{y}_i = \boldsymbol{\tau} + \boldsymbol{\Lambda} (\mathbf{I}_{m \times m} - \mathbf{B})^{-1} (\boldsymbol{\alpha} + \boldsymbol{\zeta}_i) + \boldsymbol{\varepsilon}_i \quad (2.4)$$

Finally, the mean and covariance structures are obtained by calculating the expected value and the variance-covariance matrix of the observed variables:

$$\mathbb{E}(\mathbf{y}_i) = \boldsymbol{\tau} + \boldsymbol{\Lambda} (\mathbf{I}_{m \times m} - \mathbf{B})^{-1} \boldsymbol{\alpha} = \boldsymbol{\mu}(\boldsymbol{\theta}) \quad (2.5)$$

$$\text{Cov}(\mathbf{y}_i, \mathbf{y}_i) = \boldsymbol{\Lambda} (\mathbf{I}_{m \times m} - \mathbf{B})^{-1} \boldsymbol{\Phi} [(\mathbf{I}_{m \times m} - \mathbf{B})^{-1}]^\top \boldsymbol{\Lambda}^\top + \boldsymbol{\Psi} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) \quad (2.6)$$

Since the right-hand side of Equation 2.4 is a linear combination of the normally distributed random vectors  $\boldsymbol{\varepsilon}_i$  and  $\boldsymbol{\zeta}_i$ , the distribution of the observed variables  $\mathbf{y}_i$  can be expressed in terms of SEM matrices as follows:

$$\mathbf{y}_i \stackrel{iid}{\sim} \mathcal{N} \left( \boldsymbol{\tau} + \boldsymbol{\Lambda} (\mathbf{I}_{m \times m} - \mathbf{B})^{-1} \boldsymbol{\alpha}, \boldsymbol{\Lambda} (\mathbf{I}_{m \times m} - \mathbf{B})^{-1} \boldsymbol{\Phi} [(\mathbf{I}_{m \times m} - \mathbf{B})^{-1}]^\top \boldsymbol{\Lambda}^\top + \boldsymbol{\Psi} \right) \quad (2.7)$$

Traditionally, point estimates for the free model parameters  $\boldsymbol{\theta}$  are estimated by minimizing some form of fitting function that measures the discrepancy between the sample mean and covariance and their observed model-implied counterparts as defined in Equation 2.5 and 2.6. Different fitting functions have been suggested. Some of them, like the ordinary least squares fitting function (e.g., Duncan, 1966) and the weighted least squares fitting function (Browne, 1984), are distribution-free methods. Others, such as the generalized least-squares fitting function (e.g., Browne, 1974) and the maximum likelihood fitting function (Jöreskog, 1977), assume multivariate normally distributed data. In recent times, Bayesian estimation of SEMs has become more popular (see Smid et al., 2020, for an overview). In what follows, I focus on the maximum likelihood estimation for multivariate normal data because it is most commonly used for the estimation of SEMs. Yuan and Bentler (2006) define the maximum likelihood fitting function as

$$\begin{aligned} F(\bar{\mathbf{y}}, \mathbf{S}, \boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})) &= (\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta})) + \text{tr} (\mathbf{S} \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}) \\ &\quad - \ln [\det (\mathbf{S} \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1})] - p. \end{aligned} \quad (2.8)$$

Later in Section 3.1, I will briefly outline how Equation 2.8 can be minimized to obtain parameter estimates.

### 2.2 Model Fit Evaluation

Assessing the fit of a SEM to data is crucial to prevent drawing incorrect conclusions from an inadequate model. Numerous ways of assessing the fit of SEMs have been proposed. In the following, I give an overview of the most commonly used approaches based on Kline

(2016).

Global fit statistics are arguably the most popular criteria to judge SEMs. These statistics serve as an average or overall measure, condensing the fit of a model into a single number. In the past, the likelihood-ratio test was recommended to test the null hypothesis that there are no differences between the means and covariances as predicted by the model and the means and covariances of the population. Rejecting this hypothesis says that the differences between means, covariances, or both exceed those expected by sampling error, and the model is at least partially misspecified. Conversely, if the null hypothesis is not rejected, it is assumed that the model fits the data well. The corresponding test statistic can be obtained by multiplying the value of the fitting function  $F$  in Equation 2.8 by  $N - 1$ . Under the assumption of independent and identically multivariate normally distributed data and a correctly specified model,  $(N - 1)F$  follows asymptotically a  $\chi^2$  distribution with  $p(p + 3)/2 - q$  degrees of freedom (Bollen, 1989).<sup>2</sup> This particular likelihood-ratio test is usually referred to as the  $\chi^2$  test in the SEM literature.

Methodological problems of the  $\chi^2$  test led to a gradual shift away from the exact testing of model fit to approximating model fit with various indices. The main reason for this shift is the severe dependency of the  $\chi^2$  test on sample size in the sense that even a small and, in practical terms, meaningless misspecification will lead to a rejection of the model, given a large enough sample (Hu & Bentler, 1998). As a solution to this problem, fit indices such as the root mean square error of approximation (RMSEA; Steiger, 1990), the comparative fit index (CFI; Bentler, 1990), or the standardized root mean square residual (SRMR; Bentler, 1995) have been proposed to approximate the global fit of a SEM (see Schermelleh-Engel et al., 2003; West et al., 2012, for an overview). Most of these fit indices are a function of the  $\chi^2$  test and involve a trade-off between goodness of fit and model parsimony. In practice, the fit of a model is judged as acceptable if its fit indices satisfy certain thresholds. Although fit indices are less sensitive to the sample size than the likelihood-ratio test, other problems remain. While some researchers questioned the practice of using fit indices with thresholds as if they were test statistics (Barrett, 2007; Fan & Sivo, 2005; Marsh et al., 2004), others noted that fit indices can vary dramatically with the values of incidental parameters that are unrelated with the misspecification in the model (Sarlis et al., 2009).

As the name implies, global fit statistics measure only the average fit of a model and do not inform researchers what parts of a given model are likely to be misspecified. Therefore, Kline (2016) suggests to accompany global fit statistics with more local measures of model fit and discusses three strategies. First, researchers may locate model misfit by carefully inspecting the residuals of the SEM, that is, the differences between the observed means and covariances and the estimated means and covariances as implied by the model. Second,

---

<sup>2</sup>When only the covariance structure is of interest and no mean structure is specified, the degrees of freedom are  $p(p + 1)/2$ .



## 2. Heterogeneity and Structural Equation Models

one may specify and estimate a set of slightly deviating SEMs and select the one associated with the best global fit. Third, one can employ a hypothesis testing approach. Starting with a basic model, one adds parameters to the model as long as the model fit is improved. Usually, a likelihood-ratio test or a score test is employed to judge whether the modification improves the model fit. This testing procedure requires the basic model to be nested in the modified model. This strategy is also called specification search in the SEM literature (Kaplan, 1988) and will be discussed at a later point in greater detail.

Since the procedures outlined above predominate the assessment of SEMs, it is important to note that they do not alert researchers to unaccounted heterogeneity. Jedidi et al. (1997) demonstrate this problem by simulating data for a SEM with two exogenous latent factors that predict a single endogenous latent factor. The authors generated data for two groups of equal size that differed in the signs of the regression coefficients of the latent factors and the means of the exogenous factors. Further, they ignored the group differences and fitted a SEM on the pooled data. Although Jedidi et al. obtained heavily biased estimates of the model parameters, which did not represent any of the two groups, the  $\chi^2$  test was nonsignificant and the approximate fit indices were perfect, implying that the model fitted the data well. This behavior of the global fit statistics given unaccounted heterogeneity is certainly undesirable. It stems from the fact that global fit statistics are functions of the differences between the observed means and covariances and the estimated means and covariances. Since the sample means and covariances combine information from all individuals in the sample, differences between individuals and groups are averaged out and thus cannot affect the global fit statistics. The same applies to most local procedures for assessing model fit. Thus, specialized methods are needed to address heterogeneity in SEMs.

### 2.3 Effects of Heterogeneity

Heterogeneity can affect the measurement part, the structural part, or both parts of a SEM. In the measurement part, differences between individuals or groups can occur in the intercepts  $\boldsymbol{\tau}$  of the measurement part, the factor loadings  $\boldsymbol{\Lambda}$ , or in the covariance matrix  $\boldsymbol{\Psi}$  of the measurement errors. Such differences violate the assumption of measurement invariance (see Horn & McArdle, 1992; Meredith, 1993; Millsap, 2011), implying that the hypothesized latent variables  $\boldsymbol{\eta}$  do not capture the same theoretical meaning across individuals and groups. Thus, a violation of measurement invariance entails that relationships among the latent variables are only interpretable at the level of homogeneous subgroups. Therefore, data should not be pooled across groups. Conversely, when measurement invariance can be successfully established, one can be confident that potential heterogeneity does not result from a different understanding of the latent variables across individuals and groups. Heterogeneity in the structural part may influence the intercepts  $\boldsymbol{\alpha}$  of the latent variables,

## 2. Heterogeneity and Structural Equation Models

the regression coefficients  $\mathbf{B}$ , or the covariance matrix  $\Phi$  of the latent errors. Researchers that ignore parameter heterogeneity in the structural part and analyze the pooled data may encounter biased parameter estimates (Becker et al., 2013). Importantly, these aggregated estimates may not represent any individual in the sample and can lead to type I and type II errors. For instance, if one heterogeneous parameter is zero in one group but non-zero in the remaining sample, one may conclude that the estimate is significantly non-zero in the complete sample. Moreover, sign differences in regression or covariance parameters may mask important relationships at the individual or subgroup level as nonsignificant results at the overall sample level. Consequently, it is critical to assess, identify, and account for heterogeneity to avoid drawing incorrect conclusions from a SEM.

### 2.4 Methods for Addressing Heterogeneity

In the following, I will synthesize and compare some popular methods for studying heterogeneity in SEMs. SEM methods to address heterogeneity may be divided into two types: (1) methods that employ covariates to estimate observed heterogeneity and (2) methods that aim to uncover unobserved heterogeneity and do not rely on covariates.

The first group of methods employs moderators such as grouping variables (e.g., male or female) or individual characteristics of the persons in the data set (e.g., age) to account for differences across individuals and groups. One of the first methods proposed to study heterogeneity in SEMs are multigroup structural equation models (MGSEMs; Jöreskog, 1971; Sörbom, 1974). MGSEMs are routinely applied in the exploration of measurement invariance (e.g., Millsap & Kwok, 2004; van de Schoot et al., 2012) and stimulated the development of further approaches to address heterogeneity in SEMs. The objective of MGSEMs is to disclose the effect of a grouping variable that splits the sample into two or more non-overlapping, mutually exclusive subsets. Within each of these subsets, a group-specific submodel is estimated. The values of one, multiple, or all model parameters may vary across submodels, allowing the MGSEM to adapt to group differences. MGSEMs are most suitable to examine the effect of a single grouping variable with few levels. Studying grouping variables with many categories or the joint effects of multiple grouping variables usually requires a large sample to estimate the submodels in every subset reliably. When these sample size requirements are not met, researchers may accept a certain loss of information and resort to collapsing categorical variables into fewer groups or analyze multiple variables one at a time. Another disadvantage is that MGSEMs require continuous covariates like participants' age to be split up into small categories such as young, middle-aged, and old individuals. Categorizing continuous covariates can be particularly difficult when no information about the number of groups and the best split points are available. Moreover, MGSEMs neglect the ordering inherent to continuous or ordinal covariates. For example, when the effect of age on learning speed is investigated, one would expect a

## 2. Heterogeneity and Structural Equation Models

monotonic effect, where a decrease in learning speed accelerates with age. Unfortunately, there is no way to incorporate the ordering of a covariate into a classic MGSEM analysis.

Several alternatives to MGSEMs have been put forward. SEMs with interaction terms can be used to investigate the effect of a continuous covariate on the relationship between latent variables. Marsh et al. (2013) give an overview of different approaches to specify SEMs with interaction terms. Like MGSEMs, SEMs with interactions are best suited for studying the effects of a single covariate. Disentangling the effects of two or more covariates often leads to complicated models with many parameters, which can be hard to estimate.

Merkle and Zeileis (2013) and Merkle et al. (2014) suggested testing parameter heterogeneity with a family of score-based tests. These score-based tests infer all necessary information for evaluating the effect of a covariate from an estimated SEM that does not need to include said covariate. Notably, this property makes score-based tests much simpler than MGSEMs or SEMs with interaction terms since no additional SEMs need to be specified and estimated. However, different from MGSEMs and SEMs with interaction terms, score-based tests merely provide hypothesis tests and do not directly quantify heterogeneity. Furthermore, score-based tests are limited to investigate the effects of a single covariate at a time.

Brandmaier et al. (2013) proposed SEM trees as a method aiming specifically at the joint evaluation of multiple covariates. SEM trees are built upon the model-based recursive partitioning paradigm (Strobl et al., 2009; Zeileis et al., 2008). SEM trees split the sample recursively with respect to a set of covariates, building a tree-like structure in the process until homogeneous groups are found. An important feature of the method is its interpretability. The split decision of a SEM tree can be graphically represented as a decision tree. On a statistical level, SEM trees automate MGSEMs. The method performs an exhaustive search for every possible split of all available covariates, fitting a large number of MGSEMs in the process. Unlike MGSEMs, SEM trees take the ordering inherent to continuous and ordinal covariates into account. Moreover, the method is well suited for investigating larger sets of covariates since SEM trees rank the importance of covariates by placing them into different levels of the tree. A well-documented weakness of tree methods such as SEM trees is their susceptibility to random fluctuations, where even small changes to the data can lead to very different trees (e.g., Berk, 2006; Hastie et al., 2009). As a remedy, Brandmaier et al. (2016) suggested using SEM forests to complement SEM trees. SEM forests are ensemble methods that resample SEM trees to obtain more robust results about the influence of covariates. An important drawback of SEM trees and forests is their reliance on likelihood ratios to determine splits. Obtaining these likelihood ratios is computationally demanding, often rendering SEM trees and forests infeasible. In Project 3, I introduce computationally more efficient SEM trees that are guided by score-based tests instead of likelihood ratios.

## 2. Heterogeneity and Structural Equation Models

The second group of methods does not rely on covariates to detect heterogeneity. Various clustering methods (see Hastie et al., 2009, for an overview) and approaches such as finite mixture models (Arminger et al., 1999; Jedidi et al., 1997; Lubke & Muthén, 2005; Muthén & Shedden, 1999) are designed to detect clusters of similar individuals in heterogeneous data sets. Clustering algorithms can be employed before the SEM analysis is conducted and aim to minimize within-cluster variability while maximizing between-cluster variability. On the other hand, finite mixture models are model-based procedures that allow detecting heterogeneity in model parameters (see also Magidson & Vermunt, 2002, for a comparison of finite mixture models and clustering). Finite mixture SEMs can be seen as a generalization of MGSEMs. Whereas the number of groups must be prespecified, the group membership does not need to be known beforehand and is learned from the data. Generally, mixture approaches are considered a very stringent approach to heterogeneity as they do not rely on covariates. However, they have a lower statistical power to detect heterogeneity than tests for given groups when informative covariates are available (Smit et al., 2000). Moreover, mixture models provide no straightforward interpretation of the resulting groups. Therefore, latent class analyses are usually followed up by a second step of characterizing the latent groups using covariates for interpretability (e.g., Cohen & Bolt, 2005; Maij-de Meij et al., 2008).

When the clustering of the data is known a priori, unobserved heterogeneity across clusters can be accounted for with multilevel SEMs. Typical examples for clustered data structures are students nested in classrooms or repeated observations of the same individuals. For longitudinal data, it is common to model differences in developmental patterns between individuals by specifying latent variables whose factor loadings are fixed at specific values (see Mehta & Neale, 2005). For instance, latent growth curve models (e.g., Bollen & Curran, 2005) and dynamic panel models (e.g., Hamaker et al., 2015; Zyphur, Allison, et al., 2020; Zyphur, Voelkle, et al., 2020) usually contain latent variables called random intercepts whose factor loadings are all constrained to one in order to account for stable differences between individuals over time (so-called traits). Another more flexible multilevel SEM approach addresses heterogeneity by defining separate SEMs for each level of the data (e.g., Asparouhov & Muthén, 2021; Muthén, 1994). One SEM describes how responses vary within their respective clusters, and another SEM pertains to the variation between clusters. Both multilevel SEM approaches allow quantifying heterogeneity and enable researchers to make cluster-specific predictions. However, in order to explain heterogeneity, covariates are usually added to multilevel SEMs.

### **2.5 Individual Parameter Contribution Regression and Score-Guided Structural Equation Model Trees**

In this dissertation, I develop IPC regression and score-guided SEM trees, which are two novel approaches to address heterogeneity in SEMs with covariates. IPC regression is introduced in detail in Project 1 and Project 2, and score-guided SEM trees are presented in Project 3. Importantly, IPC regression and score-guided SEM trees offer advantages over the established methods for addressing heterogeneity in SEMs as outlined above. Both approaches are very general in terms of covariates and allow the investigation of categorical, ordinal, and continuous variables within the same framework. Moreover, IPC regression and score-guided SEM trees enable researchers to study the effects of multiple covariates and their interactions simultaneously. Another aspect of both methods is their data-driven nature that allows researchers to work with a single theory-guided model, rendering the respecification and re-estimation of alternative models unnecessary. This feature can prevent specification errors and is particularly important when a theory-guided model is already complex and difficult to estimate, and it makes both methods computationally very efficient, resulting in a short runtime.

### 3 Score-Based Approaches to Model Evaluation

IPC regression and score-guided SEM trees are based on the same methodological foundation. Both methods employ the score function to detect and estimate heterogeneity in SEM parameter estimates. In the following chapter, I illuminate important properties of the score function that are used in the individual projects of this dissertation but not derived explicitly. Afterwards, I provide an overview of score-based approaches for model checking and highlight how they are related to IPC regression and score-guided SEM trees. Finally, leading on to the individual projects, I introduce the case-wise or individual score function used to quantify heterogeneity.

#### 3.1 The Score Function

Fisher (1922, 1925) defined the score function, or simply score, as the partial derivative of the log-likelihood function with respect to the parameter vector. In his early works about maximum likelihood theory, Fisher primarily used the score to obtain maximum likelihood estimates (Stigler, 2007). Later, Rao (1948) introduced the first score-based procedure for hypothesis testing, known as Rao's score test. This section will first introduce the score function and its key properties in very general terms. Then, I relate the results to SEMs.

Let  $X$  be a (possibly multivariate) random variable from the probability density function  $f(X; \boldsymbol{\theta})$ . As in the previous section,  $\boldsymbol{\theta}$  denotes a  $q$ -variate vector of unknown parameters. Further,  $\ln L(\boldsymbol{\theta}; X) = \ln f(X, \boldsymbol{\theta})$  is the log-likelihood function of  $X$ . The corresponding score function is obtained by taking the partial derivative of the log-likelihood with respect to the parameters:

$$S(\boldsymbol{\theta}; X) = \left[ \frac{\partial \ln L(\boldsymbol{\theta}; X)}{\partial \theta_1} \quad \dots \quad \frac{\partial \ln L(\boldsymbol{\theta}; X)}{\partial \theta_q} \right]^\top \quad (3.1)$$

The score measures the steepness of the log-likelihood when evaluated at a specific point of the parameter vector. While the score is a function of the parameters, it also depends on the random variable  $X$  and is therefore affected by the random character of sampling. In other words, the score itself is a random variable whose properties can be studied. In the following, I will derive the mean and variance of the score.

Under certain regularity conditions (see Serfling, 1980), it can be shown that the expected value of the score evaluated at the true parameter value  $\boldsymbol{\theta}$  is zero:

$$\begin{aligned} E[S(\boldsymbol{\theta}; X) | \boldsymbol{\theta}] &= \int_{-\infty}^{\infty} \frac{\partial \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(x; \boldsymbol{\theta}) dx \\ &= \int_{-\infty}^{\infty} \frac{\partial f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(x; \boldsymbol{\theta})^{-1} f(x; \boldsymbol{\theta}) dx \end{aligned}$$

### 3. Score-Based Approaches to Model Evaluation

$$\begin{aligned}
&= \int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}} f(x; \boldsymbol{\theta}) dx \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \int_{-\infty}^{\infty} f(x; \boldsymbol{\theta}) dx \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} 1 \\
&= \mathbf{0}_q
\end{aligned} \tag{3.2}$$

Hence, the score evaluated at the true parameters fluctuates randomly around zero. This property of the score function is crucial for different score-based hypothesis tests, as I will show later.

The variance of the score is called Fisher information and is defined as

$$\begin{aligned}
\text{Var}[S(\boldsymbol{\theta}; X) | \boldsymbol{\theta}] &= \text{E} \{ [S(\boldsymbol{\theta}; X) - \text{E}(S(\boldsymbol{\theta}; X) | \boldsymbol{\theta})] [S(\boldsymbol{\theta}; X) - \text{E}(S(\boldsymbol{\theta}; X) | \boldsymbol{\theta})]^\top | \boldsymbol{\theta} \} \\
&= \text{E} [S(\boldsymbol{\theta}; X) S(\boldsymbol{\theta}; X)^\top | \boldsymbol{\theta}] \\
&= \mathcal{I}(\boldsymbol{\theta}).
\end{aligned} \tag{3.3}$$

The Fisher information plays an important role in many statistical applications (see Ly et al., 2017). Briefly put, it measures the amount of information the random variable  $X$  carries about the unknown parameter  $\boldsymbol{\theta}$ . In maximum likelihood theory, the Fisher information is crucial for assessing the precision of the parameter estimates and is used to estimate the variance-covariance matrix and confidence intervals of the parameter estimates and to construct hypothesis tests.

Suppose the log-likelihood function is twice differentiable with respect to the parameters and certain regularity conditions (Lehmann & Casella, 1998) to hold. Then, the Fisher information is equal to the negative expected value of the Hessian matrix of the log-likelihood. More formally, let  $\mathbf{0}_{q \times q}$  be a zero matrix of dimensions  $q \times q$ , then it can be shown that

$$\begin{aligned}
\mathbf{0}_{q \times q} &= \frac{\partial}{\partial \boldsymbol{\theta}^\top} \text{E} [S(\boldsymbol{\theta}; X) | \boldsymbol{\theta}] \\
&= \frac{\partial}{\partial \boldsymbol{\theta}^\top} \int_{-\infty}^{\infty} \frac{\partial \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(x; \boldsymbol{\theta}) dx \\
&= \int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}^\top} \left[ \frac{\partial \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(x; \boldsymbol{\theta}) \right] dx \\
&= \int_{-\infty}^{\infty} \left[ \frac{\partial^2 \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} f(x; \boldsymbol{\theta}) + \frac{\partial \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \right] dx \\
&= \int_{-\infty}^{\infty} \frac{\partial^2 \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} f(x; \boldsymbol{\theta}) dx + \int_{-\infty}^{\infty} \frac{\partial \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} dx \\
&= \int_{-\infty}^{\infty} \frac{\partial^2 \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} f(x; \boldsymbol{\theta}) dx + \int_{-\infty}^{\infty} \frac{\partial \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} f(x; \boldsymbol{\theta}) dx
\end{aligned}$$

### 3. Score-Based Approaches to Model Evaluation

$$\begin{aligned}
&= \mathbb{E} \left[ \frac{\partial^2 \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \middle| \boldsymbol{\theta} \right] + \mathbb{E} \left[ \frac{\partial \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \middle| \boldsymbol{\theta} \right] \\
&= \mathbb{E} \left[ \frac{\partial^2 \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \middle| \boldsymbol{\theta} \right] + \mathbb{E} [S(\boldsymbol{\theta}; X) S(\boldsymbol{\theta}; X)^\top | \boldsymbol{\theta}].
\end{aligned} \tag{3.4}$$

Rearranging terms yields:

$$\mathbb{E} [S(\boldsymbol{\theta}; X) S(\boldsymbol{\theta}; X)^\top | \boldsymbol{\theta}] = - \mathbb{E} \left[ \frac{\partial^2 \ln f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \middle| \boldsymbol{\theta} \right] = \mathcal{I}(\boldsymbol{\theta}) \tag{3.5}$$

Thus, the Fisher information is also the expected curvature (or second derivative) of the log-likelihood. When evaluated at the maximum likelihood estimates, the Fisher information indicates how certain the model parameters are estimated. Low values of the Fisher information imply that the log-likelihood appears to be flat around the estimated parameters and many nearby parameter values are approximately as likely. Conversely, larger values of the Fisher information indicate a peaked likelihood, meaning the parameter estimates are more reliable.

Arguably, the most important area of application of the score function is parameter estimation. Let  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]^\top$  be a data matrix with  $N$  independent identically distributed,  $p$ -variate random variables with probability density functions  $f(\mathbf{y}_1, \boldsymbol{\theta}), \dots, f(\mathbf{y}_N, \boldsymbol{\theta})$ . The associated log-likelihood function of the joint probability distribution is  $\ln L(\boldsymbol{\theta}; \mathbf{Y}) = \ln \prod_{i=1}^N f(\mathbf{y}_i, \boldsymbol{\theta})$ . Maximum likelihood estimates  $\hat{\boldsymbol{\theta}}$  for the unknown parameters  $\boldsymbol{\theta}$  can be obtained by maximizing the log-likelihood function, that is,

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \ln L(\boldsymbol{\theta}; \mathbf{Y}). \tag{3.6}$$

It is easier to find the roots (the zeros) of the score than the maximum of the log-likelihood in most situations. Therefore, maximum likelihood estimates are most commonly found by solving

$$S(\hat{\boldsymbol{\theta}}; \mathbf{Y}) = \mathbf{0}_q. \tag{3.7}$$

Usually, this is achieved by applying an optimization algorithm like gradient ascent that repeatedly evaluates the score until a root is found (see Nocedal & Wright, 2006). However, the score function is not only zero at a maximum of the log-likelihood but also at a minimum. Hence, it is necessary to verify that the solution is indeed a maximum. This condition can be checked by computing the second partial derivative of the log-likelihood function. If for all  $j = 1, \dots, q$ ,

$$\frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{Y})}{(\partial \theta_j)^2} \bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \theta_j = \hat{\theta}_j} < 0, \tag{3.8}$$



### 3. Score-Based Approaches to Model Evaluation

then the log-likelihood in the neighborhood of  $\hat{\boldsymbol{\theta}}$  is convex and  $L(\hat{\boldsymbol{\theta}}; \mathbf{Y})$  is a maximum. Additionally, it needs to be verified if the maximum is global and not a local maximum. Unfortunately, there are no methods that can guarantee to identify whether the maximum is global or not. In practice, the optimization algorithm is run multiple times with different starting values to see whether the same solution for Equation 3.7 is obtained repeatedly. More details about estimation are given by Myung (2003).

The properties of the score derived above translate directly to the SEM framework. The corresponding normal-theory log-likelihood function for SEMs is

$$\ln L(\boldsymbol{\theta}; \bar{\mathbf{y}}, \mathbf{S}) = -\frac{N}{2} \ln |\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{N}{2} \text{tr} \left\{ \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \left[ \mathbf{S} + (\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta})) (\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta}))^\top \right] \right\} \quad (3.9)$$

(Yuan & Bentler, 2006). A simple relationship exists between the log-likelihood function in Equation 3.9 and the maximum likelihood fitting function  $F$  shown in Equation 2.8.  $F$  is the difference between a log-likelihood where all the elements of  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  are freely estimated and the usual log-likelihood as defined in Function 3.9. More specifically, when evaluated at  $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$  and  $\hat{\boldsymbol{\Sigma}} = \mathbf{S}$ , the log-likelihood is

$$\ln L(\bar{\mathbf{y}}, \mathbf{S}; \bar{\mathbf{y}}, \mathbf{S}) = -\frac{N}{2} \ln |\mathbf{S}| - \frac{Np}{2} \quad (3.10)$$

and the difference yields the fitting function  $F$ :

$$\frac{2}{N} [\ln L(\bar{\mathbf{y}}, \mathbf{S}; \bar{\mathbf{y}}, \mathbf{S}) - \ln L(\boldsymbol{\theta}; \bar{\mathbf{y}}, \mathbf{S})] = F(\bar{\mathbf{y}}, \mathbf{S}, \boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})), \quad (3.11)$$

Since  $\ln L(\bar{\mathbf{y}}, \mathbf{S}; \bar{\mathbf{y}}, \mathbf{S})$  is a constant, the maximum likelihood estimate  $\hat{\boldsymbol{\theta}}$  found by maximizing  $\ln L(\boldsymbol{\theta}; \bar{\mathbf{y}}, \mathbf{S})$  also minimizes  $F(\bar{\mathbf{y}}, \mathbf{S}, \boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ . Likewise, the score function and Fisher information can be equivalently obtained based on the fitting function  $F$ :

$$-\frac{\partial}{\partial \boldsymbol{\theta}} \frac{N}{2} F(\bar{\mathbf{y}}, \mathbf{S}, \boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})) = S(\boldsymbol{\theta}; \bar{\mathbf{y}}, \mathbf{S}) \quad (3.12)$$

$$\text{E} \left[ \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \frac{N}{2} F(\bar{\mathbf{y}}, \mathbf{S}, \boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})) \right] = \mathcal{I}(\boldsymbol{\theta}) \quad (3.13)$$

Closed-form solutions that express the score and Fisher information in terms of SEM vectors and matrices as defined in Section 2.1 are provided by Neudecker and Satorra (1991) and von Oertzen and Brick (2014).

### 3.2 Score-Based Model Checking

Various statistical techniques utilize the score function to determine whether a theoretical model fits the data adequately. These score-based approaches may be broadly divided into two categories: (1) procedures that consider the scores of the sample jointly and (2)

### 3. Score-Based Approaches to Model Evaluation

methods that analyze the individual or case-wise scores. A classic example of the former methods is the score test, primarily used to test parameter constraints. The latter group of methods, including IPC regression and score-guided SEM trees, is mainly applied to assess parameter instability due to differences across persons or changes over time. This section will briefly summarize some important score-based approaches that utilize the joint score function. I introduce the use of individuals scores in the next section.

The eponymous score test put forward by Rao (1948) was the first type of score-based procedure to be used extensively in practical applications and stimulated the development of other score-based approaches. The score test can be equivalently understood in terms of the Lagrange multipliers used in optimizing a function subject to restrictions, as shown by Aitchison and Silvey (1958) and Silvey (1959). Especially in econometric literature, the Lagrangian interpretation has become more commonly used after an influential paper by Breusch and Pagan (1980).

The score test, along with the likelihood-ratio test (Neyman & Pearson, 1928) and the Wald test (Wald, 1943), is one of the three classical approaches to hypothesis testing. In statistical literature, the three tests are often referred to as the Holy Trinity (e.g., Rao, 2005). All three tests are asymptotically equivalent but may differ in finite samples. Other than the likelihood-ratio test, which requires the estimation of an unrestricted and a restricted model, and the Wald test, which requires the estimation of an unrestricted model, the score test only requires the estimation of a restricted model. This property makes the score test computationally more efficient than the likelihood-ratio test and well suited to test constraints on parameters whose unconstrained maximum likelihood estimates lie close to boundary points in the parameter space. A detailed comparison of the three tests is given by Buse (1982).

More formally, let  $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 = [\boldsymbol{\theta}_r^\top, \boldsymbol{\theta}_u^\top]^\top$  be the null hypothesis to be tested, where  $\boldsymbol{\theta}_r$  represents  $k$  components of the parameter vector  $\boldsymbol{\theta}$  that are fixed at specific values.  $\boldsymbol{\theta}_u$  denotes the remaining  $q - k$  elements of  $\boldsymbol{\theta}$  that are to be freely estimated. Further, we denote the maximum likelihood estimate of  $\boldsymbol{\theta}_0$  with  $\hat{\boldsymbol{\theta}}$ . Then,

$$ST = S(\hat{\boldsymbol{\theta}}; \mathbf{Y})^\top \mathcal{I}(\hat{\boldsymbol{\theta}})^{-1} S(\hat{\boldsymbol{\theta}}; \mathbf{Y}) \quad (3.14)$$

is the score test statistic and has an asymptotic distribution of  $\chi_k^2$ , when  $H_0$  is true (Rao, 1948). Recall that the mean of the scores evaluated at the true parameters is zero (see Equation 3.2). Intuitively speaking, if the values of the restricted parameters in  $\boldsymbol{\theta}_r$  are close to the true values, the log-likelihood function is near its maximum and the score should not differ from zero by more than the sampling error. Consequently, large deviations of the score function from zero constitute evidence against the null hypothesis.

In psychometric research and especially in the SEM literature, the score test is better known as the modification index (MI), popularized by the influential LISREL software

### 3. Score-Based Approaches to Model Evaluation

(Jöreskog et al., 2016) used to estimate SEMs. Typically, the MI tests the null hypothesis that a single parameter equals zero and is usually computed separately for all SEM parameters that are fixed to zero. Thus, the MI estimates how the fit of a SEM, measured in terms of the  $\chi^2$  test statistic, would improve if a fixed parameter were added to the model and freely estimated. Different procedures have been proposed to calculate the MI. The most recent iteration of the MI (see Saris et al., 1987; Satorra, 1989; Sörbom, 1989) is equivalent to the score test statistic shown in Equation 3.14. Since the MI usually tests only parameters constrained to zero, the maximum likelihood estimates can be taken directly from the model. This property makes the MI computationally more efficient for detecting specification errors in SEMs than other methods such as the likelihood-ratio test, which would require the specification and estimation of further SEMs with the added parameters.

The use of the MI for the detection of specification errors has been repeatedly questioned. Several simulation studies show that the MI is not as accurate as one would hope (Kaplan, 1988; MacCallum, 1986; MacCallum et al., 1992; Silvia & MacCallum, 1988; Whittaker, 2012). Especially in small samples and when confronted with models with a higher number of specification errors, following the advice of the MI by adding parameters to the model does not guarantee that one arrives at the correct model (but see also Chou & Bentler, 1990, for more promising results). Moreover, as demonstrated by Saris et al. (1987) and Saris et al. (2009), the MI is often not an adequate indication of the size of the model's misspecification and may vary with the values of other incidental parameters unrelated to the specification errors. As a remedy, the authors suggested using the MI in combination with the expected parameter change (EPC). The EPC provides an estimate of the value a parameter would have, had it been freely estimated. Thus, the EPC offers additional information about the size of the misspecification that goes beyond the significance test provided by the MI. Computationally, the EPC is closely connected to the MI and can be calculated for any fixed parameter  $\theta_l$ , where  $l = 1, \dots, k$ , by dividing its MI by the partial derivative of the log-likelihood with respect to  $\theta_l$ .

Critics have also argued that model modification based on the MI or EPC is data-driven, thus susceptible to capitalize on chance characteristics of the sample (e.g., MacCallum et al., 1992). Therefore, it poses the risk of overfitting the model, and it remains highly doubtful whether a modified model will generalize to other samples or even the population. Overfitting is not only a potential pitfall for the use of the MI and EPC but a more general concern, also pertaining to IPC regression and score-guided SEM trees. I will touch upon this issue in the Discussion in Chapter 7. Despite this criticism, the MI has recently found a new application as part of the group iterative multiple model estimation algorithm that has been put forward to address heterogeneity in neuroscientific and ambulatory assessment data (e.g., Gates et al., 2017; Gates & Molenaar, 2012; Nestler & Humberg, 2021).

### 3. Score-Based Approaches to Model Evaluation

The MI and EPC are closely related to IPC regression. Although the MI and EPC aim to quantify specification errors and not heterogeneity like IPC regression, Oberski (2013) demonstrated that the MI and EPC for MGSEMs correspond to IPC regression under certain conditions. However, this equivalency ends in situations that cannot be handled by MGSEMs, such as continuous or multiple covariates, making IPC regression a much more flexible method for the investigation of heterogeneity.

#### 3.3 Individual Scores and Parameter Heterogeneity

While analyses of the joint score function are helpful to detect model misspecification by assessing parameter constraints, individual scores are helpful to determine if model parameters are invariant across persons. Individual scores of SEMs can be obtained by calculating the partial derivative of the individual log-likelihood function with respect to the parameters. Under the assumption of multivariate normality, the following individual log-likelihood only considers data from individual  $i$ :

$$\ln L(\boldsymbol{\theta}; \mathbf{y}_i) = -\frac{1}{2} \left[ (\mathbf{y}_i - \boldsymbol{\mu}(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} (\mathbf{y}_i - \boldsymbol{\mu}(\boldsymbol{\theta})) + \ln |\boldsymbol{\Sigma}(\boldsymbol{\theta})| + p \ln(2\pi) \right] \quad (3.15)$$

Summing the individual log-likelihood functions yields the full sample log-likelihood in Equation 3.9 used to estimate SEMs.

When evaluated at the parameter estimates  $\hat{\boldsymbol{\theta}}$ , the individual scores

$$S(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) = \left[ \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \theta_1} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, \theta_1=\hat{\theta}_1} \quad \cdots \quad \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \theta_q} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, \theta_q=\hat{\theta}_q} \right]^\top, \quad i = 1, \dots, N, \quad (3.16)$$

form a new matrix, where each individual  $i$  has  $q$  scores, one for each estimated parameter in  $\hat{\boldsymbol{\theta}}$ . The scores act as goodness-of-fit between persons and parameter estimates, where small values indicate good fit and large scores point towards bad fit. Importantly, this information can be used to derive insights about parameter heterogeneity. Let us assume the same parameter values hold for all individuals. In this case, all individual scores fluctuate randomly around zero, much like the residuals in a linear regression model. However, if the parameters are heterogeneous, the fluctuation of the scores is no longer entirely random. For example, say there is a difference between two groups in the sample. At an individual level, this group difference is reflected in the scores as follows: on average, scores of the first group will be more often negative than positive, and the scores of second the group will be more often positive than negative, or vice versa. Although such a pattern is a clear sign for parameter heterogeneity, it is usually too subtle for visual inspection. Therefore, methods such as IPC regression and score-guided SEM trees are needed.

IPC regression and score-guided SEM trees process the information in the scores in

### 3. Score-Based Approaches to Model Evaluation

different ways. On the one hand, IPC regression uses individual scores to approximate individual-specific parameter values called IPCs. Then, potential heterogeneity in these IPCs is modeled by means of linear regression. IPC regression estimates how parameters change as functions of covariates and provides  $p$ -values that allow inferring whether the partial effects of the covariates are significantly different from zero. The first two projects, summarized in Chapter 4 and 5, introduce IPC regression in great detail. On the other hand, score-guided SEM trees rely on a family of score-based tests to detect heterogeneity. These score-based tests aggregate the individual scores into a single test statistic to infer whether the model parameters vary with respect to a covariate. Different aggregation schemes of the individual scores are presented in the final project, summarized in Chapter 6.

## 4 Identifying Heterogeneity in Dynamic Panel Models with Individual Parameter Contribution Regression

Arnold, M., Oberski, D. L., Brandmaier, A. M., & Voelkle, M. C. (2020). Identifying heterogeneity in dynamic panel models with individual parameter contribution regression. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(4), 613-628. <https://doi.org/10.1080/10705511.2019.1667240>

***A copy of this project is attached in the Appendix on pages 47 to 62.***

Dynamic panel models (Hsiao, 2014) are popular approaches to studying longitudinal data where the same subjects are observed multiple times. Building upon the idea of Granger causality (Granger, 1969), they allow answering questions concerning the direction and strength of reciprocal relationships between repeatedly measured variables. Especially in psychological research, it is common practice to specify and estimate dynamic panel models within the SEM framework (Allison et al., 2017; Bollen & Brand, 2010; Zyphur, Allison, et al., 2020; Zyphur, Voelkle, et al., 2020).

There are many variants of dynamic panel models. An important distinguishing feature is the treatment of time. On the one hand, so-called discrete-time dynamic panel models assume that the temporal spacing between assessments is constant throughout the study (Biesanz, 2012). This assumption is often problematic because it complicates comparing estimates from studies with different sample schemes and, if violated, leads to biased parameter estimates. On the other hand, continuous-time dynamic panel models solve these issues by treating time as a continuous variable (Oud & Jansen, 2000; Voelkle et al., 2012). These models allow the time intervals to vary between measurements and across individuals and produce parameter estimates that are not affected by the sampling scheme.

A commonly encountered problem that complicates the analysis of longitudinal data is heterogeneity in the form of systematic differences across individuals or groups. For instance, individuals may show stable, or trait-like, differences in the means levels, exhibit different recovery speeds after random shocks, or differ in the coupling of the measured variables. However, methods to address heterogeneity in dynamic panel models are often complicated to use or computationally demanding. In Project 1, we demonstrate how heterogeneity in discrete and continuous-time dynamic panel models can be identified and estimated with IPC regression.

IPC regression was originally proposed by Oberski (2013) as a flexible and fast tool to detect heterogeneity in SEMs. IPC regression allows to model individual and group differences in SEM parameters as a function of covariates. In Project 1, we show by means of a Monte Carlo simulation that IPC regression provides biased estimates of heterogeneity in certain situations. To solve this issue, we introduce a novel bias correction procedure

#### 4. Identifying Heterogeneity in Dynamic Panel Models with Individual Parameter Contribution Regression

termed iterated IPC regression that produces approximately unbiased estimates at the cost of increased variability. In terms of power to detect heterogeneity, IPC regression proved to be less powerful than MGSEMs in our simulation study. Given homogeneous samples without any heterogeneity, IPC regression showed an adequate control of type I errors. Overall, IPC regression performed better when used to investigate discrete-time dynamic panel models than continuous-time dynamic panel models. On a theoretical level, we extend Oberski's work in the following ways: first, we connect IPC regression to the general maximum likelihood estimation technique. Second, we correct some errors in Oberski's derivations that lead to incorrect estimates for certain parameter types. Third, we present a theoretical example illuminating the bias of IPC regression. Last but not least, we provide users with the **ipcr** package (<https://github.com/manuelarnold/ipcr>) for the R system for statistical computing (R Core Team, 2021) that allows users to perform IPC regression with ease.

In sum, Project 1 demonstrates that IPC regression is a promising tool to study heterogeneity in dynamic panel models that is easy to use and computationally efficient. Moreover, our study provides important empirical and theoretical insights into the potential and challenges of the IPC regression framework and advances the understanding of the method.

## 5 Predicting Differences in Model Parameters with Individual Parameter Contribution Regression Using the R Package `ipcr`

Arnold, M., Brandmaier, A. M., & Voelkle, M. C. (2021). Predicting differences in model parameters with individual parameter contribution regression using the R package `ipcr`. *Psych*, 3(3), 360-385. <https://doi.org/10.3390/psych3030027>

*A copy of this project is attached in the Appendix on pages 63 to 88.*

Project 2 builds upon Project 1 and further advances the IPC regression framework. The project aims specifically at potential users of IPC regression who may wish to know how IPC regression can be employed to detect heterogeneity in their models. To this end, we describe `ipcr`, an R package which supplies functions to perform IPC regression in practice. We provide a step-by-step introduction to `ipcr` with detailed example code using the classic `HolzingerSwineford1939` data set shipped with the `lavaan` package (Rosseel, 2012).

`ipcr` is a flexible and user-friendly package that allows users to perform IPC regression with a single command and offers different ways to visualize the results of an IPC regression analysis. To a large extent, `ipcr` relies on the infrastructure provided by the `sandwich` package (Zeileis et al., 2020) mainly used for calculating robust covariance matrices of the model parameter estimates. Moreover, the `ipcr` package allows users to perform regularized IPC regression by interfacing with the `glmnet` package (Friedman et al., 2010). This functionality aims to decrease overfitting and is particularly handy when the effects of many covariates are to be investigated.

Besides these rather practical instructions for potential users, Project 2 also advances IPC regression on a methodological level. First, we expand IPC regression's application area from SEMs to a broader range of parametric models, including linear regression and mixed-effects models, and analytically derive results for linear regression models. Second, we describe four novel Monte Carlo simulations that benchmark IPC regression in situations that have not been considered by Oberski (2013) or in Project 1. These simulations investigate the performance of IPC regression for linear regression models, multiple covariates, and continuous covariates and compare the results with MGSEMs and SEMs with an interaction term. As in the previous simulations, IPC regression showed overall promising results. The observed type I error rates were close to the optimal rate in homogeneous, regardless of the types of covariates (dummy or continuous), the number of covariates, or sample size. Moreover, we found little differences in terms of bias and variance of the estimates provided by IPC regression compared to the estimates of MGSEMs or SEMs with an interaction term. However, IPC regression was consistently less powerful than the established methods.



## 6 Score-Guided Structural Equation Model Trees

Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2021). Score-guided structural equation model trees. *Frontiers in Psychology, 11*, 1-18. <https://doi.org/10.3389/fpsyg.2020.564403>

***A copy of this project is attached in the Appendix on pages 89 to 106.***

Model-based recursive partitioning (Strobl et al., 2009; Zeileis et al., 2008) is a general framework for uncovering heterogeneity in parametric models. It identifies important covariates that explain differences in model parameters by recursively splitting the sample into subgroups so that these subgroups are maximally different from each other. This way, hierarchical tree structures of covariates can be grown that describe subsets with sets of parameter estimates.

Various implementations of model-based recursive partitioning for different model classes have been proposed. The first instance of model-based recursive partitioning to investigate SEMs is the SEM tree algorithm put forward by Brandmaier et al. (2013), which is implemented in the R package `semtree`. The SEM tree algorithm selects split variables among covariates by splitting the sample into subgroups and fitting a MGSEM to each subgroup. Then, a likelihood ratio for each possible split of each covariate is calculated. If the split associated with the largest likelihood ratio improves the fit of the model significantly, the data is partitioned and the algorithm proceeds recursively.

Although straightforward, the original SEM tree algorithm suffers from two problems: first, the algorithm is computationally demanding. The calculation of every likelihood ratio requires the estimation of SEMs, which often makes growing SEM trees unfeasible. Second, the split decision suffers from a variable selection bias (see Strobl et al., 2007) that favors the selection of covariates with many unique values over covariates with few. Brandmaier et al. (2013) implemented different correction procedures for this bias to the `semtree` package.

In Project 3, we adopt the generic model-based recursive partitioning algorithm suggested by Zeileis et al. (2008) and guide the construction of SEM trees by a family of score-based tests instead of likelihood ratios. Introduced to psychometrics by Merkle and Zeileis (2013) and Merkle et al. (2014), these score-based tests monitor fluctuation in case-wise derivatives of the log-likelihood function to detect parameter differences. Unlike the classic likelihood-ratio approach, score-based tests are computationally inexpensive because they do not require refitting models for every possible split. Besides the new score-guided SEM trees, we also put forward a new version of likelihood-ratio-guided SEM trees that provides unbiased variable selection. We show by means of a Monte Carlo simulation that guiding SEM trees by score-based tests improves the runtime drastically.

## 6. Score-Guided Structural Equation Model Trees

Moreover, the score-guided and the new likelihood-ratio-guided SEM trees outperformed the classic methods proposed by Brandmaier et al. (2013) in terms of power to detect heterogeneity and provided accurate type I error rates. Finally, we implemented all new methods to the **semtree** package.

## 7 Discussion

This dissertation presents IPC regression and score-guided SEM trees, two methods for addressing heterogeneity in SEMs with covariates. After highlighting key differences between IPC regression and score-guided SEM trees, I compare both approaches to established methods for addressing heterogeneity. Finally, I devote the remainder of the dissertation to a discussion of conceptual and technical challenges and point towards possible areas of application and future research.

### 7.1 Comparison of Methods

While IPC regression and score-guided SEM trees address heterogeneity with covariates, both methods are by no means identical. In the following, I first summarize the key difference and then compare both methods to established approaches to heterogeneity in SEMs.

#### *7.1.1 Individual Parameter Contribution Regression and Score-Guided Structural Equation Model Trees*

**Objective.** The primary objective of IPC regression and score-guided SEM trees is to detect and quantify heterogeneity. Both methods approach this goal in different ways. In addition to predicting heterogeneity, score-guided SEM trees provide means to identify homogeneous subgroups of individuals with identical parameter values. This property makes the trees also well-suited for classification tasks which may even be the main objective of one's study.

**Quantifying heterogeneity.** Both methods quantify parameter heterogeneity in different ways. IPC regression estimates the effects of covariates on model parameters in the form of linear functions, one for each parameter. The results of an IPC regression analysis can be interpreted in the same way as researchers interpret linear regression models. Moreover, based on the estimates of the IPC regression equations, IPC regression allows predicting individual and group-specific parameter values for any covariate values. Non-linear relationships or interactions can be investigated by simply adding polynomial terms or product terms of the covariates to the IPC regression equations. In contrast, score-guided SEM trees address the relationships between covariates and model parameters in a non-linear way by splitting with respect to a covariate and estimating separate parameter values for the post-split samples. In theory, this non-linear approach seems advantageous as it allows the trees to explore arbitrary non-linear effects of covariates and covariate interactions by repeatedly splitting the sample. However, approximating

complicated non-linear effects will often require an enormous sample size. Moreover, the non-linear approach makes score-guided SEM trees less suited for extrapolating parameter values than IPC regression.

Another critical difference pertains to the interpretation of the estimated parameter values. IPC regression provides estimates of partial effects of covariates on model parameters. That is, the effect of a covariate when all other covariates are held constant. In contrast, score-guided SEM trees estimate conditional parameter values, meaning the parameters estimated after a split further down the tree depend on all splits selected further up the tree. As a result, the first split of a tree is immensely influential, and a different split may lead to an entirely different tree. Unconditional measures of variable importance can be obtained by resampling trees leveraging the SEM forest approach (see Brandmaier et al., 2016).

**Covariate types.** Although IPC regression and score-guided SEM trees encompass discrete, ordinal, and continuous covariates, the types of the covariates can affect the analysis. IPC regression is most parsimonious and easiest to interpret for the investigation of continuous and dichotomous covariates. Categorical or ordinal covariates with multiples levels must be split up into several dummy variables, where one level serves as a baseline. Unfortunately, this categorization approach disregards the ordering of ordinal covariates. Alternatively, ordinal covariates could also be treated as continuous variables, but this may not always be appropriate. However, since IPC regression is merely standard linear regression with IPCs, users may resort to one of the many methods proposed for linear regression models with ordinal predictors (e.g., Bürkner & Charpentier, 2020; Helwig, 2017). Conversely, score-guided SEM trees work best with categorical and ordinal variables, whereas the effects of continuous covariates are harder to interpret as the trees dichotomize them to split the sample.

**Preventing overfitting.** IPC regression and score-guided SEM trees provide several safeguards that prevent overfitting. As described in Project 2, IPC regression can be coupled with regularization techniques such as the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996, 2011). Although different ways of regularization are implemented in the `ipcr` package, there are currently no studies that indicate how well regularized IPC regression may work in practice. Besides regularization, overfitting may also be averted by adjusting the significance level to correct the testing of multiple covariates. However, as with regularized IPC regression, such corrections for IPC regression have not been studied yet. Fortunately, the situation is different for score-guided SEM trees, and different techniques to prevent overfitting were already investigated in the model-based recursive partitioning literature. For instance, Brandmaier et al. (2016) demonstrated the use of SEM forests to prevent overfitting and improve generalizability

to new samples. Moreover, in Project 3, we successfully applied a Bonferroni correction to control the testing of multiple covariates.

**Output.** Both methods provide users with different outputs. On the one hand, IPC regression provides users with a standard regression output for each model parameter. These outputs indicate if and how covariates influence model parameters and can guide the modification of the model. On the other hand, score-guided SEM trees estimate a MGSEM that accounts for heterogeneity in the sample. In theory, this MGSEM could be directly reported or may serve as a basis for modifying the original model.

**Statistical properties.** Not much is known about how both approaches compare in terms of power to detect heterogeneity and qualities of their estimates. The Monte Carlo simulations in the individual projects imply that both IPC regression and score-guided SEM trees are slightly less powerful and less precise than MGSEMs. The online supplemental material of Project 1 (<https://www.tandfonline.com/doi/suppl/10.1080/10705511.2019.1667240>) compares IPC regression and the original SEM tree implementation put forward by Brandmaier et al. (2013). Given normally distributed data, SEM trees exhibited a smaller power than IPC regression. When provided with non-normally distributed data, SEM trees were prone to committing type I errors. In contrast, IPC regression yielded robust results that were mostly unaffected by non-normality. However, the results of SEM trees would likely improve had the trees been estimated with the new score-guided split selection procedure. For instance, score-guided SEM trees clearly outperformed the original SEM tree implementation in Project 3.

### ***7.1.2 Established Methods***

Compared to the established methods for addressing heterogeneity in SEMs with covariates summarized in Section 2.4, IPC regression and score-guided SEM trees offer some advantages. In contrast to MGSEMs and SEMs with interaction terms that are used to study either discrete grouping variables or continuous covariates, IPC regression and score-guided SEM trees are very general in terms of covariates and allow the investigation of categorical, ordinal, and continuous variables within the same framework. Moreover, IPC regression and score-guided SEM trees enable researchers to study the effects of multiple covariates and their interactions simultaneously. Another useful feature is that both methods render repeated respecification and re-estimation of alternative models unnecessary. Not only will this feature save the user time, but it may also prevent specification errors and is particularly important when the theory-guided model is already complex and difficult to estimate. Furthermore, the score-based nature of both methods makes them computationally very efficient. This advantage is most noticeable when comparing the runtime of score-guided SEM trees with the conventional SEM trees suggested by

Brandmaier et al. (2013). The score implementation can often reduce runtime drastically from hours to seconds or minutes. Finally, when compared to score-based, IPC regression and score-guided SEM trees also provide estimates of parameter heterogeneity in addition to hypothesis tests.

Clearly, the main disadvantage of IPC regression and score-guided SEM trees is that the methods' performance depends on the covariates available. If none of these covariates is in any way related to individual or group differences, both methods will fail to detect any heterogeneity. Thus, IPC regression and score-guided SEM trees are not suited to confirm the homogeneity of a sample but merely confirm homogeneity with respect to available covariates. If a more stringent homogeneity test is needed, researchers may resort to latent class or finite mixture models.

## 7.2 Applications and Caveats

IPC regression and score-guided SEM trees can be applied in a variety of ways. Next, I discuss possible applications and caveats, starting with the topic of model modification.

### 7.2.1 Model Modification

Traditionally, model modification, sometimes also called specification search, refers to the process in which a model's fit is improved in a stepwise fashion (Kaplan, 1988; MacCallum, 1986; MacCallum et al., 1992; Yuan & Liu, 2021). The process typically begins by fitting a SEM to data. If the fit of the initial model is inadequate, one or more parameters are added to the model in order to improve its fit to the data. Especially in the early days of SEMs, the MI served as one of the main heuristics to guide model modification because it provides a computationally inexpensive way to assess how much the fit had improved had a new parameter been introduced to the model (MacCallum et al., 1992). However, other criteria for the goodness of fit can be used just as well. While model modification of SEMs was traditionally most often performed as a forward selection procedure that increased model complexity in the process, regularization techniques stemming originally from machine learning were proposed by Jacobucci et al. (2016) and Huang et al. (2017) and offer a backward specification search that penalizes model complexity (Yuan & Liu, 2021).

Model modification based on the MI has been criticized as a data-driven technique that is prone to capitalizing on chance characteristics of data and producing an overfitted model, which may perform well in a sample at hand but generalizes poorly to new data (MacCallum et al., 1992). Some methods have been proposed to minimize the risk of overfitting. Saris et al. (1987) and Saris et al. (2009) suggest using the MI in combination with the EPC, which approximates the value of a new parameter had it been added to the model. The authors argue that the EPC gives a direct estimate of the size of the

## 7. Discussion

misspecification that should be considered for model modification. MacCallum et al. (1992) investigate the use of cross-validating modified models and find mixed results. MacCallum et al. conclude that a conservative approach with few modifications of the original model and clear interpretability is most useful and call for researchers to acknowledge that their initial model has been subject to modification to improve its fits. Importantly, the criticism and recommendations for model modification based on the MI also provide valuable insights for model modification guided by IPC regression and score-guided SEM trees.

Model modification guided by IPC regression is complicated by multiple testing issues. IPC regression produces an estimate with a corresponding  $p$ -value for each model parameter and each covariate (ignoring possible interactions between covariates and polynomial effects). For example, when provided with a model with 20 parameters and 5 covariates, IPC regression will produce 100 estimates of how the covariates affect the model parameters. Not only might the sheer amount of information overwhelm the user, but it also increases the likelihood of false-positive findings, that is, type I errors. In Project 2, we discussed three strategies for how IPC regression results can be best translated into an improved model. In the following, I reiterate and add to these strategies. The first strategy aims to reduce the amount of positive or significant findings by either applying regularization as proposed in Project 2 or adjusting the significance level for multiple testing. If LASSO regularization is applied, I suggest treating every effect that was not set to zero as potentially important. An adjustment of the significance level, such as the well-known Bonferroni correction, can address multiple testing of covariates, model parameters, or both. In most cases, it seems reasonable to adjust only for the number of covariates to avoid overcorrection. Second, following the advice of Saris et al. (1987) and Saris et al. (2009) regarding the use of the EPC, I recommend considering the estimated effects of the covariates from a substantive point of view. Only those effects that are interpretable and large enough to matter in the substantive theory should give grounds for changing the model. Likewise, one should also consider what model parameters are found to be heterogeneous. Most likely, heterogeneity in nuisance parameters, such as error variance parameters, needs to be much more pronounced than heterogeneity in a regression parameter central to one's inquiry to justify a model modification. Third, I strongly recommend to cross-validate the augmented model. Cross-validation would be best performed on a completely new sample. Since new samples are likely not always available, researchers can split the sample into a training and a test data set if the sample is sufficiently large. The training data set is used for model modification guided by IPC regression, and the test data set is then used to assess the fit of the augmented model. Often, the sample will be too small to be reasonably split in half. In that case, one may modify the model using the complete sample and then evaluate the augmented model with  $k$ -fold cross-validation. Although the strategies outlined above follow the recommendations

regarding model modification using the MI, they have not been applied in practice or studied. Thus, future research is needed to determine how IPC regression can be best used for model modification.

In contrast to IPC regression, score-guided SEM trees provide users with a MGSEM that could, in principle, replace the original model. However, in practice, the resulting MGSEM will often be too complex, comprising too many groups and parameters, which makes it hard to interpret. Fortunately, the **semtree** R package (Brandmaier et al., 2013) offers users several options to prevent SEM trees from splitting the sample into too sparse groups. In Project 3, we successfully applied a Bonferroni adjustment of the significance level to control for the testing of multiple covariates. While this correction prevented an inflation of type I errors in our simulation study, it has the apparent disadvantage that it decreases the tree’s power to detect heterogeneity drastically if the number of covariates is large. Another strategy is to specify a minimum number of individuals within each group. By selecting a reasonable minimal group size, researchers can limit the number of splits and be assured that the SEMs in the leaf nodes (i.e., the submodels after the last splits of the tree) are stable enough to be interpreted. Besides the number of splits, the number of model parameters is also an issue. Currently, the SEM tree algorithm estimates a separate set of parameters for every group identified by a tree. On the one hand, this approach makes SEM trees flexible by detecting heterogeneity in all parts of a model, but, on the other hand, it leads most likely to an over-parametrized MGSEM. In Project 3, we investigate the use of so-called global parameter constraints that set the value of a parameter equal across all groups. We found that while these global constraints could reduce the number of parameters, they can greatly hinder a tree’s ability to detect heterogeneity when heterogeneous parameters are constrained. Therefore, we discourage specifying global constraints unless reliable prior knowledge about the homogeneity of certain parameters exists. So, what is currently missing, and could facilitate model modification, is an automated way that identifies which parameters are equal across specific groups of the tree. Alternatively, users could also be provided with more detailed information on which parameters are most heterogeneous and drive the tree’s split decisions. Finally, model modification guided by score-guided SEM trees is complicated by the well-recognized problem that tree methods tend to be unstable. Even slight changes to the original data sets (e.g., removing some individuals with outliers) may sometimes change one of the top splits, leading to an entirely different tree structure (Philipp et al., 2018). To solve this issue, Brandmaier et al. (2016) proposed SEM forests. SEM forests are an ensemble approach that are based on random forests (Breiman, 2001) and estimate the importance of covariates. SEM forests consist of a large number of SEM trees. These trees are provided with a subset of covariates, which is resampled for every tree. The resampling decorrelates covariates and allows to compute a new measure that ranks the importance of covariates for a forest. I suggest that SEM forests always



accompany SEM trees to ensure that the importance of covariates is factored into the modified model.

I close this discussion of model modification by reiterating and adding to MacCallum et al.'s (1992) call for transparency about model modification. In addition to disclosing that a theory-grounded model was changed, authors should strive to be as precise as possible in stating how they augmented their models. Brandmaier and Jacobucci (in press) highlight that researchers often fail to report the full set of analyses conducted. In particular, Brandmaier and Jacobucci identify meta-parameters (e.g., penalty terms used in regularized IPC regression or the minimum group size of SEM trees) as crucial pieces of information that are frequently not disclosed. Often, default values for these meta-parameters are employed, which can change in the developmental process of software (see Epskamp, 2019). Therefore, Brandmaier and Jacobucci (in press) encourage users to use workflows with containerized environments which guarantee that others can execute code with the same software version originally used for the analysis. An example of such a workflow is given by Peikert and Brandmaier (2021). If this is not feasible, Brandmaier and Jacobucci (in press) recommend making code publicly available so the full set of settings that were used can be examined.

### ***7.2.2 Causal Interpretation***

At various places, the language in this work suggests some sort of causal connection between covariates and parameter heterogeneity. For example, in Project 1 and 2, IPC regression is presented as a method to estimate the effects of covariates on individual or group differences, implicitly implying that the parameter differences are caused by the covariates. Likewise, the group differences found by score-guided SEM trees in Project 3 could easily be understood as estimates of causal quantities. However, this language was mainly employed for ease of illustration. Thus, and in view of the growing interest in the subject of causality and causal inferences in recent times (e.g., Pearl & Mackenzie, 2018), it seems worthwhile to clarify that the covariates' effects estimated by IPC regression and score-guided SEM trees are not guaranteed to be causal quantities. Such as other statistical methods like linear regression, which operate on the level of associations (Chen & Pearl, 2013), IPC regression and score-guided SEM trees merely predict heterogeneity. Besides the causal direction between covariate and parameter differences, IPC regression and score-guided SEM trees are also unable to rule out the possibility of spurious relationships that are caused by unobserved variables. Hence, additional reasoning and assumptions are needed to make a case for a convincing causal interpretation of the findings, especially in non-experimental data.

Importantly, this warning also pertains to demographic variables such as age, gender, or variables like personality traits. Often, it is erroneously assumed that these fixed variables can only be a cause but not an effect (Meehl, 1971; Spector & Brannick, 2011). Although

psychological or environmental factors can obviously not alter demographic variables like age, they can affect the age distribution of a sample. Spector and Brannick (2011) illustrate this argument with an example: they suppose that in a given sample, women report on average a higher job satisfaction than men. One possible explanation is that gender has caused the difference in job satisfaction. However, Spector and Brannick argue that it could also be the case that women are less tolerant to dissatisfying working conditions and therefore quit dissatisfying jobs more often than men. Thus, since less dissatisfied women work than dissatisfied men, we would expect that overall job satisfaction for women is higher than for men. Spector and Brannick also provide alternative scenarios in which job satisfaction affects the gender distribution of the sample.

In summary, I recommend being cautious with causal language when reporting the results of an IPC regression or score-guided SEM tree analysis in the same way as one would report linear regression results. Nevertheless, although IPC regression and score-guided SEM trees cannot identify and estimate causal effects on their own, they are still helpful by identifying which covariates are linked with parameter heterogeneity, which can be the first step of establishing a causal relationship (see Shadish et al., 2002).

### ***7.2.3 Predicting Interindividual Differences in Intraindividual Variation***

Especially in longitudinal data analysis, it is common to differentiate between two sources of variation: interindividual differences that occur between persons and intraindividual differences that unfold within persons over time (Schmiedek et al., 2020; Voelkle et al., 2014). Importantly, interindividual and intraindividual variation is rarely identical or even similar. One reason is that interindividual variation also reflects temporally stable differences, which per definition, does not feature in the quantification of intraindividual variation. Generally, inference about intraindividual variation from interindividual variation requires the ergodic assumption to hold (Molenaar, 2004), which implies that the sample is homogenous and its characteristics do not change over time (but see Adolf & Fried, 2019).

Interestingly, IPC regression and score-guided SEM trees can be useful tools to identify and predict interindividual differences, or heterogeneity, in intraindividual variation. However, the ability of the methods to do so depends largely on the data and model. For example, when provided with cross-sectional data, both methods are limited to predicting interindividual differences in model parameters. Generalizing these results to the intraindividual level is seldom appropriate and may lead to misleading conclusions (see Simpson's paradox; Kievit et al., 2013). However, when longitudinal data is available, the theory-guided model may explicitly account for variation on the interindividual and intraindividual level with different model parameters. Popular examples are dynamic panel models (Zyphur, Allison, et al., 2020; Zyphur, Voelkle, et al., 2020). These models are often specified with random intercepts that account for stable differences in the means of the observed variables (e.g., Hamaker et al., 2015). The variance parameters of these

random intercepts quantify the mean differences between persons. Other model parameters, such as autoregressive parameters, describe intraindividual variation by gauging how long random shocks persist in the system. In Project 1, we demonstrate how interindividual differences in parameters pertaining to intraindividual variation can be identified with IPC regression. A similar demonstration in which SEM trees are used was put forward by Brandmaier et al. (2018).

#### **7.2.4 Latent Covariates**

Throughout the dissertation, covariates were conveniently assumed to be directly observable. Unfortunately, this assumption is unrealistic for applied psychological research, where the influence of latent variables such as personality traits or intelligence is often of interest. To study parameter heterogeneity with respect to latent covariates, I propose a simple two-step procedure. First, factor scores are obtained by fitting confirmatory factor analysis models (see Hardt et al., 2019, for an overview about different approaches to obtain factor scores). Then, IPC regression or score-guided SEM trees are provided with the corresponding factor scores. Alternatively, IPCs can also be studied in SEMs that contain measurement models of the latent variables.

### **7.3 Areas of Applications**

The question remains, which psychological research areas are most promising for applications of IPC regression and score-guided SEM trees. From a methodological point of view, I highly recommend always testing SEM parameters for heterogeneity because the commonly applied measures to assess SEMs do not capture heterogeneity (see Section 2.2). Further, given that sufficient sample size and predictive covariates are needed to detect heterogeneity, large data sets rich in covariates seem especially well suited. Typical examples include longitudinal panel data sets such as the German Socio-Economic Panel Study or the growing number of ambulatory assessment data sets (Wrzus & Mehl, 2015) that usually contain plenty of measurements of a small number of individuals.

From a substantive point of view, the application of IPC regression and score-guided SEM trees is most interesting in research areas that are characterized by a large degree of heterogeneity, such as cognitive aging (Lindenberger, 2014), brain structures (Kievit et al., 2018), indicators of successful aging (Gerstorf et al., 2016), or personality traits (Chamorro-Premuzic, 2014). For instance, SEM trees have been successfully applied to study heterogeneity in well-being (Brandmaier et al., 2017), the relationship between brain and cognition (de Mooij et al., 2018; Simpson-Kent et al., 2020), and psychological disorders (Ammerman et al., 2019). Moreover, both methods seem to align well with the rationales of longitudinal research as put forward by Baltes and Nesselroade (1979), who call for the identification of interindividual differences in intraindividual change and its

causes.

### 7.4 Conclusion and Outlook

This cumulative dissertation developed two novel procedures to detect and predict heterogeneity with covariates in contemporary psychological models. Both methods approach heterogeneity by analyzing the partial derivative of the log-likelihood with respect to model parameters, also known as the score function. Based on pioneering work by Oberski (2013), Project 1 advanced the IPC regression framework and highlighted its potential to predict heterogeneity in dynamic panel models. Project 2 focussed on the software implementation of IPC regression and provided further benchmarks. Finally, project 3 introduced score-guided SEM trees by combining score-based tests (Merkle & Zeileis, 2013) with SEM trees (Brandmaier et al., 2013), solving runtime and multiple testing issues in the SEM tree framework.

The concluding discussion has shown that while IPC regression and score-guided SEM trees are promising methods, much research remains to be done. While score-guided SEM trees are much more developed, IPC regression would benefit from further theoretical research, for instance, concerning different strategies for model modification. As the work presented in this dissertation is primarily theoretical and based on simulation studies, applications with real psychological data seem to be the next step.

## 8 References

- Adolf, J. K., & Fried, E. I. (2019). Ergodicity is sufficient but not necessary for group-to-individual generalizability. *Proceedings of the National Academy of Sciences*, *116*(14), 6540–6541. <https://doi.org/10.1073/pnas.1818675116>
- Aitchison, J., & Silvey, S. D. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics*, *29*(3), 813–828. <https://doi.org/10.1214/aoms/1177706538>
- Allison, P. D., Williams, R., & Moral-Benito, E. (2017). Maximum likelihood for cross-lagged panel models with fixed effects. *Socius*, *3*, 1–17. <https://doi.org/10.1177/2378023117710578>
- Ammerman, B. A., Jacobucci, R., & McCloskey, M. S. (2019). Reconsidering important outcomes of the nonsuicidal self-injury disorder diagnostic criterion A. *Journal of Clinical Psychology*, *75*(6), 1084–1097. <https://doi.org/10.1002/jclp.22754>
- Arminger, G., Stein, P., & Wittenberg, J. (1999). Mixtures of conditional mean- and covariance-structure models. *Psychometrika*, *64*(4), 475–494. <https://doi.org/10.1007/bf02294568>
- Asparouhov, T., & Muthén, B. O. (2021). Bayesian estimation of single and multilevel models with latent variable interactions. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 314–328. <https://doi.org/10.1080/10705511.2020.1761808>
- Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 1–39). Academic Press.
- Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824. <https://doi.org/10.1016/j.paid.2006.09.018>
- Becker, J.-M., Rai, A., Ringle, C. M., & Völckner, F. (2013). Discovering unobserved heterogeneity in structural equation models to avert validity threats. *MIS Quarterly*, *37*(3), 665–694. <https://doi.org/10.25300/misq/2013/37.3.01>
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. <https://doi.org/10.1037/0033-2909.107.2.238>
- Bentler, P. M. (1995). *EQS structural equations program manual*. Mutivariate Software.
- Berk, R. A. (2006). An introduction to ensemble methods for data analysis. *Sociological Methods & Research*, *34*(3), 263–295. <https://doi.org/10.1177/0049124105283119>
- Biesanz, J. (2012). Autoregressive longitudinal models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 459–471). Guilford Press.
- Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

## 8. References

- Bollen, K. A., & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces*, *89*(1), 1–34. <https://doi.org/10.1353/sof.2010.0072>
- Bollen, K. A., & Curran, P. J. (2005). *Latent curve models*. John Wiley & Sons.
- Brandmaier, A. M., & Jacobucci, R. C. (in press). Machine-learning approaches to structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling*. Guilford Press.
- Brandmaier, A. M., Driver, C. C., & Voelkle, M. C. (2018). Recursive partitioning in continuous time analysis. In K. van Montfort, J. H. L. Oud, & M. C. Voelkle (Eds.), *Continuous time modeling in the behavioral and related sciences* (pp. 259–282). Springer.
- Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods*, *21*(4), 566–582. <https://doi.org/10.1037/met0000090>
- Brandmaier, A. M., Ram, N., Wagner, G. G., & Gerstorf, D. (2017). Terminal decline in well-being: The role of multi-indicator constellations of physical health and psychosocial correlates. *Developmental Psychology*, *53*(5), 996–1012. <https://doi.org/10.1037/dev0000274>
- Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*(1), 71–86. <https://doi.org/10.1037/a0030001>
- Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Breusch, T. S., & Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies*, *47*(1), 239–253. <https://doi.org/10.2307/2297111>
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83. <https://doi.org/10.1111/j.2044-8317.1984.tb00789.x>
- Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, *8*, 1–24. <https://doi.org/10.1002/j.2333-8504.1973.tb00197.x>
- Bürkner, P.-C., & Charpentier, E. (2020). Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 420–451. <https://doi.org/10.1111/bmsp.12195>
- Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, *36*(3), 153–157. <https://doi.org/10.2307/2683166>
- Chamorro-Premuzic, T. (2014). *Personality and individual differences* (3rd ed.). John Wiley & Sons.

## 8. References

- Chen, B., & Pearl, J. (2013). Regression and causation: A critical examination of six econometrics textbooks. *Real-World Economics Review*, *65*, 2–20. [https://ftp.cs.ucla.edu/pub/stat\\_ser/r395.pdf](https://ftp.cs.ucla.edu/pub/stat_ser/r395.pdf)
- Chou, C. P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. *Multivariate Behavioral Research*, *25*(1), 115–136. [https://doi.org/10.1207/s15327906mbr2501\\_13](https://doi.org/10.1207/s15327906mbr2501_13)
- Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148. <https://doi.org/10.1111/j.1745-3984.2005.00007>
- de Mooij, S. M. M., Henson, R. N. A., Waldorp, L. J., & Kievit, R. A. (2018). Age differentiation within gray matter, white matter, and between memory and white matter in an adult life span cohort. *Journal of Neuroscience*, *38*(25), 5826–5836. <https://doi.org/10.1523/JNEUROSCI.1627-17.2018>
- Duncan, O. D. (1966). Path analysis: Sociological examples. *American Journal of Sociology*, *72*(1), 1–16. <https://doi.org/10.1086/224256>
- Epskamp, S. (2019). Reproducibility and replicability in a fast-paced methodological world. *Advances in Methods and Practices in Psychological Science*, *2*(2), 145–155. <https://doi.org/10.1177/2515245919847421>
- Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 343–367. [https://doi.org/10.1207/s15328007sem1203\\_1](https://doi.org/10.1207/s15328007sem1203_1)
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, *222*, 309–368. <https://doi.org/10.1098/rsta.1922.0009>
- Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*(5), 700–725. <https://doi.org/10.1017/S0305004100009580>
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. <https://doi.org/10.18637/jss.v033.i01>
- Gates, K. M., Lane, S. T., Varangis, E., Giovanello, K., & Guskiewicz, K. (2017). Unsupervised classification during time-series model building. *Multivariate Behavioral Research*, *52*(2), 129–148. <https://doi.org/10.1080/00273171.2016.1256187>
- Gates, K. M., & Molenaar, P. C. M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, *63*(1), 310–319. <https://doi.org/10.1016/j.neuroimage.2012.06.026>

## 8. References

- Gerstorff, D., Hoppmann, C. A., Löckenhoff, C. E., Infurna, F. J., Schupp, J., Wagner, G. G., & Ram, N. (2016). Terminal decline in well-being: The role of social orientation. *Psychology and Aging, 31*(2), 149–165. <https://doi.org/10.1037/pag0000072>
- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*(3), 424–438. <https://doi.org/10.2307/1912791>
- Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20*(1), 102–116. <https://doi.org/10.1037/a0038889>
- Hardt, K., Hecht, M., Oud, J. H. L., & Voelkle, M. C. (2019). Where have the persons gone? – An illustration of individual score methods in autoregressive panel models. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(2), 310–323. <https://doi.org/10.1080/10705511.2018.1517355>
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- Helwig, N. E. (2017). Regression with ordered predictors via ordinal smoothing splines. *Frontiers in Applied Mathematics and Statistics, 3*(15), 1–13. <https://doi.org/10.3389/fams.2017.00015>
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3), 117–144. <https://doi.org/10.1080/03610739208253916>
- Hoyle, R. H. (Ed.). (2012). *Handbook of structural equation modeling*. Guilford Press.
- Hsiao, C. (2014). *Analysis of panel data*. Cambridge University Press.
- Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424–453. <https://doi.org/10.1037/1082-989x.3.4.424>
- Huang, P.-H., Chen, H., & Weng, L.-J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika, 82*(2), 329–354. <https://doi.org/10.1007/s11336-017-9566-9>
- Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 555–566. <https://doi.org/10.1080/10705511.2016.1154793>
- Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science, 16*(1), 39–59. <https://doi.org/10.1287/mksc.16.1.39>
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409–426. <https://doi.org/10.1007/bf02291366>
- Jöreskog, K. G. (1977). Factor analysis by least square and maximum likelihood methods. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Statistical methods for digital computers* (pp. 125–165). John Wiley & Sons.



## 8. References

- Jöreskog, K. G., Olsson, U. H., & Y. Wallentin, F. (2016). *Multivariate analysis with LISREL*. Springer.
- Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, *23*(1), 69–86. [https://doi.org/10.1207/s15327906mbr2301\\_4](https://doi.org/10.1207/s15327906mbr2301_4)
- Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A.-L., de Mooij, S. M. M., Moutoussis, M., Goodyer, I., Bullmore, E., Jones, P., Fonagy, P., the NSPN Consortium, Lindenberger, U., & Dolan, R. J. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience*, *33*, 99–117. <https://doi.org/10.1016/j.dcn.2017.11.007>
- Kievit, R. A., Frankenhuys, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology*, *4*, 1–14. <https://doi.org/10.3389/fpsyg.2013.00513>
- Kline, R. B. (Ed.). (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2. ed.). Springer.
- Lindenberger, U. (2014). Human cognitive aging: Corriger la fortune? *Science*, *346*(6209), 572–578. <https://doi.org/10.1126/science.1254403>
- Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21–39. <https://doi.org/10.1037/1082-989X.10.1.21>
- Ly, A., Marsman, M., Verhagen, J., Grasman, R. P., & Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology*, *80*, 40–55. <https://doi.org/10.1016/j.jmp.2017.05.006>
- MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*(1), 107–120. <https://doi.org/10.1037/0033-2909.100.1.107>
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. <https://doi.org/10.1037/0033-2909.111.3.490>
- Magidson, J., & Vermunt, J. K. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research*, *20*, 37–44. <https://jeroenvermunt.nl/cjmr2002.pdf>
- Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, *32*(8), 611–631. <https://doi.org/10.1177/0146621607312613>
- Marsh, H. W., Wen, Z., Hau, K.-T., & Nagengast, B. (2013). Structural equation models of latent interaction and quadratic effects. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling* (pp. 267–308). Information Age Publishing.

## 8. References

- Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. [https://doi.org/10.1207/s15328007sem1103\\_2](https://doi.org/10.1207/s15328007sem1103_2)
- Meehl, P. E. (1971). High school yearbooks: A reply to Schwarz. *Journal of Abnormal Psychology*, *77*(2), 143–148. <https://doi.org/10.1037/h0030750>
- Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259–284. <https://doi.org/10.1037/1082-989X.10.3.259>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. <https://doi.org/10.1007/bf02294825>
- Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika*, *79*(4), 569–584. <https://doi.org/10.1007/S11336-013-9376-7>
- Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika*, *78*(1), 59–82. <https://doi.org/10.1007/S11336-012-9302-4>
- Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*(1), 93–115. <https://doi.org/10.1037/1082-989x.9.1.93>
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.
- Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201–218. [https://doi.org/10.1207/s15366359mea0204\\_1](https://doi.org/10.1207/s15366359mea0204_1)
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398. <https://doi.org/10.1177/0049124194022003006>
- Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*(2), 463–469. <https://doi.org/10.1111/j.0006-341x.1999.00463.x>
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. [https://doi.org/10.1016/S0022-2496\(02\)00028-7](https://doi.org/10.1016/S0022-2496(02)00028-7)
- Nestler, S., & Humberg, S. (2021). Gimme's ability to recover group-level path coefficients and individual-level path coefficients. *Methodology*, *17*(1), 58–91. <https://doi.org/10.5964/meth.2863>
- Neudecker, H., & Satorra, A. (1991). Linear structural relations: Gradient and Hessian of the fitting function. *Statistics & Probability Letters*, *11*(1), 57–61. [https://doi.org/10.1016/0167-7152\(91\)90178-t](https://doi.org/10.1016/0167-7152(91)90178-t)

## 8. References

- Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, *20*(1), 175–240. <https://doi.org/10.2307/2331945>
- Nocedal, J., & Wright, S. J. (2006). *Numerical optimization* (2nd ed.). Springer.
- Oberski, D. L. (2013). A flexible method to explain differences in structural equation model parameters over subgroups. <http://daob.nl/wp-content/uploads/2013/06/SEM-IPC-manuscript-new.pdf>
- Oud, J. H. L., & Jansen, R. A. R. G. (2000). Continuous time state space modeling of panel data by means of SEM. *Psychometrika*, *65*(2), 199–215. <https://doi.org/10.1007/BF02294374>
- Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.
- Peikert, A., & Brandmaier, A. M. (2021). A reproducible data analysis workflow. *Quantitative and Computational Methods in Behavioral Sciences*, *1*, 1–27. <https://doi.org/10.5964/qcmb.3763>
- Philipp, M., Rusch, T., Hornik, K., & Strobl, C. (2018). Measuring the stability of results from supervised statistical learning. *Journal of Computational and Graphical Statistics*, *27*(4), 685–700. <https://doi.org/10.1080/10618600.2018.1473779>
- R Core Team. (2021). R: A language and environment for statistical computing. <https://www.R-project.org>
- Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *44*(1), 50–57. <https://doi.org/10.1017/S0305004100023987>
- Rao, C. R. (2005). Score test: Historical review and recent developments. In N. Balakrishnan, H. N. Nagaraja, & N. Kannan (Eds.), *Advances in ranking and selection, multiple comparisons, and reliability* (pp. 3–20). Birkhäuser.
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Saris, W. E., Satorra, A., & Sorbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology*, *17*, 105–129. <https://doi.org/10.2307/271030>
- Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 561–582. <https://doi.org/10.1080/10705510903203433>
- Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika*, *54*(1), 131–151. <https://doi.org/10.1007/BF02294453>
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit

## 8. References

- measures. *Methods of Psychological Research Online*, 8(2), 23–74. [https://www.dgps.de/fachgruppen/methoden/mpr-online/issue20/art2/mpr130\\_13.pdf](https://www.dgps.de/fachgruppen/methoden/mpr-online/issue20/art2/mpr130_13.pdf)
- Schmiedek, F., Lövdén, M., von Oertzen, T., & Lindenberger, U. (2020). Within-person structures of daily cognitive performance differ from between-person structures of cognitive abilities. *PeerJ*, 8(e9290), 1–28. <https://doi.org/10.7717/peerj.9290>
- Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inferences*. Houghton Mifflin Company.
- Silvey, S. D. (1959). The Lagrangian multiplier test. *The Annals of Mathematical Statistics*, 30(2), 389–407. <https://doi.org/10.1214/aoms/1177706259>
- Silvia, E. S., & MacCallum, R. C. (1988). Some factors affecting the success of specification searches in covariance structure modeling. *Multivariate Behavioral Research*, 23(3), 297–326. [https://doi.org/10.1207/s15327906mbr2303\\_2](https://doi.org/10.1207/s15327906mbr2303_2)
- Simpson-Kent, I. L., Fuhrmann, D., Bathelt, J., Achterberg, J., Borgeest, G. S., & Kievit, R. A. (2020). Neurocognitive reorganization between crystallized intelligence, fluid intelligence and white matter microstructure in two age-heterogeneous developmental cohorts. *Developmental Cognitive Neuroscience*, 41, 1–15. <https://doi.org/10.1016/j.dcn.2019.100743>
- Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(1), 131–161. <https://doi.org/10.1080/10705511.2019.1577140>
- Smit, A., Kelderman, H., & van der Flier, H. (2000). The mixed Birnbaum model: Estimation using collateral information. *Methods of Psychological Research Online*, 5(4), 31–43. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.629.4523&rep=rep1&type=pdf>
- Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, 27(2), 229–239. <https://doi.org/10.1111/j.2044-8317.1974.tb00543.x>
- Sörbom, D. (1989). Model modification. *Psychometrika*, 54(3), 371–384. <https://doi.org/10.1007/BF02294623>
- Spector, P. E., & Brannick, M. T. (2011). Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods*, 14(2), 287–305. <https://doi.org/10.1177/1094428110369842>
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25(2), 173–180. [https://doi.org/10.1207/s15327906mbr2502\\_4](https://doi.org/10.1207/s15327906mbr2502_4)

## 8. References

- Stigler, S. M. (2007). The epic story of maximum likelihood. *Statistical Science*, *22*(4), 598–620. <https://doi.org/10.1214/07-STS249>
- Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(25), 1–21. <https://doi.org/10.1186/1471-2105-8-25>
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. <https://doi.org/10.1037/a0016973>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273–282. <https://doi.org/10.1111/j.1467-9868.2011.00771.x>
- van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. <https://doi.org/10.1080/17405629.2012.686740>
- Voelkle, M. C., Brose, A., Schmiedek, F., & Lindenberger, U. (2014). Toward a unified framework for the study of between-person and within-person structures: Building a bridge between two research paradigms. *Multivariate Behavioral Research*, *49*(3), 193–213. <https://doi.org/10.1080/00273171.2014.889593>
- Voelkle, M. C., Oud, J. H. L., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods*, *17*(2), 176–192. <https://doi.org/10.1037/a0027543>
- von Oertzen, T., & Brick, T. R. (2014). Efficient Hessian computation using sparse matrix derivatives in RAM notation. *Behavior Research Methods*, *46*(2), 385–395. <https://doi.org/10.3758/s13428-013-0384-4>
- Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(3), 426–482. <https://doi.org/10.2307/1990256>
- West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). Guilford Press.
- Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education*, *80*(1), 26–44. <https://doi.org/10.1080/00220973.2010.531299>

## 8. References

- Wrzus, C., & Mehl, M. R. (2015). Lab and/or field? Measuring personality processes and their social consequences. *European Journal of Personality, 29*(2), 250–271. <https://doi.org/10.1002/per.1986>
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. <https://doi.org/10.1177/1745691617693393>
- Yuan, K.-H., & Liu, F. (2021). Which method is more reliable in performing model modification: Lasso regularization or Lagrange multiplier test? *Structural Equation Modeling: A Multidisciplinary Journal, 28*(1), 69–81. <https://doi.org/10.1080/10705511.2020.1768858>
- Yuan, K.-H., & Bentler, P. M. (2006). Structural equation modeling. In S. Sinharay (Ed.), *Handbook of statistics* (pp. 297–358). Elsevier.
- Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica, 61*(4), 488–508. <https://doi.org/10.1111/j.1467-9574.2007.00371.x>
- Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics, 17*(2), 492–514. <https://doi.org/10.1198/106186008X319331>
- Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software, 95*(1), 1–36. <https://doi.org/10.18637/jss.v095.i01>
- Zyphur, M. J., Allison, P. D., Tay, L., Voelkle, M. C., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi, A., Pierides, D. C., Koval, P., & Diener, E. (2020). From data to causes I: Building a general cross-lagged panel model (GCLM). *Organizational Research Methods, 23*(4), 651–687. <https://doi.org/10.1177/1094428119847278>
- Zyphur, M. J., Voelkle, M. C., Tay, L., Allison, P. D., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi, A., Pierides, D. C., Koval, P., & Diener, E. (2020). From data to causes II: Comparing approaches to panel data analysis. *Organizational Research Methods, 23*(4), 688–716. <https://doi.org/10.1177/1094428119847280>

## 9 Appendix

The appendix section contains the following manuscripts. Note that continuous page numbers are presented on the bottom throughout the appendices.

1. Arnold, M., Oberski, D. L., Brandmaier, A. M., & Voelke, M. C. (2020). Identifying heterogeneity in dynamic panel models with individual parameter contribution regression. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(4), 613-628. <https://doi.org/10.1080/10705511.2019.1667240>

See pages 47 to 62.

2. Arnold, M., Brandmaier, A. M., & Voelke, M. C. (2021). Predicting differences in model parameters with individual parameter contribution regression using the R package ipcr. *Psych*, 3(3), 360-385. <https://doi.org/10.3390/psych3030027>

See pages 63 to 88.

3. Arnold, M., Voelke, M. C., & Brandmaier, A. M. (2021). Score-guided structural equation model trees. *Frontiers in Psychology*, 11, 1-18. <https://doi.org/10.3389/fpsyg.2020.564403>

See pages 89 to 106.

# Identifying Heterogeneity in Dynamic Panel Models with Individual Parameter Contribution Regression

Manuel Arnold,<sup>1,2</sup> Daniel L. Oberski,<sup>3</sup> Andreas M. Brandmaier,<sup>1,2,4</sup> and Manuel C. Voelkle<sup>1,4</sup>

<sup>1</sup>Humboldt University of Berlin

<sup>2</sup>Max Planck UCL Centre for Computational Psychiatry and Ageing Research

<sup>3</sup>Utrecht University

<sup>4</sup>Max Planck Institute for Human Development

Dynamic panel models are a popular approach to study interrelationships between repeatedly measured variables. Often, dynamic panel models are specified and estimated within a structural equation modeling (SEM) framework. An endemic problem threatening the validity of such models is unmodelled heterogeneity. Recently, individual parameter contribution (IPC) regression was proposed as a flexible method to study heterogeneity in SEM parameters as a function of observed covariates. In the present paper, we derive how IPCs can be calculated for general maximum likelihood estimates and evaluate the performance of IPC regression to estimate group differences in dynamic panel models in discrete and continuous time. We show that IPC regression can be slightly biased in samples with large group differences and present a bias correction procedure. IPC regression showed generally promising results for discrete time models. However, due to highly nonlinear parameter constraints, caution is indicated when applying IPC regression to continuous time models.

**Keywords:** Autoregressive cross-lagged model, continuous time modeling, heterogeneity, structural equation modeling

## INTRODUCTION

Dynamic panel models (Hsiao, 2014) are routinely used in econometrics, psychology, and sociology to model the coupling between several repeatedly measured variables. Building upon the idea of Granger causality (Granger, 1969), dynamic models

allow answering questions concerned with the direction and strength of reciprocal relationships. Especially in psychological research, it is common practice to specify and estimate dynamic panel models within the structural equation modeling (SEM) framework (e.g., Allison, Williams, & Moral-Benito, 2017; Bollen & Brand, 2010; Zyphur, Allison et al., 2019, Zyphur, Voelkle et al., 2019).

An endemic problem that complicates the analysis of longitudinal panel data are systematic differences across individuals or groups. For instance, individuals may show stable, trait-like differences in the mean levels; a random shock might have a long-lasting effect on some persons, while its effect vanishes quickly for others; or the coupling between processes may differ across subjects. By overlooking such heterogeneity, researchers risk drawing incorrect conclusions from their data (Halaby, 2004).

Heterogeneity can often be explained through covariates such as demographic variables, biomarkers, or personality traits. Various approaches have been suggested to identify if and how covariates are linked to individual or group differences

---

Correspondence should be addressed to Manuel Arnold, Department of Psychology, Humboldt University of Berlin, Unter den Linden 6, Berlin 10099, Germany. E-mail: [arnoldmz@hu-berlin.de](mailto:arnoldmz@hu-berlin.de)

Supplemental files can be accessed [here](#).

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/hsem](http://www.tandfonline.com/hsem).

This article has been republished with minor changes. The changes do not affect the academic content of the article.

© 2019 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



in dynamic panel models. A popular way is the use of multi-level models with random effects (e.g., Singer & Willet, 2003). For instance, dynamic panel models are often specified with random intercepts to account for trait-like differences in the mean level of the observed variables (e.g., Hamaker, Kuiper, & Grasman, 2015). By regressing random effects on covariates, multilevel models can also be used to explore correlates and predictors of heterogeneity. Another popular approach to investigate heterogeneity are multi-group structural equation models (MGSEM; Sörbom, 1974) which allow the specification of panel models with different parameter values across groups. MGSEMs are particularly useful if the number of groups is small. However, using MGSEMs to disentangle the effects of many grouping variables can become tedious as multiple MGSEMs need to be specified and estimated. Fortunately, there exist approaches to perform such testing automatically, which become feasible with large sample sizes: Brandmaier, von Oertzen, McArdle, and Lindenberger (2013) and Brandmaier, Prindle, McArdle, and Lindenberger (2016) proposed a combination of MGSEMs and recursive partitioning methods to recover groups with similar parameter values. These so-called SEM trees or SEM forests fit a large number of MGSEMs to identify which grouping variables are important. Recently, Brandmaier, Driver, and Voelkle (2018) also applied these methods to dynamic panel models.

While the above methods are able to detect heterogeneity in a wide range of situations, they also come with certain drawbacks. The use of random effects to detect individual or group differences in dynamic panel models is often hindered by difficulties to specify the random effects for certain types of parameters. Whereas including random effects for intercept parameters is relatively straightforward, specifying random effects for regression and variance parameters is much more problematic and usually requires Bayesian methods (e.g., Driver & Voelkle, 2018; Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker, 2016). A drawback of MGSEM and MGSEM-based approaches like SEM trees and forests is that these methods require either categorical grouping variables or require continuous covariates to be split into meaningful grouping variables which might obscure the relationship between differences in a parameter and a continuous covariate. Furthermore, SEM trees and forests may experience difficulties when there is a clear set of target parameters of interest. Since these methods compare the group-wise likelihood, which considers differences in all parameters across all levels of the covariates jointly, the difference of interest may be masked if a larger difference is found in other parameters. This masking effect is well-known in the regression mixture literature (George et al., 2013) and may occur particularly in the case of distributional misspecification (e.g., Usami, Hayes, & McArdle, 2017). Finally, especially in large data sets, the computational burden of methods like Bayesian multilevel models, SEM trees, and SEM forests often constitutes a major impediment to implement these approaches in practice.

As an alternative approach to identify and estimate heterogeneity in dynamic panel models, we propose the use of individual parameter contribution (IPC) regression (Oberski, 2013). As we will discuss in the following, the IPC regression framework allows modeling SEM parameters as a function of covariates. Put shortly, IPC regression proceeds in three steps. First, a theory-driven (confirmatory) SEM is specified and estimated. Second, individual contributions to all model parameters are calculated using the case-wise derivative of the log-likelihood function. The resulting IPCs approximate individual-specific parameter values. Third, the IPCs are regressed on a set of categorical or continuous covariates to explain group differences or individual differences in the parameters. For instance, a researcher could regress the IPCs to one parameter on individuals' age to test whether this parameter is invariant to age differences or to estimate how the parameter changes as a function of age.

The primary advantages of IPC regression over other approaches to heterogeneity outlined above are its simplicity, flexibility, and low computational demand. IPC regression separates the estimation of the theory-driven model from the investigation of individual group differences. This separation is especially useful if the theory-driven model is complex, that is, has many observed variables and parameters. Although the underlying mathematics can be challenging, on the side of the applied researcher, basic knowledge of linear regression analysis is sufficient for successfully applying IPC regression in practice. IPC regression allows testing every type of SEM parameter (e.g., means, variances, covariances) for individual or group differences without the need for specifying random effects. Moreover, the method allows studying the effect of multiple grouping variables as well as continuous covariates and their interactions. Furthermore, IPC regression is a computationally lightweight procedure that can be performed in seconds.

IPCs are not limited to SEMs and can be derived for every type of maximum likelihood estimate. The contributions are calculated by linearizing the case-wise derivative of the log-likelihood function around the maximum likelihood estimates. The case-wise derivative of the log-likelihood function, also known as score function, has long been used to investigate the plausibility of statistical models (e.g., Zeileis, 2005; Zeileis & Hornik, 2007). Recently, score-based tests became popular in the exploration of measurement invariance in SEM (Merkle, Fan, & Zeileis, 2014; Merkle & Zeileis, 2013; Wang, Merkle, & Zeileis, 2014; Wang, Strobl, Zeileis, & Merkle, 2018). These score-based tests are used to test measurement invariance with respect to a continuous or ordinal auxiliary variable. IPC regression is different to these tests by providing estimates of how a model parameter varies as a function of covariates. Other frequently applied score-based approaches to identify misspecification in SEMs are the modification index (Sörbom, 1989) and the expected parameter change (Saris, Satorra, & Sörbom, 1987), which both test the validity of certain parameter restrictions but do not address the problem of parameter heterogeneity even though they are closely related (Oberski, 2013).

As of now, IPC regression has only been evaluated for a confirmatory factor analysis model (CFA; Brown, 2006). In a Monte Carlo simulation, Oberski (2013) reported excellent finite sample performance. We will later show that these results do not fully generalize to more complex models such as dynamic panel models. In general, large individual or group differences in one specific parameter can lead to biased IPC regression estimates for that specific parameter and also may lead to biased IPC regression estimates for other parameters. As a consequence, large differences in one parameter can increase the risk of a type I error in other constant parameters. To solve this problem, we propose a bias correction procedure termed iterated IPC regression that we recommend for dynamic panel models. The remainder of this article is organized as follows: first, we will briefly present bivariate dynamic panel models in discrete and continuous time. Second, IPC regression is formally introduced. Third, we evaluate the finite-sample properties of IPC regression for dynamic panel models in two simulation studies.

AUTOREGRESSIVE AND CROSS-LAGGED MODELS FOR PANEL DATA

The following section gives an outline of the SEM specifications for two simple dynamic panel models in discrete and continuous time that will be used throughout the present article. Readers unfamiliar with dynamic panel models are referred to Biesanz (2012). More details about the continuous-time models are given by Voelkle, Oud, Davidov, and Schmidt (2012).

Figure 1 shows a path diagram for a bivariate dynamic panel model for three waves of data. This structural model can be described with the following two equations:

$$x_{i,t} = \beta_{xx}x_{i,t-1} + \beta_{xy}y_{i,t-1} + u_{i,t} \tag{1}$$

$$y_{i,t} = \beta_{yy}y_{i,t-1} + \beta_{yx}x_{i,t-1} + v_{i,t}, \quad i = 1, \dots, n, \quad t = 2, 3 \tag{2}$$

Here,  $x_{i,t}$  and  $y_{i,t}$  are the measurements of two different variables of individual  $i$  at time point  $t$ . For sake of simplicity, we assume that  $x$  and  $y$  are free of measurement error and mean centered.

The regression coefficients  $\beta_{xx}$  and  $\beta_{yy}$  are called autoregressive parameters and they describe the stability in each  $x$  and  $y$  from one measurement occasion to the next. The regression coefficients  $\beta_{xy}$  and  $\beta_{yx}$  are referred to as cross-lagged effects and indicate how  $x$  influences  $y$  and vice versa. The initial assessments of  $x$  and  $y$  are treated as exogenous variables with zero mean and variance  $\phi_{xx}$ ,  $\phi_{yy}$  respectively, and covariance  $\phi_{yx}$ . For the remaining measurement occasions,  $u$  and  $v$  denote the dynamic error terms. The variance and covariance parameters of the dynamic error terms are symbolized by  $\psi_{xx}$ ,  $\psi_{yy}$ , and  $\psi_{yx}$  respectively.

Equations (3a)–(3c) show the SEM specification of the model in Figure 1:

$$\mathbf{y}_i = \mathbf{B}\mathbf{y}_i + \boldsymbol{\zeta}_i \tag{3a}$$

$$\begin{bmatrix} x_{i,1} \\ y_{i,1} \\ x_{i,2} \\ y_{i,2} \\ x_{i,3} \\ y_{i,3} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \beta_{xx} & \beta_{xy} & 0 & 0 & 0 & 0 \\ \beta_{yx} & \beta_{yy} & 0 & 0 & 0 & 0 \\ 0 & 0 & \beta_{xx} & \beta_{xy} & 0 & 0 \\ 0 & 0 & \beta_{yx} & \beta_{yy} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_{i,1} \\ y_{i,1} \\ x_{i,2} \\ y_{i,2} \\ x_{i,3} \\ y_{i,3} \end{bmatrix} + \begin{bmatrix} x_{i,1} \\ y_{i,1} \\ u_{i,2} \\ v_{i,2} \\ u_{i,3} \\ v_{i,3} \end{bmatrix} \tag{3b}$$

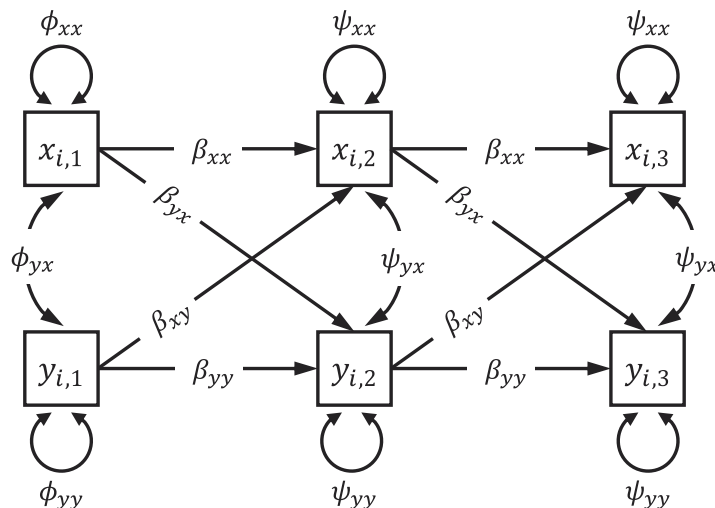


FIGURE 1 Path diagram of a bivariate autoregressive and cross-lagged panel model for three waves of data.

$$\text{Cov}(\zeta_i, \zeta_i) = \mathbf{\Phi} = \begin{bmatrix} \phi_{xx} & \phi_{yx} & 0 & 0 & 0 & 0 \\ \phi_{yx} & \phi_{yy} & 0 & 0 & 0 & 0 \\ 0 & 0 & \psi_{xx} & \psi_{yx} & 0 & 0 \\ 0 & 0 & \psi_{yx} & \psi_{yy} & 0 & 0 \\ 0 & 0 & 0 & 0 & \psi_{xx} & \psi_{yx} \\ 0 & 0 & 0 & 0 & \psi_{yx} & \psi_{yy} \end{bmatrix} \quad (3c)$$

The resulting model-implied covariance matrix of  $x$  and  $y$  is given by

$$\text{Cov}(\mathbf{y}_i, \mathbf{y}_i) = \mathbf{\Sigma}(\boldsymbol{\theta}) = (\mathbf{I}_6 - \mathbf{B})^{-1} \mathbf{\Phi} [(\mathbf{I}_6 - \mathbf{B})^{-1}]^T, \quad (4)$$

where  $\boldsymbol{\theta}$  is a vector with the model parameters and  $\mathbf{I}_6$  denotes an identity matrix of order six.

Although not explicitly stated, the temporally spacing between assessments plays an important role in the model as presented in Figure 1. The model treats time as a discrete variable that indicates the temporally ordering of the assessments and is therefore also referred to as *discrete-time* dynamic panel model. As pointed out elsewhere (e.g., Oud, 2007; Oud & Delsing, 2010; Voelkle et al., 2012), treating time as a discrete variable complicates comparing estimates from models with different sample schemes and can bias estimates if assessments are not equally spaced. A solution to these problems is treating time as a continuous variable using stochastic differential equation models (Oud & Jansen, 2000; for a recent overview of continuous-time modeling in the behavioral and related sciences, see van Montfort, Oud, & Voelkle, 2018). These *continuous-time* dynamic panel models allow estimating continuous-time parameters which can be used to extrapolate to any arbitrary time point.

Following Voelkle et al. (2012), we specify a continuous-time model by constraining the discrete-time model parameters from Figure 1 to functions of underlying continuous-time parameters  $\mathbf{A}$  and  $\mathbf{Q}$ , and the time intervals  $\Delta t_j$ . The new parameter matrix  $\mathbf{A}$  corresponds to the continuous-time version of auto- and cross-lagged effects, the drift parameters, while  $\mathbf{Q}$  contains the continuous-time version of dynamic error term variance parameters, or diffusion parameters:

$$\mathbf{A} = \begin{bmatrix} a_{xx} & a_{xy} \\ a_{yx} & a_{yy} \end{bmatrix} \quad \mathbf{Q} = \begin{bmatrix} q_{xx} & q_{yx} \\ q_{yx} & q_{yy} \end{bmatrix} \quad (5)$$

Let  $\Delta t_j$  be the time interval between the assessments  $j$  and  $j+1$ ; then the discrete-time regression coefficients are constrained as a function of  $\mathbf{A}$ :

$$\begin{bmatrix} b_{xx} & b_{xy} \\ b_{yx} & b_{yy} \end{bmatrix} = \exp(\mathbf{A} \cdot \Delta t_j), \quad (6)$$

where  $\exp$  denotes the matrix exponential function. The corresponding constraint for the variance of the dynamic error term is

$$\begin{bmatrix} \psi_{xx} & \psi_{yx} \\ \psi_{yx} & \psi_{yy} \end{bmatrix} = \text{irow} \left\{ \mathbf{A}_{\#}^{-1} [\exp(\mathbf{A}_{\#} \cdot \Delta t_j) - \mathbf{I}_4] \text{row}(\mathbf{Q}) \right\}, \quad (7)$$

where  $\mathbf{A}_{\#} := \mathbf{A} \otimes \mathbf{I}_2 + \mathbf{I}_2 \otimes \mathbf{A}$ . The operator  $\text{row}$  puts the elements of  $\mathbf{Q}$  into a column vector and the operator  $\text{irow}$  stacks the elements of a vector row-wise into a matrix.

The interpretation of the continuous-time model parameters can be facilitated by transforming them into the discrete-time parameters for an arbitrary time interval  $\Delta t_j$ . For example, plugging  $\Delta t_j = 1$  into the estimated drift parameters on the right-hand side of Equation (8) gives the discrete-time regression coefficients for a time interval of one between assessments.

## INDIVIDUAL PARAMETER CONTRIBUTION REGRESSION

In the following, we will show how heterogeneity in the parameters of dynamic panel models in discrete or continuous time can be identified and explained by IPC regression. To this end, we first motivate the derivation of IPCs for general maximum likelihood estimation. Next, we show how the contributions of SEM parameter estimates can be obtained. Then, we demonstrate that IPC regression can be biased in samples with large individual or group differences. As a solution to this problem, we present a bias correction procedure.

### IPCs to maximum likelihood estimates

Let  $\mathbf{y}_1, \dots, \mathbf{y}_n$  be a sample of independently distributed  $p$ -variate random variables with corresponding density functions  $f(\boldsymbol{\theta}_1; \mathbf{y}_1), \dots, f(\boldsymbol{\theta}_n; \mathbf{y}_n)$ . IPC regression is applicable in situations where differences between the individual-specific values of the  $q$ -variate parameter vector  $\boldsymbol{\theta}_i$  can be expressed as a function of a vector of covariates  $\mathbf{z}_i$ . For instance, differences in the parameter values of a two-group population can be estimated via IPC regression using a single dummy-coded grouping variable  $z_i$  as covariate.

For sake of illustration, we will assume that  $f$  is a multivariate normal density. The associated log-likelihood function for a single individual  $i$  is given by

$$\ln L(\boldsymbol{\theta}; \mathbf{y}_i) = -\frac{1}{2} \left\{ [\mathbf{y}_i - \boldsymbol{\mu}(\boldsymbol{\theta})]^T \mathbf{\Sigma}(\boldsymbol{\theta})^{-1} [\mathbf{y}_i - \boldsymbol{\mu}(\boldsymbol{\theta})] + \ln[\det(\mathbf{\Sigma}(\boldsymbol{\theta}))] + p \ln(2\pi) \right\} \quad (8)$$

with model-implied mean vector  $\boldsymbol{\mu}(\boldsymbol{\theta})$  and model-implied covariance matrix  $\mathbf{\Sigma}(\boldsymbol{\theta})$ . In the following, we will use  $\boldsymbol{\theta}$  to denote parameter values. True values of the parameters will be marked by a subscript, for instance  $\boldsymbol{\theta}_i$ , and the maximum likelihood estimate will be denoted by  $\hat{\boldsymbol{\theta}}$ .

The first and second derivatives of the log-likelihood function for a given person are important for computing IPCs. The first-order partial derivative of the individual log-likelihood function with respect to the parameters is the score function

$$\mathcal{S}(\boldsymbol{\theta}; \mathbf{y}_i) = \left[ \frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \theta^{(1)}} \quad \dots \quad \frac{\partial \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \theta^{(q)}} \right]^T, \quad (9)$$

where  $\theta^{(j)}$  denotes the  $j$ -th element of the parameter vector  $\boldsymbol{\theta}$ . Evaluation of the score function at specific parameter values measures to which extent an individual's log-likelihood is maximized. Note that the expected values of the score function at the true parameter values are zero, that is  $E[\mathcal{S}(\boldsymbol{\theta}; \mathbf{y}_i)] = \mathbf{0}$  holds for all individuals in the sample. The second-order partial derivative is known as Hessian matrix and will be denoted by

$$\mathcal{H}(\boldsymbol{\theta}; \mathbf{y}_i) = \frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}. \quad (10)$$

The expected value of the negative Hessian matrix evaluated at the true individual specific parameter values

$$\mathcal{I}(\boldsymbol{\theta}_i) = E \left[ - \frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}_i} \right] \quad (11)$$

is called the Fisher information matrix and plays a key role in determining standard errors and asymptotic sampling variance of the maximum likelihood estimates.

The maximum likelihood parameter estimate  $\hat{\boldsymbol{\theta}}$  can be obtained by solving the first-order conditions

$$\sum_{i=1}^n \mathcal{S}(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) = \mathbf{0}, \quad (12)$$

such that  $\hat{\boldsymbol{\theta}}$  is an extremum. In homogeneous samples, where  $\boldsymbol{\theta}_i = \boldsymbol{\theta}_0$  for  $i = 1, \dots, n$ , the resulting parameter estimate  $\hat{\boldsymbol{\theta}}$  is a consistent estimate of true parameter values  $\boldsymbol{\theta}_0$ . In heterogeneous samples,  $\hat{\boldsymbol{\theta}}$  will typically be close to the mean of the individuals' true parameter values  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ .

The idea behind the derivation of IPCs is to find the individual roots of the score function instead of finding the roots of the sum of all individual score values as shown in Equation (12). Hypothetically, solving  $\mathcal{S}(\hat{\boldsymbol{\theta}}_i; \mathbf{y}_i) = \mathbf{0}$  for every individual in the sample would yield individual parameter estimates  $\hat{\boldsymbol{\theta}}_1, \dots, \hat{\boldsymbol{\theta}}_n$ . Unfortunately, for many probability distribution such as the normal distribution, the system of equations  $\mathcal{S}(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) = \mathbf{0}$  does not have a unique solution for a single data point. However, we can approximate the individual scores by linearizing the mean of all

scores around the maximum likelihood estimate and then disaggregate the resulting expression:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \mathcal{S}(\boldsymbol{\theta}; \mathbf{y}_i) &\approx \frac{1}{n} \sum_{i=1}^n \mathcal{S}(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) \\ &+ \frac{1}{n} \sum_{i=1}^n \mathcal{H}(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \end{aligned} \quad (13)$$

Without changing the right-hand side of Equation (13), the Hessian matrix can be replaced by the estimated negative Fisher information matrix.

$$\frac{1}{n} \sum_{i=1}^n \mathcal{S}(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) - \mathcal{I}(\hat{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \quad (14)$$

In geometric terms, Equation (14) approximates the mean of scores with a tangent line at the maximum likelihood estimate. Now, we disaggregate this tangent into  $n$  individual tangents by replacing the mean of scores evaluated at the maximum likelihood estimate with the individual score values evaluated at the maximum likelihood estimate:

$$\mathcal{S}(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) - \mathcal{I}(\hat{\boldsymbol{\theta}}) (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \quad (15)$$

Finally, setting Equation (15) to zero and solving for  $\boldsymbol{\theta}$  yields a  $q$ -variate vector of individual's  $i$  contributions to the parameter estimates:

$$\mathbf{0} = \mathcal{S}(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) - \mathcal{I}(\hat{\boldsymbol{\theta}}) [\mathcal{IPC}(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) - \hat{\boldsymbol{\theta}}]$$

$$\mathcal{IPC}(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) = \hat{\boldsymbol{\theta}} + \mathcal{I}(\hat{\boldsymbol{\theta}})^{-1} \mathcal{S}(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) \quad (16)$$

The interpretation or meaning of the IPCs, and all averages or statistics based on them, follows from the interpretation of the maximum likelihood estimates  $\hat{\boldsymbol{\theta}}$ . This property is particularly important for dynamic panel models. The IPCs of autoregressive or cross-lagged parameter will only approximate the individual within-person relationship if the dynamic model separates the within-person process from stable between-person differences (Hamaker et al., 2015).

### IPCs to SEM parameter estimates

Instead of the sum of individual log-likelihoods in Equation (8), it is common to use the aggregated log-likelihood function (also called fitting function) in SEM (Voelkle, Oud, von Oertzen, & Lindenberger, 2012). The maximum likelihood fitting function for multivariate normally distributed variables is

$$\mathcal{F}(\bar{\mathbf{y}}, \mathbf{S}, \boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})) = [\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta})]^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} [\bar{\mathbf{y}} - \boldsymbol{\mu}(\boldsymbol{\theta})] + \text{tr}[\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}] - \ln[|\mathbf{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}|] - p, \quad (17)$$

with sample means  $\bar{\mathbf{y}}$  and sample covariance matrix  $\mathbf{S}$  (Yuan & Bentler, 2007). Optimizing either the sum of individual log-likelihood functions or an aggregated fitting function yields equivalent parameter estimates (Bollen, 1989).

Using the aggregated fitting function, IPCs to SEM parameter estimates are a function of the individual's data and two matrices  $\boldsymbol{\Delta}$  and  $\mathbf{V}$  that are provided by most standard SEM software packages. The first matrix  $\boldsymbol{\Delta}$  is the following Jacobian matrix

$$\boldsymbol{\Delta} = \frac{\partial[\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\sigma}(\boldsymbol{\theta})]^\top}{\partial \boldsymbol{\theta}}, \quad (18)$$

where  $\boldsymbol{\sigma}(\boldsymbol{\theta})$  denotes the half-vectorized model-implied covariance matrix.  $\boldsymbol{\Delta}$  indicates the sensitivity of the model-implied mean vector and covariance matrix to changes in the parameters. The second matrix is the weight matrix  $\mathbf{V}$  which depends on the chosen estimator (e.g., Savalei, 2014). In SEMs estimated with normal theory maximum likelihood, the corresponding weight matrix is

$$\mathbf{V} = \begin{bmatrix} \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2} \mathbf{D}_p^\top [\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \otimes \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}] \mathbf{D}_p \end{bmatrix}, \quad (19)$$

with duplication matrix  $\mathbf{D}_p$  (Magnus & Neudecker, 2019). Sample estimates of  $\boldsymbol{\Delta}$  and  $\mathbf{V}$  can be obtained by replacing  $\boldsymbol{\theta}$  with  $\hat{\boldsymbol{\theta}}$ .

Following Satorra (1989) and Neudecker and Satorra (1991), the Fisher information matrix can be expressed as  $\mathcal{I}(\boldsymbol{\theta}) = \boldsymbol{\Delta}^\top \mathbf{V} \boldsymbol{\Delta}$  and a partial derivative of the fitting function is given by

$$-\frac{1}{2} \frac{\partial \mathcal{F}(\bar{\mathbf{y}}, \mathbf{S}, \boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \boldsymbol{\Delta}^\top \mathbf{V} \left( \begin{bmatrix} \bar{\mathbf{y}} \\ \mathbf{s} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\mu}(\boldsymbol{\theta}) \\ \boldsymbol{\sigma}(\boldsymbol{\theta}) \end{bmatrix} \right). \quad (20)$$

Individual score values can be obtained by replacing the aggregated mean vector and covariance matrix in Equation (20) by the individual contributions to these sample moments. To this end, we define  $n$  vectors

$$\mathbf{d}_i := \begin{bmatrix} \mathbf{y}_i \\ \text{vech}([\mathbf{y}_i - \bar{\mathbf{y}}][\mathbf{y}_i - \bar{\mathbf{y}}]^\top) \end{bmatrix} \quad (21)$$

(Satorra, 1992), where the operator  $\text{vech}$  half-vectorizes a symmetric matrix. Note that the averaged individual contributions to the sample moments are identical to the observed sample moments, that is  $\frac{1}{n} \sum_{i=1}^n \mathbf{d}_i = [\bar{\mathbf{y}} \ \mathbf{s}]^\top$ .<sup>1</sup>

Thus, analogous to Equation (16), the individual contributions to SEM parameter estimates can be estimated by

$$\mathcal{IPC}(\hat{\boldsymbol{\theta}}; \mathbf{y}_i) = \hat{\boldsymbol{\theta}} + \left( \hat{\boldsymbol{\Delta}}^\top \hat{\mathbf{V}} \hat{\boldsymbol{\Delta}} \right)^{-1} \hat{\boldsymbol{\Delta}}^\top \hat{\mathbf{V}} \left( \mathbf{d}_i - \begin{bmatrix} \boldsymbol{\mu}(\hat{\boldsymbol{\theta}}) \\ \boldsymbol{\sigma}(\hat{\boldsymbol{\theta}}) \end{bmatrix} \right). \quad (22)$$

The above definition of the IPCs should replace that given by Oberski (2013), which yields incorrect means of the IPCs to factor loading and regression parameters.

### Predicting heterogeneity in panel models with IPC regression

The IPCs of a single individual are usually plagued by random fluctuation and will most likely be poor estimates of the true individual parameter values. However, studying the IPCs of groups of individuals or jointly modeling the IPCs of the whole sample can average out this noise. One obvious method for revealing meaningful differences in the parameters is linear regression estimated by ordinary least squares. Regressing the IPCs on a set of additional covariates  $\mathbf{z}$  allows to test and estimate if and how individual parameter values vary as a function of  $\mathbf{z}$ .

For instance, we could investigate via IPC regression whether the cross-lagged estimated effect  $\hat{\beta}_{yx}$  from  $x$  on  $y$  in the model shown in Figure 1 differs between women and men. To this end, the IPCs to  $\hat{\beta}_{yx}$  are regressed on a dummy variable  $z$  representing gender. Using women as a baseline group, the following IPC regression equation is estimated

$$\mathcal{IPC}_{i,\beta_{yx}} = \hat{\gamma}_0 + \hat{\gamma}_1 z_i + v_i, \quad (23)$$

where  $v_i$  is a random residual with mean zero. In the above equation, the IPC regression intercept  $\hat{\gamma}_0$  is the estimated value of  $\beta_{yx}$  for women and  $\hat{\gamma}_1$  denotes the estimated difference between women and men in  $\beta_{yx}$ . In other words, the IPC regression slope estimate  $\hat{\gamma}_1$  is a measure of heterogeneity in the cross-lagged effect  $\hat{\beta}_{yx}$  with respect to the covariate gender. As in standard regression analysis, a  $t$ -test could be applied to test  $\hat{\gamma}_1$ , that is, to infer whether the estimated subgroup difference between women and men in  $\hat{\beta}_{yx}$  is significantly different from zero. In this setup, Oberski (2013) showed that  $\hat{\gamma}_1$  and its Wald statistic are equivalent to the robust expected parameter change and robust modification index familiar from MGSEM (Satorra, 1989). Based on the size of the estimate and the test result, an informed decision can be made to modify the original model or not. An obvious choice of modification would be to use gender as a grouping variable in an MGSEM. The partial effects of several covariates on the parameters can be investigated using multiple linear regression analysis. To

<sup>1</sup> The biased estimate of the sample covariance is used.



investigate parameter heterogeneity in the complete model presented in Figure 1, an IPC regression equation needs to be estimated for each of the 10 model parameters:

$$\text{IPC}_{i,\beta_{xx}} = \widehat{\gamma}_{\beta_{xx}}^T \mathbf{z}_i + v_{i,\beta_{xx}} \quad (24a)$$

$$\text{IPC}_{i,\beta_{yx}} = \widehat{\gamma}_{\beta_{yx}}^T \mathbf{z}_i + v_{i,\beta_{yx}} \quad (24b)$$

⋮

$$\text{IPC}_{i,\psi_{yy}} = \widehat{\gamma}_{\psi_{yy}}^T \mathbf{z}_i + v_{i,\psi_{yy}} \quad (24j)$$

In Equations (24a)–(24j), the IPC regression estimates  $\widehat{\gamma}$  indicate the estimated effects from multiple covariates  $\mathbf{z}$  on a certain parameter estimate.

Due to its flexibility and computational efficiency, the linear regression framework offers researchers many possibilities to investigate heterogeneity by means of IPC regression. The interplay of the covariates could be studied by adding interactions to Equations (24a)–(24j). Furthermore, higher-order polynomial terms, such as quadratic or cubic terms can be easily specified to test for nonlinear relationships. If the number of covariates is large, regularization techniques like lasso (Tibshirani, 1996) could be used to aid the selection of important covariates. Finally, latent variables could be included by replacing the regression equations above with SEMs.

### Bias and inconsistency

IPC regression estimates of individual or group differences can be slightly inaccurate under certain circumstances. As shown above, IPC regression estimates are functions of maximum likelihood estimates and observed data. If an IPC regression estimate depends on a maximum likelihood estimate of a parameter that differs across individuals or groups, the IPC regression estimate will be inaccurate. As a rule of thumb, the inaccuracy increases with the amount of individual or group differences in the sample.

In the next paragraphs, we will demonstrate some properties of IPC regression estimates with the help of the exponential distribution. We chose the exponential distribution for the sake of clarity since it only has a single parameter. We will show that IPC regression estimates do not always correspond to individual- or group-specific maximum likelihood estimates, that is, with parameters estimated using homogeneous segments of the sample. Further, we will show that IPC regression estimates are not guaranteed to converge to the true individual- or group-specific parameter values and, as a result, can be inconsistent.

Consider the exponential distribution with density  $f(\lambda; y) = \lambda e^{-\lambda y}$ ,  $y \geq 0$ , and rate parameter  $\lambda > 0$ . We

assume that  $n$  individuals have been sampled in equal shares from a two-group population with different group-specific rate parameters  $\lambda_1$  and  $\lambda_2$ . The maximum likelihood estimate of  $\lambda$  for the whole sample is the reciprocal of the sample mean  $\widehat{\lambda} = \bar{y}^{-1} = n / \sum_{i=1}^n y_i$ . To recover the group differences in  $\widehat{\lambda}$ , we regress the IPCs to  $\widehat{\lambda}$  on a dummy variable  $z$  that is zero in the first group and one in the second group:

$$\text{IPC}_{i,\lambda} = \widehat{\gamma}_0 + \widehat{\gamma}_1 z_i + v_i \quad (25)$$

Next, we express the IPC regression estimates  $\widehat{\gamma}_0$  and  $\widehat{\gamma}_1$  as a function of group-specific maximum likelihood estimates  $\widehat{\lambda}_1$  and  $\widehat{\lambda}_2$  that are estimated separately in homogeneous subsamples. Intermediate steps can be found in the Appendix.

$$\widehat{\gamma}_0 = \frac{4\widehat{\lambda}_1^2 \widehat{\lambda}_2}{(\widehat{\lambda}_1 + \widehat{\lambda}_2)^2} \quad (26)$$

$$\widehat{\gamma}_1 = \frac{4\lambda_1 \lambda_2 (\widehat{\lambda}_2 - \widehat{\lambda}_1)}{(\widehat{\lambda}_1 + \widehat{\lambda}_2)^2} \quad (27)$$

Analogously to the bias of an estimator, which is the difference between an estimator’s expected value and the true value of the parameter, we may define the bias of an IPC regression estimate as the difference between an IPC regression estimate and the group-specific maximum likelihood estimate. Taking the probability limits of the resulting biases is trivial (see White, 1984) and allows us to determine whether the IPC regression estimates are consistent.

$$\widehat{\gamma}_0 - \widehat{\lambda}_1 = \frac{2\widehat{\lambda}_1^2 \widehat{\lambda}_2 - \widehat{\lambda}_1^3 - \widehat{\lambda}_1 \widehat{\lambda}_2^2}{(\widehat{\lambda}_1 + \widehat{\lambda}_2)^2} \xrightarrow{P} \frac{2\lambda_1^2 \lambda_2 - \lambda_1^3 - \lambda_1 \lambda_2^2}{(\lambda_1 + \lambda_2)^2} \neq 0 \quad (28)$$

$$\widehat{\gamma}_1 - (\widehat{\lambda}_2 - \widehat{\lambda}_1) = \frac{(\widehat{\lambda}_1 - \widehat{\lambda}_2)^3}{(\widehat{\lambda}_1 + \widehat{\lambda}_2)^2} \xrightarrow{P} \frac{(\lambda_1 - \lambda_2)^3}{(\lambda_1 + \lambda_2)^2} \neq 0 \quad (29)$$

It follows from Equations (28) and (29) that the IPC regression estimates  $\widehat{\gamma}_0$  and  $\widehat{\gamma}_1$  are systematically different from the group-specific maximum likelihood estimates. As this bias is unaffected by the sample size, the IPC regression estimates are also inconsistent. For instance, consider a sample drawn in equal shares with  $\lambda_1 = 0.5$  and  $\lambda_2 = 1.5$ . These parameter values imply that  $\widehat{\gamma}_0$  and  $\widehat{\gamma}_1$  converge to 0.375 and 0.75, respectively. Not only would IPC regression underestimate both group-specific parameter values (first group: 0.375 vs. 0.5, second group: 0.375 + 0.75 = 1.125 vs. 1.5) but also

underestimate the difference between both groups (0.75 vs. 1). In homogeneous samples, however, where  $\lambda_1 = \lambda_2$ , the IPC regression estimates are consistent as  $\widehat{\gamma}_0 - \widehat{\lambda}_1$  and  $\widehat{\gamma}_1 - (\widehat{\lambda}_2 - \widehat{\lambda}_1)$  converge in probability to zero.

Deriving the asymptotic bias for more complex models such as SEMs is challenging. However, later in the manuscript, we will demonstrate by means of Monte Carlo simulations that the results stated above generalize to dynamic panel models.

#### Iterative IPC regression: Bias correction procedure

To resolve the problems discussed in the previous paragraph, we propose an iterative algorithm similar to Fisher's scoring (e.g., Demidenko, 2013) to correct the bias of IPC regression. As discussed before, IPC regression estimates are biased if they depend on maximum likelihood estimates of parameters that differ across individuals or groups. This bias can be removed by replacing the pooled maximum likelihood estimates based on the entire sample with individual- or group-specific parameter estimates. However, instead of estimating these parameters separately, which is usually not possible for single individuals, we iteratively predict the individual- or group-specific parameters through IPC regression and re-estimate the IPC regression estimates.

Our proposed bias correction procedure, which we call iterative IPC regression, proceeds in the following way: First, an SEM is estimated and IPC regression is performed as described above. Second, the resulting IPC regression estimates are then used to predict a specific value for SEM parameter  $j$  of individual  $i$ :

$$\widetilde{\theta}_{i,j} = \mathbf{z}_i^T \widetilde{\boldsymbol{\gamma}}_j, \quad i = 1 \dots, n, \quad j = 1 \dots, q \quad (30)$$

Third, these individual-specific parameter values are used to recalculate the IPCs of each individual.

$$\widetilde{\text{IPC}}_i = \widetilde{\boldsymbol{\theta}}_i + \mathcal{I}(\widetilde{\boldsymbol{\theta}}_i)^{-1} \mathcal{S}(\widetilde{\boldsymbol{\theta}}_i; \mathbf{y}_i), \quad i = 1 \dots, n \quad (31)$$

Fourth, IPC regression estimates are re-estimated using the re-calculated IPCs for that specific parameter.

$$\widetilde{\boldsymbol{\gamma}}_j = \left( \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^T \right)^{-1} \sum_{i=1}^n \mathbf{z}_i^T \widetilde{\text{IPC}}_{i,j}, \quad j = 1 \dots, q \quad (32)$$

Re-estimating the IPC regression estimates once will reduce but not eliminate the bias. However, by iterating over the steps shown in Equations (30)–(32), the IPC regression estimates will approach unbiased and consistent estimates of individual- or group-specific differences in maximum likelihood estimates. A graphical demonstration of the bias correction procedure is presented in Figure 2.

The iterated IPC algorithm converges if the change in either the IPC regression estimates or in the log-likelihood becomes negligibly small. Unfortunately, the algorithm does not always converge. Especially, if the true individual- or group-specific value of a parameter lies close to (or at) the border of its parameter space, the algorithm might go awry. However, given strong heterogeneity in a sample, we observed across various models that the iterations often yield substantial improvement over the initial IPC regression estimates before breaking down. Therefore, the iteration with the largest log-likelihood might be preferred to the initial results.

We would like to note two more observations on the bias correction procedure. First, IPC regression estimates are unbiased in homogeneous samples and therefore cannot be further improved by updating the IPCs. If iterated IPC regression is used in a homogeneous sample, the algorithm will overfit the estimates to random fluctuation of the data. In this case, the resulting estimates can be marginally worse than the initial estimates, but the difference will be inconsequential for most practical purposes. Second, updating the IPCs comes at the cost of additional computational demands. In our experience, however, the algorithm usually converges quickly within few iterations. Even for samples with a few thousand individuals and models with more than 30 parameters, updating the IPCs took less than a minute with a standard desktop PC.

#### Software implementation

IPC regression is implemented as a package for the statistical programming language R (R Core Team, 2019), termed *ipcr*. The *ipcr* package makes it easy for researchers to study heterogeneity in the parameter estimates of an SEM fitted with the *OpenMx* package (Neale et al., 2015). The *ipcr* package performs “vanilla”, IPC regression as introduced by Oberski (2013) as well as iterated IPC regression. More information of how the *ipcr* package can be installed and used can be found under <https://github.com/manuelarnold/ipcr/>.

## MONTE CARLO SIMULATIONS

To evaluate the performance of vanilla and iterated IPC regression to detect and estimate heterogeneity in dynamic panel models in discrete and continuous time we conducted the following two Monte Carlo simulations. The first simulation aims to substantiate our theoretical considerations regarding the bias for bivariate dynamic panel models. The second simulation investigates whether IPC regression provides valid inferences and compares the power of the method with MGSEM. Additional simulations to evaluate the performance of IPC regression for non-normally distributed data, more periods, and a comparison to

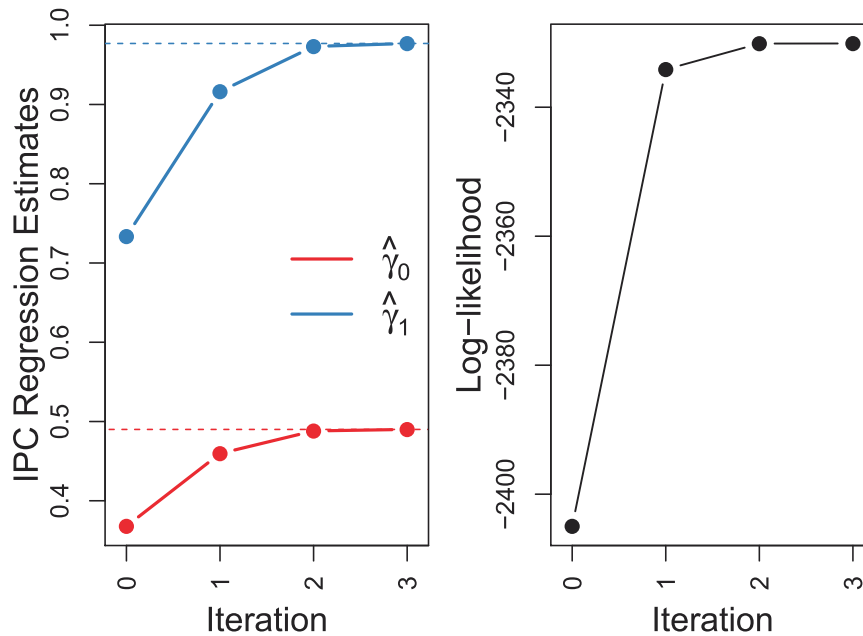


FIGURE 2 Demonstration of iterated IPC regression. 1000 individuals were sampled in equal shares from a two-group exponential distribution with group-specific rate parameters  $\lambda_1 = 0.5$  and  $\lambda_2 = 1.5$ . Iterated IPC regression with a dummy variable indicating grouping was used to estimate the group difference in the rate parameter. On the left side, initial and re-estimated IPC regression estimates are shown. Red dots are estimates of  $\lambda_1$  and blue dots are estimates of the difference  $\lambda_2 - \lambda_1$ . Dashed lines mark the corresponding maximum likelihood estimates. Clearly, the initial IPC regression estimates are biased. After just two iterations, however, the iterated IPC regression estimates approach the corresponding maximum likelihood estimates. The log-likelihood function is shown on the right side. The iterative reduction of the bias in the IPC regression estimates leads to an increase of the log-likelihood.

a multilevel model, an MGSEM, and an SEM tree are provided as [Online Supplemental Material](#).

#### Simulation I: Demonstration of the bias

In the following simulation studies, we used the discrete-time dynamic panel model depicted in [Figure 1](#) with five measurement waves as a simulation model. The data were sampled from a multivariate normal distribution with two distinct sets of parameter values. 125 observations were generated per group, resulting in a pooled sample with 250 observations in total. A discrete-time and a continuous-time dynamic panel model were fitted to the same data, ignoring the group differences. Then, we used vanilla and iterated IPC regression with a dummy variable to recover the group differences in the parameter values of the dynamic panel models. Iterated IPC regression was performed by re-estimating the IPC regression parameters until the change in all parameters was smaller than 0.0001. We repeated this procedure 10,000 times.

The discrete-time population parameter values used to generate the data are shown in the upper half of [Table 1](#), separated for both groups. For easy reference, we transformed

these parameter values into continuous time and printed them in the lower half of the table. As clearly apparent from [Table 1](#), group 1 and 2 differ substantially. The first group is characterized by strong autoregressive coefficients and no cross-lagged effects, whereas the second group exhibits substantial cross-lagged effects and smaller autoregressive coefficients. In addition, the variance of  $x$  and  $y$  was chosen twice as high for the second group as compared to the first.

We will first discuss the results for the discrete-time dynamic panel model. As expected from the theoretical example, both IPC regression methods provided accurate estimates of heterogeneity in the initial variance and covariance parameters. Further, IPC regression estimates for regression coefficients and dynamic error term variance parameters were slightly distorted. [Figure 3](#) depicts boxplots visualizing the bias of the IPC methods for regression coefficients (top graph) and dynamic error term variance parameters (lower graph). The estimates of vanilla IPC regression are printed in red and estimates after updating the IPCs are depicted in blue. Boxplots whose median lines lie close to the dotted black line indicate that the corresponding IPC regression estimates were approximately unbiased. Using the vanilla method, the intercepts (marked with the subscript 0) of the IPC regression equations were more biased than the slopes (subscript 1). Our



TABLE 1  
Group-specific Population Parameter Values for the Dynamic Panel Models in Discrete and Continuous Time

Time	$\theta$	Group 1	Group 2	$\theta$	Group 1	Group 2
Discrete	$\beta_{xx}$	0.700	0.450	$\phi_{yx}$	0.300	1.000
	$\beta_{yx}$	0.000	0.300	$\phi_{yy}$	1.000	2.000
	$\beta_{xy}$	0.000	0.300	$\psi_{xx}$	0.510	1.145
	$\beta_{yy}$	0.700	0.450	$\psi_{yx}$	0.153	0.168
	$\phi_{xx}$	1.000	2.000	$\psi_{yy}$	0.510	1.145
Continuous	$a_{xx}$	-0.357	-1.092	$\phi_{yx}$	0.300	1.000
	$a_{yx}$	0.000	0.805	$\phi_{yy}$	1.000	2.000
	$a_{xy}$	0.000	0.805	$q_{xx}$	0.713	2.760
	$a_{yy}$	-0.357	-1.092	$q_{yx}$	0.214	-1.034
	$\phi_{xx}$	1.000	2.000	$q_{yy}$	0.713	2.760

updated IPC method erased the bias in the intercepts and provided accurate estimates for all types of model parameters. Averaged over all parameters, the root mean squared error of iterated IPC regression (RMSE = 0.089) was slightly smaller than the one of the vanilla procedure (RMSE = 0.094).

The performance of the IPC regression methods for the continuous-time dynamic model was similar to the findings for the discrete-time parameters above. The estimates for the initial variance and covariance parameters provided by both IPC regression methods were near the true values, whereas estimates for the remaining model parameters were biased. Figure 4 presents the bias in the IPC regression estimates for drift and diffusion parameters. Overall, the IPC regression estimates showed more variability for the continuous-time parameters than for the discrete-time parameters. As for the discrete-time model, vanilla IPC regression exhibited a slight bias. Re-estimating the IPCs with our correction procedure reduced this bias at the cost of increased variability of the IPC regression estimates. Moreover, the iterated IPC algorithm converged only in 53.78% of the trials and fell back to the starting values or an intermediate solution in the remaining trials. Nevertheless, in terms of the RMSE averaged over all parameters, iterated IPC regression (RMSE = 0.168) slightly outperformed vanilla IPC regression (RMSE = 0.174).

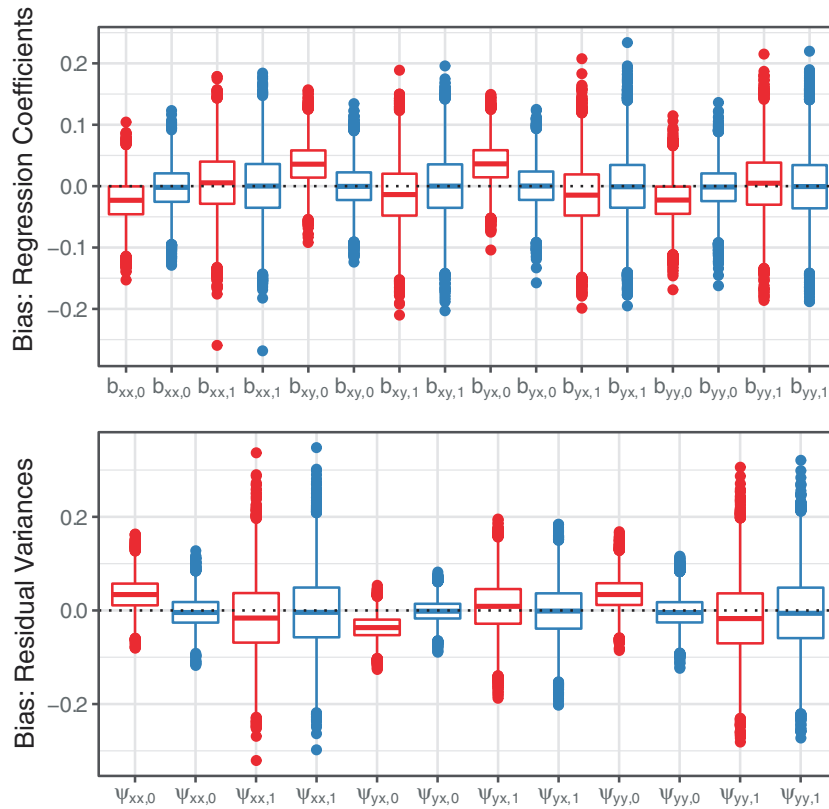


FIGURE 3 Boxplots of the bias of the IPC regression estimates for the discrete-time dynamic panel model. Red: vanilla IPC regression, blue: iterated IPC regression.

Simulation II: Statistical power and false positive rate

In the second simulation, we investigated the power of IPC regression to detect a difference in a parameter value and the false positive rate in case of homogeneous parameters. We generated multivariate normal data from bivariate dynamic panel models with five measurement occasions. We specified the population models in a way that only the cross-lagged effects from the variable  $x$  on  $y$  differed slightly between two groups. All other parameters were equal. In contrast to the previous simulation, we used different population models for the discrete- and continuous-time models. The corresponding parameter values for both population models (shown in Table 2) resulted in similar but not identical population covariance matrices. After a data set was generated, a pooled dynamic panel model was fitted, and parameter heterogeneity was tested with IPC regression (vanilla and iterated) using a dummy variable. We used the same convergence criterion for iterated IPC regression as in the previous simulation. We investigated power and false positive rate for group sizes of 100, 125, 150, 175, and 200 resulting in total sizes of

200, 250, 300, 350, and 400. For each sample size, we replicated this process 10,000 times.

As a reference, we compared the power of the IPC regression methods to the power of MGSEM. Although MGSEM lacks the flexibility and computational simplicity of IPC regression, in simple (single-variable) group comparisons with correctly specified models, standard maximum-likelihood theory suggests it should provide the uniformly most powerful test. MGSEM therefore presents a good gold standard reference for these cases. The MGSEMs were specified by letting only the cross-lagged effects of  $x$  on  $y$  differ between groups. We computed the power of the MGSEMs by conducting likelihood ratio tests that compared the fit of the MGSEMs to the fit of the pooled models.

Figure 5 shows the power of IPC regression for the discrete-time model. Depicted is the rejection rate of the null hypothesis that the cross-lagged effects from  $x$  on  $y$  are equal in both groups, plotted against the number of individuals for a significance level of 5%. Red lines refer to the power of vanilla IPC regression, blue lines to iterated IPC regression, and black lines mark the power of MGSEM. For the discrete-

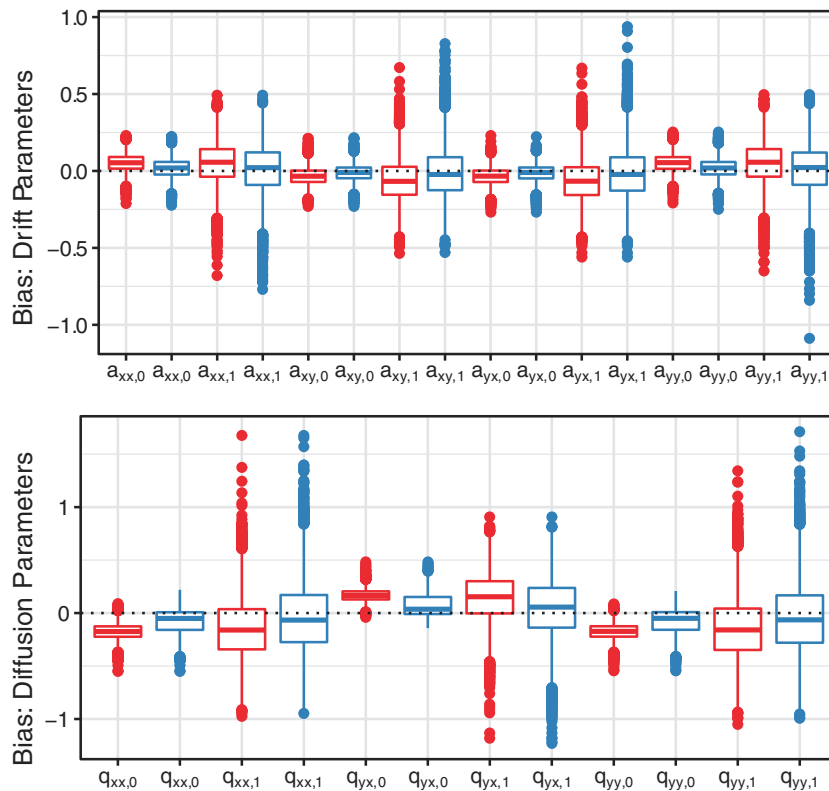


FIGURE 4 Boxplots of the bias of the IPC regression estimates for the continuous-time dynamic panel model. Red: vanilla IPC regression, blue: iterated IPC regression.

TABLE 2  
Population Parameter Values for the Dynamic Panel Models Used in Simulation II.

Discrete time				Continuous time			
$\theta$	Value	$\theta$	Value	$\theta$	Value	$\theta$	Value
$\beta_{xx}$	0.500	$\phi_{yx}$	0.300	$a_{xx}$	-0.780	$\phi_{yx}$	0.300
$\beta_{yx}$	0.200/0.300	$\phi_{yy}$	1.000	$a_{yx}$	0.424/0.546	$\phi_{yy}$	0.300
$\beta_{xy}$	0.200	$\psi_{xx}$	0.650	$a_{xy}$	0.424	$q_{xx}$	1.306
$\beta_{yy}$	0.500	$\psi_{yx}$	0.013	$a_{yy}$	-0.780	$q_{yx}$	-0.379
$\phi_{xx}$	1.000	$\psi_{yy}$	0.650	$\phi_{xx}$	1.000	$q_{yy}$	1.306

Note. That the cross-lagged effects  $\beta_{yx}$  and  $a_{yx}$  differ between the two groups.

time model, the IPC regression methods appeared to be on average 3.97 percentage points (range: [3.03, 5.30]) less powerful than MGSEM. Iterated IPC regression achieved a marginally larger power with a difference of 0.66 percentage points (range: [0.35, 0.94]). The power for the continuous-time model is presented in Figure 6. We found that the difference in power between the IPC regression methods and MGSEM were substantively larger for the continuous-time model than for the discrete-time model. On average, the power of the IPC regression was 20.68 percentage points (range: [14.25, 27.47]) smaller than the power of MGSEM. In addition, the power of IPC regression appeared to grow more slowly as a function of sample size. Again, iterated IPC appeared slightly more powerful than vanilla IPC regression (average difference: 0.28, range: [0.17, 0.44]).

Besides power, the false detection rate of the IPC regression methods is of great importance for drawing correct

conclusions from the data. We assessed the type I error rate for population parameters that are identical in the two groups for a significance level of 5%. We summarized the results by averaging the type I error rate for the three parameter types in the models (initial variance, regression coefficient/drift, dynamic error term variance/diffusion). Table 3 shows the proportions of type I errors for the discrete-time model and Table 4 for the continuous-time model. In line with simulation results from Oberski (2013), the type I error rates were close to 5% for most parameters. Iterated IPC regression committed slightly more type I errors for regression and drift parameters. These findings imply that the standard errors of iterated IPC regression for regression/drift parameters were slightly too small and could explain why iterated IPC regression appeared marginally more powerful to detect heterogeneity.

In contrast to Simulation I, there was not a single case of non-convergence of the iterated IPC regression algorithm in Simulation II. This finding suggests that the convergence problems for the continuous-time dynamic panel model were mainly driven by the larger group differences used in the previous simulation.

## DISCUSSION

The present study investigated the performance of IPC regression (Oberski, 2013) to identify and estimate parameter heterogeneity in dynamic panel models. Overall, we found that IPC regression is a promising method to identify and estimate individual or group differences. In comparison to other contemporary approaches formally addressing heterogeneity with covariates, IPC regression offers a general

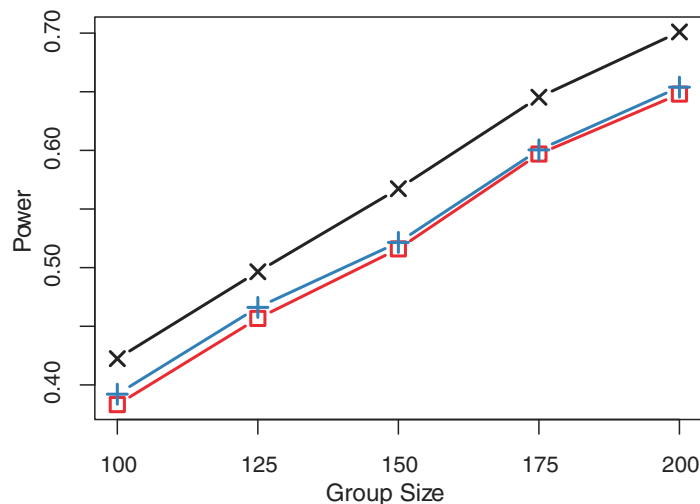


FIGURE 5 Power to detect that the population group difference in the cross-lagged effect  $\beta_{yx}$  of the discrete-time model is non-zero. Black crosses: MGSEM, red squares: vanilla IPC regression, blue pluses: iterated IPC regression.

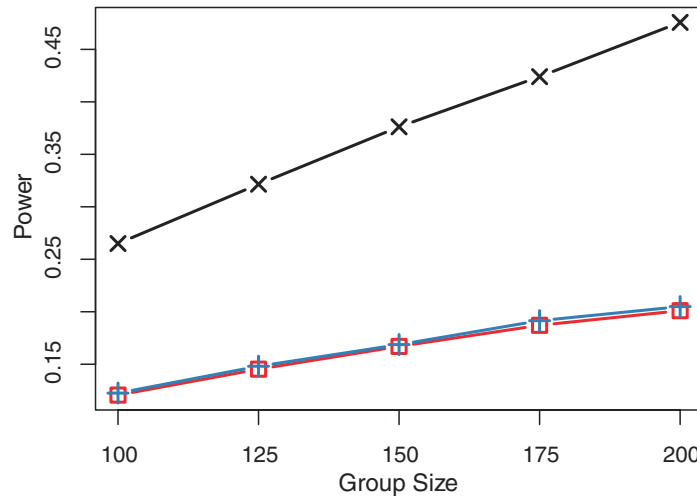


FIGURE 6 Power to detect that the population group difference in the drift parameter  $a_{yx}$  of the continuous-time model is non-zero. Black crosses: MGSEM, red squares: vanilla IPC regression, blue pluses: iterated IPC regression.

TABLE 3  
Proportions of Type I Errors for the Parameters Estimates of the Discrete-time Dynamic Panel Model

Group size	Vanilla IPC			Iterated IPC		
	$\beta$	$\phi$	$\psi$	$\beta$	$\phi$	$\psi$
100	5.483	5.047	5.353	6.097	5.047	5.437
125	5.167	5.163	5.133	5.663	5.163	5.277
150	5.037	5.143	5.010	5.500	5.143	5.207
175	5.173	4.940	5.150	5.600	4.940	5.090
200	5.093	4.900	4.850	5.443	4.900	4.983

TABLE 4  
Proportions of Type I Errors for the Parameters Estimates of the Continuous-time Dynamic Panel Model

Group size	Vanilla IPC			Iterated IPC		
	$a$	$\phi$	$q$	$a$	$\phi$	$q$
100	5.333	5.083	5.353	5.250	5.083	4.477
125	5.103	5.007	5.263	5.183	5.007	4.620
150	5.207	5.090	5.143	5.247	5.090	4.657
175	5.077	5.193	4.993	5.157	5.193	4.580
200	4.810	5.207	4.803	4.920	5.207	4.463

framework that encompasses all types of SEMs and covariates and makes identifying and explaining individual differences as simple, flexible, and fast as linear regression.

IPC regression was evaluated in terms of bias in the recovery of true group differences, the power to detect parameter heterogeneity, and the type I error rate for homogeneous parameters. By means of a theoretical

example and through Monte Carlo simulations, we demonstrated that original, “vanilla”, IPC regression estimates can be slightly biased due to large differences in regression parameters. Additional heterogeneity in variance parameters may amplify this bias. As a rule of thumb, the bias seems to affect mainly parameters connected to endogenous variables like regression and residual variance parameters, whereas the IPC regression estimates for parameters associated with exogenous variables such as the initial variance parameters remain comparatively unbiased. Hence, IPC regression may perform worse for SEMs with many directed paths such as dynamic panel models than for models with few directed paths such as CFA models. This argument would also explain why Oberski (2013) found nearly unbiased estimates of group differences in a CFA model.

To correct the bias in vanilla IPC regression, we introduced a novel updating procedure, which we termed iterated IPC regression. Iterated IPC regression produced approximately unbiased estimates of group differences in the parameters of a discrete-time dynamic panel model and outperformed vanilla IPC regression in terms of the RMSE. For the continuous-time dynamic panel model, however, iterated IPC regression corrected the bias but at the cost of adding additional variability to the estimates. Nevertheless, updating the IPCs still improved the estimates on average as indicated by a smaller RMSE.

In situations in which MGSEM could be applied as an alternative to IPC regression, we compared the power of IPC regression to that of MGSEM, which theory suggests is uniformly most-powerful in these cases. IPC regression

yielded power only slightly below that of this theoretically optimal method to detect group differences in the cross-lagged effect of a discrete-time dynamic panel model. For the continuous-time model, however, IPC regression was no more than half as powerful as MGSEM. It should be noted that MGSEM cannot be applied to all scenarios allowed by IPC regression; for example, MGSEM does not investigate partial effects of multiple covariates of model parameters. In agreement with earlier theoretical findings, both IPC regression methods did control the type I error rate accurately.

In summary, our findings demonstrate that (iterated) IPC regression is a useful tool to study heterogeneity in discrete-time dynamic panel model. For continuous-time dynamic panel models, however, our findings were mixed: high variance caused by the bias correction procedure and a small power make (iterated) IPC regression unappealing especially in smaller data sets. We believe that these problems are caused by non-linear parameter constraints and high correlation between parameter estimates of the continuous-time dynamic panel model. Considering these difficulties, IPC regression seems more appropriate for models that can be parameterized without non-linear constraints such as the discrete-time dynamic panel model or other contemporaneous models for longitudinal data such as latent growth curve models (Bollen & Curran, 2006) or latent change score models (McArdle, 2001), if these models are applicable.

Although IPC regression is a general, easy to use, and flexible approach to detect parameter heterogeneity, we want to stress that it is not always the most appropriate one. Depending on a study's objective, other methods for addressing heterogeneity should be preferred to IPC regression. For example, multilevel models are like an obvious choice in situations where it is sufficient to allow for varying parameter values between individuals and there is no interest in explaining these differences. In contrast, if a study aims to test differences between few known groups in the data (e.g., in variance parameters), MGSEM will often be the better choice. If a study's goal is to determine homogeneous groups in the data with help of additionally observed covariates, partitioning methods like SEM trees or forests often are better suited for the task, in particular if computation time is not an issue.

In the following, we will briefly touch upon some limitations of IPC regression that researchers should consider. First, the usefulness of IPC regression depends on the covariates available. If none of the additional covariates is related to individual or group differences in the parameters, IPC regression will fail to detect the source of heterogeneity. In cases of unobserved group membership, researchers may want to resort to methods like finite mixture models (Jedidi, Jagpal, & DeSarbo, 1997; Lubke & Muthén, 2005; Muthén & Shedden, 1999). Second, IPC regression is




a data-driven or exploratory procedure and therefore susceptible to capitalize on chance characteristics of the data (MacCallum, Roznowski, & Necowitz, 1992). Modifying models by blindly following the advice of IPC regression may lead to a model that works well in the observed sample but does not generalize to others. We thus recommend paying not only close attention to the *p*-value provided by IPC regression, but also to the size of the estimated individual or group difference. See also Saris, Satorra, and van der Veld (2009), for a related discussion about model modification using the modification index and expected parameter change. Third, using IPC regression to investigate the effect of a large number of covariates on complex models with many parameters will yield a large number of IPC regression estimates that can be challenging to interpret. Regularization techniques such as lasso (Tibshirani, 1996) could be used to find a subset of the most important covariates.

In summary, however, we believe that IPC regression is a useful tool to investigate parameter heterogeneity in SEMs for longitudinal data such as dynamic panel models that combines flexibility with its unique computational simplicity.

#### ACKNOWLEDGEMENT

We acknowledge support by the Open Access Publication Fund of Humboldt-Universität zu Berlin.

#### ORCID

Manuel Arnold  <http://orcid.org/0000-0002-1623-2565>  
 Daniel L. Oberski  <http://orcid.org/0000-0001-7467-2297>  
 Andreas M. Brandmaier  <http://orcid.org/0000-0001-8765-6982>  
 Manuel C. Voelkle  <http://orcid.org/0000-0001-5576-8103>

#### REFERENCES

- Allison, P. D., Williams, R., & Moral-Benito, E. (2017). Maximum likelihood for cross-lagged panel models with fixed effects. *Socius*, 3, 1–17. doi:10.1177/2378023117710578
- Biesanz, J. C. (2012). Autoregressive longitudinal models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 459–471). New York, NY: Guilford.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Bollen, K. A., & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces*, 89, 1–34. doi:10.1353/sof.2010.0072

- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: John Wiley & Sons.
- Brandmaier, A. M., Driver, C. C., & Voelkle, M. C. (2018). Recursive partitioning in continuous time analysis. In K. van Montfort, J. H. L. Oud, & M. C. Voelkle (Eds.), *Continuous time modeling in the behavioral and related sciences* (pp. 259–282). New York, NY: Springer.
- Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods, 21*, 566–582. doi:10.1037/met0000090
- Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods, 18*, 71–86. doi:10.1037/a0030001
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.
- Demidenko, E. (2013). *Mixed models: Theory and applications with R* (2nd ed.). Hoboken, NJ: Wiley.
- Driver, C. C., & Voelkle, M. C. (2018). Hierarchical Bayesian continuous time dynamic modeling. *Psychological Methods, 23*, 774–779. doi:10.1037/met0000168
- George, M. R. W., Yang, N., Jaki, T., Feaster, D. J., Lamont, A. E., Wilson, D. K., & van Horn, M. L. (2013). Finite mixtures for simultaneously modelling differential effects and non-normal distributions. *Multivariate Behavioral Research, 48*, 816–844. doi:10.1080/00273171.2013.830065
- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*, 424–438. doi:10.2307/1912791
- Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology, 30*, 507–544. doi:10.1146/annurev.soc.30.012703.110629
- Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20*, 102–116. doi:10.1037/a0038889
- Hsiao, C. (2014). *Analysis of panel data* (3rd ed.). Cambridge, UK: University Press.
- 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. doi:10.1287/mksc.16.1.39
- Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21–39. doi:10.1037/1082-989X.10.1.21
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490–504. doi:10.1037/0033-2909.111.3.490
- Magnus, J. R., & Neudecker, H. (2019). *Matrix differential calculus with applications in statistics and econometrics* (3rd ed.). New York, NY: Wiley. doi:10.1002/9781119541219
- McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analysis. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling* (pp. 341–380). Lincolnwood, IL: Scientific Software International.
- Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika, 79*, 569–584. doi:10.1007/s11336-013-9376-7
- Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika, 78*, 59–82. doi:10.1007/s11336-012-9302-4
- Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*, 463–469. doi:10.1111/j.0006-341X.1999.00463.x
- Neale, M. C., Hunter, M. D., Pritkin, J., Zahery, M., Brick, T. R., Kirkpatrick, R. M., ... Boker, S. M. (2015). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika, 81*, 535–549. doi:10.1007/s11336-014-9435-8
- Neudecker, H., & Satorra, A. (1991). Linear structural relations: Gradient and Hessian of the fitting function. *Statistics & Probability Letters, 11*, 57–61. doi:10.1016/0167-7152(91)90178-T
- Oberski, D. L. (2013). *A flexible method to explain differences in structural equation model parameters over subgroups*. Retrieved from <http://daob.nl/wp-content/uploads/2013/06/SEM-IPC-manuscript-new.pdf>
- Oud, J. H. L. (2007). Continuous time modeling of reciprocal relationships in the cross-lagged panel design. In S. M. Boker & M. J. Wenger (Eds.), *Data analytic techniques for dynamic systems in the social and behavioral sciences* (pp. 87–129). Mahwah, NJ: Erlbaum.
- Oud, J. H. L., & Delsing, M. J. M. H. (2010). Continuous time modeling of panel data by means of SEM. In K. van Montfort, J. H. L. Oud, & A. Satorra (Eds.), *Longitudinal research with latent variables* (pp. 201–244). New York, NY: Springer.
- Oud, J. H. L., & Jansen, R. A. R. G. (2000). Continuous time state space modeling of panel data by means of SEM. *Psychometrika, 65*, 199–215. doi:10.1007/BF02294374
- R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology, 17*, 105–129. doi:10.2307/271030
- Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*, 561–582. doi:10.1080/10705510903203433
- Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika, 54*, 131–151. doi:10.1007/BF02294453
- Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. *Sociological Methodology, 22*, 249–278. doi:10.2307/270998
- Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling, 21*, 149–160. doi:10.1080/10705511.2013.824793
- Schuurman, N. K., Ferrer, E., de Boer-Sonnenschein, M., & Hamaker, E. L. (2016). How to compare cross-lagged associations in a multilevel autoregressive model. *Psychological Methods, 21*, 206–221. doi:10.1037/met0000062
- Singer, J. D., & Willet, J. B. (2003). *Applied longitudinal data analysis*. Oxford, UK: Oxford University Press.
- Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematics and Statistical Psychology, 27*, 229–239. doi:10.1111/j.2044-8317.1974.tb00543.x
- Sörbom, D. (1989). Model modification. *Psychometrika, 54*, 371–384. doi:10.1007/BF02294623
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, 58*, 267–288. doi:10.1111/rssb.1996.58.issue-1
- Usami, S., Hayes, T., & McArdle, J. (2017). Fitting structural equation model trees and latent growth curve mixture models in longitudinal designs: The influence of model misspecification. *Structural Equation Modeling, 24*, 585–598. doi:10.1080/10705511.2016.1266267
- van Montfort, K., Oud, J. H. L., & Voelkle, M. C. (2018). *Continuous time modeling in the behavioral and related sciences*. New York, NY: Springer.
- Voelkle, M. C., Oud, J. H. L., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods, 17*, 176–192. doi:10.1037/a0027543
- Voelkle, M. C., Oud, J. H. L., von Oertzen, T., & Lindenberger, U. (2012). Maximum likelihood dynamic factor modeling for arbitrary N and T using SEM. *Structural Equation Modeling, 19*, 329–350. doi:10.1080/10705511.2012.687656



Wang, T., Merkle, E. C., & Zeileis, A. (2014). Score-based tests of measurement invariance: Use in practice. *Frontiers in Psychology*, *5*, 438. doi:10.3389/fpsyg.2014.00438

Wang, T., Strobl, C., Zeileis, A., & Merkle, E. C. (2018). Score-based tests of differential item functioning via pairwise maximum likelihood estimation. *Psychometrika*, *83*, 132–155. doi:10.1007/s11336-017-9591-8

White, H. (1984). *Asymptotic theory for econometricians*. Orlando, FL: Academic Press.

Yuan, K.-H., & Bentler, P. M. (2007). Structural equation modeling. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 297–358). Amsterdam, Netherlands: North Holland.

Zeileis, A. (2005). A unified approach to structural change tests based on ML scores, F statistics, and OLS residuals. *Econometric Reviews*, *24*, 445–466. doi:10.1080/07474930500406053

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*, 488–508. doi:10.1111/stan.2007.61.issue-4

Zyphur, M. J., Allison, P. D., Tay, L., Voelkle, M. C., Preacher, K. J., Zhang, Z., ... Diener, E. (2019). From data to causes I: Building a general cross-lagged panel model (GCLM). *Organizational Research Methods*. Advance online publication. doi:10.1177/1094428119847278

Zyphur, M. J., Voelkle, M. C., Tay, L., Allison, P. D., Preacher, K. J., Zhang, Z., ... Diener, E. (2019). From data to causes II: Comparing approaches to panel data analysis. *Organizational Research Methods*. Advance online publication. doi:10.1177/1094428119847280

### APPENDIX EXPRESSING IPC REGRESSION ESTIMATES WITH GROUP-SPECIFIC ESTIMATES

In the following, we express the IPC regression estimates  $\hat{\gamma}_0$  and  $\hat{\gamma}_1$  in terms of the group-specific maximum likelihood estimates  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$  as shown in Equations (26) and (27). Note that  $\hat{\gamma}_0$  and  $\hat{\gamma}_1$  are simple ordinary least squares estimates given by  $\hat{\gamma}_1 = s_{IPC,z}/s_z^2$  and  $\hat{\gamma}_0 = \overline{IPC} - \hat{\gamma}_1 \bar{z}$ , where  $s_{IPC,z}$  is the sample covariance between the IPCs and the covariate  $z_i$ ,  $s_z^2$  is the sample variance of the covariate, and  $\overline{IPC}$  and  $\bar{z}$  are the sample means of the IPCs and the covariate, respectively.

Following Equation (16), the IPC of individual  $i$  is given by

$$IPC(\hat{\lambda}; y_i) = \hat{\lambda} + I(\hat{\lambda})^{-1} S(\hat{\lambda}; y_i) = \hat{\lambda} + \hat{\lambda}^2 \left( \frac{1}{\hat{\lambda}} - y_i \right) = 2\hat{\lambda} - \hat{\lambda}^2 y_i.$$

Next, we express the pooled maximum likelihood estimate  $\hat{\lambda}$  as a function of the group-specific maximum likelihood estimates:

$$\hat{\lambda} = \left( \frac{1}{n} \sum_{i=1}^n y_i \right)^{-1} = \left[ \frac{1}{n} \left( \sum_{i=1}^{n_1} y_i + \sum_{i=n_1+1}^{n_2} y_i \right) \right]^{-1} = \left[ \frac{1}{n} \left( \frac{n_1}{\hat{\lambda}_1} + \frac{n_2}{\hat{\lambda}_2} \right) \right]^{-1} \stackrel{n_1=n_2=n}{=} \left( \frac{1}{2\hat{\lambda}_1} + \frac{1}{2\hat{\lambda}_2} \right)^{-1} = \frac{2\hat{\lambda}_1 \hat{\lambda}_2}{\hat{\lambda}_1 + \hat{\lambda}_2}$$

Using both equations from above, the IPC regression slope  $\hat{\gamma}_1$  can be written in terms of the group-specific maximum likelihood estimates:

$$\begin{aligned} \hat{\gamma}_1 &= \frac{s_{IPC,z}}{s_z^2} \\ &= \frac{\sum_{i=1}^n z_i IPC(\hat{\lambda}; y_i) - \frac{1}{n} \sum_{i=1}^n z_i \sum_{i=1}^n IPC(\hat{\lambda}; y_i)}{\sum_{i=1}^n z_i^2 - \frac{1}{n} \left( \sum_{i=1}^n z_i \right)^2} \\ &= \frac{\sum_{i=n_1+1}^n \left( 2\hat{\lambda} - \hat{\lambda}^2 y_i \right) - \frac{n_2}{n} \sum_{i=1}^n \left( 2\hat{\lambda} - \hat{\lambda}^2 y_i \right)}{\frac{n_2 n - n_2^2}{n} - \frac{-n\hat{\lambda}^2 \sum_{i=n_1+1}^n y_i + n_2 \hat{\lambda}^2 \sum_{i=1}^n y_i}{n_2 n - n_2^2}} \\ &\stackrel{n=n_1+n_2}{=} \frac{\hat{\lambda}^2 \left( -n_1 \sum_{i=n_1+1}^n y_i - n_2 \sum_{i=n_1+1}^n y_i + n_2 \sum_{i=1}^n y_i \right)}{n_1 n_2} \\ &= \hat{\lambda}^2 \left( \frac{1}{n_1} \sum_{i=1}^{n_1} y_i - \frac{1}{n_2} \sum_{i=n_1+1}^n y_i \right) = \hat{\lambda}^2 \left( \frac{1}{\hat{\lambda}_1} - \frac{1}{\hat{\lambda}_2} \right) \\ &= \left( \frac{2\hat{\lambda}_1 \hat{\lambda}_2}{\hat{\lambda}_1 + \hat{\lambda}_2} \right)^2 \frac{\hat{\lambda}_2 - \hat{\lambda}_1}{\hat{\lambda}_1 \hat{\lambda}_2} = \frac{4\hat{\lambda}_1 \hat{\lambda}_2 (\hat{\lambda}_2 - \hat{\lambda}_1)}{(\hat{\lambda}_1 + \hat{\lambda}_2)^2} \end{aligned}$$

Finally, we can derive the IPC regression intercept  $\hat{\gamma}_0$  in the same way

$$\begin{aligned} \hat{\gamma}_0 &= \overline{IPC} - \hat{\gamma}_1 \bar{z} \\ &= \frac{1}{n} \sum_{i=1}^n IPC(\hat{\lambda}; y_i) - \hat{\gamma}_1 \frac{1}{n} \sum_{i=1}^n z_i \\ &= \frac{1}{n} \sum_{i=1}^n \left( 2\hat{\lambda} - \hat{\lambda}^2 y_i \right) - \hat{\lambda}^2 \left( \frac{1}{n_1} \sum_{i=1}^{n_1} y_i - \frac{1}{n_2} \sum_{i=n_1+1}^n y_i \right) \frac{n_2}{n} \\ &= 2\hat{\lambda} + \hat{\lambda}^2 \left( -\frac{1}{n} \sum_{i=1}^n y_i - \frac{n_2}{n_1 n} \sum_{i=1}^{n_1} y_i + \frac{1}{n} \sum_{i=n_1+1}^n y_i \right) \\ &= 2\hat{\lambda} + \hat{\lambda}^2 \left( -\frac{1}{n} \sum_{i=1}^{n_1} y_i - \frac{n_2}{n_1 n} \sum_{i=1}^{n_1} y_i \right) \\ &\stackrel{n=n_1+n_2}{=} 2\hat{\lambda} - \frac{\hat{\lambda}^2}{n_1} \sum_{i=1}^{n_1} y_i = 2\hat{\lambda} - \frac{\hat{\lambda}^2}{\hat{\lambda}_1} \\ &= 2 \frac{2\hat{\lambda}_1 \hat{\lambda}_2}{\hat{\lambda}_1 + \hat{\lambda}_2} - \frac{1}{\hat{\lambda}_1} \left( \frac{2\hat{\lambda}_1 \hat{\lambda}_2}{\hat{\lambda}_1 + \hat{\lambda}_2} \right)^2 = \frac{4\hat{\lambda}_1^2 \hat{\lambda}_2}{(\hat{\lambda}_1 + \hat{\lambda}_2)^2} \end{aligned}$$

Article

# Predicting Differences in Model Parameters with Individual Parameter Contribution Regression Using the R Package *ipcr*

Manuel Arnold <sup>1,2,\*</sup> , Andreas M. Brandmaier <sup>2,3</sup>  and Manuel C. Voelkle <sup>1</sup> 

- <sup>1</sup> Psychological Research Methods, Department of Psychology, Humboldt-Universität zu Berlin, 12489 Berlin, Germany; manuel.voelkle@hu-berlin.de
- <sup>2</sup> Max Planck UCL Centre for Computational Psychiatry and Ageing Research, 14195 Berlin, Germany; brandmaier@mpib-berlin.mpg.de
- <sup>3</sup> Center for Lifespan Psychology, Max Planck Institute for Human Development, 14195 Berlin, Germany
- \* Correspondence: arnoldmz@hu-berlin.de

**Abstract:** Unmodeled differences between individuals or groups can bias parameter estimates and may lead to false-positive or false-negative findings. Such instances of heterogeneity can often be detected and predicted with additional covariates. However, predicting differences with covariates can be challenging or even infeasible, depending on the modeling framework and type of parameter. Here, we demonstrate how the individual parameter contribution (IPC) regression framework, as implemented in the R package *ipcr*, can be leveraged to predict differences in any parameter across a wide range of parametric models. First and foremost, IPC regression is an exploratory analysis technique to determine if and how the parameters of a fitted model vary as a linear function of covariates. After introducing the theoretical foundation of IPC regression, we use an empirical data set to demonstrate how parameter differences in a structural equation model can be predicted with the *ipcr* package. Then, we analyze the performance of IPC regression in comparison to alternative methods for modeling parameter heterogeneity in a Monte Carlo simulation.

**Keywords:** heterogeneity; individual differences; linear regression; R; structural equation modeling; latent variables



**Citation:** Arnold, M.; Brandmaier, A.M.; Voelkle, M.C. Predicting Differences in Model Parameters with Individual Parameter Contribution Regression Using the R Package *ipcr*. *Psych* **2021**, *3*, 360–385. <https://doi.org/10.3390/psych3030027>

Academic Editor: Alexander Robitzsch

Received: 3 June 2021  
Accepted: 28 July 2021  
Published: 6 August 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

A fundamental assumption of parametric modeling is that the model parameters represent all individuals in the sample. However, populations investigated in the behavioral sciences and related fields are rarely homogeneous and instead are often characterized by substantial differences between individuals or groups. For example, heterogeneity may reflect age differences in cognitive functioning (e.g., [1]) and differences in the prevalence of depression between women and men (e.g., [2]). Overlooking such heterogeneity results in an incomplete model with potentially biased parameter estimates and may lead to false-positive or false-negative findings [3,4]. Conversely, discovering and explaining heterogeneity can substantially support theory building.

Individual and group differences can often be accounted for with additional covariates such as contextual or background variables. In practice, researchers routinely probe the effects of potentially important covariates by using them as moderator variables or by estimating group-specific parameter values using pre-defined grouping variables. However, researchers are often limited to investigate only certain types of model parameters, depending on the choice of modeling technique. For instance, in regression models such as linear regression, generalized linear regression, and linear mixed models, differences in regression coefficients can be studied by simply adding covariates with main and interaction effects. On the contrary, uncovering heterogeneity in the variance parameter of the regression errors (or in the different variance parameters of a mixed model) with covariates is much more challenging. For example, in the statistical programming language R [5],



neither the built-in function `lm` (for fitting linear regression models) or the functions of the widely used `lme4` package (for fitting mixed models; [6]) allow for ways to explain individual and group differences in the variance parameters in the model with covariates. Other modeling frameworks, such as structural equation modeling (SEM; [7,8]), offer the user more flexibility to incorporate covariates and investigate their effects on all model parameters. One popular method for investigating the effect of a discrete covariate is the use of multigroup structural equation models (MGSEMs, [9]). MGSEMs allow estimating SEMs with different parameter values across the levels of a grouping variable. If the covariate under investigation is continuous, one possible way to assess its effect is to include it into the SEM as a moderator variable by specifying a SEM with an interaction term (e.g., [10]). MGSEMs and SEMs with interactions are best suited in situations with a priori knowledge about the sources of heterogeneity and a clear set of target parameters. However, both approaches can quickly become impractical in an exploratory analysis of the potential effects of a larger number of covariates on multiple model parameters. As MGSEMs and SEMs with interactions are usually conducted by looking at the effects of one covariate at a time, many SEMs need to be specified, estimated, and evaluated. Moreover, exploring potential interactions between covariates may result in MGSEMs with too sparse groups or overly complex interaction models with too many parameters to be stably estimated.

As an alternative procedure to exploring the effects of covariates by adding them to the model, we introduced individual parameter contribution (IPC) regression [11,12]. IPC regression allows determining whether and how the parameters of a model vary as functions of covariates. IPC regression proceeds in three steps. First, the model is fitted to data. This model represents a given scientific theory, encompassing all variables that pertain to this theory and for which hypothesized relationships can be specified. Second, individual contributions to the model parameters are calculated. These IPCs are rough approximations of individual-specific parameter values. Third and last, the IPCs are regressed on a set of covariates such as discrete grouping variables or continuous variables to investigate whether any of these covariates predict differences in the parameters. The result of an IPC regression analysis may be used to explore predictors of parameter heterogeneity, generate new hypotheses about predictors of individual or group differences in model parameters, and may ideally help revise the substantive theory. To perform IPC regression in practice, we provide the R package `ipcr`.

IPC regression is a general, flexible, simple, and computationally efficient method. In principle, IPC regression allows investigating all types of model parameters of a parametric model. For instance, IPC regression makes it feasible to predict differences in certain types of model parameters, such as the residual variance parameters in regression models, which would not be possible by adding covariates directly to the model. Furthermore, IPC regression encompasses both discrete grouping variables as well as continuous covariates. This flexibility regarding the measurement level of the covariates appears particularly advantageous when working with SEMs. For SEMs, the effects of discrete and continuous covariates are often studied with different techniques, that is, either with MGSEMs (for discrete variables) or SEMs with interactions (for continuous variables). With IPC regression, we can leverage the same approach for studying covariates with different levels of measurement. Although the derivation of the IPCs is a little more involved, a basic understanding of linear regression is sufficient to perform IPC regression in practice. A final key feature of IPC regression is a clear separation between model estimation and the investigation of parameter differences. This separation is especially advantageous when the model is complex and hard to estimate (either in the sense of computation time or convergence problems) because assessing parameter heterogeneity does not involve repeated re-specification and re-estimation.

IPCs are calculated by transforming the partial derivative of an objective function with respect to the parameters. Objective functions such as the well-known log-likelihood function are used to estimate model parameters. Various statistical procedures analyze the derivative of an objective function to assess the fit of a statistical model. In the SEM

framework, the modification index (MI; [13]) uses the derivative to approximate how much the fit of a model would change after a new parameter is added to the model. As an extension to the MI, which is merely a test statistic, the expected parameter change (EPC; [14,15]) has been put forward for obtaining a direct estimate of the added parameter. Although the MI and EPC aim to quantify specification errors, Oberski [11] demonstrated that the MI and EPC for MGSEMs correspond to IPC regression under certain conditions. However, this equivalency ends in situations that cannot be handled by MGSEMs, such as continuous or multiple covariates, making IPC regression a much more flexible method for the investigation of heterogeneity. Other methods that analyze partial derivatives are structural change tests (e.g., [16,17]). Originally used in the detection of change points in time series analysis [18], structural change tests have been recently popularized by Merkle and Zeileis [19] and Merkle et al. [20] to uncover parameter heterogeneity in psychometric models. The difference between structural change tests and IPC regression is that structural change tests provide formal tests whether the parameters of a model are invariant with respect to a covariate. In contrast, IPC regression seeks to model the relationships between parameters and covariates by means of linear regression.

The purpose of this study is threefold. First, we introduce the `ipcr` package, which offers functions to perform IPC regression in R. Second, we expand upon earlier research, which focused on predicting differences in SEM parameters, by discussing how IPC regression can be used to investigate a much broader range of parametric models. Third, we compare the performance of IPC regression to established methods in four small Monte Carlo simulations. The remainder is structured as follows: Section 2 illustrates IPC regression by means of an instructive example. Section 3 deals with the theory behind IPC regression. This section is optional and addresses readers that are interested in more technical details. Section 4 gives an overview of the `ipcr` package. Section 5 illustrates how the `ipcr` package is applied in practice based on data from the `lavaan` package [21]. Section 6 presents the simulation results. Finally, this study concludes in Section 7 with a discussion of the method and the simulation results.

## 2. Introductory Example

In the following, we illustrate the rationale behind IPC regression with a simple regression example. Let us assume we want to predict an outcome variable  $y$  as a linear function of a single explanatory variable  $x$ . The corresponding simple linear regression model can be defined as

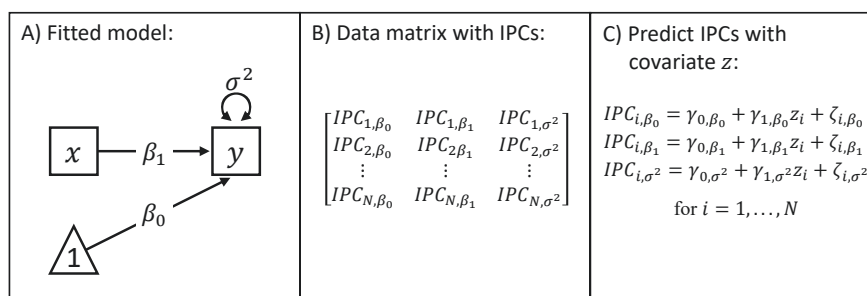
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, N. \quad (1)$$

This regression model contains the parameters  $\beta_0$ ,  $\beta_1$ , and  $\sigma^2$ . The regression coefficients  $\beta_0$  and  $\beta_1$  denote the intercept and slope of the regression line.  $\sigma^2$  is the variance of the regression errors  $\varepsilon$ ; that is, the part of the variability in the outcome  $y$  that is not explained by the explanatory variable  $x$ . Unbiased and efficient parameter estimates can be obtained by employing the ordinary least squares (OLS) estimation method.

Further, we suspect that our model parameters could vary with respect to a covariate  $z$ . One way to assess the influence of  $z$  would be to simply add the covariate to our model by specifying a direct effect from  $z$  on  $y$  and an interaction effect between  $z$  and  $x$ . The direct and interaction effects would then tell us how  $\beta_0$  and  $\beta_1$  are influenced by  $z$ . However, adding the covariate to the model does not enable us to quantify its influence on the variance parameter  $\sigma^2$ . On the other hand, IPC regression offers a way to estimate the effect of  $z$  on  $\sigma^2$ .

An IPC regression analysis involves three steps. In the first step, the simple linear regression model defined in Equation (1) is fitted to data. Generally, models investigated with IPC regression consist solely of variables important to one's scientific theory. Therefore, covariates are usually not included in the first-step model but are used later as predictor variables in step three. In the second step, a new data matrix of IPCs is calculated. Every row of this data matrix corresponds to one individual and every column to one parameter.

For example, row  $i$  contains the values of the three model parameters specific to individual  $i$ . The IPCs of a single individual are usually too noisy to be useful estimates of individual-specific parameters. However, regressing the IPCs on covariates averages out this noise and may reveal meaningful differences in the parameters. Figure 1 illustrates the three steps of IPC regression.



**Figure 1.** The three steps of IPC regression. Panel A shows the model under investigation. It is not necessary to include any covariates in the model. Panel B visualizes the structure of the new data matrix containing the IPCs. Panel C shows the three IPC regression equations for the respective model parameters. Note that we model the IPCs to each of the three model parameters in a separate regression model.

The IPC regression coefficients are interpreted as in any linear regression model. The IPC regression slope  $\gamma_{1,\beta_0}$  represents the change in the intercept  $\beta_0$  of the first-step model for a one-unit change in the covariate  $z$ . Further, the IPC regression slope  $\gamma_{1,\beta_1}$  indicates how  $z$  moderates the effect of the explanatory variable  $x$  on the outcome  $y$ . Finally,  $\gamma_{1,\sigma^2}$  represents the relationship between  $z$  and the variance of the regression errors  $\sigma^2$ . A positive relationship indicates that individuals with smaller covariate values exhibit a smaller error variance and are therefore closer to the regression line, whereas individuals with larger values show more unexplained variability.

As in standard regression analysis, a  $t$ -test can be carried out to test the null hypotheses  $\gamma_0 = 0$  and  $\gamma_1 = 0$ . Testing  $\gamma_1 = 0$  can be seen as a test of parameter homogeneity with respect to the covariate  $z$ , and rejecting it implies that the parameter under investigation is not constant across individuals. It should be noted that the errors  $\zeta_i$  of the IPC regression equations are usually not normally distributed, even if the original data used for model fitting are normal. Therefore, the standard errors of the IPC regression estimates may be inaccurate in small samples.

Figure 1 shows a simplistic example with a single covariate. In practice, one might be interested in investigating the effects of multiple covariates and their interactions or estimating non-linear relationships between the model parameters and covariates by adding quadratic and cubic terms. All these techniques work exactly as in standard linear regression.

### 3. Derivation and Properties of Individual Parameter Contributions

In the following, we will first derive the IPCs in very general terms and then give more specific results for linear regression models. Further derivations for SEMs are provided by Arnold et al. [12]. Readers uninterested in technical details may skip this section.

#### 3.1. Calculation of the Individual Parameter Contributions

In brief, IPCs are calculated by transforming the partial derivative of a case-wise objective function with respect to the parameters. More formally, let  $f(\theta)$  be an objective function used to estimate a  $q$ -variate vector of model parameters  $\theta$  based on data from  $N$  individuals. Typical examples for such an objective function are the sum of squared residuals (SSE) for linear regression models or the normal-theory maximum-likelihood

fitting function for SEMs. The model parameter estimates  $\hat{\theta}$  are obtained by minimizing the objective function, that is,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} f(\theta). \quad (2)$$

In practice,  $f$  is minimized by finding the roots (the zeros) of the partial derivative of the objective function. Thus, parameter estimates  $\hat{\theta}$  can be found by solving the first-order condition

$$S(\hat{\theta}) = \mathbf{0}, \quad (3)$$

where

$$S(\theta) = \left[ \frac{\partial f(\theta)}{\partial \theta^{(1)}} \quad \dots \quad \frac{\partial f(\theta)}{\partial \theta^{(q)}} \right]^{\top} \quad (4)$$

is the partial derivative of the objective function  $f(\theta)$  with respect to the parameters  $\theta$ . Additional steps may be required to ensure that the parameter estimates are associated with the global minimum, depending on the type of objective function. In maximum likelihood estimation, the partial derivative defined in Equation (4) is known as the score. We will use this term to refer to the derivatives of all kinds of objective functions in the following.

The idea behind the derivation of the IPCs is simple. Instead of employing a score  $S(\theta)$ , which takes information from all individuals into account to estimate a single set of parameters, we specify individual-specific scores to find individual-specific parameter estimates. In greater detail, let  $S_i(\theta)$  denote the score that uses only data from individual  $i$ . Hypothetically, solving  $S_i(\hat{\theta}_i) = \mathbf{0}$  would yield a vector of parameter estimates  $\hat{\theta}_i$  specific to individual  $i$ . Unfortunately, the system of equations  $S_i(\hat{\theta}) = \mathbf{0}$  is undetermined for most objective functions because there are more unknown parameters than individual-specific data. However, we can approximate the individual scores by combining individual data with information from the model.

Individual scores can be approximated through linearization. The equation for the linearization of the score  $S(\theta)$  at the model parameter estimates  $\hat{\theta}$  is:

$$S(\theta) \approx S(\hat{\theta}) + H(\hat{\theta})(\theta - \hat{\theta}) \quad (5)$$

The first term, which is the intercept of the linear approximation, is just the score itself, evaluated at the parameter estimates. From Equation (3), we can see that this intercept is always zero. The function

$$H(\theta) = \frac{\partial^2 f(\theta)}{\partial \theta \partial \theta^{\top}} \quad (6)$$

is the Hessian matrix of second-order partial derivatives of the objective function with respect to the parameters. In the equation above,  $H(\theta)$  is the slope of the linearized score.

In the next step, we individualize this linear approximation of the score by replacing the sample score  $S(\hat{\theta})$  with an individual-specific score  $S_i(\hat{\theta})$ . As a result, the intercepts of the individual score approximations are no longer zero but fluctuate around zero. In contrast to the sample score  $S(\hat{\theta})$ , the sample Hessian matrix  $H(\hat{\theta})$  cannot be individualized as that would destabilize the approximation. Therefore, we are borrowing the Hessian matrix from the complete sample. As a result, the individualized approximated scores

$$S_i(\hat{\theta}) + \frac{1}{N} H(\hat{\theta})(\theta - \hat{\theta}), \quad \text{for } i = 1, \dots, N, \quad (7)$$

all have the same slope.

Finally, finding the roots of these individualized scores yields the IPCs.

$$\begin{aligned} \mathbf{0} &= S_i(\hat{\boldsymbol{\theta}}) + \frac{1}{N}H(\hat{\boldsymbol{\theta}}) \left[ IPC_i(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\theta}} \right] \\ IPC_i(\hat{\boldsymbol{\theta}}) &= \hat{\boldsymbol{\theta}} - \left[ \frac{1}{N}H(\hat{\boldsymbol{\theta}}) \right]^{-1} S_i(\hat{\boldsymbol{\theta}}) \end{aligned} \quad (8)$$

The mean and variance of the IPCs correspond to well-known quantities. It follows directly from Equation (3) that the mean of the IPCs equals the parameter estimates; that is,  $1/N \sum_{i=1}^N IPC_i(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}}$ . Calculating the variance of the IPCs yields the following sample estimate of the so-called sandwich estimator of the covariance matrix of the parameters (see [22]):

$$\widehat{\text{Var}}[IPC_i(\hat{\boldsymbol{\theta}})] = \left[ \frac{1}{N}H(\hat{\boldsymbol{\theta}}) \right]^{-1} \frac{1}{N} \sum_{i=1}^N S_i(\hat{\boldsymbol{\theta}})S_i(\hat{\boldsymbol{\theta}})^\top \left\{ \left[ \frac{1}{N}H(\hat{\boldsymbol{\theta}}) \right]^{-1} \right\}^\top = \widehat{\text{Var}}(\hat{\boldsymbol{\theta}}) \quad (9)$$

Next, we will give more explicit examples by deriving the IPCs to the regression coefficients of the standard multiple linear regression model, using the SSE objective function. The following model

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, N, \quad (10)$$

extends the simple linear regression model in Equation (1) by having  $p$  regressors  $\mathbf{x}$  with a corresponding  $p$ -variate vector of regression coefficients  $\boldsymbol{\beta}$ .

To calculate the IPCs to the estimated regression coefficients  $\hat{\boldsymbol{\beta}}$ , we need the first-order and second-order partial derivatives of the objective function. The SSE objective function minimizes the sum of the squared differences between the observed values of the outcome variable  $y$  and the predicted values based on the regression model. Formally, the SSE objective function is defined as

$$f(\boldsymbol{\beta}) = \sum_{i=1}^N (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2. \quad (11)$$

The first-order and second-order derivatives of the objective function are:

$$S(\boldsymbol{\beta}) = 2 \sum_{i=1}^N (-\mathbf{x}_i y_i + \mathbf{x}_i \mathbf{x}_i^\top \boldsymbol{\beta}) \quad (12)$$

$$H(\boldsymbol{\beta}) = 2 \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \quad (13)$$

Thus, the IPCs of individual  $i$  to the estimated regression coefficients  $\hat{\boldsymbol{\beta}}$  are given by

$$\begin{aligned} IPC_i(\hat{\boldsymbol{\beta}}) &= \hat{\boldsymbol{\beta}} - \left[ \frac{1}{N}H(\hat{\boldsymbol{\theta}}) \right]^{-1} S_i(\hat{\boldsymbol{\beta}}) \\ &= \hat{\boldsymbol{\beta}} - \left( \frac{2}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} (-2\mathbf{x}_i y_i + 2\mathbf{x}_i \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}) \\ &= \hat{\boldsymbol{\beta}} + \left( \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \mathbf{x}_i y_i - \left( \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \mathbf{x}_i \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} \\ &= \left( \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \mathbf{x}_i y_i + \left[ \mathbf{I}_p - \left( \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \mathbf{x}_i \mathbf{x}_i^\top \right] \hat{\boldsymbol{\beta}}, \end{aligned} \quad (14)$$

where  $\mathbf{I}_p$  denotes an identity matrix of order  $p$ .

Equation (14) allows us to illustrate the behavior of IPCs from groups of the sample. Let  $G$  be the index set of a group of  $n_G$  individuals with the same values in their covariates. Consider the following group-specific mean:

$$\begin{aligned} \frac{1}{n_G} \sum_{j \in G} IPC_j(\hat{\beta}) &= \left( \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \frac{1}{n_G} \sum_{j \in G} \mathbf{x}_j y_j \\ &+ \left[ \mathbf{I}_p - \left( \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \frac{1}{n_G} \sum_{j \in G} \mathbf{x}_j \mathbf{x}_j^\top \right] \hat{\beta} \end{aligned} \quad (15)$$

In the above equation, the outer product of the regressors  $\mathbf{x}\mathbf{x}^\top$  plays an important role. If the group-specific average of the outer product is identical to the outer product averaged over the complete sample, then the first term of Equation (15) equals the OLS estimate of the group-specific regression coefficients and the second term is zero. Therefore, the mean of the IPCs from this group will be an unbiased estimator of the group-specific regression coefficients. However, if the group-specific average of the outer product differs from the outer product averaged over the complete sample, then the first term will not be identical to the OLS estimate and the second term will not drop out. Thus, group differences in the outer product bias the IPCs. The size of this bias is proportional to the size of the differences in the outer products. These insights also apply to other model classes. Generally speaking, the bias of the IPCs increases with the amount of heterogeneity in the data [12]. In homogeneous samples, the IPCs are guaranteed to be unbiased.

### 3.2. Bias Correction Procedure

Arnold et al. [12] suggested a correction procedure termed iterated IPC regression that removes bias in the IPCs in a stepwise fashion. Iterated IPC regression proceeds as follows: first, the model is fitted, and standard IPC regression is performed. Second, individual-specific model parameters are predicted using the coefficients of the IPC regression equations. Third, these individual-specific model parameters are used to re-estimate the IPCs. The third step reduces some bias in the IPCs as the re-estimated IPCs are closer to the true individual-specific parameter values. By iterating over steps 2 and 3, the bias of the IPCs usually vanishes completely and the IPC regression estimates converge to unbiased estimates of the individual parameters. In a simulation study, Arnold et al. [12] compared both IPC regression approaches by predicting group differences in dynamic panel models. In contrast to standard IPC regression, whose estimates were slightly biased, the iterated algorithm provided nearly unbiased but slightly less precise estimates. Iterated IPC regression is implemented in the **ipcr** package. However, the iterated algorithm is limited to SEMs at present.

## 4. The ipcr Package: Overview and Installation

The **ipcr** package supplies functions for calculating IPCs and performing IPC regression. To a large extent, **ipcr** relies on infrastructure provided by the **sandwich** package [23]. **sandwich** allows users to estimate a wide variety of model-robust covariance estimators. The **sandwich** package contains the generic functions `estfun()` and `bread()` used as the building blocks to calculate the IPCs. `estfun()` returns the case-wise partial derivatives of the objective function with respect to the model parameters (see Equation (4) for the sum of the case-wise scores). `bread()` extracts an estimator for the expectation of the negative derivative of the objective function; usually the Hessian as defined in Equation (6).

Since the **ipcr** package was first introduced by Arnold et al. [12] to investigate parameter differences in SEMs fitted with the **OpenMx** package [24], support for various parametric models has been added. Now, **ipcr** can also be used to investigate SEMs fitted with the **lavaan** package [21]. Other than SEMs, linear and generalized regression models fitted by R's built-in `lm()` and `glm()` functions are supported. Moreover, by using some extractor functions provided by the **merDeriv** package [25], **ipcr** can also be applied to study

parameter heterogeneity in Gaussian linear mixed models fitted with the **lme4** package [6]. Other notable new features are better plotting capabilities and regularized IPC regression.

**ipcr** can be downloaded from GitHub <https://github.com/manuelarnold/ipcr> (accessed on 5 August 2021). The easiest way to obtain **ipcr** is via the **devtools** package [26]. To install and load **ipcr** within R, simply use the following commands:

```
R> library("devtools")
R> install_github("manuelarnold/ipcr")
R> library("ipcr")
```

## 5. Application

This section demonstrates how parameter differences can be detected and predicted with the **ipcr** package, using the HolzingerSwineford1939 data set included in the **lavaan** package.

### 5.1. Data Overview

HolzingerSwineford1939 is a classic data set that has often been used to illustrate different SEM applications. It consists of mental ability test scores of 301 seventh-grade and eighth-grade children from two different schools and additional variables indicating the children's sex, age, school, and grade. For the sake of simplicity, we limit this demonstration to a subset of the variables. We will use three visual ability test scores to fit a confirmatory factor analysis (CFA) model. Then, we use sex, age, school, and grade as IPC regression predictor variables to predict differences in the model. The variables are briefly described in Table 1.

**Table 1.** Selected variables from the HolzingerSwineford1939 data set.

Variable Name	Description	Level of Measurement
<i>Model data:</i>		
x1	Visual perception	Interval ( $M = 4.94$ , $SD = 1.17$ )
x2	Cubes	Interval ( $M = 6.09$ , $SD = 1.18$ )
x3	Lozenges	Interval ( $M = 2.25$ , $SD = 1.13$ )
<i>Covariates:</i>		
sex	Gender	Nominal (48.3% female, 51.7% male)
ageyr	Age, year part	Interval ( $M = 13.00$ , $SD = 1.05$ )
agemo	Age, month part	Interval ( $M = 5.38$ , $SD = 3.46$ )
school	School (Pasteur or Grant-White)	Nominal (52% Pasteur, 48% Grant-White)
grade	Grade	Ordinal (52.3% grade 7, 47.7% grade 8)

Note: *M*: mean, *SD*: standard deviation.

### 5.2. Data Pre-Processing

Below, we explain how missing data can be handled and show some preparation steps that facilitate the interpretation of IPC regression. First, we load the **lavaan** package that contains the data set. Then, we store the data set under a new variable name:

```
R> library("lavaan")
R> HS_data <- HolzingerSwineford1939
```

The students' age is stored in two different variables: `ageyr` measures the age in completed years, and the variable `agemo` contains the additional months. Next, we combine this information into a single variable that measures the students' age in months:

```
R> HS_data$age_months <- 12 * HS_data$ageyr + HS_data$agemo
```

Currently, the **ipcr** package requires the data used to fit the model and covariates to be complete, that is, without any missing values. In the cases of the HolzingerSwineford1939 data set, there is just a single missing value in the covariate `grade`. In order to keep this



illustration simple, we will exclude the individual with the missing value from the data set by deleting the corresponding row in the `data.frame`. However, it has long been known that list-wise deletion is not optimal, and users with incomplete data sets might consider more sophisticated methods such as multiple imputation (see [27,28]) to deal with missing values. Here, we delete the incomplete row with the following command:

```
R> HS_data <- HS_data[complete.cases(HS_data), ]
```

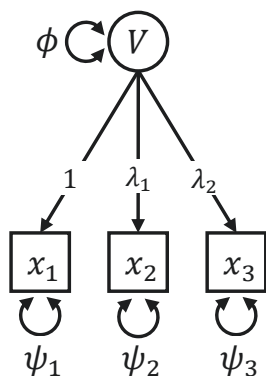
As in standard linear regression, the interpretation of the intercept can often be facilitated by centering the predictor variables at their means. After centering the covariates, a regression intercept represents the estimated parameter value for individuals with an average value on all covariates. The effects of categorical covariates, such as grouping variables, can be studied by encoding them as dummy variables. In the following, we center the covariate months and change the coding of all categorical covariates to either 0 or 1. We store the resulting covariates in a new `data.frame` called `covariates`:

```
R> covariates <- data.frame(sex = HS_data$sex - 1,
+                           months = scale(HS_data$age_months,
+                                           center = TRUE,
+                                           scale = FALSE),
+                           school = as.numeric(HS_data$school) - 1,
+                           grade = HS_data$grade - 7)
```

### 5.3. Fitting the Model

After the data have been prepared, we fit a CFA model using the `cfa` function from **lavaan**. The CFA consists of a latent variable, indicating the students' visual ability, measured by three manifest variables  $x_1$ ,  $x_2$ , and  $x_3$ . Figure 2 shows a graphical representation of the model. The corresponding **lavaan** syntax for specifying this model is as follows:

```
R> model_visual <- 'visual =~ x1 + x2 + x3'
```



**Figure 2.** Path diagram of the CFA model. The parameters  $\lambda_1$  and  $\lambda_2$  denote factor loadings,  $\phi$  represents the variance of the latent variable  $V$ , and  $\psi_1$ ,  $\psi_2$ , and  $\psi_3$  are measurement error variances.

We can now fit the model:

```
R> fit_visual <- cfa(model = model_visual, data = HS_data)
```

The model contains 6 free parameters: 2 factor loadings (the factor loading of  $x_1$  is fixed at 1 and the loadings of  $x_2$  and  $x_3$  are estimated), 1 latent variance parameter, and 3 residual variance parameters of the manifest variables. Since the model has zero degrees of freedom and is just identified, we cannot assess its fit.



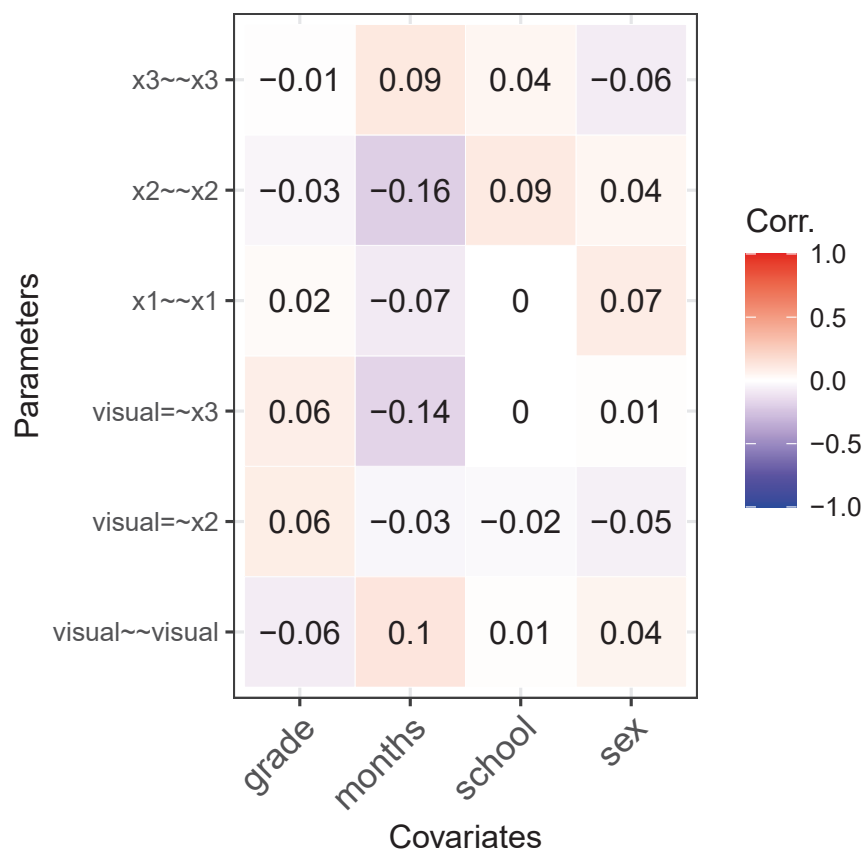
#### 5.4. Individual Parameter Contribution Regression

The core of the **ipcr** package is the `ipcr()` function. `ipcr()` calculates IPCs for the parameters of a fitted model and, when provided with a `data.frame` of covariates, performs IPC regression. Other than the `ipcr()` function, the **ipcr** package offers the convenience functions `get_ipcs()` and `get_scores()`, which return a `data.frame` of IPCs or scores, respectively.

In the following, we call the `ipcr()` function to investigate whether and how the parameters of the CFA model for visual abilities vary with respect to the covariates:

```
R> library(ipcr)
R> ipcr_visual <- ipcr(fit = fit_visual, covariates = covariates)
```

`ipcr()` automatically regresses the IPCs of all model parameter estimates on all covariates in the provided `data.frame` and returns an object of class `ipcr`. The usual R accessor functions such as `coef()`, `fitted()`, `plot()`, `predict()`, `print()`, and `summary()` can be used to extract various information from the `ipcr` object. To get a first overview, we can plot a heatmap using `plot(ipcr_visual)` that shows the correlations between the IPCs and covariates (see Figure 3). By specifying the argument `print_corr = TRUE`, the correlation coefficients are printed on the heatmap. Note that the heatmap depicts the zero-order correlations between IPCs and covariates, whereas IPC regression estimates the partial effects of the covariates on the IPCs. Zero-order correlations and partial effects might differ, especially if some of the covariates are correlated.



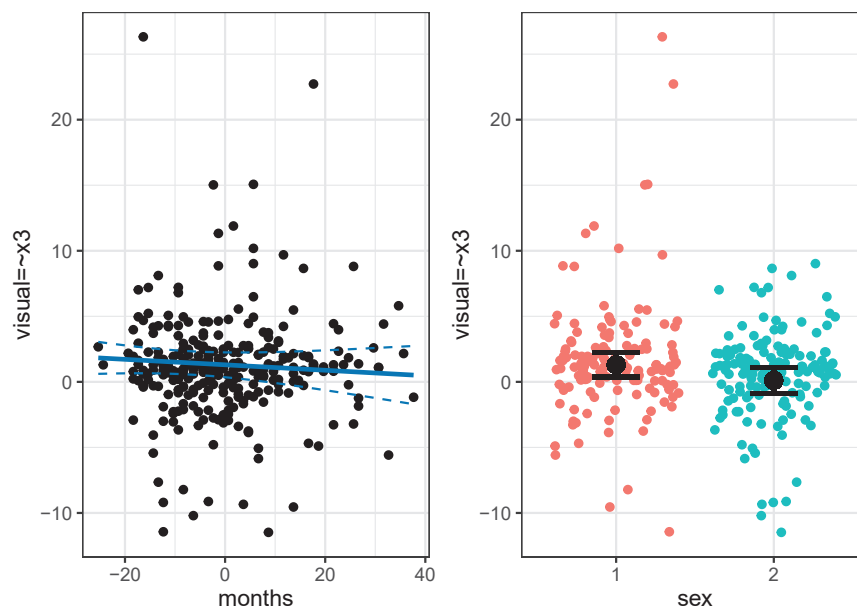
**Figure 3.** Correlations between covariates (x-axis) and the IPCs of the model parameters (y-axis) obtained with `plot(ipcr_visual, print_corr = TRUE)`. Parameter names are given in **lavaan** syntax.

More detailed information can be obtained with the `summary()` function. `summary()` prints the IPC regression results for each estimated model parameter:

```
R> summary(ipcr_visual)
```

```
Standard IPC regression coefficients:
Parameter  Covariate Estimate Std..Error t.value Pr...t..
1 visual=~x2 (Intercept)  0.565    0.326  1.732  0.084
2 visual=~x2      sex    -0.253    0.313 -0.809  0.419
3 visual=~x2    months  -0.027    0.016 -1.712  0.088
4 visual=~x2    school   0.037    0.321  0.114  0.909
5 visual=~x2    grade    0.676    0.365  1.853  0.065
6 visual=~x3 (Intercept)  1.301    0.484  2.690  0.008
7 visual=~x3      sex   -1.192    0.464 -2.566  0.011
8 visual=~x3    months  -0.021    0.023 -0.897  0.371
9 visual=~x3    school   0.152    0.477  0.320  0.749
10 visual=~x3    grade   0.722    0.541  1.333  0.184
11 x1~~x1 (Intercept)  1.046    0.288  3.634  0.000
12 x1~~x1      sex    -0.253    0.276 -0.917  0.360
13 x1~~x1    months   0.015    0.014  1.065  0.288
14 x1~~x1    school  -0.069    0.284 -0.242  0.809
15 x1~~x1    grade  -0.093    0.322 -0.287  0.774
16 x2~~x2 (Intercept)  1.305    0.229  5.695  0.000
17 x2~~x2      sex   -0.603    0.220 -2.739  0.007
18 x2~~x2    months   0.003    0.011  0.251  0.802
19 x2~~x2    school   0.325    0.226  1.438  0.151
20 x2~~x2    grade  -0.195    0.257 -0.759  0.448
21 x3~~x3 (Intercept)  0.255    0.269  0.947  0.344
22 x3~~x3      sex    0.341    0.259  1.317  0.189
23 x3~~x3    months  -0.014    0.013 -1.084  0.279
24 x3~~x3    school   0.240    0.266  0.905  0.366
25 x3~~x3    grade   0.149    0.302  0.493  0.622
26 visual~~visual (Intercept)  0.577    0.310  1.864  0.063
27 visual~~visual      sex   0.611    0.297  2.055  0.041
28 visual~~visual    months  0.028    0.015  1.914  0.057
29 visual~~visual    school  -0.108    0.305 -0.353  0.725
30 visual~~visual    grade  -0.650    0.347 -1.875  0.062
```

The output is structured as follows: the first column contains the names of the model parameters. The second column shows the names of the covariates, where (Intercept) simply refers to the intercept of the corresponding IPC regression equation. The remaining columns display the estimates, standard errors,  $t$ -values, and  $p$ -values of the IPC regression parameters as in a standard regression output. Using a significance level of  $\alpha = 0.05$ , we infer that there are three significant sex differences: the factor loading of the indicator  $x_3$ , the error variance of the indicator  $x_2$ , and the variance of the latent factor differs significantly between girls and boys. The other covariates do not significantly affect the remaining model parameters. The partial effects of the IPC regression equations can be visualized with the `plot_differences()` function. Figure 4 shows the estimated factor loading of the indicator  $x_3$  for different values of the covariates months and sex when all other covariates are zero.



**Figure 4.** Estimated conditional differences in the factor loading of the indicator x3. Left plot: IPCs to the factor loading are plotted on the students' age in months. The solid blue line marks the estimated factor loading as a function of months when all other covariates are zero. The dashed blue lines represent the upper and lower limit of the 95% pointwise confidence bands. Right plot: IPCs to the factor loading are plotted on the students' gender. The big dots mark the estimated factor loadings for girls and boys when all other covariates are zero. The error bars indicate the 95% confidence intervals of the estimated gender-specific factor loading.

By default, the accessor functions extract information about all model parameters. The output can be limited to a single or a subset of parameters of interest by specifying the parameter argument. For instance, the function call `coef(ipcr_visual, parameter = "visual~visual")` shows only the IPC regression estimates for the variance parameter of the latent factor.

### 5.5. Non-Linear Effects and Interactions

Since IPC regression is linear regression at its heart, studying non-linear effects or interactions between covariates follows the known logic of adding extra terms to the regression model that are functions of the original covariates. For example, curvilinear relationships between model parameters and covariates can be modeled by adding polynomial terms to the `data.frame` provided to the `covariate` argument. Similarly, we can probe interactions between covariates by including product terms. To avoid multicollinearity, we recommend centering all non-dummy covariates used to generate new polynomial or product terms. It is important to remember that more complex models may seem to explain a larger part of the observed variation than simple models unless model selection is performed using approaches that adjust for model complexity or use independent data. A detailed account of modeling non-linear and interaction effects in linear regression models is given by Cohen et al. [29].

Besides exploring relationships between model parameters and covariates, IPC regression also provides means to discover non-linear relationships in the model. For example, consider the simple linear regression model in Equation (1). This model postulates a linear effect of the explanatory variable  $x$  on the outcome  $y$  denoted by the regression coefficient  $\beta_1$ . By regressing the IPCs to  $\beta_1$  on the squared variable  $x$ , we can estimate and test the explanatory variable's potential quadratic effect on the outcome  $y$ .

### 5.6. Bias Correction

IPC regression may provide biased estimates of parameter differences in heterogeneous samples (see [12]). The `ipcr` package provides a correction procedure called iterated IPC regression that removes the bias in the IPC regression estimates at the cost of adding additional variance. One can perform iterative instead of standard IPC regression by calling the `ipcr()` function with the additional argument `iterate` set to `TRUE`. By default, `iterate` is set to `FALSE`, and standard IPC regression is performed. The iterated IPC regression algorithm tries to remove bias in the IPCs in a stepwise fashion and updates the IPC regression estimates as long as the change in any parameter between iterations is smaller than a numerical threshold (by default, the absolute difference needs to become smaller than  $1 \times 10^{-4}$ ). This stopping criterion can be changed via the `conv` argument of the `ipcr()` function. Moreover, the user can also specify the maximum number of iterations by changing the `max_it` argument. By default, the algorithm terminates after 50 iterations. We will later compare the performance of both standard and iterated IPC regression in a series of simulation studies.

### 5.7. Regularization

If the number of covariates becomes large, there is a risk that we overfit the regression models. A solution to this problem is to make the IPC regression output sparser by coupling IPC regression with the least absolute shrinkage and selection operator (LASSO) [30,31]. LASSO is a widely applied form of regularization that adds a penalty term to the SSE objective function. This penalty term shrinks regression parameters towards zero, thus setting some of them to zero, and, as a result, produces sparser models and reduces overfitting. Note that this approach sets out to minimize prediction error at the cost of regression coefficients becoming biased estimates (see also [32]).

The `ipcr` package performs LASSO regularized IPC regression by interfacing the `glmnet` package [33]. More specifically, by setting the argument `regularization` to `TRUE`, `ipcr()` calls the `cv.glmnet()` function to estimate regularized linear regression models, using  $k$ -fold cross-validation. The settings of `cv.glmnet()` can be changed by providing `ipcr()` with the specific arguments. By default, 10-fold cross-validation with 100 different values for the penalty term is carried out.

To perform LASSO regularized IPC regression, all covariates need to be standardized beforehand. Standardization prevents LASSO from preferring variables with more variability over variables with less variability. Without standardization, covariates with smaller variances will be penalized more severely than covariates with larger variances. The following command standardizes all covariates:

```
R> covariates_std <- scale(covariates, center = FALSE, scale = TRUE)
```

Then, we perform regularized IPC regression and inspect the results with `summary()`:

```
R> ipcr_visual_reg <- ipcr(fit = fit_visual, covariates = covariates_std,
+                          regularization = TRUE)
R> summary(ipcr_visual_reg)
```

Regularized standard IPC regression coefficients:

Parameter	Covariate	Estimate
1	visual=~x2 (Intercept)	0.6
2	visual=~x2 sex	-0.142
3	visual=~x2 months	-0.274
4	visual=~x2 school	.
5	visual=~x2 grade	0.403
6	visual=~x3 (Intercept)	1.449
7	visual=~x3 sex	-0.543
8	visual=~x3 months	.
9	visual=~x3 school	.

```

10 visual=~x3      grade      0.071
11 x1~~x1 (Intercept) 0.835
12 x1~~x1      sex          .
13 x1~~x1      months      .
14 x1~~x1      school      .
15 x1~~x1      grade      .
16 x2~~x2 (Intercept) 1.211
17 x2~~x2      sex        -0.329
18 x2~~x2      months      .
19 x2~~x2      school      0.131
20 x2~~x2      grade      .
21 x3~~x3 (Intercept) 0.627
22 x3~~x3      sex          .
23 x3~~x3      months      .
24 x3~~x3      school      .
25 x3~~x3      grade      .
26 visual~~visual (Intercept) 0.527
27 visual~~visual      sex          .
28 visual~~visual      months      .
29 visual~~visual      school      .
30 visual~~visual      grade      .

```

The `summary()` function shows by default the regularized parameter estimates for the penalty term associated with the minimum mean cross-validated error. Parameter estimates for specific penalty terms can be obtained by specifying the `s` argument in the `ipcr()` function call and can then be displayed via `summary()`. In the above output, coefficients that were set to zero are marked with a dot, resulting in a sparse coefficients matrix with the most important coefficients. Standard errors,  $t$ -values, and  $p$ -values are not shown in the output as they are not provided by `glmnet`.

## 6. Simulation Studies

The following series of simulation studies showcases different applications of IPC regression and illustrates the method's statistical properties. In Simulation I, we revisit the introductory example and predict group differences in the variance of the regression errors of a simple linear regression model. In the subsequent three simulation studies, we turn to SEMs. In Simulation II, we investigate the effect of multiple covariates on the type I error rate. Then, we compare the performance of IPC regression with MGSEMs (Simulation III) and SEMs with interactions (Simulation IV) as established benchmark methods.

All simulations were carried out with R (version 4.0.3). We used R's built-in `lm` function to estimate the simple linear regression model in Simulation I. All SEMs were fitted with the `lavaan` package (version 0.6–7). We used a developmental snapshot of the `ipcr` package to perform IPC regression (commit: <https://github.com/manuelarnold/ipcr/commit/d4132c73b0e05ced1da6be71119d846424a1a3d6> (accessed on 6 August 2021). d4132c7). Throughout all simulations, we set the significance level for all hypothesis tests to 5%. We replicated all experimental conditions 1000 times. The simulation scripts and complete simulation results can be found in our online Supplementary Material (<https://osf.io/p5xrk> accessed on 6 August 2021).

### 6.1. Simulation I: Simple Linear Regression Model

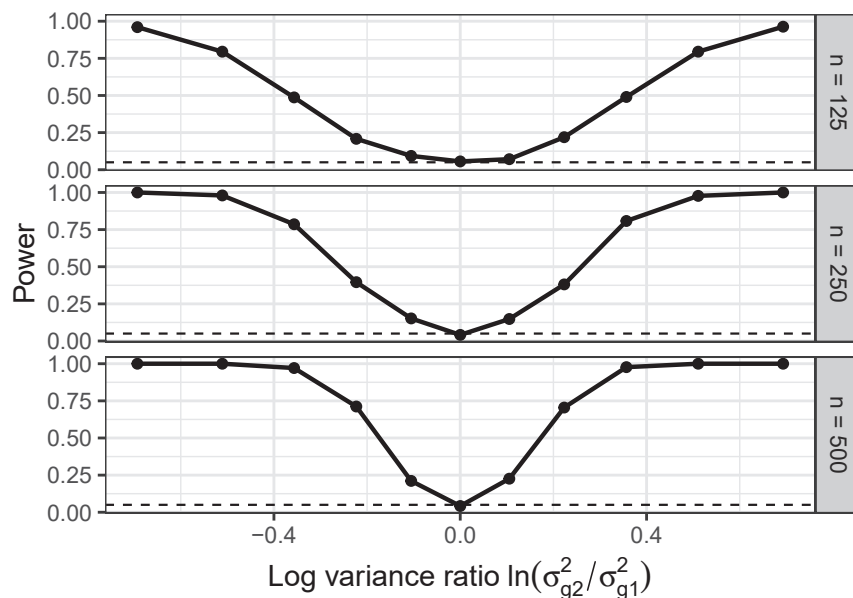
Simulation I assessed how well IPC regression could predict group differences in a simple linear regression model such as the one shown in Equation (1). We sampled the data from a two-group population with a group difference in the error variance  $\sigma^2$ . This simulation setup corresponds to a situation in which the reliability of the outcome variable differs between groups. For example, such group differences are often encountered in cross-cultural samples (e.g., [34]). If overlooked, the group differences would bias standard

errors and render the OLS estimator inefficient. The remaining model parameters, the intercept  $\beta_0$  and the slope  $\beta_1$ , did not vary between groups and were set to 1 and 0.5, respectively. In every simulation replication, we simulated data, fitted the linear regression model while ignoring the group difference, and then recovered the group difference with IPC regression using a dummy variable which was 0 in the first group and 1 in the second group. We considered only the performance of standard IPC regression in Simulation I because iterated IPC regression for linear regression models is not yet implemented in the `ipcr` package. We investigated the following simulation conditions:

- *Group-specific value of  $\sigma^2$* : The error variance of the first group  $\sigma_{g1}^2$  was set to 1 in all simulation conditions. In the second group, the error variance  $\sigma_{g2}^2$  varied across the following values: 5/10, 6/10, 7/10, 8/10, 9/10, 10/10, 10/9, 10/8, 10/7, 10/6, 10/5. We chose the values so that the absolute value of the log-variance ratio  $|\ln(\sigma_{g2}^2/\sigma_{g1}^2)|$  was the same for the most extreme conditions (5/10 and 10/5). Note that the 10/10 condition resulted in a homogeneous sample without group differences.
- *Sample size*: The sample size per group  $n$  was either 125, 250, or 500. The total sample size  $N$ , therefore, equaled 250, 500, or 1000.

#### 6.1.1. Power

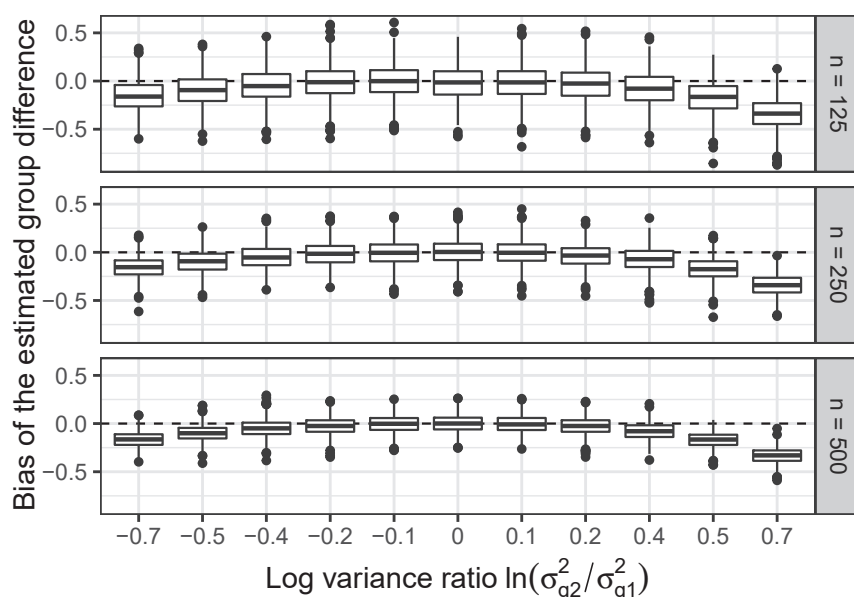
First, we report the observed statistical power of IPC regression to detect the group difference in the error variance. Statistical power is the proportion of  $t$ -tests that correctly rejected the null hypothesis of equal error variance in both groups if there truly was a difference. Figure 5 relates the power to the size of the group difference and sample size. We plotted the natural logarithm of the ratio of the error variance in the second group to the error variance in the first group on the  $x$ -axis so that the distances between consecutive ratios were similar. As expected, the power increased with the size of the group difference and sample size. When the error variance in the second group was either half or twice as large as in the first group, the power approached 100%, regardless of sample size. In the absence of any group differences, when the error variance was equal to 1 in both groups, the proportions of the rejected test approached the significance level of 5%.



**Figure 5.** The power of IPC regression to detect a group difference in the error variance  $\sigma^2$ . The dashed line marks a power of 5%.

### 6.1.2. Estimated Group Difference

Besides the statistical power to detect a group difference, the quality of the estimated group-specific parameters is also crucial. Figure 6 depicts the bias of the estimates of the group difference in the error variance provided by IPC regression. IPC regression yielded nearly unbiased estimates when the true group difference was small or zero. However, larger group differences were negatively biased. This negative bias implied that IPC regression either overestimated or underestimated the group difference, depending on whether the error variance of the second group (which varied across simulation conditions) was smaller or larger than the error variance of the first group (which was fixed at 1 in all experimental conditions). When the error variance of the second group was smaller than the error variance of the first group, IPC regression overestimated the group difference. Conversely, when the error variance was larger in the second group than in the first group, IPC regression underestimated the group difference. The sample size had no apparent effect on the bias but decreased the variability of the estimates.

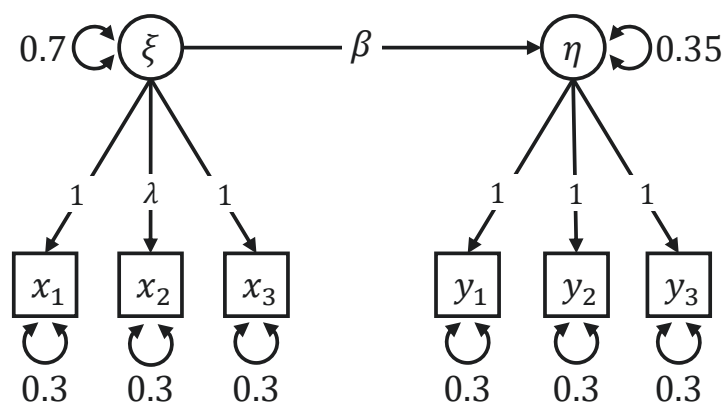


**Figure 6.** Boxplots with the bias of the estimated group difference in the error variance  $\sigma^2$ .

### 6.2. Simulation II: Type I Error Rate

Simulation II aimed to assess the type I error rate of IPC regression, that is, the proportions of false-positive findings that a parameter differs with respect to a covariate although it is constant in the population. Ideally, the rate of false positives should approach the significance level of 5%.

The population model used in Simulation II and the remaining simulations is presented in Figure 7. Depicted is a regression model with two latent variables, each measured by three indicators. The values of the factor loading  $\lambda$  and the regression coefficient  $\beta$  varied across groups and individuals in Simulation III and IV but were constant in Simulation II with  $\lambda = 1$  and  $\beta = 0.5$ . Thus, the population used in Simulation II was homogeneous. After generating multivariate normally distributed data from the population, we fitted the model by fixing the first indicator of each latent variable (that is,  $x_1$  and  $y_1$ ) to 1 and estimating the remaining 13 parameters (4 factor loadings, 6 residual variance parameters, 2 latent variance parameters, and 1 regression parameter) freely. Then, standard and iterated IPC regression were performed.



**Figure 7.** Path diagram of the simulation model. The factor loading  $\lambda$  varied between two groups in Simulation III and the regression parameter  $\beta$  differed across individuals in Simulation IV.

We investigated the following experimental factors:

- *Number of covariates:* The IPC regression algorithm was provided either with 1, 2, or 3 covariates. These covariates did not predict any parameter differences.
- *Type of covariates:* The covariates were either dummy or standard normally distributed variables.
- *Sample size (N):* The simulated samples contained either 250, 500, or 1000 individuals.

#### 6.2.1. Type I Error Rate

The observed type I error rates were close to the optimal rate of the 5% significance level. The average type I error rate was 4.92% for standard IPC regression with little variation across model parameters and simulation conditions. The rate of iterated IPC regression was slightly larger and was 5.25%. Iterated IPC regression rejected the null hypothesis more often than expected in smaller samples. When averaged over all model parameters and the other simulation conditions, iterated IPC regression exhibited a type I error rate of 5.58% in samples with 250 individuals, 5.40% in samples with 500, and 5.09% in samples with 1000 individuals.

Additional information about the convergence of the iterated IPC regression algorithm is given in Appendix A.1.

#### 6.3. Simulation III: Group Difference in the Measurement Part

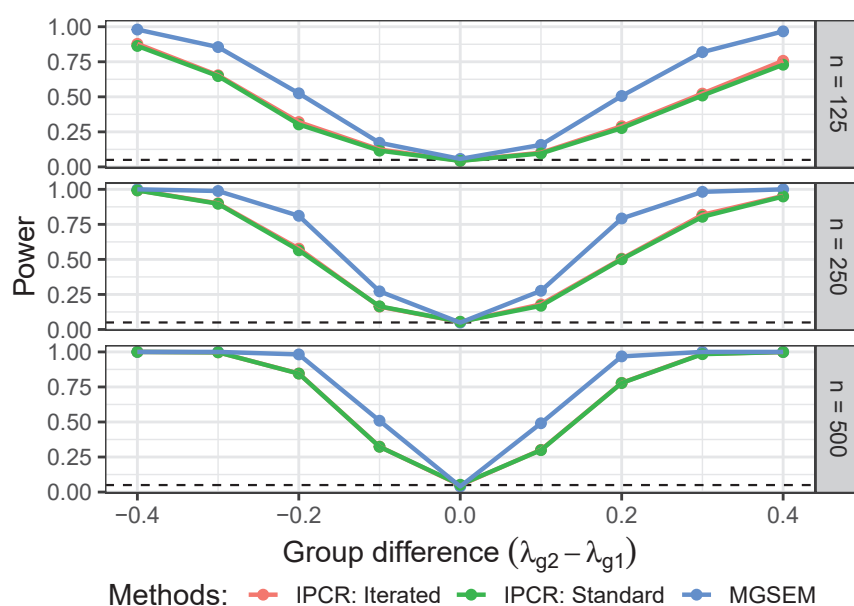
In Simulation III, we introduced heterogeneity to the measurement part of the model shown in Figure 7 by letting the factor loading  $\lambda$  vary between two groups. The regression parameter  $\beta$  was constant and was set to 0.5. IPC regression was provided with a grouping variable to predict the group difference. As a reference, we compared the performance of IPC regression with MGSEMs. The power to detect group differences of MGSEMs was assessed by performing likelihood-ratio tests between a MGSEM where  $\lambda$  was allowed to vary between groups and a single-group SEM where  $\lambda$  was constrained to be equal between groups. The remaining parameters of the MGSEMs were constrained to be equal between groups. The following simulation conditions were investigated:

- *Group-specific value of  $\lambda$ :* The value of the factor loading in the first group  $\lambda_{g1}$  was set to 1. For the second group, the value of  $\lambda_{g2}$  varied across 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, and 1.4.
- *Sample size:* The sample size per group  $n$  was either 125, 250, 500. Therefore, the total sample size  $N$  was 250, 500, or 1000.



### 6.3.1. Power

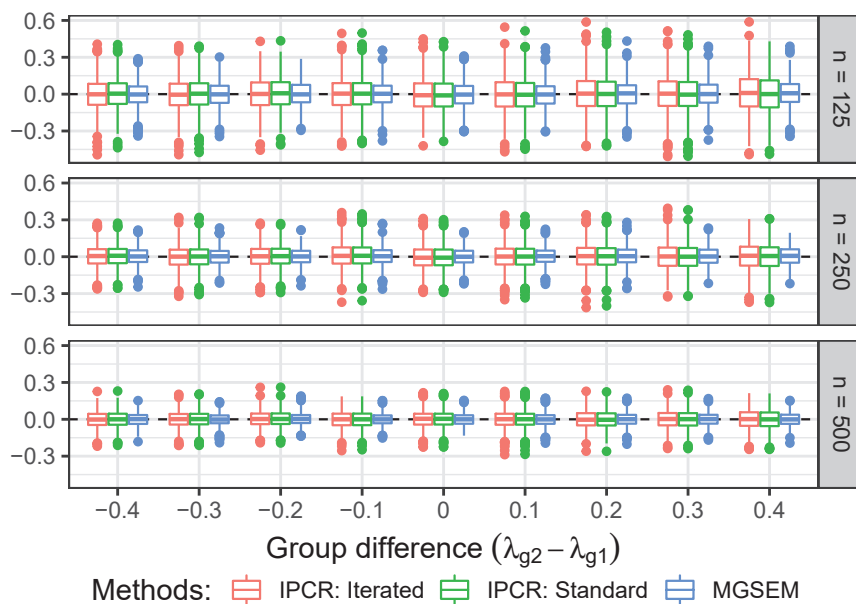
Figure 8 shows the power of standard and iterated IPC regression and the MGSEMs to detect the differences in the factor loading  $\lambda$ . The power of standard and iterated IPC regression was nearly identical, resulting in overlapping lines. As expected, the power of all methods grew substantively with the sample size and the size of the group differences. MGSEMs were consistently more powerful than the IPC regression methods in detecting the group difference. For medium-sized group differences, when  $\lambda_{g2}$  was either 0.8 or 1.2, we found MGSEMs on average to outperform the IPC regression methods by roughly 21 percentage points. When  $\lambda_{g2} = 1$ , the factor loading was identical in both groups, resulting in a homogeneous sample. In this case, the proportions of significant tests approached the significance level of 5% for all methods.



**Figure 8.** The power of standard and iterated IPC regression and MGSEMs to detect a group difference in the factor loading  $\lambda$ . The dashed line marks a power of 5%.

### 6.3.2. Estimated Group Difference

Figure 9 presents boxplots of the bias of the estimated group difference in the factor loading, provided by standard and iterated IPC regression and MGSEMs. All methods produced nearly unbiased estimates. The estimates of the group difference provided by MGSEMs exhibited less variability than the IPC regression estimates in all experimental conditions. In terms of the root mean squared error (RMSE) averaged over the two group-specific estimates of the factor loading, MGSEMs (RMSE: 0.067) were marginally more accurate than standard (RMSE: 0.076) and iterated (RMSE: 0.077) IPC regression.



**Figure 9.** Boxplots with the bias of the estimated group difference in the factor loading  $\lambda$  ( $y$ -axis) for different true group differences and sample sizes.

#### 6.4. Simulation IV: Individual Differences in the Structural Part

In Simulation IV, we compared the performance of standard and iterated IPC regression to a SEMs with an interaction term. The data were generated by letting the regression parameter  $\beta$  vary across individuals and fixing the factor loading  $\lambda$  at 1. More specifically, the individual values of  $\beta$  were a linear function of a standard normally distributed covariate  $z$ :

$$\beta_i = 0.5 + \gamma z_i \quad (16)$$

The value of the coefficient  $\gamma$  determined the relationship between covariate and regression parameter and was of primary interest in this simulation. We investigated the following simulation conditions:

- *Value of  $\gamma$ :* The dependency of the individual regression parameter values on the covariate was either  $-0.2$ ,  $-0.15$ ,  $-0.1$ ,  $-0.05$ ,  $0$ ,  $0.05$ ,  $0.1$ ,  $0.15$ , or  $0.2$ . Note that the zero condition corresponds to a homogeneous sample with a constant regression parameter  $\beta$ .
- *Sample size ( $N$ ):* The simulated samples contained either 250, 500, or 1000 individuals.

There are several ways to specify SEMs with interactions (see [10] for an overview). Following Marsh et al. [35], we contrasted IPC regression with a SEM containing a product term consisting of the indicators of the exogenous latent variable  $\zeta$  and the covariate  $z$ . We used a parcel of the indicators  $x_1$ ,  $x_2$ , and  $x_3$  to create this product term, which we denote with  $z\bar{x}$ . Note that  $\bar{x}$  refers to the row-wise means and not the mean of a single variable. We also added the covariate to the model and specified a main effect of the covariate on the endogenous latent variable  $\eta$ . Moreover, we let all exogenous predictor variables covary. Finally, to avoid having to specify a mean structure, we used double-mean centering (see [36]); that is, we centered all manifest variables, computed the product term using the centered variables, and then centered the product term again. A path diagram of the interaction model is shown in Figure 10. The direct effect of the product term on the endogenous variable (i.e., the regression parameter  $\beta_{\eta, z\bar{x}}$ ) is an estimate of the coefficient  $\gamma$

in Equation (16). We assessed the power of the SEMs with a product term by testing the hypothesis  $\beta_{\eta,z\bar{x}} = 0$  via z-tests.

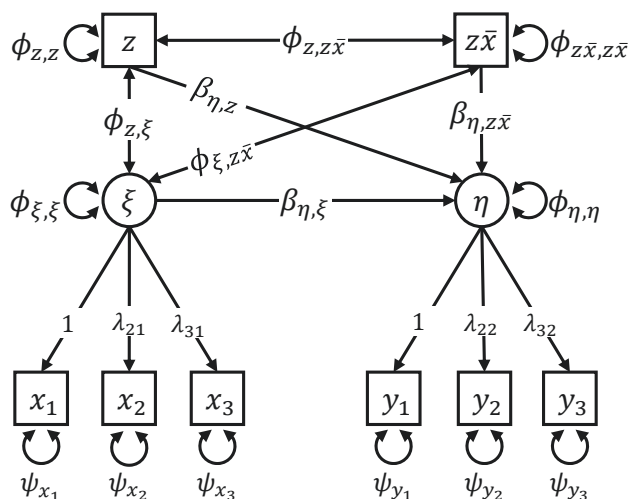


Figure 10. Path diagram of the interaction model with a product term.

6.4.1. Power

Figure 11 presents the observed power of the IPC regression methods and the interaction SEMs to detect that the regression parameter  $\beta$  depended on the covariate  $z$ . As in Simulation II, the power of standard and iterated IPC regression was nearly the same. The SEMs with the product term were slightly more powerful than the IPC regression methods. The largest difference in power was found when the sample consisted of 500 individuals and the absolute value of the coefficient  $\gamma$  was 0.1. Under this condition, the SEMs with a product term outperformed the IPC regression methods by approximately 15 percentage points. For  $\gamma = 0$ , when the regression parameter  $\beta$  did not depend on the covariate  $z$ , the proportions of significant tests of all methods approached the significance level of 5%.

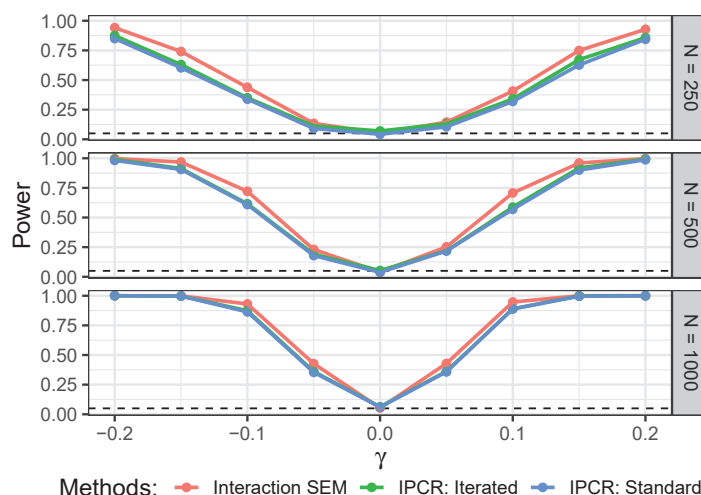
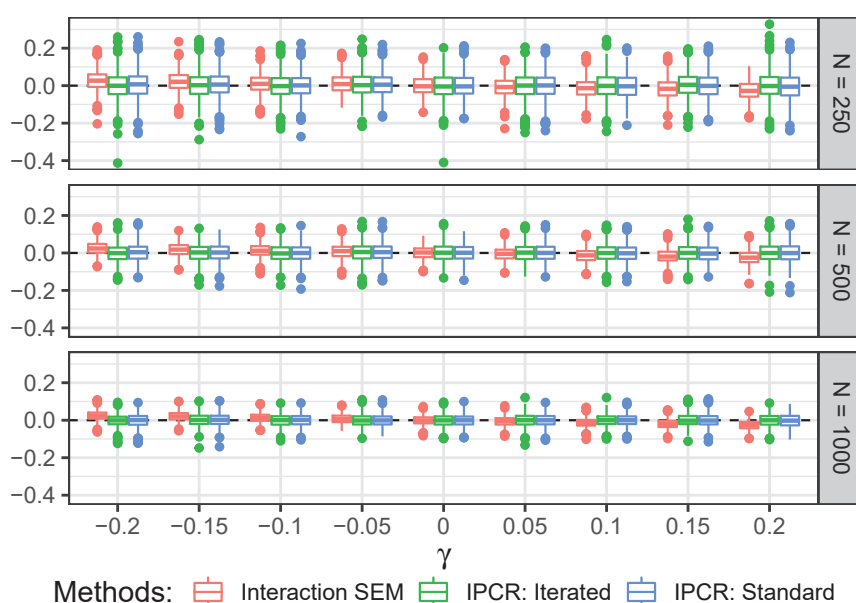


Figure 11. The power of standard and iterated IPC regression and SEMs with a product term to detect the dependency of the regression parameter  $\beta$  on the covariate  $z$ . The dashed line marks a power of 5%.

#### 6.4.2. Estimated Interaction

Figure 12 shows boxplots with the bias of the estimates of the coefficient  $\gamma$  provided by standard and iterated IPC regression and the SEMs with a product term. SEMs with a product term underestimated the absolute value of the dependency slightly. In comparison, IPC regression yielded mostly unbiased but also more volatile estimates. Especially in smaller samples with 250 observations, iterated IPC regression produced some outliers that were not found in the estimates provided by standard IPC regression. In terms of RMSE averaged over all simulation conditions, standard (RMSE: 0.051) and iterated (RMSE: 0.051) IPC regression were marginally outperformed by SEMs with a product term (RMSE: 0.042).

Additional information about the convergence of the iterated IPC regression algorithm is given in Appendix A.2.



**Figure 12.** Boxplots with the bias of the estimates of the dependency  $\gamma$  ( $y$ -axis) for different values of  $\gamma$  and sample sizes.

## 7. Discussion

The present study showed how differences in model parameters can be predicted with IPC regression using the `ipcr` package. We showcased the functionality of the `ipcr` package using a classic data set and provided detailed R code. We expanded upon earlier research about predicting parameter differences in SEMs and demonstrated that IPC regression can also predict parameter differences in linear regression models. Moreover, we presented novel simulation results that illuminated the method's performance for linear regression models and contrasted it with SEMs with an interaction term.

We see IPC regression primarily as a diagnostic tool that indicates whether there is any predictive potential for parameter heterogeneity in a set of covariates and provides researchers with hints on how these covariates can be integrated into their models. As such, IPC regression offers some advantages compared to other procedures for exploring heterogeneity with covariates. One of these advantages is the flexibility of the method: IPC regression allows researchers to investigate all types of model parameters. We demonstrated this flexibility by predicting group differences in the error variance of a linear regression model. Another merit of IPC regression is its simplicity. Especially for model classes where model specification and estimation can become complicated such as SEMs, researchers can obtain results much quicker with IPC regression via a single call of the

`ipcr()` function than, for instance, from specifying and estimating MGSEMs and SEMs with interaction terms. One can also argue that this simplicity makes IPC regression less prone to specification errors and may prove useful if a complex model is hard to estimate. Finally, IPC regression can guide the selection of important covariates among a larger set of covariates, particularly in combination with LASSO regularization as implemented in the `ipcr` package.

The question remains as to how we can best translate the results of IPC regression into an improved parametric model. A key problem of data-driven methods like IPC regression is their susceptibility to capitalize on chance, that is, to overfit to noise in the data (see [37,38] for a discussion of the problem). In other words, IPC regression may indicate model modifications that increase the model's fit in the sample at hand but generalize poorly to new data. The risk of overfitting rises with the complexity of the model and the number of covariates. For example, given a model with 10 parameters and a set of 10 covariates, IPC regression will produce 100 IPC regression estimates with associated  $p$ -values. Using the conventional 5% level of significance and assuming homogeneous data (i.e., no parameter differences at all), one would expect to find five false-positive effects of a covariate on a model parameter on average. We recommend three different strategies to minimize the risk of overfitting the model. First, adopting a machine learning perspective, we encourage researchers to apply regularized IPC regression, which penalizes each regression for model complexity. Second, if the sample size is sufficiently large, one should additionally split the sample into a training and a test data set. The training data set is used for model estimation with IPC regression. After modifying the model following the results of IPC regression, the fit of the new augmented model is then independently evaluated in the test data set. Third, we suggest considering the estimated change in model parameters in addition to  $p$ -values to determine how to modify a given theory-based model. For instance, a statistically significant group difference in a parameter may be just too small to be of scientific interest and would not necessarily warrant model modification. In most cases, such thresholds for scientific relevancy will be different for different types of parameters. Most likely, heterogeneity in nuisance parameters such as an error variance needs to be much more pronounced than heterogeneity in a regression parameter central to one's inquiry to justify a model modification.

The results of our simulations were overall promising. In line with previous simulation studies reported by Oberski [11] and Arnold et al. [12], we found that IPC regression provided adequate control of type I errors under the null hypothesis of homogeneous parameters. Moreover, we found little differences in terms of bias and variance of the estimates provided by IPC regression and MGSEMs. However, as in the previous simulation studies, IPC regression was consistently less powerful in detecting parameter heterogeneity than MGSEMs. We expanded upon previous studies that only discussed IPC regression for SEMs and predicted a group difference in the error variance of a linear regression model. The estimates of the group differences were slightly biased. In another simulation study, we compared IPC regression to SEMs with a product interaction term suggested by Marsh et al. [10]. Interestingly, IPC regression was more accurate but less precise than SEMs with an interaction term and exhibited a lower power. As part of our simulation studies, we also compared standard IPC regression with iterated IPC regression. Iterated IPC regression was suggested by Arnold et al. [12] as an unbiased alternative to standard IPC regression. Unfortunately, these comparisons were limited to SEMs as iterated IPC regression has not yet been implemented for other model classes. Both IPC regression methods yielded almost identical results in terms of power and bias. However, iterated IPC regression showed some cases of non-convergence, especially in smaller samples and when provided with continuous covariates. Furthermore, the type I error rate of iterated IPC regression was marginally larger than the significance level in smaller samples. Similar performance of the IPC regression methods was to be expected because the parameter differences tested were too small to bias standard IPC regression.

Although the `ipcr` package makes performing IPC regression for various parametric models straightforward in most situations, it is still under ongoing development, and we would like to note some of its current limitations. The biased estimates obtained for the linear regression model underline the need to generalize the iterated IPC regression algorithm and develop a version for regression models. Moreover, the sometimes poor convergence rate of iterated IPC regression highlights the necessity for an improved algorithm that is more reliable in smaller samples and when provided with continuous covariates. Another limitation is that IPC regression currently requires complete data sets without missing values. Especially for researchers working with SEMs who routinely employ full information maximum likelihood to deal with missingness, this limitation might be an issue. We plan to implement support for SEMs fitted on incomplete data in the future. In the meantime, we suggest using an imputation technique such as multiple imputation as a workaround [27,28]. Finally, even though the combination of IPC regression and regularization seems like a natural fit to handle larger sets of covariates, more research is needed to understand this interplay better and determine better default settings.

IPC regression may not always be the best choice for exploring parameter heterogeneity with covariates. Besides adding covariates directly to the model, other techniques are available for testing and estimating the effects of covariates. If the sample is clustered into known groups, mixed-effects or multilevel models (e.g., [39,40]) are an obvious alternative and allow parameters to vary across groups. Unlike IPC regression, the use of mixed-effects models is usually motivated by a prior belief that certain data segments differ. In contrast, IPC regression is a more exploratory or data-driven procedure to identify potentially important covariates. Further, IPC regression does not rely on a clustered data structure. Of course, IPC regression can also be used to study heterogeneity in mixed-effects models. Other methods for exploring parameter heterogeneity are structural change tests (e.g., [16,18]). These tests have been recently introduced to psychometrics and are applied to discover parameter differences with respect to a covariate [19]. Especially if one wants to explore the effect of ordinal covariates, which can be hard to incorporate in the regression framework, structural change tests may be a well-suited alternative (see [20]). However, unlike IPC regression, structural change tests are limited to testing if a parameter changes but do not estimate how it changes with respect to a covariate. Another established approach to investigate heterogeneity with covariates are model-based recursive partitioning techniques (e.g., [41,42]). These methods divide a data set into homogeneous subgroups by finding covariates that predict parameter differences. For SEMs, model-based recursive partitioning was popularized as SEM trees and gained attention in the past years [43–45]. Finally, it is important to note that the performance of IPC regression depends primarily on the available covariates. If there are parameter differences, but these differences are unrelated to the covariates, IPC regression will fail to detect them. That is, covariates need to be sufficiently informative and reliable to be usable in the context of IPC regression. Consequently, IPC regression is not a global test of heterogeneity. Methods such as finite mixture models (e.g., [46–48]) that do not depend on covariates provide a more thorough test of parameter homogeneity than IPC regression.

Having these alternatives to IPC regression and limitations of the `ipcr` package in mind, we believe that IPC regression can be applied in many situations to uncover heterogeneity and gross model misspecification. Furthermore, in many cases in which additional covariates are available, it makes sense to run IPC regression as part of exploration and model checking procedures. Note that IPC regression should then be labeled as exploratory analyses and run only after all confirmatory actions were performed.

**Supplementary Materials:** The following are available at <https://osf.io/p5xrk>.

**Author Contributions:** Conceptualization, M.A., A.M.B. and M.C.V.; methodology, M.A.; software, M.A.; validation, M.A.; formal analysis, M.A.; investigation, M.A.; resources, A.M.B.; data curation, M.A.; writing—original draft preparation, M.A.; writing—review and editing, M.A., A.M.B. and

M.C.V.; visualization, M.A.; supervision, A.M.B and M.C.V.; project administration, M.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The simulated data presented in this study can be found in the following online repository: <https://osf.io/p5xrk> (accessed on 5 August 2021). The developmental snapshot of the `ipcr` package can be downloaded here: <https://github.com/manuelarnold/ipcr/commit/d4132c73b0e05ced1da6be71119d846424a1a3d6> (accessed on 5 August 2021).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Convergence of Iterated IPC Regression

The iterated IPC algorithm was not always able to find a satisfactory solution. We counted all simulation trials in which the algorithm did not converge after 50 iterations or aborted prematurely as a non-converged trial. In the following, we report the percentages of non-converged trials for Simulation II and IV. All attempts converged in Simulation III. Iterated IPC regression was not evaluated in Simulation I as it has not yet been implemented for linear regression models.

### Appendix A.1. Simulation II

Table A1 presents the percentages of non-converged trials for continuous covariates. The instability of the iterated IPC regression algorithm was driven by the sample size and the number of continuous covariates. Smaller samples and more continuous covariates led to a larger number of non-converged trials. However, when provided with dummy variables, the iterated IPC algorithm was much more stable, and there were only four instances of non-convergence for smaller samples with 250 individuals and three covariates.

**Table A1.** Non-Convergence of the Iterated IPC Regression Algorithm.

N	Number of Covariates	NC <sup>a</sup>
250	1	3.10
500	1	0.20
1000	1	0.00
250	2	9.90
500	2	0.60
1000	2	0.00
250	3	24.40
500	3	3.20
1000	3	0.00

<sup>a</sup> Percentages of the non-converged simulation trials.

### Appendix A.2. Simulation IV

Table A2 summarizes the percentages of trials in which the iterated IPC regression algorithm did not converge, separated for different sample sizes. Convergence problems were mainly limited to smaller samples with 250 individuals, whereas iterated IPC regression found a solution almost every time in larger samples.



**Table A2.** Non-Convergence of the Iterated IPC Regression Algorithm.

N	NC <sup>a</sup>
250	4.2
500	0.3
1000	0

<sup>a</sup> Percentages of the non-converged simulation trials.

## References

- Lindenberger, U. Human cognitive aging: Corriger la fortune? *Science* **2014**, *346*, 572–578. [CrossRef]
- Kuehner, C. Why is depression more common among women than among men? *Lancet Psychiatry* **2017**, *4*, 146–158. [CrossRef]
- Jedidi, K.; Jagpal, H.S.; DeSarbo, W.S. Finite-Mixture Structural Equation Models for Response-Based Segmentation and Unobserved Heterogeneity. *Mark. Sci.* **1997**, *16*, 39–59. [CrossRef]
- Becker, J.M.; Rai, A.; Ringle, C.M.; Völckner, F. Discovering Unobserved Heterogeneity in Structural Equation Models to Avert Validity Threats. *MIS Q.* **2013**, *37*, 665–694. [CrossRef]
- R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing, Vienna, Austria, 2020.
- Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **2015**, *67*. [CrossRef]
- Bollen, K.A. *Structural Equations with Latent Variables*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1989; [CrossRef]
- Kline, R.B. *Principles and Practice of Structural Equation Modeling*, 4th ed.; Methodology in the Social Sciences; The Guilford Press: New York, NY, USA; London, UK, 2016.
- Sörbom, D. A general method for studying differences in factor means and factor structure between groups. *Br. J. Math. Stat. Psychol.* **1974**, *27*, 229–239. [CrossRef]
- Marsh, H.W.; Wen, Z.; Nagengast, B.; Hau, K.-T. Structural equation models of latent interaction. In *Handbook of Structural Equation Modeling*; Hoyle, R.H., Ed.; Guilford Press: New York, NY, USA, 2012; pp. 436–458.
- Oberski, D.L. A Flexible Method to Explain Differences in Structural Equation Model Parameters over Subgroups. Available online: <http://daob.nl/wp-content/uploads/2013/06/SEM-IPC-manuscript-new.pdf> (accessed on 5 August 2021)
- Arnold, M.; Oberski, D.L.; Brandmaier, A.M.; Voelkle, M.C. Identifying Heterogeneity in Dynamic Panel Models with Individual Parameter Contribution Regression. *Struct. Equ. Model. Multidiscip. J.* **2020**, *27*, 613–628. [CrossRef]
- Sörbom, D. Model modification. *Psychometrika* **1989**, *54*, 371–384. [CrossRef]
- Saris, W.E.; Satorra, A.; Sorbom, D. The Detection and Correction of Specification Errors in Structural Equation Models. *Sociol. Methodol.* **1987**, *17*, 105. [CrossRef]
- Saris, W.E.; Satorra, A.; van der Veld, W.M. Testing Structural Equation Models or Detection of Misspecifications? *Struct. Equ. Model. Multidiscip. J.* **2009**, *16*, 561–582. [CrossRef]
- Hjort, N.L.; Koning, A. Tests For Constancy Of Model Parameters Over Time. *J. Nonparametric Stat.* **2002**, *14*, 113–132. [CrossRef]
- Zeileis, A.; Hornik, K. Generalized M-fluctuation tests for parameter instability. *Stat. Neerl.* **2007**, *61*, 488–508. [CrossRef]
- Andrews, D.W.K. Tests for Parameter Instability and Structural Change With Unknown Change Point. *Econometrica* **1993**, *61*, 821. [CrossRef]
- Merkle, E.C.; Zeileis, A. Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika* **2013**, *78*, 59–82. [CrossRef]
- Merkle, E.C.; Fan, J.; Zeileis, A. Testing for measurement invariance with respect to an ordinal variable. *Psychometrika* **2014**, *79*, 569–584. [CrossRef]
- Rosseel, Y. Lavaan: An R Package for Structural Equation Modeling. *J. Stat. Softw.* **2012**, *48*, 1–36. [CrossRef]
- Stefanski, L.A.; Boos, D.D. The Calculus of M-Estimation. *Am. Stat.* **2002**, *56*, 29–38. [CrossRef]
- Zeileis, A.; Köll, S.; Graham, N. Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. *J. Stat. Softw.* **2020**, *95*, 1–36. [CrossRef]
- Neale, M.C.; Hunter, M.D.; Pritikin, J.N.; Zahery, M.; Brick, T.R.; Kirkpatrick, R.M.; Estabrook, R.; Bates, T.C.; Maes, H.H.; Boker, S.M. OpenMx 2.0: Extended Structural Equation and Statistical Modeling. *Psychometrika* **2016**, *81*, 535–549. [CrossRef] [PubMed]
- Wang, T.; Strobl, C.; Zeileis, A.; Merkle, E.C. Score-Based Tests of Differential Item Functioning via Pairwise Maximum Likelihood Estimation. *Psychometrika* **2018**, *83*, 132–155. [CrossRef]
- Wickham, H.; Hester, J.; Chang, W. Devtools: Tools to Make Developing R Packages Easier. [Computer software manual]. (R package version 2.3.2). 2020. Available online: <https://cran.r-project.org/web/packages/devtools/> (accessed on 6 August 2021).
- van Buuren, S. *Flexible Imputation of Missing Data*, 2nd ed.; CRC Press: Boca Raton, FL, USA; London, UK; New York, NY, USA, 2018.
- Meinfielder, F. Multiple Imputation: An attempt to retell the evolutionary process. *ASTA Wirtsch. Sozialstat. Arch.* **2014**, *8*, 249–267. [CrossRef]
- Cohen, J.; Cohen, P.; West, S.G.; Aiken, L.S. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed.; Routledge Taylor & Francis Group: New York, NY, USA; London, UK; Mahwah, NJ, USA, 2003.



30. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [[CrossRef](#)]
31. Tibshirani, R. Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B* **2011**, *73*, 273–282. [[CrossRef](#)]
32. Shmueli, G. To Explain or to Predict? *Stat. Sci.* **2010**, *25*, 289–310. [[CrossRef](#)]
33. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1. [[CrossRef](#)]
34. Boynton, P.M.; Wood, G.W.; Greenhalgh, T. Reaching beyond the white middle classes. *BMJ* **2004**, *328*, 1433–1436. [[CrossRef](#)]
35. Marsh, H.W.; Wen, Z.; Hau, K.T. Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychol. Methods* **2004**, *9*, 275–300. [[CrossRef](#)]
36. Lin, G.C.; Wen, Z.; Marsh, H.; Lin, H.S. Structural Equation Models of Latent Interactions: Clarification of Orthogonalizing and Double-Mean-Centering Strategies. *Struct. Equ. Model. Multidiscip. J.* **2010**, *17*, 374–391. [[CrossRef](#)]
37. MacCallum, R.C.; Roznowski, M.; Necowitz, L.B. Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychol. Bull.* **1992**, *111*, 490–504. [[CrossRef](#)]
38. Yarkoni, T.; Westfall, J. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspect. Psychol. Sci. J. Assoc. Psychol. Sci.* **2017**, *12*, 1100–1122. [[CrossRef](#)]
39. Raudenbush, S.W.; Bryk, A.S. Hierarchical linear models: Applications and data analysis methods. In *Advanced Quantitative Techniques in the Social Sciences*, 2nd ed.; Sage Publ: Thousand Oaks, CA, USA, 2002; Volume 1.
40. Singer, J.D.; Willett, J.B. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*; Oxford University Press: Oxford, UK, 2003.
41. Zeileis, A.; Hothorn, T.; Hornik, K. Model-Based Recursive Partitioning. *J. Comput. Graph. Stat.* **2008**, *17*, 492–514. [[CrossRef](#)]
42. Strobl, C.; Malley, J.; Tutz, G. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* **2009**, *14*, 323–348. [[CrossRef](#)]
43. Brandmaier, A.M.; von Oertzen, T.; McArdle, J.J.; Lindenberger, U. Structural equation model trees. *Psychol. Methods* **2013**, *18*, 71–86. [[CrossRef](#)]
44. Brandmaier, A.M.; Prindle, J.J.; McArdle, J.J.; Lindenberger, U. Theory-guided exploration with structural equation model forests. *Psychol. Methods* **2016**, *21*, 566–582. [[CrossRef](#)] [[PubMed](#)]
45. Arnold, M.; Voelkle, M.C.; Brandmaier, A.M. Score-Guided Structural Equation Model Trees. *Front. Psychol.* **2021**, *11*, 3913. [[CrossRef](#)] [[PubMed](#)]
46. Lubke, G.H.; Muthén, B. Investigating population heterogeneity with factor mixture models. *Psychol. Methods* **2005**, *10*, 21–39. [[CrossRef](#)]
47. Muthén, B.; Shedden, K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **1999**, *55*, 463–469. [[CrossRef](#)]
48. McLachlan, G.J.; Peel, D. *Finite Mixture Models*; Wiley Series in Probability and Statistics Applied Probability and Statistics Section; Wiley: New York, NY, USA, 2000.



# Score-Guided Structural Equation Model Trees

Manuel Arnold<sup>1,2\*</sup>, Manuel C. Voelkle<sup>1</sup> and Andreas M. Brandmaier<sup>2,3</sup>

<sup>1</sup> Psychological Research Methods, Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany, <sup>2</sup> Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany, <sup>3</sup> Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany

Structural equation model (SEM) trees are data-driven tools for finding variables that predict group differences in SEM parameters. SEM trees build upon the decision tree paradigm by growing tree structures that divide a data set recursively into homogeneous subsets. In past research, SEM trees have been estimated predominantly with the R package *semtree*. The original algorithm in the *semtree* package selects split variables among covariates by calculating a likelihood ratio for each possible split of each covariate. Obtaining these likelihood ratios is computationally demanding. As a remedy, we propose to guide the construction of SEM trees by a family of score-based tests that have recently been popularized in psychometrics (Merkle and Zeileis, 2013; Merkle et al., 2014). These score-based tests monitor fluctuations in case-wise derivatives of the likelihood function to detect parameter differences between groups. Compared to the likelihood-ratio approach, score-based tests are computationally efficient because they do not require refitting the model for every possible split. In this paper, we introduce score-guided SEM trees, implement them in *semtree*, and evaluate their performance by means of a Monte Carlo simulation.

**Keywords:** exploratory data analysis, heterogeneity, model-based recursive partitioning, parameter stability, structural change tests, structural equation modeling

## OPEN ACCESS

### Edited by:

Jin Eun Yoo,  
Korea National University  
of Education, South Korea

### Reviewed by:

Carolin Strobl,  
University of Zurich, Switzerland  
Achim Zeileis,  
University of Innsbruck, Austria

### \*Correspondence:

Manuel Arnold  
arnoldmz@hu-berlin.de

### Specialty section:

This article was submitted to  
Quantitative Psychology  
and Measurement,  
a section of the journal  
Frontiers in Psychology

**Received:** 21 May 2020

**Accepted:** 15 December 2020

**Published:** 28 January 2021

### Citation:

Arnold M, Voelkle MC and  
Brandmaier AM (2021) Score-Guided  
Structural Equation Model Trees.  
Front. Psychol. 11:564403.  
doi: 10.3389/fpsyg.2020.564403

## INTRODUCTION

Structural equation models (SEMs; Bollen, 1989; Kline, 2016) are a widely applied technique in social and psychological research to model the relationships between multiple variables. SEMs are especially useful when some of the variables under investigation are latent (not directly observable) or contain measurement errors. Various statistical procedures such as factor analysis, ANOVA, linear regression, mediation models, growth curve models, and dynamic panel models can be specified within the SEM framework.

A major challenge that complicates the specification and interpretation of SEMs are potential differences between subgroups of the sample. Group differences can pertain to various aspects of a SEM. For instance, in a latent growth curve model, we may find differences in how people change over time, or in a factor analysis model, the factor structure may vary across groups. By neglecting such instances of sample heterogeneity, SEM parameter estimates may not represent any individual in the sample, and researchers risk drawing incorrect conclusions from their data (e.g., Kievit et al., 2013). This makes identifying group differences in SEM parameters an important task.

One popular strategy is to detect heterogeneity in SEMs with the help of covariates. Multi-group structural equation models (MGSEMs; Sörbom, 1974) allow estimating different parameter values for the levels of a grouping variable, such as males and females or treated versus non-treated.

By comparing the fit of a single-group SEM to the fit of a MGSEM, equality constraints on parameters across groups can be tested with the likelihood-ratio test. Multi-group structural equation modeling excels as a confirmatory tool to test a limited number of hypotheses about group differences. As part of exploratory data analysis, however, the method can often become tedious in large data sets. With many potentially important grouping variables, many MGSEMs need to be specified and estimated. Moreover, since MGSEMs require discrete grouping variables, numeric and ordinal covariates such as age or socioeconomic status need to be discretized, which often leads to a loss of information (but see Hildebrandt et al., 2016).

SEM trees, as first presented by Brandmaier et al. (2013b), can be seen as an extension of MGSEMs for exploring parameter heterogeneity in SEMs. SEM trees are a data-driven approach that automatically searches through all available covariates to identify partitions of the full sample that differ with respect to SEM parameter estimates. SEM trees build upon the model-based recursive partitioning paradigm (for an overview, see Zeileis et al., 2008; Strobl et al., 2009). One key feature of SEM trees is their interpretability: SEM trees provide a graphical representation of how covariates and covariate interactions predict non-linear differences in SEM parameters. The building blocks of SEM trees are called nodes, each containing a SEM fitted to a distinct subsample. The SEM tree algorithm forms a binary tree structure by hierarchically splitting these nodes. Each node of the SEM tree has either two successors (daughter nodes) and is called an *inner node* or no successors and is called a *leaf* (or terminal node). The first node of the tree is called the *root* and has no parent nodes. The inner nodes of the tree represent split decisions. Each split decision involves a covariate (e.g., age of the observed individuals) and a cut point in the covariate (e.g., divide the sample into individuals younger and older than 45 years). A leaf of a tree contains a partition of the sample that is best described with a set of SEM parameters. All leaves taken together exhaustively partition the original sample and can be thought of as a MGSEM with potentially many groups. An important difference to conventional multi-group structural equation modeling is that the group membership in a SEM tree is not pre-specified but learned from the data.

There are currently two software packages for the statistical programming language R that allow fitting SEM trees. One is the `semtree` package (Brandmaier et al., 2013b) that has been widely applied in the literature (Brandmaier et al., 2013a, 2016, 2017, 2018; Jacobucci et al., 2017; Usami et al., 2017, 2019; de Mooij et al., 2018; Ammerman et al., 2019; Serang et al., 2020; Simpson-Kent et al., 2020). The other software implementation is the `partykit` package (Hothorn and Zeileis, 2015). Unlike `semtree`, `partykit` is not limited to a specific model class such as SEMs but provides the infrastructure for general recursive partitioning across various model classes. Among other features, `partykit` provides the generic MOB algorithm for model-based recursive partitioning that has been used to study heterogeneity in M-estimators (Zeileis et al., 2008), Bradley-Terry models (Strobl et al., 2011), Rasch models (Strobl et al., 2015; Komboz et al., 2018), multinomial processing trees (Wickelmaier and Zeileis, 2018),

generalized linear mixed-effects models (Fokkema et al., 2018), network models (Jones et al., 2020), and circular regression models (Lang et al., 2020). Moreover, MOB is also used in more specialized recursive partitioning packages such as `psychotree` (Zeileis et al., 2020). Recently, Zeileis (2020) demonstrated on his blog how MOB can be coupled with the SEM software `lavaan` (Rosseel, 2012) to estimate SEM trees. Outside of the R ecosystem, SEM trees have also been fitted in `Mplus` (Serang et al., 2020).

SEM trees are estimated by recursively selecting the covariate that best partitions the sample into different subgroups. Thus, the evaluation of potential splits is the central aspect of the algorithm. The `semtree` package uses a procedure that transforms all non-dichotomous covariates (that is, covariates with more than two values) into a set of dichotomous split candidates. Then, the tree growing algorithm computes the likelihood ratio between a single SEM (fitted on the complete sample of the current node) and MGSEMs (representing the model after the split) for every split candidate and selects the candidate associated with the largest likelihood ratio. The number of MGSEMs needed to calculate these likelihood ratios is directly related to the number of possible splits of the covariate. For instance, evaluating a numeric covariate such as age with many different values will require more MGSEMs to be estimated than evaluating a discrete covariate such as handedness. The reliance of the `semtree` package on likelihood ratios has the apparent drawback that the computational burden becomes large to excessive if there are many covariates and the covariates have many unique values. Another problem of the current `semtree` package is that the standard approach to split evaluation (called *naïve* selection approach in `semtree`) is biased by favoring the selection of covariates with many unique values over covariates with few unique values (Brandmaier et al., 2013b). The `semtree` package offers a correction procedure (*fair* selection approach) for this selection bias (also known as attribute selection error; Jensen and Cohen, 2000). However, this correction procedure is heuristic and comes at the price of decreased statistical power to detect group differences. To solve this problem, we suggest to use a well-known method for likelihood-ratio-guided covariate selection that does not suffer from a selection bias while retaining full statistical power. We implemented this method into the `semtree` package.

In contrast to the `semtree` package, model-based recursive partitioning in the `partykit` package uses so-called score-based or structural change tests (e.g., Zeileis and Hornik, 2007) for assessing whether the values of one or more parameters depend on a covariate. Score-based tests are obtained by cumulating the case-wise gradients of the log-likelihood function evaluated at the parameter estimates. Unlike the likelihood-ratio test, score-based tests do not require the estimation of group-specific models for the evaluation of each split. This property leads to two advantages that make score-based tests highly attractive for model-based recursive partitioning. First, they are computationally efficient, as only the pre-split model needs to be estimated once. Second, when subgroups become small, fitting multi-group models to obtain likelihood ratios

may become unstable. We propose using the advantages of score-based tests and added SEM trees guided by score-based tests to the `semtree` package. Our implementation of score-guided trees differs in some points from the generic MOB algorithm from the `partykit` package. MOB uses score-based tests to select a covariate, and it locates the optimal cut point in this covariate by comparing likelihood ratios. In contrast, our `semtree` implementation uses a score-based cut point localization, which is computationally more efficient. Moreover, MOB is currently limited to a single score-based test statistic, whereas `semtree` offers a broader selection of different test statistics that recently became popular in the exploration of measurement invariance in SEMs (Merkle and Zeileis, 2013; Merkle et al., 2014; Wang et al., 2014, 2018).

The present study assesses a wide range of variable selection techniques in a Monte Carlo simulation study using the `semtree` package. We implemented an optimal likelihood-ratio-based method to improve the statistical properties of the likelihood-ratio-based split selection in `semtree` and added a family of score-based tests as a computationally efficient alternative. We evaluated the performance of these new methods next to the classical *naïve* and *fair* methods. Moreover, we explored two techniques offered by `semtree` that allow testing specific hypotheses and incorporating *a priori* knowledge about group differences. The remainder of this manuscript is organized as follows: first, we reiterate the basic principles of SEM trees. Second, the existing likelihood-ratio-based implementation is outlined in detail and complemented with an unbiased method for selecting covariates. Third, we recapitulate a family of score-based tests and show how they can be used to guide the split decision of SEM trees. Fourth, the simulation setup and results are shown. The study concludes with a discussion of the simulation results and recommendations for future research.

## INTRODUCTORY EXAMPLE

In the following, we illustrate the rationale behind SEM trees with an instructive example. Readers familiar with SEM trees may skip this section.

Let us assume a researcher estimated a confirmatory factor analysis (CFA; Brown, 2015) model that explains the scores of three ability tests of 600 male and female test takers of different ages with a single common latent factor and test-specific error terms. The data were collected at two different testing facilities. The researcher wonders if the parameter values of her CFA model differ with respect to the sites, the test takers' age, and gender. She investigates this question with the help of a SEM tree.

The data for this fictional example were simulated such that the factor loading of the first ability test for individuals older than 45 years was smaller (0.6) than for younger individuals (0.8). This represents a violation of measurement invariance; that is, differences among individuals' responses to an item are not only due to differences in the latent factor but also due to the item functioning differently across groups and being measured with different precision. Further, we lowered all factor loadings

of older individuals tested at the second site by 0.1, imposing another form of violation of metric invariance. The covariate gender had no impact on the parameters of the CFA model and served as a noise variable.

**Figure 1** shows the resulting SEM tree for the simulated data set. The SEM tree consists of 5 nodes depicted as ovals, each of them containing a CFA model. Node 1 is the root node of the SEM tree and contains the CFA model fitted on the full data set with  $N = 600$  individuals. In this illustrative example, the SEM tree algorithm concluded that the fit of the model in the root node could be improved most by splitting the data into a group of 300 individuals younger than 45 years (Node 2) and a group of 300 individuals older than 45 years (Node 3). Node 2 and 3 are said to be the daughters of Node 1. After splitting the sample associated with Node 1, the algorithm proceeds recursively with Node 2 and 3. Whereas the fit of the model for younger individuals (Node 2) could not be improved any further, the SEM tree algorithm split the group of older individuals (Node 3) into two subgroups with 150 older individuals tested at site 1 (Node 4) and 150 older individuals tested at site 2 (Node 5). After this split, the SEM tree algorithm terminated as no further split would significantly improve any of the submodels' fit. Nodes 2, 4, and 5 are the leaves of the SEM tree, and individuals within these nodes were found to be homogeneous with respect to the covariates. As expected, the SEM tree algorithm did not select the covariate gender for splitting because this covariate was not associated with any group differences in the simulated data set.

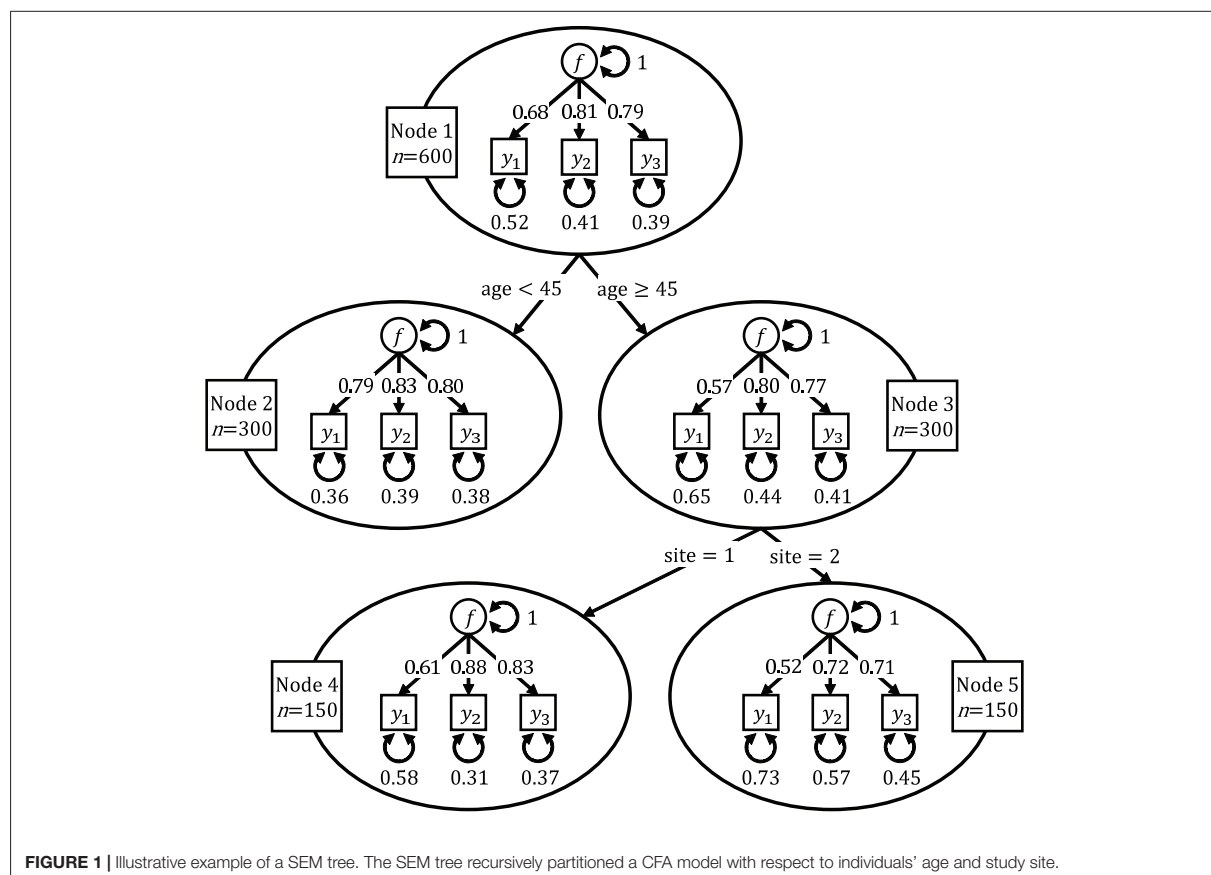
It is important to note that the structure of the SEM tree shown in **Figure 1** is not specified *a priori* but learned top-down in an exploratory way. The algorithm only requires a pre-specified template SEM (in the example, the CFA model) and a data set including covariates that serve as split candidates to identify homogeneous groups. The selection of covariates and the identification of optimal cut points are then learned from the data. Throughout the tree, the structure of the template SEM remains the same, and only the values of the parameter estimates change as the model is fitted recursively on different subsamples.

## STRUCTURAL EQUATION MODEL TREES

The general SEM tree algorithm can be described in four steps:

1. Specify a template SEM.
2. Fit the template SEM to all observations in the current node.
3. Assess whether the SEM parameter estimates are constant or vary with respect to the covariate.
4. Choose the covariate that is associated with the largest group differences. If the group difference exceeds a threshold, split the node into two daughter nodes, and repeat the procedure with Step 2 for both daughter nodes. Otherwise terminate.

Likelihood-ratio-guided and score-guided SEM trees differ in how Step 3 of the general SEM tree algorithm is implemented. In other words, the procedures use different approaches to



**FIGURE 1** | Illustrative example of a SEM tree. The SEM tree recursively partitioned a CFA model with respect to individuals' age and study site.

evaluate heterogeneity and to search for optimal split points in covariates. The following section outlines Steps 1–4 for likelihood-ratio-guided SEM trees before introducing score-guided SEM trees afterward.

### Step 1: Specification of the Template Model

The starting point for growing a SEM tree is the specification of a template SEM. A template model reflects hypotheses about the data by specifying relations among observed variables and latent constructs and is determined by the research question. The template model is fitted on all subsamples associated with the nodes of the SEM tree. It is important to note that the structure of the template model stays the same in the entire SEM tree (but see Brandmaier et al., 2013a for trees with multiple models). Hence, parameters fixed to a constant (e.g., zero or one) in the template model are fixed to the same constant in all submodels of the tree. Only parameters freely estimated in the template model are allowed to differ across groups and contribute to the assessments of splits.

Fixing many parameters of the template model to constants can hinder the SEM tree algorithm from identifying group differences. Usually, some parameters are fixed to ensure the

identification of the SEM. In some model classes, additional constraints are specified to model specific relationships or trajectories. For instance, in latent growth curve models (see McArdle, 2012 for an overview), the factor loadings of a latent random slope variable are often fixed to model a specific growth pattern such as linear or quadratic growth. By fixing these loadings, a SEM tree will not be able to estimate different growth patterns between groups and, as a result, may overlook heterogeneity. In this case, estimating the factor loadings as free parameters may improve the SEM tree's flexibility to adapt to subgroup-specific trajectories.

By default, SEM trees estimate all non-fixed parameters freely in each submodel, and every parameter contributes to the evaluation of split candidates. This behavior is suboptimal if there is a clear set of target parameters that are of interest to investigate a given theory. As a solution, the `semtree` package offers the option to specify a set of so-called focus parameters. By declaring focus parameters, the SEM tree will only consider heterogeneity in these parameters when assessing split candidates. Thus, focus parameters are useful for testing parameter-specific hypotheses about group differences. For instance, if one wants to test measurement invariance, one could specify the measurement model's parameters as focus parameters



and disregard heterogeneity in the structural model. Besides focus parameters, the `semtree` package allows constraining specific parameters to be equal across all submodels of the tree. This is done by estimating these parameters in the full sample once and using the resulting values throughout the tree. Such equality constraints allow incorporating prior knowledge about homogeneous parameters and can increase the power to detect heterogeneity in the remaining parameters. We will later demonstrate the use of focus parameters and equality constraints in two short simulation studies.

### Step 2: Model Estimation

Various estimation techniques for SEMs have been discussed. In principle, SEM trees can operate with any estimation method that provides a fit statistic and are not necessarily limited to a multivariate normal distribution. At present, however, only maximum likelihood estimation for multivariate normal data is implemented in the `semtree` package. Therefore, `semtree` is currently less suited for investigating models fitted on non-normal data such as SEMs with categorical outcomes. In the following, we will focus on maximum likelihood estimation for multivariate normal data.

SEMs are usually specified by expressing the structure of a mean vector and a covariance matrix as a function of a  $q$ -variate vector  $\theta$  with model parameters. These parameters are estimated by minimizing a fitting function  $F$  that measures the discrepancy between the observed means  $\bar{y}$  and the model-implied means  $\mu(\theta)$  as well as the discrepancy between the observed covariance matrix  $S$  and the model-implied covariance matrix  $\Sigma(\theta)$ . Several fitting functions have been proposed. The following maximum likelihood fitting function is widely used as it yields efficient parameter estimates under the assumption of multivariate normally distributed data:

$$F_{ML}[\bar{y}, S, \mu(\theta), \Sigma(\theta)] = [\bar{y} - \mu(\theta)]^T \Sigma(\theta)^{-1} [\bar{y} - \mu(\theta)] + \text{tr}[S\Sigma(\theta)^{-1}] - \ln \{ \det [S\Sigma(\theta)^{-1}] \} - p \tag{1}$$

In the equation above,  $p$  denotes the number of observed variables in the SEM. A fitting function also provides a test of overall model fit. Evaluated at the parameter estimates  $\hat{\theta}$ ,  $(N-1)F$  asymptotically follows a  $\chi^2$  distribution with  $q$  degrees of freedom under the null hypothesis of a correctly specified model, where  $N$  refers to the sample size. A detailed account of SEM estimation can be found in the textbooks by Bollen (1989) and Kline (2016).

### Step 3: Split Evaluation

The original SEM tree algorithm suggested by Brandmaier et al. (2013b) compares the fit of a single-group model to the fit of a MGSEM, which consists of all submodels in the current leaves, to decide whether to split a node according to a covariate. For the sake of simplicity, we assume that all covariates are dichotomous and discuss non-dichotomous covariates afterward.

Let  $M_F$  represent the model associated with the root node (that contains the full data set) and let  $\hat{\theta}_F$  denote the corresponding parameter estimates. Further, we mark the observed mean vector

of the full data set as  $\bar{y}_F$  and the observed covariance matrix as  $S_F$ . To evaluate a candidate covariate for a specific node, we split the node into two daughter nodes according to the covariate. Then, group-specific SEM parameters  $\theta_j, j = 1, \dots, J$ , are estimated for all subsamples associated with the  $J$  current leaf nodes. Since the subsamples associated with the current leaves are non-overlapping, the submodels can be joined into a MGSEM, which we from now on refer to as  $M_{SUB}$ . As  $M_F$  is nested within  $M_{SUB}$ , we can test the following null hypothesis of parameter homogeneity with respect to the covariate under evaluation:

$$H_0 : \theta_j = \theta_0, \quad \forall j = 1, \dots, J \tag{2}$$

Rejecting Equation 2 implies that the model parameters vary with respect to the covariate. Brandmaier et al. (2013b) suggested using the following log-likelihood ratio between  $M_F$  and  $M_{SUB}$  as a test statistic for Equation 2:

$$LR = (N - 1) \left\{ F_{ML}[\bar{y}_F, S_F, \mu(\hat{\theta}_F), \Sigma(\hat{\theta}_F)] - \sum_{j=1}^J \frac{n_j}{N} F_{ML}[\bar{y}_j, S_j, \mu(\hat{\theta}_j), \Sigma(\hat{\theta}_j)] \right\} \tag{3}$$

Under the null hypothesis that there is no influence of the covariate under scrutiny,  $LR$  asymptotically follows a  $\chi^2$  distribution with  $(J-1)q$  degrees of freedom.

This testing procedure provides a powerful and efficient solution for dichotomous covariates. However, evaluating a categorical, an ordinal, or a continuous covariate that has more than two unique values requires an additional step of locating the optimal cut point. Brandmaier et al. (2013b) suggested to compute the likelihood ratio in Equation 3 for every meaningful partition of the covariate and then to select the cut point associated with the maximum likelihood ratio. For categorical covariates, the best partition is found by splitting them into a set of dichotomous variables applying a one-against-the-rest scheme for all possible combinations of categories. For ordinal and continuous covariates, the ordering inherent to these covariates allows applying a procedure known as exhaustive split search (Quinlan, 1993) to find the optimal cut point. Given a covariate with  $m$  unique values, this procedure tests  $m-1$  potential partitions to locate the maximum of the likelihood ratios. For continuous covariates, it is also necessary to omit a certain fraction of the data associated with the smallest and largest values of the covariate in order to obtain a sufficiently large sample to estimate the SEMs in both partitions. From the above, it is clear that the computational demand of SEM trees grows with the number of covariates with many unique values as every potential cut point requires the estimation of SEMs.

Locating the optimal cut point in categorical, ordinal, and continuous covariates with the maximum of the likelihood ratios has important implications for the test statistic shown in Equation 3. By choosing the maximum of a set of statistics (one for each possible partition), the resulting distribution is no longer the same as the distribution of the individual statistics. Thus, a maximally selected likelihood-ratio test statistic does not

follow a  $\chi^2$  distribution under the null hypothesis of parameter homogeneity. The deviation from the  $\chi^2$  distribution is directly related to the number of potential cut points. With a growing number of possible cut points, the maximum of the likelihood-ratio values will be increased purely by random fluctuations. Consequently, using the  $\chi^2$  test for the evaluation of covariates will artificially inflate the probability of type I errors and favors the selection of covariates with many potential cut points over the selection of covariates with few.

Brandmaier et al. (2013b) discuss different correction procedures for this selection bias that are used in the `semtree` package. The default method, labeled *naïve* in `semtree`, uses the  $\chi^2$  distribution for evaluating covariates and simply ignores the resulting selection bias. To reduce this bias, `semtree` offers the option to use the *naïve* method in combination with a Bonferroni correction for multiple testing within the same covariate by dividing the  $p$ -value obtained from the likelihood-ratio test in Equation 3 by the number of potential cut points. However, this Bonferroni adjustment can lead to overcorrection and decreases the probability of selecting covariates with many possible cut points, as demonstrated by Brandmaier et al. (2013b). We will refer to this Bonferroni adjusted *naïve* method simply as the *naïve* method from now on. Besides the Bonferroni correction, different cross-validation methods are implemented in the `semtree` package. Cross-validation separates the estimation of SEMs from the testing of a potential cut point (e.g., Jensen and Cohen, 2000). SEM trees can be grown with a two-stage approach (Loh and Shih, 1997; Shih, 2004; Brandmaier et al., 2013b) that splits the sample associated with a node in half. One half of the sample is used to find the optimal cut point for every covariate. The other half is used to evaluate only the best cut points via the likelihood-ratio test. This method is called *fair* in the `semtree` package. Since the *fair* method uses only half of the sample for split selection, its power for detecting heterogeneity can be expected to be considerably lower than the power of methods that employ the whole sample. A much simpler and more elegant way of avoiding the selection bias and correction procedures altogether is to use the correct distribution of the maximally selected likelihood-ratio test statistic (*maxLR*). Andrews (1993) showed that the asymptotic distribution of *maxLR* is the supremum of a certain tied-down Bessel process from which  $p$ -values can be obtained (see Zeileis et al., 2008; Merkle and Zeileis, 2013). We now implemented the *maxLR* statistic into the `semtree` package to provide a more efficient and robust likelihood-ratio-based covariate selection.

#### Step 4: Covariate Selection

To select a single covariate from a set of candidate covariates, the likelihood ratio for the optimal cut point is computed for every covariate, and the covariate associated with the smallest  $p$ -value is chosen. If the  $p$ -value is smaller than a pre-specified threshold, determined by the desired probability of a type I error, splitting is continued recursively. One should keep in mind that testing several covariates will artificially inflate the type I error probability. One of several solutions to this problem is the use of Bonferroni adjusted  $p$ -values. Given a large number of covariates,

however, the Bonferroni correction will reduce the power of the SEM tree drastically and will produce sparse trees. In such cases, one may resort to unadjusted  $p$ -values for the selection of covariates and, if needed, can limit the size of the SEM tree with additional stopping criteria like a minimum number of individuals per node.

## SCORE-GUIDED SEM TREES

Using likelihood-ratio tests to grow SEM trees can become computationally burdensome if not infeasible as the evaluation of a covariate requires the estimation of MGSEMs for every potential cut point. Furthermore, when subgroups become small, fitting MGSEMs may become unstable. Alternatively, SEM trees can be guided by score-based tests that do not require the estimation of MGSEMs to evaluate a split at all. This makes score-based tests computationally efficient and often more stable as compared to likelihood-ratio tests. In the following, we will first introduce the general notion behind score-based tests and then introduce a family of score-based test statistics for covariates with different levels of measurement.

### Score-Based Tests

Score-based tests originated in econometrics, where they are primarily employed to detect parameter instability in time series models (e.g., Hansen, 1992; Andrews, 1993). Score-based tests can be summarized in three steps: first, the case-wise derivatives of the log-likelihood function with respect to the model parameters are computed. These case-wise derivatives, also called scores, indicate how well the model parameters represent an individual. The larger the score, the larger the misfit of a given model parameter for a given individual. Second, the scores are sorted with respect to a covariate for which we want to test parameter homogeneity. Third, the scores are aggregated into a test statistic that allows testing of the null hypothesis of homogeneous parameters (see Equation 2).

Score-based tests have been derived for general M-estimators that encompass popular estimation techniques such as least-squares methods and maximum likelihood as special cases (Zeileis and Hornik, 2007). For the sake of simplicity, we limit ourselves to maximum likelihood estimation for multivariate normally distributed data. The associated log-likelihood function for a single individual  $i$  is given by

$$\ln L(\theta; y_i) = \frac{1}{2} \left\{ [y_i - \mu(\theta)]^T \Sigma(\theta)^{-1} [y_i - \mu(\theta)] + \ln [\det(\Sigma(\theta))] + p \ln(2\pi) \right\}. \quad (4)$$

Equation 4 is the normal theory log-likelihood function for a single individual  $i$  and yields identical parameter estimates to  $F_{ML}$  shown in Equation 1 if summed over individuals and maximized.

The individual scores are calculated by taking the partial derivative of the log-likelihood function with respect to the

parameters and evaluating the expression at the estimates:

$$S(\hat{\theta}; \mathbf{y}_i) = \left[ \frac{\partial \ln L(\theta; \mathbf{y}_i)}{\partial \theta_1} \Big|_{\theta=\hat{\theta}} \cdots \frac{\partial \ln L(\theta; \mathbf{y}_i)}{\partial \theta_q} \Big|_{\theta=\hat{\theta}} \right]^T \quad (5)$$

The scores assess the extent to which an individual’s log-likelihood is maximized by one of the  $q$  parameters. Values close to zero indicate a good fit between model and individual, whereas large scores point toward a strong misfit. Note that by definition, the scores evaluated at the maximum likelihood estimates  $\hat{\theta}$  sum up to zero; that is,  $\sum_{i=1}^N S(\hat{\theta}; \mathbf{y}_i) = 0$ .

For the construction of a test statistic, the scores are cumulated according to the order induced by a covariate under scrutiny. For instance, if parameter homogeneity is assessed with respect to age, the first row consists of scores from the youngest individual. For the second row, scores of the youngest and second youngest individuals are summed up, and so forth. More formally, the cumulative score process is defined as

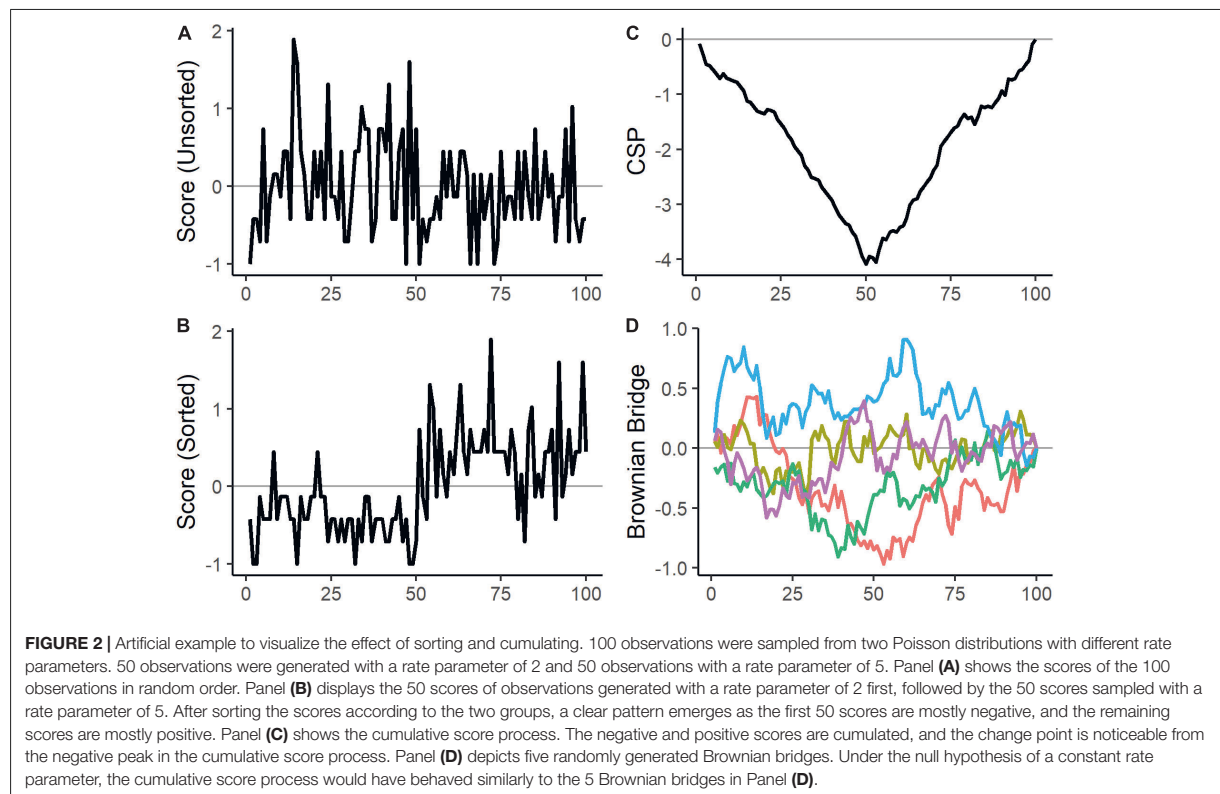
$$CSP(\hat{\theta}; s) = \frac{1}{\sqrt{N}} I(\hat{\theta})^{-1/2} \sum_{h=1}^s S(\hat{\theta}; \mathbf{y}_h), \quad (6)$$

where the index  $s$  denotes the number of sorted individuals entering the equation, and the index  $h$  selects the sorted individuals until  $h = s$ . Furthermore,  $I(\hat{\theta})^{-1/2}$  is the estimated

half-squared inverse of the Fisher information matrix. Pre-multiplying with  $I(\hat{\theta})^{-1/2}$  decorrelates the scores so that the  $q$  cumulative score processes are unrelated to each other. In the following, we place the values of the cumulative score process row-wise into an  $N \times q$  matrix that we denote with **CSP** and refer to the cumulative sum from the first  $s$ -th individuals of the  $k$ -th parameter as  $CSP_{s,k}$ . The plots in Panel (A–C) in **Figure 2** illustrate how sorting and cumulating scores make parameter heterogeneity visible.

Hjort and Koning (2002) show that under mild conditions and constant parameters, each column of the cumulative score process matrix **CSP** converges in distribution to a univariate Brownian bridge. A Brownian bridge is a stochastic process that is pinned to zero at the start and end and exhibits the most variability in the middle. Thus, the null hypothesis of parameter homogeneity in Equation 2 can be tested by comparing the observed cumulative score process to the analogous statistic of a Brownian bridge. Panel (C) and (D) in **Figure 2** illustrate the difference between the cumulative score process of a heterogeneous parameter and the Brownian bridge.

Test statistics can be obtained by aggregating the cumulative score process matrix into a single scalar. Critical values and  $p$ -values for these test statistics can be found by applying the same aggregation to the asymptotic Brownian bridge (Zeileis and Hornik, 2007). Different ways of aggregating the cumulative





scores will produce test statistics that will be sensitive to different patterns of parameter heterogeneity. The choice of a test statistic also depends on the level of measurement of the covariate.

Merkle and Zeileis (2013) proposed three different test statistics for continuous covariates:

$$DM = \max_{s=1, \dots, N} \left[ \max_{k=1, \dots, q} (|CSP_{s,k}|) \right] \quad (7)$$

$$CvM = \frac{1}{N} \sum_{s=1}^N \sum_{k=1}^q CSP_{s,k}^2 \quad (8)$$

$$\max LM = \max_{s=\underline{s}, \dots, \bar{s}} \left\{ \left[ \frac{s}{N} \left( 1 - \frac{s}{N} \right) \right]^{-1} \sum_{k=1}^q CSP_{s,k}^2 \right\} \quad (9)$$

Equations 7–9 show the double maximum (*DM*), Cramér-von-Mises (*CvM*), and maximum Lagrange multiplier (*maxLM*) test statistics. *DM* is the simplest test statistic and rejects the null hypothesis if, at any point, the maximum of any of the *q* processes strays too far away from zero. However, Merkle and Zeileis (2013) note that considering only the maximum of the *q* processes wastes power because the *DM* statistic ignores heterogeneity in other parameters. Furthermore, even for the same parameter, smaller peaks before and after the maximum are not considered, which may lead to a loss of power if the parameter changes its values across more than two groups. Using sums instead of maxima solves these problems. The *CvM* statistic sums the squared values over all parameters and individuals and is therefore well suited for detecting multiple group differences in several parameters. If one suspects that a single change point will manifest in several parameters, the *maxLM* statistic that considers the maximum values of all parameters at a single point is more appropriate. Unlike the other test statistics for continuous covariates, the *maxLM* statistic contains a scaling term  $\frac{s}{N} \left( 1 - \frac{s}{N} \right)$ , which increases sensitivity for peaks before and after the middle of the processes. A disadvantage of this scaling is that individuals with very small and very large values of the covariate need to be omitted to stabilize the test statistic. Therefore, one has to specify an interval  $[\underline{s}, \dots, \bar{s}]$  with a lower and upper threshold of the covariate. Parameter shifts outside of these boundaries are not considered. The *maxLM* statistic is asymptotically equivalent to the *maxLR* statistic from the previous section (Andrews, 1993).

For ordinal and categorical covariates, Merkle et al. (2014) suggested test statistics that focus on bins of individuals at each level of the covariates:

$$WDM = \max_{l=1, \dots, m-1} \left\{ \left[ \frac{n_l}{N} \left( 1 - \frac{n_l}{N} \right) \right]^{-1/2} \max_{k=1, \dots, q} |CBSP_{l,k}| \right\} \quad (10)$$

$$\max LM_O = \max_{l=1, \dots, m-1} \left\{ \left[ \frac{n_l}{N} \left( 1 - \frac{n_l}{N} \right) \right]^{-1} \sum_{k=1}^q CBSP_{l,k}^2 \right\} \quad (11)$$

Equations 10 and 11 present the weighted double maximum (*WDM*) and the maximum Lagrange multiplier statistics for ordinal covariates (*maxLM<sub>O</sub>*). For both test statistics, we first group the individuals into *m*–1 bins associated with the first *m*–1 levels of the covariate. Then, we sum the scores in each bin and cumulate the sums, yielding a (*m*–1) × *q* matrix *CBSP*

of cumulative bins of scores. In the equations above, we denote the cumulative bin of scores associated with the *l*-th level of the covariate and the *k*-th parameter with *CBSP<sub>l,k</sub>*. Both statistics are scaled by  $\frac{n_l}{N} \left( 1 - \frac{n_l}{N} \right)$ , where *n<sub>l</sub>* represents the cumulative number of individuals per bin. The main difference is that the *maxLM<sub>O</sub>* statistic considers heterogeneity in all parameters, whereas the *WDM* only considers the most heterogeneous parameter.

Categorical covariates do not possess a natural ordering that can be used to construct a test statistic. Alternatively, a test statistic can be obtained by summing the squared differences in the sum of scores across bins of individuals associated with a different level of the covariate (Hjort and Koning, 2002). In the following Lagrange multiplier (*LM*) statistic

$$LM = \sum_{l=1}^m \sum_{k=1}^q (BSP_{l,k} - BSP_{l-1,k})^2, \quad (12)$$

*BSP<sub>l,k</sub>* denotes the sum of the scores of the *k*-th parameter from individuals associated with the *l*-th level of the covariate. *B<sub>0,k</sub>*, *k* = 1, ..., *q*, is not associated with any of the 1, ..., *m* levels of the covariate and is set to zero.

To apply the test statistics outlined above in practice, critical values and *p*-values are needed in order to compare split points across covariates. Analytic solutions are available for the *DM*, *maxLR*, *WDM*, and *LM* statistic. For the remaining test statistics, critical values and *p*-values can be obtained through repeated simulation of Brownian bridges. Different strategies for obtaining critical values and *p*-values for the *DM*, *maxLR*, and *CvM* statistics are discussed by Merkle and Zeileis (2013) and for the *WDM*, *maxLM<sub>O</sub>*, and *LM* statistics by Merkle et al. (2014).

### SEM Trees Guided by Score-Based Tests

Score-guided SEM trees can be obtained by replacing the evaluation of covariates in Step 3 of the general SEM tree algorithm with score-based tests instead of the likelihood-ratio test. Because score-based tests operate like an omnibus test for all possible cut points in a covariate, a single best cut point needs to be located after the selection of a covariate. Cut points can be obtained by identifying which of the unique values of the covariate maximizes the respective score-based test statistic. Omitting the outer sums or maxima in the Equations 7–11 pairs every unique value of the covariate with a specific value of the partially summed test statistic. Then, a cut point can be determined by splitting the sample after the observation associated with the maximum of these partially summed test statistics. Due to its scaling term, the respective *maxLM* statistic for ordinal and continuous covariates appears to be particularly well suited for identifying the optimal cut points. We implemented this fully score-based cut point localization procedure in the *semtree* package. Alternatively, the optimal cut point can be determined by maximizing the partitioned log-likelihood (that is, the sum of the log-likelihood for all observations to the left and the sum for all observations to the right of the cut point) over all conceivable values of the covariate. Since this approach requires the estimation of a sequence of SEMs, it will be slower than a purely score-based cut point identification. However, it will still be faster than a SEM tree purely guided by likelihood ratios because only the localization

of a cut point but not the selection of the covariate requires the estimation of additional SEMs. This hybrid strategy is currently applied by the generic MOB algorithm from the `partykit` package, which uses the `maxLM` statistic for selecting covariates and the `maxLR` statistic for locating cut points.

## SIMULATION STUDY

We conducted four Monte Carlo simulations to evaluate SEM trees in different settings. The first two simulations compare the original SEM tree split selection methods with the newly proposed SEM trees guided by the `maxLR` statistic and score-based tests. The first simulation aims at illustrating the performance of the different SEM trees under the null hypothesis of parameter homogeneity. The second simulation investigates power, the precision of cut point estimation, and group recovery for a heterogeneous population consisting of two groups. The third and fourth simulations demonstrate the use and common pitfalls of SEM trees with focus parameters and equality constraints.

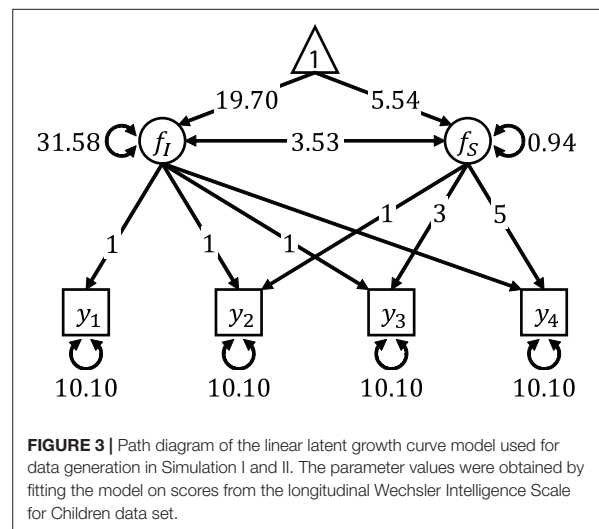
All simulations were carried out with the statistical programming language R. SEM trees were fitted with the `semtree` package. `semtree` interfaces the `OpenMx` package (Neale et al., 2016) for the estimation of SEMs. To grow score-guided SEM trees, we linked `semtree` to the `strucchange` package (Zeileis et al., 2002). `strucchange` offers a unified framework for implementing score-based tests for a wide range of models. All features used in this simulation are available in the `semtree` package. Our simulations were performed using R 4.0.2, `OpenMx` 2.18.1, `strucchange` 1.5-2, and a developmental snapshot of the `semtree` package<sup>1</sup>. The simulation scripts and results are provided as Online Supplemental Material<sup>2</sup>.

In all simulations, we aimed at ensuring an optimal type I error rate; that is, we tried limiting the proportions of false-positive splits to the significance level of 5%. To achieve this, we adjusted the  $p$ -values of the likelihood-ratio and score-based tests with the Bonferroni procedure to correct for the multiple testing of several covariates. Besides the Bonferroni correction, we used the default settings of the `semtree` package throughout our simulation studies. The score-based tests were performed by applying the default settings of the `strucchange` package. The data used to fit the SEMs were drawn from a multivariate normal distribution. All experimental conditions were replicated 10,000 times.

### Simulation I: Type I Error Rate and Runtime

Simulation I assessed the type I error rate under the null hypothesis of constant parameters and the runtime for a different number of noise variables and sample sizes. The simulated data was homogeneous without any group differences.

**Figure 3** shows the linear latent growth curve model used in Simulation I and II. Model specification and parameter values



**FIGURE 3** | Path diagram of the linear latent growth curve model used for data generation in Simulation I and II. The parameter values were obtained by fitting the model on scores from the longitudinal Wechsler Intelligence Scale for Children data set.

were taken from McArdle and Epstein (1987), who modeled the scores of 204 young children from the Wechsler Intelligence Scale for Children over four repeated occasions of measurement at 6, 7, 9, and 11 years of age (see Brandmaier et al., 2013b for a SEM tree analysis of these data). In both simulation studies, we generated multivariate normal data, using the mean vector and covariance matrix implied by the model presented in **Figure 3**.

After generating the data, the linear latent growth curve model presented in **Figure 3** was estimated, serving as a template model for the SEM trees. The model was defined by six free parameters: the mean and the variance of the random intercept  $f_I$ , the mean and the variance of the random slope  $f_S$ , the covariance between the random intercept and the random slope, and the residual error variance that was constrained to be equal for all four measurements of the observed variable  $y$ .

The following experimental factors were varied:

- *Level of measurement of the noise variables:* We provided the SEM trees with randomly generated noise variables. The noise variables were either continuous (standard normal), ordinal with 6 levels (with an equal number of observations per level), or dichotomous (with an equal number of observations in both classes). For a given condition, all noise variables had the same level of measurement.
- *Number of noise variables:* Either 1, 3, or 5 noise variables were generated.
- *Sample size (N):* The simulated samples contained either 504 or 1,008 observations. The odd numbers resulted from the necessity to be divisible by 6 to allow for an equal number of observations per level of the ordinal noise variables.

First, we will inspect the type I error rates of the different SEM tree approaches and compare their computation time afterward.

<sup>1</sup><https://github.com/brandmaier/semtree/commit/30ca7500e43ca99975dfe6b8917ef8f293beaeb3>

<sup>2</sup><https://osf.io/k82y3/>

### Percentage of Type I Errors

Every tree consisting of more than one node was counted as a type I error. Ideally, the proportion of type I errors should approach 5%. **Table 1** shows the empirical type I error rates of the different SEM tree approaches. The results are sorted with respect to the level of measurement of the noise variables. To get a better understanding of the simulated error rates, we printed results for methods that fell inside a 95% confidence interval around the optimal rate of 5% for 10,000 replications (CI: [4.573; 5.427]) in bold. For ordinal and dichotomous noise variables, all SEM tree implementations yielded error rates mostly close to the desired 5%. For continuous covariates, however, only *fair*, *maxLR*, *CvM*, and *maxLM* trees had satisfactory type I error rates. *DM* trees exhibited slightly too few type I errors. As predicted by Brandmaier et al. (2013b), *naïve* trees that were provided with continuous noise variables over-adjusted and produced error rates that were too small by a factor of 10. Increasing the sample sizes amplified this overcorrection. For the remaining methods, varying the number of noise variables and the sample size did not systematically influence the error rates.

### Runtime

We recorded the computation time for the different SEM trees in seconds. As a matter of course, the runtime varies widely depending on the computing platform. However, comparing the runtime of the different methods allows for relative comparisons and provides estimates for current standard computing platforms. Necessarily, the absolute estimates will become outdated soon. The simulation was conducted with an Intel® Xeon® CPU E5-2670 processor using a single core.

**Table 2** presents the median of the computation time in seconds. The median runtime for ordinal and dichotomous noise variables was small. The score-guided trees (*WDM*, *maxLM<sub>O</sub>*, and *LM*) showed a minor speed advantage over the likelihood-ratio-guided trees (*naïve*, *fair*, and *maxLR*). For continuous noise variables, however, for which many possible cut points needed to be evaluated, the runtime of likelihood-ratio-guided SEM trees was excessively larger than the computation time of score-guided SEM trees. For instance, given the larger sample size and five noise variables, likelihood-ratio-guided SEM trees needed several minutes to compute, whereas score-guided SEM trees (*DM*, *CvM*, and *maxLM*) were performed in fractions of a second. The runtime of the *fair* trees was roughly half as long as the runtime of *naïve* and *maxLR* trees, most likely because the *fair* method tests only half of the possible cut points for continuous variables. As expected, a larger sample size and more noise variables led to an increase in computation time of the likelihood-ratio-guided SEM trees. In contrast, the runtime of score-guided SEM trees remained virtually the same. This implies that even for larger samples consisting of larger numbers of individuals and many covariates, score-guided SEM trees can be computed in short time.

### Simulation II: Power, Cut Point Estimation, and Group Recovery

Simulation II evaluated the performance of likelihood-ratio and score-guided SEM trees in heterogeneous samples consisting of two subgroups.

**TABLE 1** | Empirical type I error rates.

Nr. noise	N	Continuous						Ordinal					Dichotomous		
		<i>Naïve</i>	<i>Fair</i>	<i>maxLR</i>	<i>DM</i>	<i>CvM</i>	<i>maxLM</i>	<i>Naïve</i>	<i>Fair</i>	<i>maxLR</i>	<i>WDM</i>	<i>maxLM<sub>O</sub></i>	<i>Naïve</i>	<i>Fair</i>	<i>LM</i>
1	504	0.56	<b>5.15</b>	<b>5.27</b>	3.85	<b>5.01</b>	<b>5.05</b>	4.50	5.71	<b>5.31</b>	<b>5.08</b>	<b>5.16</b>	<b>5.15</b>	<b>5.21</b>	<b>4.89</b>
3	504	0.60	<b>5.01</b>	5.43	4.17	<b>5.05</b>	5.57	<b>4.99</b>	5.78	<b>5.41</b>	<b>5.38</b>	5.59	<b>5.18</b>	5.45	<b>4.76</b>
5	504	0.51	<b>5.26</b>	<b>5.25</b>	4.18	<b>5.11</b>	5.62	<b>5.10</b>	5.73	5.66	5.59	5.69	<b>5.07</b>	<b>5.39</b>	4.55
1	1,008	0.25	<b>4.71</b>	<b>5.39</b>	3.93	<b>4.86</b>	<b>5.17</b>	3.93	<b>5.27</b>	<b>4.71</b>	<b>4.95</b>	<b>4.76</b>	<b>5.17</b>	<b>5.31</b>	<b>4.99</b>
3	1,008	0.35	<b>4.95</b>	<b>5.04</b>	3.91	4.57	<b>5.01</b>	<b>4.80</b>	5.67	<b>5.23</b>	5.60	<b>5.24</b>	<b>5.13</b>	<b>4.97</b>	<b>4.92</b>
5	1,008	0.35	<b>5.23</b>	<b>5.37</b>	<b>4.61</b>	<b>5.13</b>	5.48	<b>4.95</b>	5.60	<b>5.27</b>	<b>5.25</b>	5.68	<b>4.81</b>	<b>5.04</b>	<b>4.61</b>

*Nr. noise* = number of noise variables, *N* = sample size. Error rates within the 95% confidence interval around the optimal rate of 5% are printed in bold.

**TABLE 2** | Median runtime in seconds.

Nr. noise	N	Continuous						Ordinal					Dichotomous		
		<i>Naïve</i>	<i>Fair</i>	<i>maxLR</i>	<i>DM</i>	<i>CvM</i>	<i>maxLM</i>	<i>Naïve</i>	<i>Fair</i>	<i>maxLR</i>	<i>WDM</i>	<i>maxLM<sub>O</sub></i>	<i>Naïve</i>	<i>Fair</i>	<i>LM</i>
1	504	35.0	15.8	34.7	0.2	0.2	0.2	0.6	0.7	1.4	0.2	0.2	0.4	0.4	0.2
3	504	105.6	48.1	105.7	0.2	0.2	0.2	1.3	1.5	1.9	0.2	0.2	0.4	0.7	0.2
5	504	179.3	81.0	179.3	0.2	0.2	0.2	2.1	2.5	2.6	0.2	0.2	0.7	1.0	0.2
1	1,008	72.7	34.5	72.7	0.2	0.2	0.2	0.5	0.6	1.3	0.2	0.2	0.3	0.4	0.2
3	1,008	222.7	105.0	222.7	0.2	0.2	0.2	1.3	1.5	1.9	0.2	0.2	0.5	0.7	0.2
5	1,008	374.7	175.4	366.7	0.2	0.2	0.2	2.1	2.5	2.7	0.3	0.3	0.8	1.1	0.2

*Nr. noise* = number of noise variables, *N* = sample size.

We varied the following experimental factors:

- *Level of measurement of the covariate:* The SEM tree was provided with a single covariate that was either a continuous variable (standard normal), an ordinal variable with 6 levels, or a dichotomous variable.
- *Group differences:* We tested two types of group differences. Either the fixed slope of the linear latent growth curve model shown in **Figure 3** or all random effects varied between groups. **Table 3** presents the values used for the heterogeneous parameters. Note that in the fixed slope condition, only a single parameter varied between groups, whereas in the random effects condition, three parameters varied. The values of the remaining homogeneous parameters are shown in **Figure 3**.
- *Noise variable:* In the noise condition, the SEM tree algorithm was provided with a noise variable in addition to the informative covariate. In the no-noise condition, only the informative covariate was given to the tree. The noise variable was independent of the group differences and randomly selected to be a continuous variable (standard normal), an ordinal variable with 6 levels, or a dichotomous variable.
- *Cut point location:* We tested three different positions of the optimal cut point in the informative covariate. The cut points were either central, partitioning the sample into two groups of equal size, moderately non-central, resulting in a larger subgroup consisting of 66.67% of the observations and a smaller subgroup with 33.33% of the observations, or strongly non-central with 83.33% of the observations in the larger subgroup and 16.67% of the observations in the smaller subgroup. We counterbalanced the non-central cut points so that moderately non-central cut points occurred either after the  $1/3$ - or after the  $2/3$ -quantile of the covariate and strongly non-central cut points either after the  $1/6$ - or the  $5/6$ -quantile.
- *Sample size (N):* The sample consisted either of 504 or 1,008 observations.

We evaluated each method in terms of statistical power to detect heterogeneity, the precision of the estimated cut points, group recovery, and runtime. For each condition, the results of the best-performing method are printed in bold in the following tables. Due to space constraints, we report only the most important simulation results. The complete simulation results are provided as Online Supplemental Material<sup>2</sup>.

**TABLE 3 |** Parameter differences used in Simulation II.

Parameter	Fixed effects		Parameter	Random effects	
	Group 1	Group 2		Group 1	Group 2
$E(f_i)$	5.389	5.695	$Var(f_i)$	25.137	38.023
			$Var(f_s)$	2.808	4.247
			$Cov(f_i, f_s)$	0.745	1.127

### Power

We define statistical power as the percentage of SEM trees that correctly selected the covariate as a split at any cut point and any level of the tree.

**Table 4** shows the estimated power of the different SEM trees. We will first compare the overall performance of the original *naïve* and *fair* trees with the newly implemented *maxLR* and score-guided trees. With respect to power, we found that *naïve* trees performed roughly as well as the newly implemented methods for ordinal and dichotomous covariates but poorly for continuous covariates. The other classical method, *fair* trees, showed overall the lowest power of all methods under investigation. As expected, the likelihood-ratio-guided *maxLR* trees yielded similar results as the score-guided *maxLM* trees but were consistently slightly more powerful. Among the experimental conditions, the type of group differences and the cut point location impacted the rank order of the methods the most. *DM* and *WDM* trees were the most powerful methods for detecting heterogeneity in the fixed slope parameter. In contrast, *maxLR*, *CvM*, *maxLM*, and *maxLM<sub>O</sub>* trees proved to be the more powerful methods for detecting heterogeneity in the random effects. We expected this behavior because the *DM* and *WDM* test statistics focus on heterogeneity in a single parameter, whereas all other methods monitor group differences in multiple parameters. Overall, the likelihood-ratio-based test statistic *maxLR* and the score-based test statistics with a scaling term (that is, *maxLM*, *WDM*, and *maxLM<sub>O</sub>*) were more sensitive for non-central cut points but less sensitive for central cut points than the *DM* and *CvM* statistics for continuous covariates that do not use any scaling. As an optimal baseline, we compared the power of the SEM trees with MGSEMs, denoted as *MG* in **Table 4**. Like the SEM trees, the MGSEMs were specified by letting all parameters vary between groups. In contrast to SEM trees, MGSEMs were unaffected by noise variables and were informed about the true cut point. Therefore, the MGSEMs present the upper limit achievable in terms of statistical power. Not surprisingly, MGSEMs were more powerful than all SEM tree methods, given continuous and ordinal covariates, but equally powerful in conditions with dichotomous covariates and without noise variables, where cut points did not need to be learned from the data.

The presence of a noise variable (not shown in **Table 4**) approximately halved the power of all tree methods but affected *naïve* trees most severely. The pronounced effect of noise variables on *naïve* trees was mainly driven by continuous noise variables, which led to severely over-adjusted *p*-values. Providing *naïve* trees with ordinal or dichotomous noise variables led to a decrease of power that was comparable to the decrease in other methods. Increasing the sample size had an approximately uniform effect and raised the power of all methods substantively.

### Precision of Estimated Cut Points

The estimation of the optimal cut point in the covariate is crucial for recovering the true grouping of individuals. The approaches for locating cut points differed between likelihood-ratio and score-guided SEM trees. Likelihood-ratio-guided trees found cut

**TABLE 4 |** Power to detect group differences.

N	CL	Continuous							Ordinal					Dichotomous				
		Naive	Fair	maxLR	DM	CvM	maxLM	MG	Naive	Fair	maxLR	WDM	maxLM <sub>O</sub>	MG	Naive	Fair	LM	MG
<b>Group difference in the fixed slope</b>																		
504	1/2	10.0	14.5	36.8	<b>46.7</b>	41.7	35.4	52.8	35.8	17.4	38.4	<b>44.2</b>	37.0	52.4	<b>52.8</b>	27.0	51.8	52.8
	1/3	7.7	13.1	31.6	<b>35.4</b>	32.4	30.5	47.1	30.4	15.7	32.6	<b>37.7</b>	31.6	46.1	<b>45.9</b>	24.0	45.2	45.9
	1/6	3.2	8.1	<b>16.8</b>	9.8	13.0	16.4	29.7	17.8	10.4	19.4	<b>21.2</b>	19.0	29.8	<b>30.1</b>	17.0	29.3	30.1
1,008	1/2	30.6	29.8	72.9	<b>84.9</b>	76.3	72.1	87.2	73.8	37.6	75.7	<b>83.3</b>	75.0	86.4	<b>85.5</b>	51.4	85.2	85.5
	1/3	24.3	26.1	66.1	<b>74.9</b>	64.8	65.6	82.0	65.7	32.3	68.0	<b>77.2</b>	67.4	81.3	<b>81.5</b>	47.6	81.2	81.5
	1/6	8.7	15.6	<b>38.5</b>	25.9	26.2	<b>38.5</b>	58.2	39.5	19.9	42.0	<b>48.7</b>	41.4	57.9	58.6	30.4	<b>58.7</b>	58.6
<b>Group differences in the random effects</b>																		
504	1/2	18.2	19.4	51.8	48.3	<b>56.5</b>	49.9	69.4	52.1	24.5	<b>54.2</b>	46.4	53.2	68.5	<b>68.8</b>	36.0	67.4	68.8
	1/3	13.7	17.1	44.1	35.6	<b>44.6</b>	42.8	62.8	45.0	21.0	<b>47.5</b>	39.9	45.9	63.1	<b>62.9</b>	32.8	61.0	62.9
	1/6	5.3	10.4	22.9	12.0	18.1	<b>23.8</b>	40.6	24.5	13.4	26.4	22.8	<b>26.8</b>	40.9	<b>41.1</b>	21.0	38.0	41.1
1,008	1/2	53.3	45.1	88.0	85.7	<b>89.6</b>	87.6	95.9	89.4	55.0	<b>90.4</b>	84.2	90.1	95.7	<b>95.7</b>	69.1	95.5	95.7
	1/3	43.9	37.5	<b>82.7</b>	74.4	80.0	80.9	93.3	84.2	47.7	<b>85.6</b>	77.5	84.3	93.1	<b>93.4</b>	63.3	93.0	93.4
	1/6	17.3	21.0	<b>52.8</b>	26.6	36.4	49.9	73.4	55.3	28.3	<b>57.4</b>	47.3	53.6	73.4	<b>73.3</b>	39.8	69.8	73.3

N = sample size, CL = cut point location, 1/2 = central cut point location, 1/3 = moderately non-central, 1/6 = strongly non-central, MG = MGSEM. Best-performing methods are printed in bold.

points by maximizing a partitioned log-likelihood, and score-guided SEM trees determined cut points by searching through a disaggregated maxLM statistic. We limit ourselves to discuss cut points estimated by maxLR and maxLM trees in the following. We used only trees that selected the covariate for the initial split of the data, ignoring possible further splits. Since noise variables had no visible effect, we will discuss only simulation trials without additional noise variables. Also, we did not evaluate dichotomous covariates because there is only a single trivial cut point.

Table 5 presents bias, standard deviation, and root mean squared error (RMSE) of the estimated cut points. Both approaches produced similar cut points that were nearly unbiased. Overall, cut points estimated by maxLR were slightly more precise in terms of RMSE. Interestingly, group differences in the random effects led to slightly biased cut point estimates provided by maxLM trees, which was not observed for maxLR trees. Estimates for non-central cut points showed more variability than for central cut points in both methods. A larger sample size of 1,008 observations increased both methods' precision and reduced the bias of cut points estimated by maxLM.

**Group Recovery**

We used the adjusted Rand index (ARI; Hubert and Arabie, 1985; Milligan and Cooper, 1986) to measure how well the true groups are recovered by each SEM tree method. The ARI is widely used to measure the similarity between two partitions and is adjusted for agreement by chance. A large ARI value up to the maximum of 1 indicates a high agreement between the partitioning estimated by a tree and the true partitioning, while smaller values imply a lower degree of similarity. Particularly, an ARI of 0 is obtained if a tree fails to detect any group differences and does not split the sample.

The ARI of the different tree methods is shown in Table 6. In our simulation setup, the ARI of a SEM tree method seemed to be mainly determined by its power to detect heterogeneity as a

similar rank order as for the statistical power emerged. Given a continuous covariate, score-guided DM and CvM trees showed the largest ARI for central cut points, maxLM and maxLR trees showed the largest ARI for non-central cut points, while the original likelihood-ratio-guided naive and fair trees performed poorly. As with power, DM trees had a higher ARI for a difference in the slope, and the ARI of the other score-guided and maxLR trees was higher for differences in the random effects. Naive trees performed better when provided with an ordinal or dichotomous covariate. For ordinal covariates, WDM trees exhibited the largest ARI if the fixed slope differed between groups, whereas the ARI of maxLR and maxLM<sub>O</sub> trees was higher for differences in the random effects. For dichotomous covariates, naive trees showed a slightly higher ARI than score-guided LM trees. However, if provided with an additional noise variable, naive trees showed a more pronounced decrease in the ARI than LM trees (not shown in Table 6). This effect was mainly driven by continuous noise variables, which led to overcorrected p-values of naive trees. Non-central cut points generally reduced the ARI of all trees, affecting DM and CvM trees without a scaling term the most. The ARI of all tree methods improved substantially for larger samples with 1,008 simulated individuals without drastically changing the rank order.

**Runtime**

The computation time of the SEM trees in Simulation II was in line with the observed runtime in Simulation I. Overall, the median computation time of score-guided SEM trees was 0.50 s with little variability across the simulation conditions. In contrast, the runtime of likelihood-ratio-guided trees varied considerably according to the level of measurement of the covariate and noise variable. The overall median computation time for simulation conditions with ordinal or dichotomous covariate and noise variable was 0.88 s for naive trees, 0.89 s for fair trees, and 1.40 s for maxLR trees. However, if either the covariate or the noise



**TABLE 5 |** Estimated cut points.

N	CL	Continuous						Ordinal					
		maxLR			maxLM			maxLR			maxLM <sub>O</sub>		
		B	SD	RMSE	B	SD	RMSE	B	SD	RMSE	B	SD	RMSE
<b>Group difference in the fixed slope</b>													
504	1/2	0.013	0.388	0.389	0.017	0.408	0.408	-0.003	0.798	0.797	0.011	0.750	0.750
	1/3	0.014	0.430	0.430	0.014	0.440	0.440	0.022	0.955	0.956	0.013	0.878	0.878
	1/6	0.011	0.694	0.694	-0.022	0.686	0.686	0.010	1.347	1.347	0.015	1.312	1.312
1,008	1/2	-0.002	0.273	0.273	0.000	0.282	0.282	-0.004	0.546	0.545	-0.010	0.459	0.459
	1/3	-0.006	0.310	0.310	-0.007	0.311	0.311	-0.004	0.603	0.603	-0.006	0.494	0.494
	1/6	0.004	0.491	0.491	-0.001	0.499	0.499	0.004	0.933	0.933	0.007	0.836	0.836
<b>Group differences in the random effects</b>													
504	1/2	0.014	0.345	0.345	0.183	0.348	0.393	0.003	0.699	0.699	0.255	0.790	0.831
	1/3	0.011	0.381	0.381	0.177	0.401	0.439	0.018	0.769	0.769	0.246	0.882	0.916
	1/6	0.013	0.611	0.611	0.098	0.602	0.610	0.003	1.187	1.187	0.138	1.260	1.267
1,008	1/2	0.015	0.224	0.225	0.085	0.240	0.255	0.001	0.415	0.415	0.118	0.542	0.555
	1/3	0.011	0.245	0.245	0.092	0.268	0.283	0.010	0.457	0.457	0.117	0.600	0.611
	1/6	0.004	0.387	0.387	0.076	0.421	0.428	-0.001	0.728	0.728	0.102	0.943	0.949

N = sample size, CL = cut point location, 1/2 = central cut point location, 1/3 = moderately non-central, 1/6 = strongly non-central, B = bias, SD = standard deviation, RMSE = root mean squared error.

**TABLE 6 |** Adjusted Rand index.

N	CL	Continuous						Ordinal					Dichotomous		
		Naïve	Fair	maxLR	DM	CvM	maxLM	Naïve	Fair	maxLR	WDM	maxLM <sub>O</sub>	Naïve	Fair	LM
<b>Group difference in the fixed slope</b>															
504	1/2	0.069	0.085	0.243	<b>0.377</b>	0.323	0.234	0.277	0.120	0.296	<b>0.361</b>	0.294	<b>0.527</b>	0.270	0.518
	1/3	0.055	0.075	0.208	<b>0.261</b>	0.215	0.204	0.230	0.108	0.246	<b>0.306</b>	0.248	<b>0.459</b>	0.240	0.452
	1/6	0.022	0.044	0.101	0.036	0.036	<b>0.105</b>	0.131	0.065	0.141	<b>0.159</b>	0.142	<b>0.301</b>	0.170	0.293
1,008	1/2	0.248	0.209	0.557	<b>0.726</b>	0.636	0.556	0.647	0.313	0.662	<b>0.751</b>	0.675	<b>0.855</b>	0.514	0.852
	1/3	0.196	0.179	0.500	<b>0.604</b>	0.490	0.504	0.575	0.267	0.593	<b>0.697</b>	0.607	<b>0.815</b>	0.476	0.812
	1/6	0.067	0.098	0.275	0.137	0.097	<b>0.290</b>	0.335	0.159	0.353	<b>0.430</b>	0.358	0.586	0.304	<b>0.587</b>
<b>Group differences in the random effects</b>															
504	1/2	0.136	0.125	0.361	0.378	<b>0.441</b>	0.341	0.428	0.187	<b>0.444</b>	0.361	0.413	<b>0.688</b>	0.360	0.674
	1/3	0.102	0.109	0.309	0.252	<b>0.312</b>	0.293	0.371	0.160	<b>0.388</b>	0.308	0.356	<b>0.629</b>	0.328	0.610
	1/6	0.038	0.063	0.150	0.049	0.066	<b>0.170</b>	0.192	0.091	0.205	0.175	<b>0.206</b>	<b>0.411</b>	0.210	0.380
1,008	1/2	0.449	0.339	0.711	0.714	<b>0.763</b>	0.701	0.818	0.479	<b>0.827</b>	0.747	0.800	<b>0.957</b>	0.691	0.955
	1/3	0.370	0.281	<b>0.664</b>	0.585	0.632	0.645	0.769	0.415	<b>0.780</b>	0.686	0.744	<b>0.934</b>	0.633	0.929
	1/6	0.142	0.149	<b>0.406</b>	0.141	0.171	0.398	0.494	0.243	<b>0.509</b>	0.404	0.457	<b>0.733</b>	0.398	0.698

N = sample size, CL = moderately non-central, 1/2 = central cut point location, 1/3 = moderately non-central, 1/6 = strongly non-central, MG = MGSEM. Best-performing methods are printed in bold.

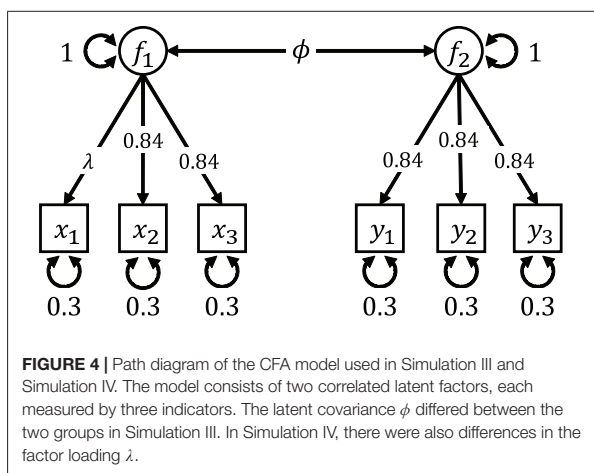
variable were continuous, the overall median runtime increased drastically. For instance, for conditions with a continuous covariate, a continuous noise variable, and a sample size of 504, the computation time of *naïve*, *fair*, and *maxLR* trees increased to 70.18, 32.85, and 71.05 s, respectively. For samples with 1,008 individuals, the runtime increased to 149.91, 72.23, and 280.58 s, respectively.

### Simulation III: Focus Parameters

The goal of Simulation III was to demonstrate how specific hypotheses about parameter heterogeneity, such as certain types

of measurement invariance, can be tested with the use of SEM trees with focus parameters. By default, SEM trees split with respect to differences in any model parameter. At times, researchers may be interested in finding group differences only for a subset of parameters that are referred to as focus parameters in the *semtree* package. When focus parameters are given, SEM trees will only assess heterogeneity in these parameters and ignore group differences in the remaining parameters to evaluate a split.

Figure 4 shows the population model used to generate data in Simulation III and IV. Depicted is a confirmatory factor model



**FIGURE 4** | Path diagram of the CFA model used in Simulation III and Simulation IV. The model consists of two correlated latent factors, each measured by three indicators. The latent covariance  $\phi$  differed between the two groups in Simulation III. In Simulation IV, there were also differences in the factor loading  $\lambda$ .

with two correlated factors, each measured by three indicators. A two group-population was simulated, where both factors were uncorrelated in the first group ( $\phi_{g1} = 0$ ) and covaried with  $\phi_{g2} = 0.471$  in the second group. The factor loading  $\lambda$  did not vary in Simulation III and was set to 0.837 in both groups, corresponding to a latent factor accounting for 70% of the variance in the observed variables. In each simulation replication, we generated 250 individuals per group, resulting in a total sample size of 500. The model shown in **Figure 4** was then estimated with a common identification constraint; specifically, by fixing the variances of the two factors  $f_1$  and  $f_2$  to one and estimating the latent covariance, all factor loadings, and all residual variances freely. We did not specify a mean structure. To recover the group difference, we provided maxLR and maxLM trees with a standard normally distributed informative covariate. The true cut point in the covariate point was central.

We explored the following three scenarios:

- **Testing measurement invariance:** We ignored heterogeneity in the latent covariance  $\phi$  and tested only the (homogeneous) measurement part of the model by treating all factor loadings and residual variances as focus parameters. This scenario can be seen as an exploration of which covariates predict violations of strict measurement invariance.
- **Testing the latent covariance:** Only the (heterogeneous) latent covariance  $\phi$  was treated as a focus parameter, and the remaining (homogeneous) parameters did not contribute to the assessment of potential splits. This scenario is akin to ignoring the covariates' information on violations of measurement invariance and, instead, investigating differences on the latent level only.
- **No focus parameters:** No focus parameters were specified, and all parameters contributed to the evaluation of a potential split. This scenario served as a baseline.

**Table 7** shows the percentage of maxLR and maxLM trees that split the sample and rejected the specific null hypothesis for a significance level of 5%. In the measurement invariance scenario,

**TABLE 7** | Type I error and power to detect group differences.

Scenario	maxLR	maxLM
Testing measurement invariance	4.86	4.85
Testing the latent covariance	99.10	99.18
No focus parameters	82.97	81.21

The first row shows the type I error rates and the second and third row the statistical power to detect group differences.

the SEM trees tested the null hypothesis that all parameters of the measurement model are homogeneous. Both maxLR and maxLM trees yielded error rates that were close to the optimal rate of 5%. In other words, the SEM tree methods successfully ignored the group difference in the covariance structure of the latent variables. Without focus parameters, the SEM trees tested the standard null hypothesis of complete parameter equivalency across groups. The maxLR and the maxLM trees rejected the null hypothesis in over 80% of the replications. If only the latent covariance  $\phi$  was declared as a focus parameter, the power of both SEM trees to detect the group difference rose substantially and almost approached one. This finding highlights that the sensitivity of SEM trees for heterogeneity in a specific set of target parameters can be significantly enhanced by specifying focus parameters if differences with respect to the non-focus parameters can be safely ignored.

### Simulation IV: Global Equality Constraints

Simulation IV aimed at investigating the utility of SEM trees with equality constraints and pointing out common pitfalls. Equality constraints are useful to incorporate prior knowledge about the homogeneity of specific parameters into a SEM tree. By constraining a parameter to equality, a so-called global constraint, this parameter is estimated once in the full sample, and the resulting estimate is used in all submodels. Constraining parameters increases a SEM tree's sensitivity for group differences in the remaining parameters and might stabilize estimation. However, by erroneously constraining parameters to equality that are actually different in certain groups, a SEM tree can be severely misspecified.

We investigated the following conditions:

- **Group differences:** We tested two types of group differences. Either the latent covariance differed between groups ( $\phi_{g1} = 0, \phi_{g2} = 0.471$ ) and the factor loading was homogeneous ( $\lambda_{g1/g2} = 0.837$ ) or the latent covariance was homogeneous ( $\phi_{g1/g2} = 0.471$ ) and the factor loading differed ( $\lambda_{g1} = 0.837, \lambda_{g2} = 0.640$ ). All other values were as shown in **Figure 4**. We generated 250 individuals per subgroup and provided the SEM trees with a standard normal covariate with a central cut point.
- **Equality constraints:** Either the heterogeneous parameter (the latent covariance  $\phi$  or the factor loading  $\lambda$ ), all factor loadings and residual variances of the factor  $f_2$ , or no parameters were constrained to equality.

**TABLE 8** | Power to detect group differences.

Equality constraints	maxLR	maxLM
<b>Group difference in the latent covariance <math>\phi</math></b>		
None	82.52	81.13
Latent covariance $\phi$	5.71	5.52
Measurement model of $f_2$	91.25	90.40
<b>Group difference in the factor loading <math>\lambda</math></b>		
None	83.20	81.44
Factor loading $\lambda$	24.89	22.65
Measurement model of $f_2$	91.52	90.92

The empirical power of maxLR and maxLM trees for detecting heterogeneity for a significance level of 5% is shown in **Table 8**. Without equality constraints, both SEM tree methods showed a power of slightly above 80%. As expected, constraining a heterogeneous parameter reduced the power significantly, but the exact effect depended on the specific parameter. After constraining the heterogeneous latent covariance  $\phi$ , there was no possibility for the trees to detect a difference between groups. The latent covariance  $\phi$  was the only parameter associated with the correlation between the factors, and the group difference had no other parameter left to manifest. As a result, the power of maxLR and maxLM trees reduced to the type I error level. However, after constraining the heterogeneous factor loading  $\lambda$ , the trees still found significant group differences in 24.89% (maxLR) and 22.65% (maxLM) of the replications. This finding implies that the group difference in the factor loading  $\lambda$  were picked up by other unconstrained parameters of the measurement model. Finally, constraining the homogeneous parameters of the measurement model of factor  $f_2$  increased the power to detect differences in the latent covariance or the factor loading by roughly 10 percentage points.

The results of Simulation IV suggest that constraining homogeneous parameters to equality can increase the power of SEM trees but also involves the risk of introducing severe misspecification. Constraining a heterogeneous parameter leads to a distorted picture of group differences as the SEM tree might or might not find significant heterogeneity in other parameters that are homogeneous across subgroups. Thus, it seems generally advisable to fully explore differences in all parameters or to use focus parameters rather than taking the risk of misspecifying trees by using inadequate equality constraints.

## Summary

In our simulation studies, the likelihood-ratio-guided *naïve* trees showed mixed results, confirming the known weaknesses of the approach. When provided with ordinal or dichotomous covariates, *naïve* trees showed an adequate control of type I errors and were among the best-performing methods in terms of power to detect heterogeneity and group recovery. However, with continuous covariates, *naïve* trees were overly conservative, resulting in too few type I errors and low power. The likelihood-ratio-guided *fair* trees showed overall the lowest power of all methods, resulting in a poor group recovery. In contrast to *naïve* trees, the type I error rate of *fair* trees

was close to optimal, regardless of the measurement level of the provided covariates. Therefore, *fair* trees may be useful in very large samples where low power is less of an issue. The likelihood-ratio-guided maxLR trees, that we implemented in the `semtree` package, resolved many of the weaknesses of the classical SEM tree methods *naïve* and *fair* and positioned themselves slightly above the score-guided maxLM trees in terms of power, group recovery, and cut point precision. maxLR trees and the score-guided maxLM (for continuous covariates), and maxLM<sub>O</sub> trees (for ordinal covariates) exceeded other split selection approaches in conditions with group differences in multiple parameters associated with the random effects and non-central cut points. SEM trees guided by the score-based *DM* (for testing continuous covariates) and *WDM* (ordinal covariates) test statistics outperformed other methods in terms of power and group recovery when group differences were to be found in a single parameter describing the fixed slope. Different from *DM* trees, *WDM* trees were also sensitive to non-central cut points. Score-guided *CvM* trees performed better than other methods in detecting heterogeneity in the random effects when the cut point was central. Finally, the score-based *LM* trees for categorical covariates were slightly less powerful than the *naïve* method. Although maxLM<sub>O</sub> and *LM* trees were roughly on par with *naïve* trees, the score-based methods clearly outperformed *naïve* trees when provided with an additional continuous noise variable. Overall, all score-guided trees and the newly implemented maxLR trees showed a satisfactory control of type I errors. The most striking difference between likelihood-ratio and score-guided SEM trees was the runtime. Whereas the runtime of all likelihood-ratio-based methods was excessive if one of the covariates under evaluation was continuous, score-guided trees were computed quickly. In summary, all newly implemented methods (maxLR and the score-based methods) outperformed the original *naïve* and *fair* methods. Moreover, no single method under evaluation performed best across all situations, and all of the new methods had some unique advantages which may justify their use given certain conditions.

Regarding focus parameters and equality constraints, we found that both can successfully be applied to increase the power of SEM trees to detect group differences if there is either a clear set of target parameters or prior knowledge about homogeneous parameters available. Still, we discourage the use of equality constraints in favor of focus parameters, which allow exploring the effects of selected parameters without incurring misspecifications during the split evaluation.

## DISCUSSION

In the present study, we introduced score-guided SEM trees as a fast and efficient way for growing SEM trees. Along with score-guided SEM trees, we also implemented a new likelihood-ratio-guided split selection based on the maxLR statistic that solved many of the shortcomings of the original likelihood-ratio-guided SEM trees (Brandmaier et al., 2013b). We evaluated and compared the newly implemented and the original SEM tree approaches in a Monte Carlo simulation study. We investigated



those cases in which users want to adjust the type I error rate for multiple testing of covariates. Overall, we conclude that the new split selection procedures are superior to the original split selection because they have higher statistical power and are unbiased in the selection of covariates that predict group differences in SEM parameters. Among the SEM tree methods, score-guided trees stand out due to their computational efficiency, making the use of SEM trees in large data sets feasible.

Our simulation studies evaluated three different likelihood-ratio-guided SEM tree approaches and five different score-guided SEM trees. The score-guided SEM trees were based on test statistics recently popularized in psychometrics by Merkle and Zeileis (2013) and Merkle et al. (2014) for studying heterogeneity in SEM parameters. When provided with continuous covariates, we found that guiding SEM trees with score-based tests significantly reduced the runtime of the trees. If solely provided with ordinal and categorical variables, score-guided SEM trees performed as well as likelihood-ratio-guided SEM trees. The large difference in the runtime of both approaches for continuous covariates can be attributed to the fact that the evaluation of a covariate by a likelihood-ratio-guided SEM tree requires the estimation of a MGSEM for every unique value of covariate. This leads to a large number of MGSEMs to be estimated as most values of continuous covariates are usually unique. However, score-guided SEM trees do not require the estimation of any MGSEMs for the evaluation of a covariate and, therefore, can be computed in little time.

Score-guided SEM trees and likelihood-ratio-guided trees based on the newly implemented *maxLR* statistic also proved to be more powerful in detecting group differences than the original SEM tree methods if one of the covariates provided to the SEM tree were continuous. The low statistical power of the original SEM tree implementation is a side effect of a suboptimal correction of the selection bias. The low power of the *naïve* method for continuous covariates can be explained by the overcorrection of the Bonferroni adjusted *p*-values due to too many possible cut points in the continuous variables. The low power that the *fair* method displayed throughout all simulation conditions was because the *fair* selection method uses only half of the sample for selecting the best covariate.

Besides the evaluation of the original and newly proposed SEM tree methods, we also demonstrated the utility and pitfalls of trees with focus parameters and equality constraints. We showed that focus parameters are well suited to investigate specific hypotheses about parameter heterogeneity, such as different types of measurement invariance. Specifying equality constraints for homogeneous parameters increased the power of SEM trees for detecting group differences in the remaining parameters. We also demonstrated that misspecified equality constraints can obscure group differences and thus discourage this approach. As the effect of misspecified equality constraints can be hard to predict for a user, we recommend to explore differences in all parameters or to use focus parameters rather than risk to misspecify trees by using inadequate equality constraints.

The faster runtime of score-based tests is a major advantage for practical use and enables the wider adoption of SEM trees. The slow runtime of likelihood-ratio tests had made SEM trees unattractive if not impossible to run with large data sets on

desktop computers. The runtime improvement may become even more important if one wishes to complement SEM tree inferences with resampling methods such as SEM forests (Brandmaier et al., 2016). SEM forests are a more robust alternative to single SEM trees if the overall importance of variables is of primary interest because small variations in the sample often lead to different trees. As SEM forests are based on hundreds if not thousands of trees, they will profit dramatically from the score-guided strategy.

The question remains which of the newly implemented methods should be used to estimate SEM trees. Our simulation results imply that all of the new methods have their unique strengths. However, in practice, when it is usually unknown how many of the model parameters are heterogeneous or if the subgroups are roughly equal in size, the advantages of the *DM*, *WDM*, and *CvM* statistics seem hard to exploit. Instead, the *maxLR* (if computational feasible), *maxLM*, *maxLM<sub>O</sub>*, and *LM* trees statistics are best suited for situations without *a priori* knowledge about potential group differences. Moreover, if one is only interested in change in a specific parameter, specifying a focus parameter may represent an excellent alternative to the *DM* and *WDM* statistics.

Although SEM trees are a powerful and flexible method for investigating heterogeneity in SEMs, we want to stress that they are not always the most appropriate one. It is important to note that the performance of the SEM trees depends on the covariates available. If none of these covariates is in any way related to group differences, SEM trees will fail to detect any heterogeneity. In situations without informative covariates, researchers may resort to latent class or finite mixture models (Jedidi et al., 1997; Muthén and Shedden, 1999; Lubke and Muthén, 2005) for detecting heterogeneity. Latent class approaches automatically test for differences between all possible groups of individuals without requiring covariates. A disadvantage of these methods is that the number of subgroups needs to be pre-specified by the user. Another disadvantage of SEM trees is that they provide only sparse information about how a parameter changes with respect to a covariate. Recently, Arnold et al. (2019) suggested a framework called individual parameter contribution regression that allows modeling SEM parameter estimates as a linear function of covariates.

There are several limitations of our study. First, we focused narrowly on the *semtree* package for growing SEM trees and did not evaluate SEM trees estimated by the generic MOB algorithm from the *partykit* package. Ideally, a future study should aim to replicate our findings using MOB. Second, most of our simulations were performed using a linear latent growth curve model with only two types of group differences. Likely, different types of SEMs or parameter differences could have changed the performance of some of the methods under investigation. However, we would expect the general pattern of results to hold for other models as well. Third, for the sake of simplicity, we tested only a small number of uncorrelated covariates and did not test any covariate interactions. Fourth, we did not assess the influence of non-normally distributed data and model misspecification on the SEM trees. These remain topics for future research.

In summary, we found score-guided SEM trees to be fast, flexible, and powerful tools for investigating heterogeneity in

SEM parameters. Based on our work, we suggest that score-guided split selection should become the new standard for estimating SEM trees and forests.

## DATA AVAILABILITY STATEMENT

The simulated data presented in this study can be found in the following online repository: <https://osf.io/k82y3/>.

## AUTHOR CONTRIBUTIONS

MA programmed the score-guided SEM tree implementation with the support of AB. MA designed and carried out the

simulation study and wrote the manuscript with the support of MV and AB. MV revised the presentation of score-based tests and supervised the project. AB conceived the original idea and provided crucial code to link the score-guided SEM tree implementation to the existing `semtree` R package. All authors discussed the results and commented on the manuscript.

## ACKNOWLEDGMENTS

We acknowledge support by the German Research Foundation (DFG) and the Open Access Publication Fund of Humboldt-Universität zu Berlin.

## REFERENCES

- Ammerman, B. A., Jacobucci, R., and McCloskey, M. S. (2019). Reconsidering important outcomes of the nonsuicidal self-injury disorder diagnostic criterion. *A. J. Clin. Psychol.* 75, 1084–1097. doi: 10.1002/jclp.22754
- Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* 61, 821–856. doi: 10.2307/2951764
- Arnold, M., Oberski, D. L., Brandmaier, A. M., and Voelkle, M. C. (2019). Identifying heterogeneity in dynamic panel models with individual parameter contribution regression. *Struct. Equ. Modeling* 27, 613–628. doi: 10.1080/10705511.2019.1667240
- Bollen, K. A. (1989). *Structural Equation with Latent Variables*. New York, NY: Wiley, doi: 10.1002/9781118619179
- Brandmaier, A. M., Driver, C. C., and Voelkle, M. C. (2018). “Recursive partitioning in continuous time analysis,” in *Continuous Time Modeling in the Behavioral and Related Sciences*, eds K. van Montfort, J. H. L. Oud, and M. C. Voelkle (Cham: Springer), 259–282. doi: 10.1007/978-3-319-77219-6\_11
- Brandmaier, A. M., Oertzen, T., von McArdle, J. J., and Lindenberger, U. (2013a). “Exploratory data mining with structural equation model trees,” in *Quantitative methodology series. Contemporary issues in Exploratory Data Mining in the Behavioral Sciences*, eds J. J. McArdle and G. Ritschard (London: Routledge), 96–127.
- Brandmaier, A. M., Oertzen, T., von McArdle, J. J., and Lindenberger, U. (2013b). Structural equation model trees. *Psychol. Methods* 18, 71–86. doi: 10.1037/a0030001
- Brandmaier, A. M., Prindle, J. J., McArdle, J. J., and Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychol. Methods* 21, 566–582. doi: 10.1037/met0000090
- Brandmaier, A. M., Ram, N., Wagner, G. G., and Gerstorf, D. (2017). Terminal decline in well-being: the role of multi-indicator constellations of physical health and psychosocial correlates. *Dev. Psychol.* 53, 996–1012. doi: 10.1037/dev0000274
- Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Researchers*, 2nd Edn. New York, NY: Guilford.
- de Mooij, S. M. M., Henson, R. N. A., Waldorp, L. J., and Kievit, R. A. (2018). Age differentiation within gray matter, white matter, and between memory and white matter in an adult life span cohort. *J. Neurosci.* 38, 5826–5836. doi: 10.1523/JNEUROSCI.1627-17.2018
- Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., and Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behav. Res. Methods* 50, 2016–2034. doi: 10.3758/s13428-017-0971-x
- Hansen, B. E. (1992). Testing for parameter instability in linear models. *J. Policy Model.* 14, 517–533. doi: 10.1016/0161-8938(92)90019-9
- Hildebrandt, A., Lüdtke, O., Robitzsch, A., Sommer, C., and Wilhelm, O. (2016). Exploring factor model parameters across continuous variables with local structural equation models. *Multivar. Behav. Res.* 51, 257–258. doi: 10.1080/00273171.2016.1142856
- Hjort, N. L., and Koning, A. (2002). Tests for constancy of model parameters over time. *J. Nonparametr. Stat.* 14, 113–132. doi: 10.1080/10485250211394
- Hothorn, T., and Zeileis, A. (2015). partykit: a modular toolkit for recursive partitioning in R. *J. Mach. Learn. Res.* 16, 3905–3909.
- Hubert, L., and Arabie, P. (1985). Comparing partitions. *J. Classif.* 2, 193–218. doi: 10.1007/BF01908075
- Jacobucci, R., Grimm, K. J., and McArdle, J. J. (2017). A comparison of methods for uncovering sample heterogeneity: structural equation model trees and finite mixture models. *Struct. Equ. Modeling* 24, 270–282. doi: 10.1080/10705511.2016.1250637
- Jedidi, K., Jagpal, H. S., and DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Mark. Sci.* 16, 39–59. doi: 10.1287/mksc.16.1.39
- Jensen, D. D., and Cohen, P. R. (2000). Multiple comparison in induction algorithms. *Mach. Learn.* 38, 309–338. doi: 10.1023/A:1007631014630
- Jones, P. J., Mair, P., Simon, T., and Zeileis, A. (2020). Network trees: a method for recursively partitioning covariance structures. *Psychometrika* doi: 10.1007/s11336-020-09731-4 Online ahead of print
- Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., and Borsboom, D. (2013). Simpson’s paradox in psychological science: a practical guide. *Front. Psychol.* 4:513. doi: 10.3389/fpsyg.2013.00513
- Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling*, 4th Edn. New York, NY: Guilford.
- Komboz, B., Strobl, C., and Zeileis, A. (2018). Tree-based global model tests for polytomous Rasch models. *Educ. Psychol. Meas.* 78, 128–166. doi: 10.1177/0013164416664394
- Lang, M. N., Schlosser, L., Hothorn, T., Mayr, G. J., Stauffer, R., and Zeileis, A. (2020). Circular regression trees and forests with an application to probabilistic wind direction forecasting. *J. R. Stat. Soc. C* 69, 1357–1374. doi: 10.1111/rssc.12437
- Loh, W.-Y., and Shih, Y.-S. (1997). Split selection methods for classification trees. *Stat. Sin.* 7, 815–840.
- Lubke, G. H., and Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychol. Methods* 10, 21–39. doi: 10.1037/1082-989X.10.1.21
- McArdle, J. J. (2012). “Latent curve modeling of longitudinal growth data,” in *Handbook of Structural Equation Modeling*, ed. R. H. Hoyle (New York, NY: Guilford Press), 547–570.
- McArdle, J. J., and Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Dev.* 58:110. doi: 10.2307/1130295
- Merkle, E. C., Fan, J., and Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika* 79, 569–584. doi: 10.1007/S11336-013-9376-7
- Merkle, E. C., and Zeileis, A. (2013). Tests of measurement invariance without subgroups: a generalization of classical methods. *Psychometrika* 78, 59–82. doi: 10.1007/S11336-012-9302-4
- Milligan, G. W., and Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multiv. Behav. Res.* 21, 441–458. doi: 10.1207/s15327906mbr2104\_5

- Muthén, B., and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* 55, 463–469. doi: 10.1111/j.0006-341X.1999.00463.x
- Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., et al. (2016). Openmx 2.0: Extended structural equation and statistical modeling. *Psychometrika* 81, 535–549. doi: 10.1007/s11336-014-9435-8
- Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02
- Serang, S., Jacobucci, R., Stegmann, G., Brandmaier, A. M., Cuijans, D., and Grimm, K. J. (2020). Mplus trees: structural equation model trees using Mplus. *Struct. Equ. Modeling* doi: 10.1080/10705511.2020.1726179 [Epub ahead of print].
- Shih, Y.-S. (2004). A note on split selection bias in classification trees. *Comput. Stat. Data Anal.* 45, 457–466. doi: 10.1016/S0167-9473(03)00064-1
- Simpson-Kent, I. L., Fuhrmann, D., Bathelt, J., Achterberg, J., Borgeest, G. S., and Kievit, R. A. (2020). Neurocognitive reorganization between crystallized intelligence, fluid intelligence and white matter microstructure in two age-heterogeneous developmental cohorts. *Dev. Cogn. Neurosci.* 41:100743. doi: 10.1016/j.dcn.2019.100743
- Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *Br. J. Math. Stat. Psychol.* 27, 229–239. doi: 10.1111/j.2044-8317.1974.tb00543.x
- Strobl, C., Kopf, J., and Zeileis, A. (2015). Rasch trees: a new method for detecting differential item functioning in the Rasch model. *Psychometrika* 80, 289–316. doi: 10.1007/S11336-013-9388-3
- Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* 14, 323–348. doi: 10.1037/a0016973
- Strobl, C., Wickelmaier, F., and Zeileis, A. (2011). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *J. Educ. Behav. Stat.* 36, 135–153. doi: 10.3102/1076998609359791
- Usami, S., Hayes, T., and McArdle, J. (2017). Fitting structural equation model trees and latent growth curve mixture models in longitudinal designs: the influence of model misspecification. *Struct. Equ. Modeling* 24, 585–598. doi: 10.1080/10705511.2016.1266267
- Usami, S., Jacobucci, R., and Hayes, T. (2019). The performance of latent growth curve model-based structural equation model trees to uncover population heterogeneity in growth trajectories. *Comput. Stat.* 34, 1–22. doi: 10.1007/s00180-018-0815-x
- Wang, T., Merkle, E. C., and Zeileis, A. (2014). Score-based tests of measurement invariance: use in practice. *Front. Psychol.* 5:438. doi: 10.3389/fpsyg.2014.00438
- Wang, T., Strobl, C., Zeileis, A., and Merkle, E. C. (2018). Score-based tests of differential item functioning via pairwise maximum likelihood estimation. *Psychometrika* 83, 132–155. doi: 10.1007/s11336-017-9591-8
- Wickelmaier, F., and Zeileis, A. (2018). Using recursive partitioning to account for parameter heterogeneity in multinomial processing tree models. *Behav. Res. Methods* 50, 1217–1233. doi: 10.3758/s13428-017-0937-z
- Zeileis, A. (2020). *Structural equation model trees with partykit and lavaan*. Available online at: <https://eeecon.uibk.ac.at/~zeileis/news/lavaantree/> (accessed December 17, 2020).
- Zeileis, A., and Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Stat. Neerl.* 61, 488–508. doi: 10.1111/j.1467-9574.2007.00371.x
- Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *J. Comput. Graph. Stat.* 17, 492–514. doi: 10.1198/106186008X319331
- Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). strucchange: an R package for testing for structural change in linear regression models. *J. Stat. Softw.* 7, 1–38. doi: 10.18637/jss.v007.i02
- Zeileis, A., Strobl, C., Wickelmaier, F., Komboz, B., and Kopf, J. (2020). *Psychotree: Recursive Partitioning Based on Psychometric Models (Version 0.15-3)* [Computer software]. Available online at: <https://cran.r-project.org/web/packages/psychotree/index.html> (accessed December 17, 2020).

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Arnold, Voelke and Brandmaier. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.