Computational Statistics and Data Analysis 86 (2015) 65–80

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Tree-based varying coefficient regression for longitudinal ordinal responses

Reto Bürgin *, Gilbert Ritschard

National Center of Competence in Research LIVES, Switzerland
Institute of Demographic and Life Course Studies, University of Geneva, Switzerland

ARTICLE INFO

Article history:
Received 11 June 2014
Received in revised form 5 January 2015
Accepted 5 January 2015
Available online 13 January 2015

Keywords:
Recursive partitioning
Varying coefficient models
Mixed models
Generalized linear models
Longitudinal data analysis
Ordinal regression
Statistical learning

ABSTRACT

A tree-based algorithm for longitudinal regression analysis that aims to learn whether and how the effects of predictor variables depend on moderating variables is presented. The algorithm is based on multivariate generalized linear mixed models and it builds piecewise constant coefficient functions. Moreover, it is scalable for many moderators of possibly mixed scales, integrates interactions between moderators and can handle nonlinearities. Although the scope of the algorithm is quite general, the focus is on its usage in an ordinal longitudinal regression setting. The potential of the algorithm is illustrated by using data derived from the British Household Panel Study, to show how the effect of unemployment on self-reported happiness varies across individual life circumstances.[1]

1. Introduction

Regression analysis for longitudinal responses addresses a wide range of applications, particularly in medical and social sciences. Siddall et al. (2003), for example, analyze long-term effects of injuries on repeatedly measured pain. Likewise, Oesch and Lipps (2013) use repeatedly measured well-being to examine the impact of the transition from employment to unemployment.

When carrying out longitudinal regression analysis, researchers are specifically interested in the impact of moderator variables on selected regression coefficients in order to enhance insights on the studied relation and/or to control for confounding variables. For example, the effect of an injury could depend on age, while that of unemployment could vary across social groups. Herein, we propose a method to learn such moderation in longitudinal data. The method combines a mixed model approach with a regression tree approach. Although the proposed method applies generally in the multivariate generalized linear mixed model (MGLMM) setting, we focus on its usage with longitudinal ordinally scaled responses such as pain or well-being.

The remainder of the article is organized as follows. Sections 1.1 and 1.2 introduce the framework used in the present study and related works. Section 2 explains the method in detail. Section 3 illustrates its potential by using an empirical example and simulation studies and, finally, Section 4 concludes, including addressing the limitations of the proposed method and the software implementation.

* Correspondence to: University of Geneva (CH), Bd du Pont d'Arve 40, Switzerland. Tel.: +41 22 379 98 72.
E-mail addresses: Reto.Buergin@unige.ch (R. Bürgin), Gilbert.Ritschard@unige.ch (G. Ritschard).

[1] R-codes and datasets are available online as supplementary files (see Appendix B).

http://dx.doi.org/10.1016/j.csda.2015.01.003
0167-9473/

1.1. Framework

The proposed algorithm extends multivariate generalized linear mixed models (e.g. Tutz and Hennevogl, 1996) by allowing the fixed coefficients to vary as nonparameterized functions of some moderator variables $Z_1, \ldots, Z_L$. Let $\mathbf{y}_{it}$ denote the $R \times 1$ response vector of individual i at time t, $i = 1, \ldots, N$, $t = 1, \ldots, N_i$. Denote by $\mathbf{X}_{it}$ and $\mathbf{W}_{it}$ the $Q \times P_{\boldsymbol{\beta}}$ and $Q \times P_{\mathbf{b}_i}$ design matrices associated with fixed coefficients $\boldsymbol{\beta}$ and (individual-specific) random coefficients $\mathbf{b}_i$, respectively. Further, denote by $\mathbf{z}_{it}$ the $L \times 1$ vector of moderators, also called *effect modifiers* in the literature (e.g. Hastie and Tibshirani, 1993). MGLMMs link the $Q \times 1$ predictor vector $\boldsymbol{\eta}_{it}$ with the conditional expectation $\boldsymbol{\mu}_{it} = E(\mathbf{y}_{it}|\mathbf{b}_i; \mathbf{X}_{it}, \mathbf{W}_{it}, \mathbf{z}_{it})$ via $\boldsymbol{\mu}_{it} \in \mathbb{R}^R \mapsto \boldsymbol{\eta}_{it} = \mathbf{g}(\boldsymbol{\mu}_{it}) \in \mathbb{R}^Q$, where $\mathbf{g}$ is a known link function. We aim to fit predictor functions of the form

$$\mathcal{M} : \boldsymbol{\eta}_{it} = \mathbf{X}_{it}\boldsymbol{\beta}(\mathbf{z}_{it}) + \mathbf{W}_{it}\mathbf{b}_i, \quad \mathbf{b}_i \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_b). \tag{1}$$

The fixed coefficients $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot), \ldots, \beta_{P_{\boldsymbol{\beta}}}(\cdot))^\top$ of $\mathcal{M}$ are *varying coefficients* that state that the linear effects of the elements of matrix $\mathbf{X}_{it}$ on the expectation of $\mathbf{y}_{it}$ are nonparameterized functions of $\mathbf{z}_{it}$. In the predictor function $\mathcal{M}$, the intercept coefficients are included in $\boldsymbol{\beta}(\cdot)$. Such *varying intercepts* are functions of $\mathbf{z}_{it}$ and estimate the direct effects of $\mathbf{z}_{it}$ on $E(\mathbf{y}_{it}|\cdot)$. In contrast to fixed coefficients, the individual-specific random coefficients $\mathbf{b}_i$ do not depend on $\mathbf{z}_{it}$ in $\mathcal{M}$. Such random coefficients are used to take into account the correlation between repeated responses and could include individual-specific intercepts or slopes over time. As stated in (Eq. (1)), we assume here that the random coefficients are normally, identically and independently distributed with $\mathrm{E}(\mathbf{b}_i) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{b}_i) = \boldsymbol{\Sigma}_b$.

MGLMMs include models with density functions of the multivariate exponential family that, with random coefficients $\mathbf{b}_i$, have the general form

$$f(\mathbf{y}_{it}|\mathbf{b}_i; \boldsymbol{\beta}, \phi) = \exp\left\{\frac{\mathbf{y}_{it}^\top\boldsymbol{\theta}_{it} - b(\boldsymbol{\theta}_{it})}{\phi} + c(\mathbf{y}_{it}, \phi)\right\}, \tag{2}$$

with ϕ the dispersion parameter and $b(\cdot)$ and $c(\cdot)$ family-specific functions. $\boldsymbol{\theta}_{it}$ is the so-called vector of natural parameters. It is here defined as $\boldsymbol{\theta}_{it} = \mathbf{d}(\boldsymbol{\mu}_{it}) = \mathbf{d}(\mathbf{g}^{-1}(\mathbf{X}_{it}\boldsymbol{\beta}(\mathbf{z}_{it}) + \mathbf{W}_{it}\mathbf{b}_i))$, with $\mathbf{d}(\cdot)$ a known, vector-valued function. MGLMMs include, for instance, several univariate models such as the (Gaussian) linear mixed model or the Poisson mixed model. Here, we restrict the consideration of specific models to that of the cumulative logit mixed model, which really requires the multivariate form above.

The cumulative logit mixed model (CLMM). The cumulative logit model (e.g. McCullagh, 1980) is a popular and conceptually simple model for ordinal response variables Y taking ordered categorical values r in $\{1, \ldots, R\}$. It is motivated (e.g. Tutz, 2012) by assuming that Y is a coarse version of a latent continuous variable $Y^* = f(\cdot) + \varepsilon$, with $f(\cdot)$ a function of predictors and ε the error with distribution $\varepsilon \overset{i.i.d.}{\sim} \mathrm{Logistic}(0, 1)$. The connection between the observed ordinal and the latent variable is defined as: $Y = r \Leftrightarrow \theta_{r-1} < Y^* \leq \theta_r$; with $-\infty = \theta_0 < \theta_1 < \cdots < \theta_R = \infty$ the *threshold* coefficients.

The cumulative logit mixed model has been introduced by Hedeker and Gibbons (1994), and Tutz and Hennevogl (1996) exemplified it as a special case of MGLMMs. Here, the CLMM with varying coefficients is defined as follows: Let $\mathbf{y}_{it} = (y_{it1}, \ldots, y_{itR})^\top$ be the response vector of individual i at time t, which is coded as $y_{itr} = 1$ if $Y_{it} = r$ and $y_{itr} = 0$ if $Y_{it} \neq r$. Assume that $\mathbf{y}_{it}$ is an outcome of a multinomial distribution with the conditional probabilities $E(\mathbf{y}_{it}|\mathbf{b}_i; \mathbf{x}_{it}, \mathbf{w}_{it}, \mathbf{z}_{it}) = \boldsymbol{\pi}_{it}$, with $\mathbf{x}_{it}$ and $\mathbf{w}_{it}$ the predictor vectors to be incorporated into the design matrices $\mathbf{X}_{it}$ and $\mathbf{W}_{it}$. The CLMM links the predictor $\boldsymbol{\eta}_{it}$ with the conditional probabilities $\boldsymbol{\pi}_{it}$ via $\eta_{itq} = g_q(\boldsymbol{\pi}_{it}) = \log((\pi_{it1} + \cdots + \pi_{itq})/(1 - \pi_{it1} - \cdots - \pi_{itq})) = \mathrm{logit}(P(Y_{it} \leq q))$ for $q = 1, \ldots, Q = R - 1$. The predictor function is defined as

$$\mathcal{M}_{\mathrm{CLMM}} : \begin{bmatrix} \eta_{it1} \\ \vdots \\ \eta_{itQ} \end{bmatrix} = \begin{bmatrix} 1 & & \mathbf{x}_{it}^\top \\ & \ddots & \mathbf{x}_{it}^\top \\ & 1 & \mathbf{x}_{it}^\top \end{bmatrix} \boldsymbol{\beta}(\mathbf{z}_{it}) + \begin{bmatrix} 1 & \mathbf{w}_{it}^\top \\ \vdots & \vdots \\ 1 & \mathbf{w}_{it}^\top \end{bmatrix} \mathbf{b}_i, \tag{3}$$

where the qth row determines the logits of responding with $\{1, \ldots, q\}$ rather than with $\{q + 1, \ldots, R\}$. The first Q elements of $\boldsymbol{\beta}(\cdot)$ are the varying intercepts, or *varying thresholds* $\theta_1(\cdot), \ldots, \theta_{R-1}(\cdot)$ in terms of the latent variable motivation, that take into account the direct effects of the moderators $\mathbf{z}_{it}$. In order to maintain the order $P(Y_{it} \leq 1) \leq \cdots \leq P(Y_{it} \leq Q)$, these intercepts must satisfy $\beta_1(\mathbf{z}_{it}) \leq \cdots \leq \beta_Q(\mathbf{z}_{it})\ \forall\ (i, t)$. Further, stacking the vectors $\mathbf{x}_{it}^\top$ and $(1, \mathbf{w}_{it}^\top)$ in the design matrices constraints the corresponding effects to be identical for all Q cumulative logits. This constraint, which considerably simplifies the model, is commonly called the *proportional odds assumption* (e.g. McCullagh, 1980) or *parallelism*. For the direct effects of $\mathbf{z}_{it}$, the proportional odds assumption is relaxed in $\mathcal{M}_{\mathrm{CLMM}}$ since the corresponding varying intercepts are logit-specific. Therefore, $\mathcal{M}_{\mathrm{CLMM}}$ can be seen as a *partial proportional odds model* (e.g. Tutz, 2012, Chap. 9.1.3). Note that if $R = 2$, $\mathcal{M}_{\mathrm{CLMM}}$ is equivalent to a logistic mixed model.

The unknown varying coefficients $\boldsymbol{\beta}(\cdot)$ of the predictor function $\mathcal{M}$ (Eq. (1)) are proposed to be approximated by a piecewise constant function, based on *model-based recursive partitioning*, which is conceptually similar to *regression trees* (e.g. Breiman et al., 1984). These two approaches can be distinguished by their aims: regression trees attempt to discover differences in the mean, while model-based recursive partitioning aims to discover differences in the model coefficients. While recursive partitioning has certain drawbacks, particularly that it is a heuristic and may be instable regarding small changes in the data, its advantages for statistical learning are hardly covered by the alternative methods to date (cf. Hastie

et al., 2001, Sec. 10.7). Recursive partitioning is conceptually simple, can handle many inputs (moderators), nonlinearities and interactions, treats inputs of different scales (nominal, continuous etc.) uniformly and yields easily readable outcomes in the form of decision trees.

The algorithm proposed in this study builds on the so-called MOB algorithm of Zeileis et al. (2008), which provides a unified design for splitting and tree size selection based on M-estimation and which has been extended to various models (e.g. Rusch and Zeileis, 2012; Strobl et al., forthcoming). We aim to redesign MOB to fit $\mathcal{M}$ (Eq. (1)) while preserving the algorithm's statistical properties. This redesign involves two adjustments relative to MOB. The first adjustment allows us to include time-varying moderators while maintaining the random effect component. Because MOB fits a tree with unconnected models at the terminal nodes, a split by a time-varying moderator can render impossible the connection between observations of the same individual. Inspired by the algorithms of Hajjem et al. (2011) and Sela and Simonoff (2012), our algorithm builds a closed model that consists of a tree-structured fixed effect component and a global random effect component. By doing so, the observations of an individual are connected with the single set of corresponding random coefficients, regardless of in which nodes these observations fall. The second adjustment tailors the coefficient constancy tests for the variable and tree size selection of MOB to our algorithm.

1.2. Related work

Literature on longitudinal varying coefficient regression refers primarily to spline or kernel regression techniques for modeling the fixed coefficients as functions of time. For example, Tutz and Kauermann (2003) and Zhang (2004) develop generalized linear mixed models with time-varying fixed coefficients, based on local polynomial regression, and Kauermann (2000) proposes an implementation for the marginal cumulative logit model. The tree-based approach for varying coefficients originates from combining linear models and regression trees, e.g., see Quinlan (1992) or Alexander et al. (1996). Wang and Hastie (2014) formalize their tree-based algorithm most explicitly as an approach for varying coefficient regression and provide an in-depth comparison of the tree-based and spline/kernel methods. One of the rare explicit tree-based techniques for longitudinal varying coefficient regression is that of Su et al. (2011), focusing on moderation on a single predictor.

Our research also intersects with the recent discussion on longitudinal regression trees based on mixed models. The first implementation may be that of Abdolell et al. (2002), fitting unconnected linear mixed models for subspaces of a single variable. The MERT (Hajjem et al., 2011) and RE-EM Tree (Sela and Simonoff, 2012) algorithms aim to approximate general fixed effect components by a piecewise constant function. Similar to our approach, these algorithms fit closed models where only the fixed effect component is built algorithmically. Hajjem (2010) extends MERT for generalized linear mixed models. Eo and Cho (2014) propose with MELT an implementation focusing on trends over time and building on the splitting procedure of GUIDE (Loh, 2002). Specifically, they fit a tree with unconnected linear mixed models that specify polynomials of time in the fixed effect component. Our contribution is extending the scope of longitudinal regression trees based on mixed models to general varying coefficient regression and proposing a new splitting procedure based on MOB (Zeileis et al., 2008). MERT and RE-EM Tree focus, in our terminology, on the case where only varying intercepts are specified and all covariates are assigned to vector $\mathbf{z}_{it}$. MELT, in turn, focuses on the case where $\mathbf{X}_{it}$ represents a polynomial expansion of time and the remaining covariates are assigned to vector $\mathbf{z}_{it}$. Unlike the above tree approaches, our algorithm does not include auto-correlated errors and, unlike MELT, it does not fit separate random coefficients for every terminal node.

2. Method

2.1. Piecewise constant approximation for varying coefficients

The algorithm approximates the varying coefficients $\boldsymbol{\beta}(\cdot)$ of $\mathcal{M}$ (Eq. (1)) by using a vectorial piecewise constant function. Consider a partition of the value space $\mathcal{Z}_1 \times \cdots \times \mathcal{Z}_L$ of the L moderators $Z_1, \ldots, Z_L$ into M (terminal) nodes $\{\mathcal{B}_1, \ldots, \mathcal{B}_M\}$. The approximating predictor function is

$$\widehat{\mathcal{M}} : \boldsymbol{\eta}_{it} = \sum_{m=1}^{M} 1(\mathbf{z}_{it} \in \mathcal{B}_m)\mathbf{X}_{it}\boldsymbol{\beta}_m + \mathbf{W}_{it}\mathbf{b}_i. \tag{4}$$

The right-hand side of $\widehat{\mathcal{M}}$ is linear and states that the elements of $\boldsymbol{\beta}(\cdot)$ may vary across nodes $\mathcal{B}_m$, but that they remain constant within nodes. The total vector of unknown coefficients of $\widehat{\mathcal{M}}$ is $\boldsymbol{\gamma} := (\boldsymbol{\beta}^\top, \mathrm{vec}(\boldsymbol{\Sigma}_b^{1/2})^\top)^\top$, with length $P_{\boldsymbol{\gamma}} := MP_{\boldsymbol{\beta}} + P_{\mathbf{b}_i}(P_{\mathbf{b}_i} + 1)/2$. For some MGLMMs, there is also an additional dispersion parameter ϕ.

Estimation. The predictor function $\widehat{\mathcal{M}}$ can be estimated by using the techniques for generalized linear mixed models (e.g. Tutz, 2012, Chap. 14.3). We focus on the direct maximization of the marginal log-likelihood equation by using numeric integration. To simplify the integral of that equation, the random coefficients $\mathbf{b}_i$ are standardized as $\mathbf{a}_i = \boldsymbol{\Sigma}_b^{-1/2}\mathbf{b}_i$ so that $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_Q)$. As a consequence, the Cholesky decomposition $\boldsymbol{\Sigma}_b^{1/2}$ is estimated instead of $\boldsymbol{\Sigma}_b$. The marginal likelihood with standardized random coefficients is

$$\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{N} \log L_i(\boldsymbol{\gamma}) = \sum_{i=1}^{N} \log \int \prod_{t=1}^{N_i} f(\mathbf{y}_{it}|\mathbf{a}_i; \boldsymbol{\gamma})\phi(\mathbf{a}_i)\, d\mathbf{a}_i, \tag{5}$$

where $f(\mathbf{y}_{it}|\mathbf{a}_i;\boldsymbol{\gamma})$ is the family-specific conditional density of $\mathbf{y}_{it}$ and $\phi(\cdot)$ is the multivariate normal density function. To maximize $\ell(\boldsymbol{\gamma})$ of (Eq. (5)), we solve the score equations $\sum_{i=1}^{N}\frac{\partial \log L_i(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} := \sum_{i=1}^{N}\mathbf{u}_i(\boldsymbol{\gamma}) = \mathbf{0}$ for $\boldsymbol{\gamma}$, where

$$\mathbf{u}_i(\boldsymbol{\gamma}) = \frac{1}{L_i(\boldsymbol{\gamma})}\int \sum_{t=1}^{N_i}\frac{\frac{\partial}{\partial\eta}f(\mathbf{y}_{it}|\mathbf{a}_i;\boldsymbol{\gamma})\frac{\partial}{\partial\boldsymbol{\gamma}}\boldsymbol{\eta}_{it}}{f(\mathbf{y}_{it}|\mathbf{a}_i;\boldsymbol{\gamma})}\prod_{t=1}^{N_i}f(\mathbf{y}_{it}|\mathbf{a}_i;\boldsymbol{\gamma})\phi(\mathbf{a}_i)\,d\mathbf{a}_i \tag{6}$$

is a $P_{\boldsymbol{\gamma}}\times 1$ vector. Our software solves these equations by using Fisher's scoring algorithm with Gauss–Hermite quadrature to approximate the integral in (Eq. (6)). Note that the score equations can be expressed as the sum of the observation-scores, $\sum_{i=1}^{N}\mathbf{u}_i(\boldsymbol{\gamma}) = \sum_{i=1}^{N}\sum_{t=1}^{N_i}\mathbf{u}_{it}(\boldsymbol{\gamma})$, where

$$\mathbf{u}_{it}(\boldsymbol{\gamma}) = \frac{1}{L_i(\boldsymbol{\gamma})}\int \frac{\frac{\partial}{\partial\eta}f(\mathbf{y}_{it}|\mathbf{a}_i;\boldsymbol{\gamma})\frac{\partial}{\partial\boldsymbol{\gamma}}\boldsymbol{\eta}_{it}}{f(\mathbf{y}_{it}|\mathbf{a}_i;\boldsymbol{\gamma})}\prod_{t=1}^{N_i}f(\mathbf{y}_{it}|\mathbf{a}_i;\boldsymbol{\gamma})\phi(\mathbf{a}_i)\,d\mathbf{a}_i. \tag{7}$$

Estimating CLMMs. For the CLMM, the conditional density of $\mathbf{y}_{it}$ is $f(\mathbf{y}_{it}|\mathbf{a}_i;\boldsymbol{\gamma}) = \prod_{r=1}^{R}\pi_{itr}^{y_{itr}} = \prod_{r=1}^{R}(\frac{e^{\eta_{itr}}}{1+e^{\eta_{itr}}} - \frac{e^{\eta_{it,r-1}}}{1+e^{\eta_{it,r-1}}})^{y_{itr}}$, where $\eta_{it0} = -\infty$ and $\eta_{itR} = \infty$. The used fitting function `olmm` of the R package **vcrpart** (Bürgin, 2015) solves the score equations (Eq. (6)) by using initial values that respect the order $P(Y_{it}\leq 1)\leq\cdots\leq P(Y_{it}\leq Q)$. This procedure generally works well, but problems could arise if some response categories occur very rarely. Fahrmeir and Tutz (2001) mention that the procedure can fit inadmissible thresholds if these are very similar. To avoid this, they propose a reparameterization that could be considered to improve `olmm`. Further, Kosmidis (2014) points out that coefficients can diverge to infinity and therefore proposes an improved estimator. Ad hoc solutions for both problems could be to merge response categories or to specify a sufficiently large minimum node size (see Section 2.4). Finally, the number of quadrature points for approximating the integral in (Eq. (6)) can impact the accuracy of the fit. Higher numbers increase the accuracy, however, at the cost of computational time. `olmm` allows to control the number of points manually and uses a default of seven.

2.2. Algorithm

The predictor function $\widehat{\mathcal{M}}$ (Eq. (4)) is a good approximation for $\mathcal{M}$ (Eq. (1)) if the true coefficient functions are fairly constant within the nodes. To find such nodes, we propose a breath-first search algorithm (e.g. Russell and Norvig, 2003) that in each iteration splits one of the current M nodes into two. Splitting requires three selections in each step: a node; a moderator; and a split in the selected variable. The constraint is the maintenance of the global random coefficients, on the basis of which a closed model including all observations must be fitted at any stage.

Algorithm 1: Fitting tree-based varying coefficients in MGLMMs

Input: $\alpha\in[0,1]$, e.g., $\alpha = 0.05/L$

Initialize $\mathcal{B}_1\leftarrow\mathcal{Z}_1\times\cdots\times\mathcal{Z}_L$ and $M\leftarrow 1$

repeat

1 Fit the MGLMM with the predictor function

$$\boldsymbol{\eta}_{it} = \sum_{m=1}^{M}1(\mathbf{z}_{it}\in\mathcal{B}_m)\mathbf{X}_{it}\boldsymbol{\beta}_m + \mathbf{W}_{it}\mathbf{b}_i.$$

2 Test for the constancy of the coefficients $\boldsymbol{\beta}_m$ separately for each variable Z_l, $l = 1,\ldots,L$, in each node $\mathcal{B}_m$, $m = 1,\ldots,M$. This yields $L\times M$ p-values, $p_{11},\ldots,p_{LM}$, for rejecting coefficient constancy.

if $p_{min} := \min(p_{11},\ldots,p_{LM})\leq\alpha$ **then**

3 Select the variable Z_l and node $\mathcal{B}_s$ where $p_{ls} = p_{min}$ **foreach** *unique candidate split* $\triangle_k$ *in* $\{z_{lit} : \mathbf{z}_{it}\in\mathcal{B}_s\}$ *dividing* $\mathcal{B}_s$ *into two nodes* $\mathcal{B}_{sk1}$ *and* $\mathcal{B}_{sk2}$ **do**

4 Compute $\hat{\ell}_{\triangle_k} = \max_{\boldsymbol{\gamma}}\ell_{\triangle_k}(\boldsymbol{\gamma})$ of the MGLMM

$$\boldsymbol{\eta}_{it} = \sum_{m\neq s}^{M}1(\mathbf{z}_{it}\in\mathcal{B}_m)\mathbf{X}_{it}\boldsymbol{\beta}_m + \sum_{m=1}^{2}1(\mathbf{z}_{it}\in\mathcal{B}_{skm})\mathbf{X}_{it}\boldsymbol{\beta}_{sm} + \mathbf{W}_{it}^{\top}\mathbf{b}_i.$$

end

5 Split $\mathcal{B}_s$ by $\hat{\triangle} = \arg\max_{\triangle_k}\hat{\ell}_{\triangle_k}$ and set $M\leftarrow M+1$.

end

until $p_{min} > \alpha$

Algorithm 1 summarizes the proposed algorithm. Varying coefficients are fitted separately on an increasing number of small nodes until the tests in Step 2 accept coefficient constancy, for all moderators in all nodes. These tests are also used in each step to select the node and variable simultaneously, while the split in the variable is selected by using exhaustive search.

In Section 2.3 it turns out that the constancy tests for Step 2 must be adjusted, while splitting entirely based on exhaustive search (e.g. Wang and Hastie, 2014) could be applied straightforwardly. We implement these tests for statistical and computational reasons. Statistically, the variable selection based on these tests is not biased towards moderators with many splits, as it is with exhaustive search (cf. Hothorn et al., 2006). Computationally, the advantage is that with such tests the algorithm must refit the model for the splits in the selected variable and node only. By contrast, full exhaustive search requires refitting the model for the splits in all moderators and all nodes, the number of which increases in each iteration.

2.3. Coefficient constancy tests for variable, node and tree size selection

Coefficient constancy tests have been studied extensively in econometrics (e.g. Nyblom, 1989; Andrews, 1993). Although these tests, often called *structural change tests*, have been developed to examine coefficient constancy over time, they naturally extend to other variables. For our purposes, it is computationally convenient to focus on score-based tests, such as the *M-fluctuation* tests of Zeileis and Hornik (2007), which merely require us to estimate the model under the H_0 hypothesis of coefficient constancy. Specifically, we want to use the observation-scores $\hat{\mathbf{u}}_{it} := \mathbf{u}_{it}(\hat{\boldsymbol{\beta}})$ of (Eq. (7)), which allows testing fixed coefficient constancy with respect to both time-varying and time-invariant moderators. Thereby, the remaining coefficients $\boldsymbol{\Sigma}_b$ and ϕ are treated as nuisance parameters. In the following, we summarize the M-fluctuation tests for multivariate generalized linear models (without random coefficients) and introduce two preparatory steps for their use in Algorithm 1. The first step linearly transforms the observation-scores $\hat{\mathbf{u}}_{it}$ to remove intra-individual correlations. The second step extracts and mean-centers the subsets of these scores to apply the tests nodewise. The aim of both steps is to ensure that the transformed observation-scores have approximately the same first two moments and covariances as have the scores of models without random coefficients. While asymptotic aspects are not considered, a comprehensive simulation study is presented in Section 3.2.

2.3.1. Coefficient constancy tests for multivariate generalized linear models

For a complete description of these M-fluctuation tests, see Zeileis and Hornik (2007). Here, we summarize the M-fluctuation for multivariate generalized linear models. Let $\mathbf{y}_i$, $i = 1, \ldots, N$ be the $R \times 1$ response vectors and $\mathbf{X}_i$ the corresponding $Q \times P_\beta$ design matrices. Assume that $\mathbf{y}_i$ given $\mathbf{X}_i$ follows a distribution of the multivariate exponential family, and that the conditional expectation is determined by $\mathbf{g}(E(\mathbf{y}_i|\mathbf{X}_i)) = \mathbf{X}_i\boldsymbol{\beta}_i$, with $\mathbf{g}$ a known link function. In particular, we test $H_0 : \boldsymbol{\beta}_i = \boldsymbol{\beta}_1$ for all i against the alternative that the coefficients $\boldsymbol{\beta}_i$ change with the values of a variable Z. Using M-fluctuation tests for this approach requires estimating the model under H_0, namely maximizing the likelihood or solving the score equations $\sum_{i=1}^{N} \boldsymbol{\psi}_i(\boldsymbol{\beta}_1) = \mathbf{0}$ for $\boldsymbol{\beta}_1$. By using the fitted model, the cumulative process of the estimated scores along the values of Z,

$$\boldsymbol{\Psi}_N(\tau) = \frac{1}{\sqrt{N}} \sum_{i=1}^{\lfloor \tau N \rfloor} \hat{\boldsymbol{\psi}}_{\sigma(z_i)} \quad (0 \le \tau \le 1), \tag{8}$$

is examined for divergences from its expectation $\mathbf{0}$. The $\hat{\boldsymbol{\psi}}_i = \boldsymbol{\psi}_i(\hat{\boldsymbol{\beta}}_1)$ are the estimated scores and $\sigma(z_i)$ is the ordering permutation giving the antirank of observation z_i in vector $(z_1, \ldots, z_N)$. $\boldsymbol{\Psi}_N$ is computed as a P_β-dimensional sequence of length $N + 1$ that starts and ends with zero. Assuming that under H_0: (i) $\mathrm{E}(\hat{\boldsymbol{\psi}}_i) = \mathbf{0}\ \forall\ i$; (ii) $\mathrm{Var}(\hat{\boldsymbol{\psi}}_i) = \mathrm{Var}(\hat{\boldsymbol{\psi}}_1)\ \forall\ i$; and (iii) $\mathrm{Cov}(\hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_{i'}) = \mathrm{Cov}(\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_2)\ \forall\ i \ne i'$; which requires that the predictors are *stationary* over the tested variable (cf. Hjort and Koning, 2002), it can be derived (see Appendix A.1) that $\mathrm{Cov}(\hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_{i'}) = -\frac{1}{N-1}\mathrm{Var}(\hat{\boldsymbol{\psi}}_1)$ for $i \ne i'$ and, consequently, $\mathrm{Cov}(\boldsymbol{\Psi}_N(\tau_1), \boldsymbol{\Psi}_N(\tau_2)) = \frac{\lfloor N\tau_1 \rfloor (N - \lfloor N\tau_2 \rfloor)}{N(N-1)}\mathrm{Var}(\hat{\boldsymbol{\psi}}_1)$ for $\tau_1 < \tau_2$. Moreover, under regularity conditions, $\boldsymbol{\Psi}_N$ can be shown (e.g. Zeileis and Hornik, 2007) to converge under H_0 to a limit process $\boldsymbol{\Psi}^0$ as N tends to infinity. This limit process has covariance $\mathrm{Cov}(\boldsymbol{\Psi}^0(\tau_1), \boldsymbol{\Psi}^0(\tau_2)) = \tau_1(1 - \tau_2)\mathrm{Var}(\boldsymbol{\psi}(\boldsymbol{\beta}_1))$, where $\mathrm{Var}(\boldsymbol{\psi}(\boldsymbol{\beta}_1))$ is the variance of scores at the true $\boldsymbol{\beta}_1$. In other words, $\boldsymbol{\Psi}_N$ converges to a linear transformation of P_β independent Brownian bridges $\mathbf{B}^0$. Likewise, the *standardized cumulative score process*

$$\ddot{\boldsymbol{\Psi}}_N(\tau) = \hat{\mathbf{J}}^{-1/2}\boldsymbol{\Psi}_N(\tau) \quad (0 \le \tau \le 1), \tag{9}$$

where $\hat{\mathbf{J}}$ is an estimate for $\mathrm{Var}(\boldsymbol{\psi}(\boldsymbol{\beta}_1))$, typically $\hat{\mathbf{J}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{\psi}}_i \hat{\boldsymbol{\psi}}_i^\top$, can be shown to converge to P_β independent Brownian bridges. To construct a test, a suitable scalar statistic $\lambda(\ddot{\boldsymbol{\Psi}}_N)$ is applied, the H_0 distribution of which is simply the limiting distribution of $\lambda(\mathbf{B}^0)$.

Test statistics. Our algorithm adopts the statistics used in MOB (Zeileis et al., 2008). For continuous variables, we use the Lagrange multiplier statistic "supLM" of Andrews (1993), which is designed to capture coefficient shifts at a single, unknown cutpoint. It is defined as

$$\lambda_{supLM}\left(\ddot{\boldsymbol{\Psi}}_N\right) = \max_{i=\underline{i},\ldots,\bar{i}} \left(\frac{i}{N} \cdot \frac{N-i}{N}\right)^{-1} \left\| \ddot{\boldsymbol{\Psi}}_N\,(i/N) \right\|_2^2, \tag{10}$$

i.e., as the maximum of the squared L_2 norm of $\ddot{\boldsymbol{\Psi}}_N$ in interval $[\underline{i}, \bar{i}]$ (e.g., $[\lceil N/10 \rceil,\ N - \lceil N/10 \rceil]$). Asymptotically, λ_{supLM} is distributed as the supremum of a squared, P_β-dimensional tied-down Bessel process $\sup_\tau (\tau(1-\tau))^{-1} \left\| \mathbf{B}^0(\tau) \right\|_2^2$.

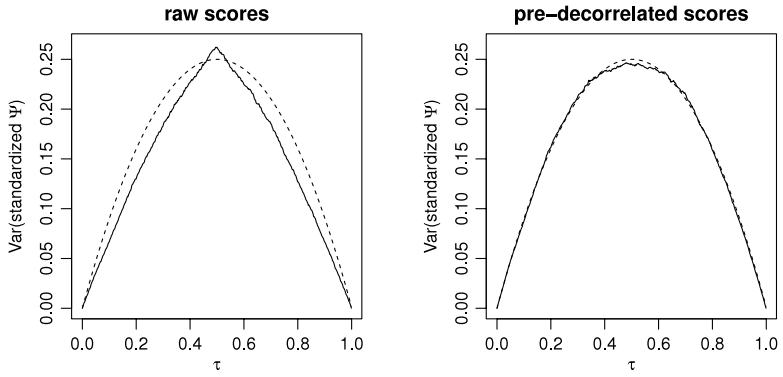

Fig. 1. Case example: variance of standardized cumulative score processes $\breve{\Psi}$. *Solid lines*, variance of simulated processes based on the raw scores (left panel) and based on the pre-decorrelated scores (right panel); *dashed lines*, the variance of a Brownian bridge. In the right panel the lines cover each other.

For categorical variables, we use the χ^2-type statistic of Hjort and Koning (2002), which is designed to capture overall between-category coefficient variation. For variables with categories c in $\{1, \ldots, C\}$, it is defined as

$$\lambda_{\chi^2}\left(\breve{\boldsymbol{\Psi}}_N\right) = \sum_{c=1}^{C} \frac{1}{N_c N} \left\| \Delta_c\left(\breve{\boldsymbol{\Psi}}_N(i/N)\right)\right\|_2^2, \tag{11}$$

where N_c is the number of observations in category c and $\Delta_c(\breve{\boldsymbol{\Psi}}_N)$ is the increment of $\breve{\boldsymbol{\Psi}}_N$ over the observations of category c. Under H_0, λ_{χ^2} is χ^2-distributed with $P_{\boldsymbol{\beta}}(C-1)$ degrees of freedom.

2.3.2. Pre-decorrelating the observation-scores of MGLMMs

Substituting the scores $\hat{\boldsymbol{\psi}}_i$ in (Eq. (8)) with the scores $\hat{\mathbf{u}}_{it}$ of (Eq. (7)) is misleading. While the $\hat{\boldsymbol{\psi}}_i$'s depend on each other only via the constraint $\sum_i \hat{\boldsymbol{\psi}}_i = \mathbf{0}$, the $\hat{\mathbf{u}}_{it}$'s are additionally intra-individually linked via (Eq. (7)). Therefore, $\mathrm{Cov}(\hat{\mathbf{u}}_{it}, \hat{\mathbf{u}}_{it'})$ for $t \neq t'$ is hardly equal to $\mathrm{Cov}(\hat{\mathbf{u}}_{it}, \hat{\mathbf{u}}_{i't'})$ for $i \neq i'$. To illustrate an outcome from using the *raw* scores $\hat{\mathbf{u}}_{it}$ in M-fluctuation tests and to motivate the *pre-decorrelation* transformation below, we consider the following case example: We repeatedly (5,000 times) generated responses $\mathbf{y}_{it}$ with $i = 1, \ldots, 50$ and $t = 1, \ldots, 10$, from the logistic mixed model: $\mathcal{M}_{ex} : \mathrm{logit}(P(Y_{it} = 1)) = \beta_0 + b_i$, with $\beta_0 = 0$ and $b_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$; fitted the true model $\mathcal{M}_{ex}$ on these data; and computed $\breve{\Psi}$ of (Eq. (9)) from the raw scores $u_{it}(\hat{\beta}_0)$ and from the pre-decorrelated scores $u^*_{it}(\hat{\beta}_0)$. Specifically, to compute $\breve{\Psi}$, we first cumulated scores with indices $t = 1, \ldots, 5$, and then scores with indices $t = 6, \ldots, 10$. Since we fit the true model on the data, the computed processes $\breve{\Psi}$ should be distributed as a Brownian bridge.

Fig. 1 compares the variance of a Brownian bridge with the variance of the simulated processes $\breve{\Psi}$, based on the raw scores (left) and the pre-decorrelated scores (right). The plots suggest that the processes based on the pre-decorrelated scores are distributed as a Brownian bridge, but not the processes based on the raw scores. Further experiments revealed that the variance pattern of processes from raw scores depends on the cumulative order, and that the triangular pattern above is a special case.

The proposed pre-decorrelated scores $\hat{\mathbf{u}}^*_{it}$ are computed by using the linear within-individual transformation

$$\hat{\mathbf{u}}^*_{it} = \hat{\mathbf{u}}_{it} + \mathbf{T} \sum_{t'=1, t' \neq t}^{N_i} \hat{\mathbf{u}}_{it'}, \tag{12}$$

where $\mathbf{T}$ is the $MP_{\boldsymbol{\beta}} \times MP_{\boldsymbol{\beta}}$ transformation matrix, such that under H_0

$$\mathrm{E}(\hat{\mathbf{u}}^*_{it}) = \mathbf{0} \quad \forall\, i, t, \tag{13}$$

$$\mathrm{Var}(\hat{\mathbf{u}}^*_{it}) = \mathrm{Var}(\hat{\mathbf{u}}^*_{11}) \quad \forall\, i, t \quad \text{and} \tag{14}$$

$$\mathrm{Cov}(\hat{\mathbf{u}}^*_{it}, \hat{\mathbf{u}}^*_{it'}) = \mathrm{Cov}(\hat{\mathbf{u}}^*_{it}, \hat{\mathbf{u}}^*_{i't''}) = -\frac{1}{\sum_{i=1}^{N} N_i - 1} \mathrm{Var}(\hat{\mathbf{u}}^*_{11}), \tag{15}$$

for all $(i, t) \neq (i, t')$ and $(i, t) \neq (i', t'')$. The transformation forces the expectation, the variance and the covariance of $\mathbf{u}^*_{it}$'s to comply with those of the $\boldsymbol{\psi}_i$'s, see assumptions (i)–(iii). Therefore, if such a matrix $\mathbf{T}$ exists, we can assume that the covariance of processes $\breve{\boldsymbol{\Psi}}_N$ (Eq. (9)) based on the $\mathbf{u}^*_{it}$'s is the same as that based on the $\boldsymbol{\psi}_i$'s.

Balanced data. For balanced data where $N_i = N_1 \,\forall\, i$, the scores are symmetrical in the sense that every score $\hat{\mathbf{u}}_{it}$ relates to $N_1 - 1$ "internal" counterparts $\{\hat{\mathbf{u}}_{it'} : t \neq t'\}$ via (Eq. (7)) and the constraint $\sum_{i,t} \hat{\mathbf{u}}_{it} = 0$; and to $(N-1)N_1$ "external" counterparts only via $\sum_{i,t} \hat{\mathbf{u}}_{it} = 0$. Therefore, we assume that under H_0 (iv) $\mathrm{E}(\hat{\mathbf{u}}_{it}) = \mathbf{0} \;\forall\; (i,t)$; (v) $\mathrm{Var}(\hat{\mathbf{u}}_{it}) = \mathrm{Var}(\hat{\mathbf{u}}_{11}) \;\forall\; (i,t)$; (vi) $\mathrm{Cov}(\hat{\mathbf{u}}_{it}, \hat{\mathbf{u}}_{it'}) = \mathrm{Cov}(\hat{\mathbf{u}}_{11}, \hat{\mathbf{u}}_{12}) \;\forall\; i,\ t \neq t'$; and (vii) $\mathrm{Cov}(\hat{\mathbf{u}}_{it}, \hat{\mathbf{u}}_{i't'}) = \mathrm{Cov}(\hat{\mathbf{u}}_{11}, \hat{\mathbf{u}}_{21}) \;\forall\; t, t',\ i \neq i'$. Under these assumptions, $\mathbf{T}$ is found by solving $\widehat{\mathrm{Cov}}(\hat{\mathbf{u}}^*_{11}, \hat{\mathbf{u}}^*_{12}) - \widehat{\mathrm{Cov}}(\hat{\mathbf{u}}^*_{11}, \hat{\mathbf{u}}^*_{21}) = \mathbf{0}$, see Appendix A.2 for details. The resulting multiple quadratic equation depends on N_1, $\widehat{\mathrm{Var}}(\hat{\mathbf{u}}_{11})$, $\widehat{\mathrm{Cov}}(\hat{\mathbf{u}}_{11}, \hat{\mathbf{u}}_{12})$ and $\widehat{\mathrm{Cov}}(\hat{\mathbf{u}}_{11}, \hat{\mathbf{u}}_{21})$ and can be solved numerically, e.g., with Newton's method.

Unbalanced data. For unbalanced data, the scores $\hat{\mathbf{u}}_{it}$ are not symmetrical in the sense above. Therefore, the assumptions (v)–(vii) would hardly hold. To use the solution for $\mathbf{T}$ for balanced data, we construct a balanced score matrix by recomputing the scores of individuals i with $N_i < N_{max} = \max_{i'} N_{i'}$ under the inclusion of $N_{max} - N_i$ imputed values. The imputation procedure is described in Appendix A.3, where the crucial point is the generation of response values by means of the model under H_0. Denote by $\hat{\tilde{\mathbf{u}}}_{i1}, \ldots, \hat{\tilde{\mathbf{u}}}_{iN_i}$ the recomputed scores for individual i and by $\hat{\tilde{\mathbf{u}}}_{i,N_i+1}, \ldots, \hat{\tilde{\mathbf{u}}}_{iN_{max}}$ the scores corresponding to the imputed observations. The pre-decorrelation (Eq. (12)) for *incomplete* individuals yields

$$\hat{\mathbf{u}}^*_{it} = \hat{\tilde{\mathbf{u}}}_{it} + \mathbf{T}\left(\sum_{t'=1,t'\neq t}^{N_i} \hat{\tilde{\mathbf{u}}}_{it'} + \sum_{t'=N_i+1}^{N_{max}} \hat{\tilde{\mathbf{u}}}_{it'}\right). \tag{16}$$

Matrix $\mathbf{T}$ is based on the raw scores $\hat{\mathbf{u}}_{it}$ of individuals with $N_i = \max_{i'} N_{i'}$ and the recomputed scores $\hat{\tilde{\mathbf{u}}}_{it}$ of individuals with $N_i < \max_{i'} N_{i'}$.

The proposed solution for unbalanced data perturbs the tests because of the randomness involved in the imputation. To account for this, we repeat the entire test procedure (e.g., five times) and use the average of the resulting p-values.

2.3.3. Nodewise tests

Step 2 in Algorithm 1 processes the coefficient constancy tests separately for each variable $Z_1, \ldots, Z_L$ in each node $\mathcal{B}_1, \ldots, \mathcal{B}_M$. The nodewise implementation has two advantages: (i) it is computationally convenient to select the node to split, and (ii) it eliminates the dependency between the node predictor "$1(\mathbf{z}_{it} \in \mathcal{B}_m)$" and the variables $Z_1, \ldots, Z_L$ that violate the *stationarity* assumption.

The procedure for testing coefficient constancy regarding a variable Z_l in a node $\mathcal{B}_m$ involves five steps. First, we compute the $N_T = \sum_{i=1}^{N} N_i$ pre-decorrelated scores $\hat{\mathbf{u}}^*_{it}$. Second, we extract from the obtained $N_T \times MP_\beta$ score matrix $\hat{\mathbf{U}}^*$ and the $N_T \times 1$ vector $\mathbf{z}_l$ the N_m observations corresponding to node $\mathcal{B}_m$. Third, to ensure that the sum of scores is zero and that the tests are independent across nodes (see Appendix A.4), we mean-center the score matrix by column. Fourth, we compute $\breve{\boldsymbol{\Psi}}_{N_m}$ of (Eq. (9)) by substituting the $\hat{\boldsymbol{\psi}}_i$'s of Section 2.3.1 with the elements of the column-centered node score matrix. Finally, we extract the P_β columns of $\breve{\boldsymbol{\Psi}}_{N_m}$ corresponding to $\boldsymbol{\beta}_m$ and compute the test statistic and the p-value.

2.4. Further details

Splitting. Step 4 of Algorithm 1 cycles through the unique candidate splits in the values of the selected moderator Z_l in the selected node $\mathcal{B}_s$. Splits for ordinal or continuous moderators are based on rules of the form $\{\text{is } z_{lit} \leq \zeta_k?\}$, with ζ_k the unique values in the set $\{z_{lit} : \mathbf{z}_{it} \in \mathcal{B}_s\}$. For nominal moderators, we use rules of the form $\{\text{is } z_{lit} \in \zeta_k?\}$ where the ζ_k's are groupings of the categories in $\{z_{lit} : \mathbf{z}_{it} \in \mathcal{B}_s\}$. Thereby, to have sufficient observations to estimate the nodewise coefficients, we evaluate by default only those splits that yield nodes with a minimum size of 50 observations. For computational efficiency, our software also implements the split reduction techniques of Wang and Hastie (2014) that provide control on the maximum number of evaluated splits at each iteration.

Tree size. The significance threshold α is the principal tuning parameter to control the tree size. Conventionally, this parameter is interpreted as the probability of a type I error, i.e., the probability of falsely rejecting coefficient constancy in a node. To account for the multiple test setting, a nodewise Bonferroni correction is applied, and for a 5% value probability a value of 0.05 divided by the number of moderators would be used. An alternative to determine α, which is not investigated in more detail here, is the use of cross-validation (e.g. Hastie et al., 2001, Sec. 7).

Alternative specifications. Alternative fixed effect components to those in $\mathcal{M}$ (Eq. (1)) can be specified by means of simple modifications. For instance, single fixed coefficients — without moderation — can be integrated by omitting them from the splitting procedure. In the latter case, the component $\mathbf{X}_{it}\boldsymbol{\beta}(\mathbf{z}_{it})$ of $\mathcal{M}$ would be decomposed as $\mathbf{X}_{1it}\boldsymbol{\beta}_1(\mathbf{z}_{it}) + \mathbf{X}_{2it}\boldsymbol{\beta}_2$. This approach can be useful to define a non-zero mean for a random slope.

Time-varying moderators. Time-varying covariates such as the education level are common in longitudinal studies, e.g., see the empirical example below (Table 1). To allow such time-varying covariates, we use a closed model approach that ensures that the random effect component is maintained when splitting by time-varying moderators. However, the inclusion of time-varying moderators may raise interpretability problems. For example, when focusing on varying trends over time as do Eo and Cho (2014), splits in time-varying variables mean that individuals can switch between different static trends, which could be difficult to communicate. In such cases, it may be better to omit the time-varying moderators, or to summarize them as time-invariant variables, e.g., see Eo and Cho (2014, Sec. 2.5).

Table 1
Moderator variables for the analysis of the effect of unemployment on happiness. Abbreviations: ti = time-invariant; tv = time-varying.

	Name	Label	Description	
1	Gender	GENDER	ti	0, female; 1, male
2	Age	AGE	tv	16, …, 64 years
3	Education	EDU	tv	0, lower; 1, upper; 2, tertiary
4	Lives with spouse	SPINHH	tv	0, no; 1, yes
5	Household income	HHINC	tv	0.55, …, 4.65 (equivalence scale)
6	Time unemployed	TUE	tv	−3, …, 2 years
7	Regional unemp.	UEREG	tv	0.05, …, 10.2%
8	Sectoral unemp.	UESEC	tv	0, …, 13.6%
9	Financial situation	FISIT	tv	0, finding it very difficult; …; 4, living comfortably
10	Spouse has job	SPJB	tv	0, no partner; 1, no; 2, yes
11	Marital status	MASTAT	tv	0, never married; 1, married; …; 5, separated
12	Head of household	HOH	tv	0, no, 1; yes
13	Number of children	NCHILD	tv	0, …, 7
14	Resp. for child < 16	RACH16	tv	0, no; 1, yes

3. Results

3.1. Empirical example

To illustrate the scope of the method, we study the effect of the transition from employment to unemployment on self-reported *happiness* (on a scale of 1 = "Much less", 2 = "Less so", 3 = "Same as usual" and 4 = "More than usual") by using data derived from the British Household Panel Survey (Taylor et al., 2010). Specifically, we extracted a subset of cases from the first 18 yearly waves (1991–2008). This subset includes those respondents who experienced at least one switch from (self-) employment to unemployment between two consecutive waves. To isolate the effect of the transition, we consider for each retained respondent a single trajectory formed by the up-to-three-year employment period before the unemployment spell and the up-to-three-year unemployment spell that followed employment. The individual periods therefore include between two and six observations. For example, the period of a respondent who was first a student, then worked for two years, then was unemployed for a year, and then found another job would consist of the two years of employment and the year of unemployment. Alternatively, the period of a respondent that worked for 12 years before being unemployed for five would consist of the last three years of employment and the first three of unemployment. If a respondent experienced multiple transitions, only the first was retained. The used data include 1487 respondents and a total of 5054 observations.

To estimate the effect of the transition to unemployment, we use cumulative logit mixed models for *happiness*, Y, including the dummy coded fixed coefficient predictor *unemployed*, UE, and, to take into account intra-individual correlation, respondent-specific random intercepts.

We use our algorithm to select and incorporate variables that moderate the effect of *unemployed* and/or have a direct effect on *happiness*. Following Oesch and Lipps (2013, see below) and own considerations, we retained the 14 variables listed in Table 1. First, we consider that all the 14 variables potentially moderate the effect of *unemployed* and/or affect *happiness* directly. This leads us to the varying coefficients CLMM

$$\mathcal{M}_1 : \text{logit}(P(Y_{it} \leq q)) = \beta_q(\mathbf{z}_{it}) + \text{UE}_{it}\beta_4(\mathbf{z}_{it}) + b_i, \quad b_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_b),$$

for $q = 1, 2, 3$, where $\mathbf{z}_{it} = (\text{GENDER}_{it}, \ldots, \text{RACH16}_{it})^\top$ is the 14×1 vector of moderators. In $\mathcal{M}_1$, the direct effects of the moderators are estimated by the varying intercepts $\beta_1(\cdot)$, $\beta_2(\cdot)$ and $\beta_3(\cdot)$ and the moderation effects by the varying coefficient $\beta_4(\cdot)$. We fitted $\mathcal{M}_1$ by using $\alpha = 0.05$ plus the Bonferroni correction. The computation time was 84 s with a 3.5 GHz processor.

Fig. 2 shows the fitted tree structure and the nodewise coefficients of the fit for model $\mathcal{M}_1$. The node panels report the nodewise estimates for the varying coefficients and the corresponding z-values, where $z = \hat{\beta}/\widehat{\text{Sd}}(\hat{\beta}_{mp})$. The estimated standard errors $\widehat{\text{Sd}}(\hat{\beta}_{mp})$ are based on the expected Fisher information matrix and do not account for the error of the model selection procedure. The plots on the bottom show the relative difference between the nodewise coefficients and the corresponding sample-average coefficients. The estimated variance of the random intercepts, which is not shown in Fig. 2, is $\hat{\Sigma}_b = 1.31$. The algorithm selects 3 of the 14 considered variables and partitions the data into 5 nodes. After the root node, it splits successively the nodes 5, 6, and 2. In the root node, all variables except *regional unemployment*, *sectoral unemployment*, *head of household* and *number of childs* show Bonferroni-corrected p-values below 0.05 in the coefficient constancy tests.

The fitted model for $\mathcal{M}_1$ can be analyzed by using Fig. 2 or the predicted distributions (conditional on $b_i = 0$) in Fig. 3. Here, we focus on Nodes 3 and 4 that include respondents in awkward *financial situations* and where the effect of *unemployed* is considerably moderated by *gender*. For females (Node 3), the cumulative logits are estimated to increase by 0.8 at the transition (corresponding to an odds ratio of $e^{0.8} = 2.2$), while for males (Node 4) the cumulative logits are estimated to increase by 0.1 (odds ratio of $e^{0.1} = 1.1$). This finding does not mean that the male respondents are happier than females; rather, the high intercepts in Node 4 indicate that the corresponding respondents are generally less happy than others, whether employed or not (direct effect).

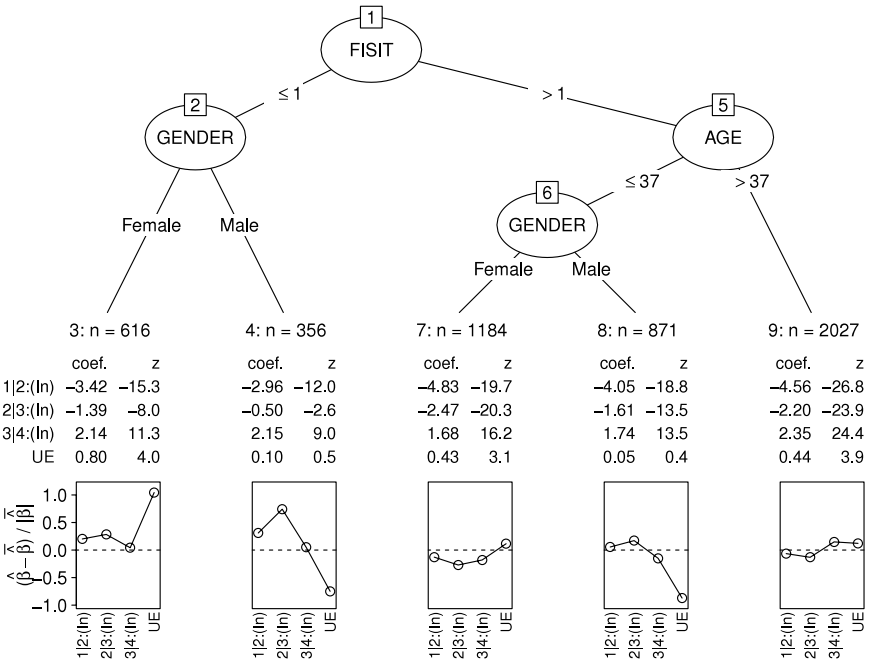

Fig. 2. *Top* fitted tree structure; *middle*, nodewise coefficients β_{m1} to β_{m4} with the corresponding z-values; *bottom*, relative differences between nodewise coefficients and the associated node-size weighted average coefficients $\bar{\hat{\boldsymbol{\beta}}} = (-4.28, -1.95, 2.05, 0.39)^\top$. The selected moderators are *financial situation* (FISIT), *gender* and *age*.

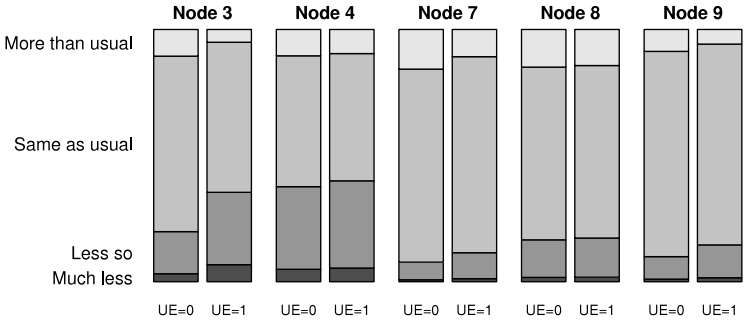

Fig. 3. Fitted model for $\mathcal{M}_1$: Predicted conditional distributions (with $b_i = 0$) of happiness during employment (UE = 0) and unemployment (UE = 1).

Predictive performance. To evaluate the performance of the algorithm in this application, we compare the negative log-likelihood prediction errors of fits of $\mathcal{M}_1$ and fits of two reference cumulative logit random intercept models: $\mathcal{M}_2$, a basis model and $\mathcal{M}_3$, a sophisticated model. The prediction errors of fits for $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ were estimated by using cluster bootstrap (Field and Welsh, 2007). We generated 250 bootstrap samples $(\mathcal{D}_1^*, \ldots, \mathcal{D}_{250}^*)$ from the total data $\mathcal{D}$, and fitted each model on each bootstrap sample. The bootstrap samples were drawn by randomly selecting 1487 respondents with replication from $\mathcal{D}$, and retaining the repeated observations corresponding to the selected respondents. Let $\widehat{\mathcal{M}}_{jk}^*$ be a fit for model $\mathcal{M}_j, j = 1, 2, 3$, based on the bootstrap sample $\mathcal{D}_k^*, k = 1, \ldots, 250$. Let $f_{\widehat{\mathcal{M}}_{jk}^*}(\mathbf{y}_{it}|b_i = 0)$ be the conditional density of $\mathbf{y}_{it}$ in $\widehat{\mathcal{M}}_{jk}$ with b_i set to its expected value 0. The negative log-likelihood prediction error of $\widehat{\mathcal{M}}_{jk}^*$ is computed as

$$\mathrm{err}(\widehat{\mathcal{M}}_{jk}^*) = \frac{1}{N_{\mathcal{D}_k^*}} \sum_{i,t \in \{\mathcal{D} \setminus \mathcal{D}_k^*\}} -\log f_{\widehat{\mathcal{M}}_{jk}}(\mathbf{y}_{it}|b_i = 0), \tag{17}$$

where $\{\mathcal{D} \setminus \mathcal{D}_k^*\}$ is the set of observations of $\mathcal{D}$ that does not appear in $\mathcal{D}_k^*$ and $N_{\mathcal{D}_k^*}$ is the number of observations of $\mathcal{D}_k^*$. Below, we will examine the pairwise differences $\mathrm{err}(\widehat{\mathcal{M}}_{1k}^*) - \mathrm{err}(\widehat{\mathcal{M}}_{2k}^*)$ and $\mathrm{err}(\widehat{\mathcal{M}}_{1k}^*) - \mathrm{err}(\widehat{\mathcal{M}}_{3k}^*)$.

$\mathcal{M}_2$: The basis model. First we compare the prediction error of fits for $\mathcal{M}_1$ and fits for the simple cumulative logit model

$$\mathcal{M}_2\text{: } \mathrm{logit}(P(Y_{it} \leq q)) = \beta_q + \mathrm{UE}_{it}\beta_4 + b_i.$$

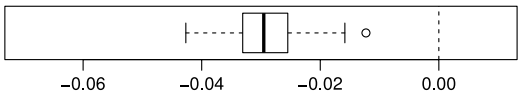

Fig. 4. Boxplot for 250 pairwise differences $\text{err}(\widehat{\mathcal{M}}^*_{1k}) - \text{err}(\widehat{\mathcal{M}}^*_{2k})$ comparing the prediction error of fits for $\mathcal{M}_1$ and $\mathcal{M}_2$.

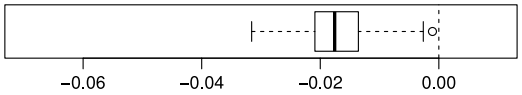

Fig. 5. Boxplot for 250 pairwise differences $\text{err}(\widehat{M}^*_{1k}) - \text{err}(\widehat{M}^*_{3k})$ comparing the prediction error of fits for $\mathcal{M}_1$ and $\mathcal{M}_3$.

$\mathcal{M}_2$ keeps the varying coefficients of $\mathcal{M}_1$ constant and, thus, ignores the variables of Table 1. Therefore, the comparison of $\mathcal{M}_2$ with $\mathcal{M}_1$ evaluates the ability of our algorithm to learn moderation or direct effects.

Fig. 4 shows the boxplot of the computed differences $\text{err}(\widehat{\mathcal{M}}^*_{1k}) - \text{err}(\widehat{\mathcal{M}}^*_{2k})$ in a boxplot. It can be seen that fits for $\mathcal{M}_1$ outperform, without exception, fits for $\mathcal{M}_2$, indicating that the algorithm significantly improves the model in this application.

$\mathcal{M}_3$: A linear CLMM with direct and moderation effects. We also wanted to compare fits for model $\mathcal{M}_1$ with fits for a more sophisticated model. Inspired by the study of Oesch and Lipps (2013), we consider the CLMM

$$\begin{aligned}\mathcal{M}_3\colon \text{logit}(P(Y_{it} \le q)) = {} & \beta_q + \text{GENDER}_{it}\beta_4 + \sum_{j=0}^{1} 1(\text{GENDER}_{it} = j) \\ & \times \Big[\overline{\text{AGE}}_{it}\beta_{5,j} + \overline{\text{AGE}}^2_{it}\beta_{6,j} + 1(\text{EDU}_{it} = 1)\beta_{7,j} + 1(\text{EDU}_{it} = 2)\beta_{8,j} + \text{SPINHH}_{it}\beta_{9,j} \\ & + \log \text{HHINC}_{it}\beta_{10,j} + \text{UE}_{it}\beta_{11,j} + 1(\text{TUE}_{it} = -1)\beta_{12,j} + \text{UEREG}_{it}\beta_{13,j} \\ & + (\text{UE}_{it} \times \text{UEREG}_{it})\beta_{14,j} + \text{UESEC}_{it}\beta_{15,j} + (\text{UE}_{it} \times \text{UESEC}_{it})\beta_{16,j}\Big] + b_i.\end{aligned}$$

In their study of the effect of unemployment on well-being, Oesch and Lipps estimate separate models for females and males. Equivalently, we specify in $\mathcal{M}_3$ the interaction between *gender* and all included covariates. For *age* (standardized, linear and squared), *education* (dummies for levels 1 and 2), *lives with spouse* (SPINHH) and the logarithm of *household income* we include only direct effects. Because Oesch and Lipps assume that well-being is different in the year before becoming unemployed, we add the dummy "$1(\text{TUE}_{it} = -1)$". For *regional unemployment* (UEREG) and *sectoral unemployment* (UESEC), we specify direct and interaction effects with *unemployment* (UE). Doing so integrates the hypothesis of Oesch and Lipps that suggests that "unemployment hurts less if there is more of it around". Although there remain differences between $\mathcal{M}_3$ and the model of Oesch and Lipps, the predictor functions of the two models are fairly comparable.

The 250 computed differences $\text{err}(\widehat{\mathcal{M}}^*_{1k}) - \text{err}(\widehat{\mathcal{M}}^*_{3k})$ shown in Fig. 5 demonstrate that fits for $\mathcal{M}_1$ outperform fits for $\mathcal{M}_3$. The median of -0.018 is here lower that the median (-0.03) observed for the difference between $\mathcal{M}_1$ and $\mathcal{M}_2$. Moreover, when measuring the complexity of $\mathcal{M}_1$ by the median of the number of coefficients plus the number of splits, and that for $\mathcal{M}_3$ by the (constant) number of coefficients, $\mathcal{M}_1$ is with 25 vs 29 also less complex. The superiority of $\mathcal{M}_1$ over $\mathcal{M}_3$ can be explained as follows: first, the (subjective) *financial situation*, which is a good predictor (cf. Fig. 2), is not included in $\mathcal{M}_3$. Second, our algorithm can benefit from technical differences, e.g., it incorporates the direct effects of moderators via the logit-specific varying intercepts rather than via the proportional odds effects. When incorporating the *financial situation* variable into $\mathcal{M}_3$ as a predictor with logit-specific effects, the median difference changes to 0.007 in favor of $\mathcal{M}_3$, the latter being however much more complex with a total of 41 coefficients. Anyway, the comparison made here does in no way invalidate hypothesis-driven model building, it just demonstrates that our algorithm is able to select relevant variables and builds parsimonious, understandable and competitive models.

3.2. Simulation studies

The following simulation studies focus exclusively on the implemented coefficient constancy tests. Because the remaining parts of the algorithm, including the likelihood-based exhaustive search, do not fundamentally differ from other tree-based algorithms such as MOB, they are not studied here. The most important conclusions from the simulation studies are as follows:

- Under coefficient constancy, the implemented tests achieve fairly accurate type I errors. Specifically, the type I errors obtained with pre-decorrelation are more accurate than those without. This finding indicates that the variable selection process of the algorithm is approximately unbiased.
- As expected, the power of the implemented tests increases with increasing moderation strengths and number of observations. The imputation for unbalanced data only slightly deteriorates the power of the tests.

Table 2
Relative frequencies of Type I errors in coefficient constancy tests for a nominal level of 5%. Values in brackets correspond to tests without pre-decorrelating the scores. Abbreviations: ii-cor = intra-individual correlation; cont = continuous, cat = categorical.

N/N_i	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
ii-cor scale	0 cont	0 cat	∼2/3 cont	∼2/3 cat	1 cont	1 cat
50/5	0.042 (0.042)	0.051 (0.054)	0.038 (0.027)	0.044 (0.038)	0.036 (0.018)	0.039 (0.020)
50/10	0.050 (0.050)	0.042 (0.041)	0.042 (0.024)	0.044 (0.030)	0.036 (0.014)	0.040 (0.010)
100/5	0.039 (0.039)	0.056 (0.054)	0.046 (0.032)	0.053 (0.038)	0.039 (0.018)	0.050 (0.020)
100/10	0.050 (0.047)	0.050 (0.048)	0.054 (0.024)	0.047 (0.027)	0.046 (0.018)	0.041 (0.010)
500/5	0.054 (0.054)	0.044 (0.044)	0.057 (0.036)	0.050 (0.036)	0.054 (0.024)	0.056 (0.018)
500/10	0.053 (0.052)	0.042 (0.044)	0.052 (0.030)	0.045 (0.028)	0.061 (0.023)	0.052 (0.012)

Table 3
Type I errors for the nodewise coefficient constancy tests for a nominal level of 5%. Values within brackets correspond to the tests without pre-decorrelating the scores.

N/N_i	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
50/5	0.038 (0.040)	0.048 (0.047)	0.036 (0.033)	0.041 (0.038)	0.020 (0.023)	0.042 (0.036)
100/5	0.050 (0.049)	0.048 (0.048)	0.045 (0.038)	0.038 (0.037)	0.044 (0.028)	0.040 (0.029)
500/5	0.040 (0.040)	0.050 (0.048)	0.054 (0.042)	0.049 (0.042)	0.052 (0.028)	0.063 (0.036)

- The power for variable selection of the implemented tests seems to be lower than that of the (slower) likelihood-based grid search approach. By contrast, they are more powerful than the M-fluctuation tests for a model that ignores intra-individual correlation.

The examined scenarios consider the coefficient constancy tests for six moderators, namely $Z_1, \ldots, Z_6$, that can be distinguished by their degree of intra-individual correlation (uncorrelated vs correlated vs time-invariant) and their scale (continuous vs categorical). Each scenario was repeated 2000 times. As explained in Section 2.3.1, the testing procedure is based on the "supLM" statistic of Andrews (1993) for continuous moderators and on the χ^2-type statistic of Hjort and Koning (2002) for categorical moderators.

Generating the simulation data. First, the values of the six moderators are generated by $z_{lit} = g_l(\tilde{z}_{1i} + \tilde{z}_{2it})$, where $\tilde{Z}_{1i} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_1)$ and $\tilde{Z}_{2it} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_2)$. For Z_1, Z_3, and Z_5, g_l is the identity function, while for Z_2, Z_4, and Z_6, g_l divides the values into four nominal categories $\{A, B, C, D\}$ based on their sample quartiles. For Z_1 and Z_2, we use $\sigma_1 = 0,\ \sigma_2 = 1$, (time-varying, uncorrelated); for Z_3 and Z_4, we use $\sigma_1 = 1,\ \sigma_2 = 1/2$ (time-varying, correlated); and for Z_5 and Z_6, we use $\sigma_1 = 1,\ \sigma_2 = 0$ (time-invariant). Second, the values x_{it} are generated by $X_{it} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Finally, the generated predictor and moderators are used to draw responses y_{it} with values in $\{1, 2, 3\}$ by using the model $\mathcal{M}_{\text{sim}}$

$$\mathcal{M}_{\text{sim}}\colon \text{logit}(P(Y_{it} \leq q)) = \beta_q + x_{it}[\delta \cdot 1_{(z_{lit} \in \mathcal{B}_l)}] + b_i, \quad b_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1),$$

with $\beta_1 = -1$ and $\beta_2 = 1$. Model $\mathcal{M}_{\text{sim}}$ states that the coefficient of x_{it} is an indicator function with an amplitude δ for one of the six moderators. The node $\mathcal{B}_l$ is defined as $\mathcal{B}_l = \mathbb{R}^+$ for the continuous moderators Z_1, Z_3, and Z_5 and as $\mathcal{B}_l = \{C, D\}$ for the nominal moderators Z_2, Z_4, and Z_6.

3.2.1. Type I errors

Root node tests. First, we set $\delta = 0$ (no moderation) and use $\mathcal{M}_{\text{root}} : \text{logit}(P(Y_{it} \leq q)) = \beta_q + x_{it}\delta + b_i$ as the model under H_0. Table 2 reports the resulting type I errors for a nominal level of 5%, for different numbers of individuals and observations per individual, with and without pre-decorrelation. The test errors based on these decorrelated scores are close to the theoretical 5%, particularly for large N's. By comparison, the errors of the naive tests based on the raw scores (values in brackets) are systematically too small for moderators that have high intra-individual correlation.

Nodewise tests. Now, we set $\delta = 1$ and test the influencing variable Z_l within node $\mathcal{B}_l$, by using $\mathcal{M}_{\text{sim}}$ as the model under H_0. The simulation is performed for $Z_1, \ldots, Z_6$ as the influencing variable in $\mathcal{M}_{\text{sim}}$. The tests should accept H_0 because the coefficient of x is constantly $\delta = 1$ within $\mathcal{B}_l$.

Table 3 shows the observed type I errors for a nominal level of 5%, for varying N's and a fixed $N_i = 5\ \forall\ i$. The results are similar to those in Table 2, confirming that nodewise testing works. The effect of the small N's is more pronounced than that in Table 2 because the nodes $\mathcal{B}_l$ enclose only about half the data.

3.2.2. Power and comparisons

To evaluate the power of our test implementation, we generate data from $\mathcal{M}_{\text{sim}}$ for varying moderation strengths $\delta = \{0, 0.1, \ldots, 0.5\}$. All tests use $\mathcal{M}_{\text{root}} : \text{logit}(P(Y_{it} \leq q)) = \beta_q + x_{it}\delta + b_i$ as the model under H_0.

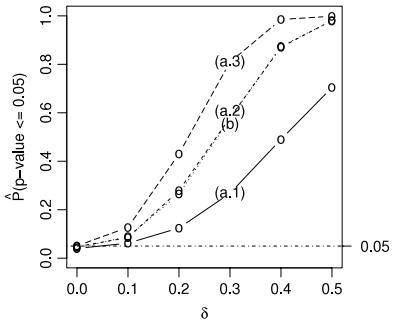

Fig. 6. Power of tests on Z_3 for increasing moderation strengths δ. The figure shows the relative frequencies for p-values below 0.05 for scenarios (a) and (b); (a) uses balanced data where $N_i = 5$ and (a.1) $N = 50$, (a.2) $N = 100$, and (a.3) and $N = 150$; (b) uses unbalanced data where N_i is 3 or 5 and $N = 150$.

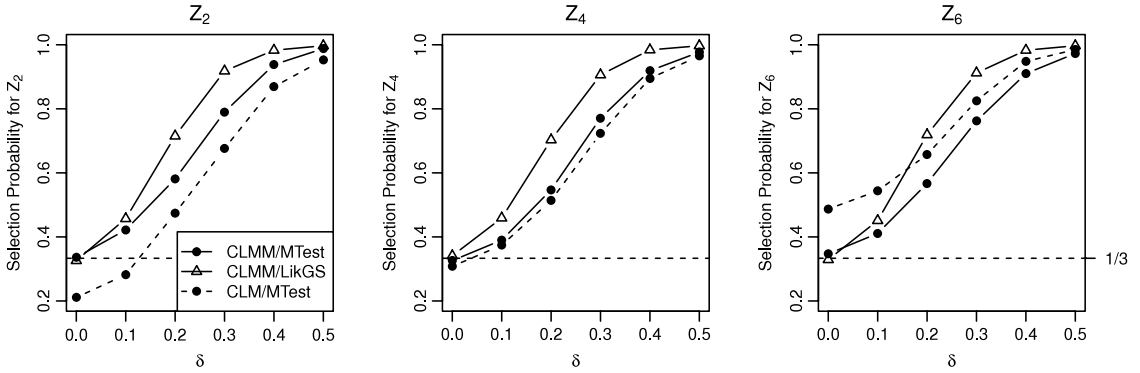

Fig. 7. Relative frequencies for selecting the true moderator among Z_2, Z_4, and Z_6 for varying moderation strengths δ. This selection is based on: *solid line, circle*, our test implementation; *solid line, triangle*, exhaustive search; and *dotted line, circle*, M-fluctuation tests with a model without random effects.

Power for balanced and unbalanced data. First, we use Z_3 (correlated, continuous) as the influencing variable to generate the data and as a moderator in the tests. The power is evaluated for scenarios (a) and (b). (a) uses balanced data where $N_i = 5\ \forall\, i$ and N varies between (a.1) $N = 50$, (a.2) $N = 100$, and (a.3) $N = 150$. (b) uses unbalanced data where $N = 140$ and $N_i = 5$ for individuals $i = 1, \ldots, 40$ and $N_i = 3$ for individuals $i = 41, \ldots, 140$. For (b), a single imputation is used to adjust the pre-decorrelation.

Fig. 6 shows the moderation strength δ against the relative frequency of p-values below 0.05. As expected, the power of the tests increases as δ increases and as the number of individuals N increases. Scenario (b) can be compared with (a.2), which also includes 500 observations. Moreover, (a.2) and (b) virtually overlap, indicating that the imputation for unbalanced data only sightly deteriorates the power of the tests.

Variable selection. In this last scenario, we use our tests to select between the moderators Z_2, Z_4, and Z_6, where (alternately) one of these moderates β_3. For the comparison, we also use the likelihood-based exhaustive search and M-fluctuation tests with the cumulative logit model without random coefficients to select $\mathcal{M}_{\text{clm}} : \text{logit}(P(Y_{it} \leq q)) = \beta_q + x_{it}\delta$ under H_0. In all scenarios, $N = 100$ and $N_i = 5\ \forall\, i$.

Fig. 7 shows the frequencies of selecting the true moderator. Both selection schemes based on CLMMs are unbiased. The exhaustive search is unbiased because all three moderators have the same number of splits. This selection method performs best, followed by our test implementation. The tests based on the model $\mathcal{M}_{\text{clm}}$ without random coefficients have lower power and they are biased towards the intra-individually correlated moderator Z_6.

4. Conclusion

The present study proposed a new tree-based algorithm for learning moderated relations in longitudinal (ordinal) regression analysis, by building on MGLMMs and the MOB algorithm of Zeileis et al. (2008). The main innovations relative to MOB are (i) similar to the approaches of Hajjem et al. (2011) and Sela and Simonoff (2012), the proposed algorithm can maintain random coefficients across nodes, meaning that observations of the same individual falling into different nodes are not treated as independent, and (ii) the coefficient constancy tests used to select the moderators and tree size are extended for testing based on observation scores. In addition, our algorithm extends the scope of longitudinal regression trees based on mixed models, which include the algorithms of Hajjem et al. (2011) and Sela and Simonoff (2012), to general longitudinal varying coefficient regression. As exemplified by examining the varying effect of unemployment, the resulting models are simple to read and therefore easily accessible to practitioners.

Although this study focused on CLMMs, the algorithm can be implemented more or less straightforwardly for other models of the MGLMM family. Further research could be directed towards improving the numerically challenging components of the algorithm. For example, alternative ways to direct marginal maximum likelihood estimation combined with Gauss–Hermite quadrature could be considered (e.g. Tutz, 2012, Chap. 14.3). Moreover, optimization by using Newton's method for the pre-decorrelation matrix has a tendency to fail for high dimensions and therefore this could be improved. Finally, the statistical power of the coefficient constancy tests could be enhanced by deriving the distribution of the partial sum processes of the raw rather than the pre-decorrelated scores. At present, we investigate the extension of building for each varying coefficient an individual tree. Such an extension allows to deduce which variable moderates which coefficient from the fitted tree structures, instead of from comparing the nodewise coefficients.

The proposed algorithm was implemented in the R (R Core Team, 2014) package **vcrpart**. The function `tvcolmm` fits tree-based varying coefficient CLMMs, with the presented methodology and corresponding methods, such as `plot` or `predict`, thereby allowing the diagnosis of the fitted model. Moreover, to overcome the potential instability of the algorithm mentioned in Section 1.1 and improve its accuracy, the package provides with `fvcolmm` a random forest implementation (Breiman, 2001) for the proposed algorithm.

Acknowledgments

This publication results from research work carried out within the framework of the Swiss National Center of Competence in Research LIVES, which is financed by the Swiss National Science Foundation. The authors are grateful to the Swiss National Science Foundation for its financial support.

Appendix A. Additional details on coefficient constancy tests

A.1. Covariance of $\boldsymbol{\Psi}_N(\cdot)$ (Section 2.3.1)

The constraint $\sum_{i=1}^{N} \hat{\boldsymbol{\psi}}_i(\hat{\boldsymbol{\beta}}_1) = \sum_{i=1}^{M} \hat{\boldsymbol{\psi}}_i = \mathbf{0}$ implies that

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{N} \hat{\boldsymbol{\psi}}_i\right) = \mathbf{0} &= \sum_{i=1}^{N}\sum_{i'=1}^{N} \mathrm{Cov}(\hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_{i'}) \\ &= \sum_{i=1}^{N} \mathrm{Var}(\hat{\boldsymbol{\psi}}_i) + \sum_{i=1}^{N}\sum_{i'=1}^{N} 1_{(i\neq i')}\mathrm{Cov}(\hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_{i'}). \end{aligned} \tag{A.1}$$

Under H_0, we assume that (ii) $\mathrm{Var}(\hat{\boldsymbol{\psi}}_i) = \mathrm{Var}(\hat{\boldsymbol{\psi}}_1)$ and (iii) $\mathrm{Cov}(\hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_{i'}) = \mathrm{Cov}(\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_2), \forall\, i \neq i'$. Based on these two assumptions and (Eq. (A.1)),

$$\mathrm{Cov}(\hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_{i'}) = -\frac{1}{N-1}\mathrm{Var}(\hat{\boldsymbol{\psi}}_1), \quad \forall i \neq i'. \tag{A.2}$$

In consequence, the covariance of the process $\boldsymbol{\Psi}_N(\tau)$ (Eq. (8)) is

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\Psi}_N(\tau_1), \boldsymbol{\Psi}_N(\tau_2)) &= \mathrm{Cov}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{\lfloor N\tau_1\rfloor} \hat{\boldsymbol{\psi}}_{\sigma(v_i)}, \frac{1}{\sqrt{N}}\sum_{i'=1}^{\lfloor N\tau_2\rfloor} \hat{\boldsymbol{\psi}}_{\sigma(v_{i'})}\right) \\ &\overset{\tau_1\leq\tau_2}{=} \frac{1}{N}\sum_{i=1}^{\lfloor N\tau_1\rfloor} \mathrm{Var}(\hat{\boldsymbol{\psi}}_i) + \sum_{i=1}^{\lfloor N\tau_1\rfloor}\sum_{i'=1}^{\lfloor N\tau_1\rfloor} \mathrm{Cov}(\hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_{i'}) + \sum_{i=1}^{\lfloor N\tau_1\rfloor}\sum_{i'=\lfloor N\tau_1\rfloor+1}^{\lfloor N\tau_2\rfloor} \mathrm{Cov}(\hat{\boldsymbol{\psi}}_i, \hat{\boldsymbol{\psi}}_{i'}) \\ &= \frac{1}{N}\left[\lfloor N\tau_1\rfloor \mathrm{Var}(\hat{\boldsymbol{\psi}}_1) - \frac{\lfloor N\tau_1\rfloor[\lfloor N\tau_1\rfloor - 1]}{N-1}\mathrm{Var}(\hat{\boldsymbol{\psi}}_1) - \frac{\lfloor N\tau_1\rfloor[\lfloor N\tau_2\rfloor - \lfloor N\tau_1\rfloor]}{N-1}\mathrm{Var}(\hat{\boldsymbol{\psi}}_1)\right] \\ &= \frac{\lfloor N\tau_1\rfloor(N - \lfloor N\tau_2\rfloor)}{N[N-1]}\mathrm{Var}(\hat{\boldsymbol{\psi}}_1). \end{aligned} \tag{A.3}$$

A.2. Pre-decorrelation of scores (Section 2.3.2)

Here we derive the pre-decorrelated observations scores $\hat{\mathbf{u}}^*_{it}$ and the computation of the transformation matrix $\mathbf{T}$ of Section 2.3.2. First, we consider balanced data where $N_i = N_1 \forall\, i$. In these cases, we assume that

$$\mathrm{Var}(\hat{\mathbf{u}}_{it}) = \mathrm{Var}(\hat{\mathbf{u}}_{11}) := \boldsymbol{\Delta}, \quad \forall\, i, t, \tag{A.4}$$

$$\mathrm{Cov}(\hat{\mathbf{u}}_{it}, \hat{\mathbf{u}}_{it'}) = \mathrm{Cov}(\hat{\mathbf{u}}_{11}, \hat{\mathbf{u}}_{12}) := \boldsymbol{\Omega}, \quad \forall\, i \text{ and } t \neq t' \quad \text{and} \tag{A.5}$$

$$\mathrm{Cov}(\hat{\mathbf{u}}_{it}, \hat{\mathbf{u}}_{i't'}) = \mathrm{Cov}(\hat{\mathbf{u}}_{11}, \hat{\mathbf{u}}_{21}) := \boldsymbol{\Psi}, \quad \forall\, t \text{ and } i \neq i', \tag{A.6}$$

where $N_T = \sum_{i=1}^{N} N_i$. Since $\mathrm{E}(\hat{\mathbf{u}}_{it}) = \mathbf{0}$ and $\mathrm{Var}(\sum_{i,t} \hat{\mathbf{u}}_{it}) = \mathbf{0}$, these matrices can be estimated by

$$\hat{\boldsymbol{\Delta}} = \frac{1}{N_T} \sum_{i=1}^{N} \sum_{t=1}^{N_1} \hat{\mathbf{u}}_{it} \hat{\mathbf{u}}_{it}^{\top}, \tag{A.7}$$

$$\hat{\boldsymbol{\Omega}} = \frac{1}{NN_1(N_1-1)} \left[\sum_{i=1}^{N} \left[\sum_{t=1}^{N_1} \hat{\mathbf{u}}_{it} \right] \left[\sum_{t=1}^{N_1} \hat{\mathbf{u}}_{it} \right]^{\top} - N_T \hat{\boldsymbol{\Delta}} \right] \quad \text{and} \tag{A.8}$$

$$\hat{\boldsymbol{\Psi}} = -\frac{1}{N_T^2 - N_T - NN_1(N_1-1)} \sum_{i=1}^{N} \left[\sum_{t=1}^{N_1} \hat{\mathbf{u}}_{it} \right] \left[\sum_{t=1}^{N_1} \hat{\mathbf{u}}_{it} \right]^{\top}. \tag{A.9}$$

It follows that the *intra-individual* covariance of the $\hat{\mathbf{u}}_{it}^{*}$'s is

$$\begin{aligned}
\mathrm{Cov}(\hat{\mathbf{u}}_{it}^{*}, \hat{\mathbf{u}}_{it'}^{*}) &\overset{t \neq t'}{=} \mathrm{Cov}\left(\hat{\mathbf{u}}_{it} + \mathbf{T} \sum_{t''=1, t'' \neq t}^{N_1} \hat{\mathbf{u}}_{it''}, \hat{\mathbf{u}}_{it'} + \mathbf{T} \sum_{t''=1, t'' \neq t'}^{N_1} \hat{\mathbf{u}}_{it''} \right) \\
&= \mathrm{Cov}(\hat{\mathbf{u}}_{it}, \hat{\mathbf{u}}_{it'}) + \sum_{t''=1, t'' \neq t'}^{N_1} \mathrm{Cov}(\hat{\mathbf{u}}_{it}, \hat{\mathbf{u}}_{it''}) \mathbf{T}^{\top} + \mathbf{T} \sum_{t''=1, t'' \neq t}^{N_1} \mathrm{Cov}(\hat{\mathbf{u}}_{it''}, \hat{\mathbf{u}}_{it'}) \\
&\quad + \mathbf{T} \sum_{t''=1, t'' \neq t}^{N_1} \sum_{t'''=1, t''' \neq t'}^{N_1} \mathrm{Cov}(\hat{\mathbf{u}}_{it''}, \hat{\mathbf{u}}_{it'''}) \mathbf{T}^{\top} \\
&= \cdots \\
&= \boldsymbol{\Delta}\mathbf{T}^{\top} + \mathbf{T}\boldsymbol{\Delta}^{\top} + [N_1 - 2]\mathbf{T}\boldsymbol{\Delta}\mathbf{T}^{\top} + \boldsymbol{\Omega} + [N_1 - 2]\boldsymbol{\Omega}\mathbf{T}^{\top} \\
&\quad + [N_1 - 2]\mathbf{T}\boldsymbol{\Omega}^{\top} + [[N_1 - 1]^2 - [N_1 - 2]]\mathbf{T}\boldsymbol{\Omega}\mathbf{T}^{\top},
\end{aligned} \tag{A.10}$$

and the *inter-individual* covariance of the $\hat{\mathbf{u}}_{it}^{*}$'s is

$$\begin{aligned}
\mathrm{Cov}(\hat{\mathbf{u}}_{it}^{*}, \hat{\mathbf{u}}_{i't'}^{*}) &\overset{i \neq i'}{=} \mathrm{Cov}\left(\hat{\mathbf{u}}_{it} + \mathbf{T} \sum_{t''=1, t'' \neq t}^{N_1} \hat{\mathbf{u}}_{it''},\ \hat{\mathbf{u}}_{i't'} + \mathbf{T} \sum_{t''=1, t'' \neq t'}^{N_1} \hat{\mathbf{u}}_{it''} \right) \\
&= \mathrm{Cov}(\hat{\mathbf{u}}_{it}, \hat{\mathbf{u}}_{i't'}) + \sum_{t''=1, t'' \neq t'}^{N_1} \mathrm{Cov}(\hat{\mathbf{u}}_{it}, \hat{\mathbf{u}}_{i't''}) \mathbf{T}^{\top} + \mathbf{T} \sum_{t''=1, t'' \neq t}^{N_1} \mathrm{Cov}(\hat{\mathbf{u}}_{it''}, \hat{\mathbf{u}}_{i't'}) \\
&\quad + \mathbf{T} \sum_{t''=1, t'' \neq t}^{N_1} \sum_{t'''=1, t''' \neq t'}^{N_1} \mathrm{Cov}(\hat{\mathbf{u}}_{it''}, \hat{\mathbf{u}}_{i't'''}) \mathbf{T}^{\top} \\
&= \cdots \\
&= \boldsymbol{\Psi} + [N_1 - 1]\boldsymbol{\Psi}\mathbf{T}^{\top} + [N_1 - 1]\mathbf{T}\boldsymbol{\Psi}^{\top} + [N_1 - 1][N_1 - 1]\mathbf{T}\boldsymbol{\Psi}\mathbf{T}^{\top}.
\end{aligned} \tag{A.11}$$

The goal is to determine the $MP_{\beta} \times MP_{\beta}$ matrix $\mathbf{T}$ such that

$$\mathrm{Cov}(\hat{\mathbf{u}}_{it}^{*}, \hat{\mathbf{u}}_{it'}^{*}) = \mathrm{Cov}(\hat{\mathbf{u}}_{it}^{*}, \hat{\mathbf{u}}_{i't''}^{*}) = -\frac{1}{\sum_{i=1}^{N} N_1 - 1} \mathrm{Var}(\hat{\mathbf{u}}_{11}^{*}), \tag{A.12}$$

for all $(i, t) \neq (i, t')$ and $(i, t) \neq (i', t'')$. $\mathbf{T}$ is found by solving

$$\begin{aligned}
\mathbf{0} &= \mathrm{Cov}(\hat{\mathbf{u}}_{it}^{*}, \hat{\mathbf{u}}_{it'}^{*}) - \mathrm{Cov}(\hat{\mathbf{u}}_{it}^{*}, \hat{\mathbf{u}}_{i't''}^{*}) \\
&= \left[\boldsymbol{\Delta}\mathbf{T}^{\top} + \mathbf{T}\boldsymbol{\Delta}^{\top} + [N_1 - 2]\mathbf{T}\boldsymbol{\Delta}\mathbf{T}^{\top} + \boldsymbol{\Omega} + [N_1 - 2]\boldsymbol{\Omega}\mathbf{T}^{\top} \right. \\
&\quad \left. + [N_1 - 2]\mathbf{T}\boldsymbol{\Omega}^{\top} + [[N_1 - 1]^2 - [N_1 - 2]]\mathbf{T}\boldsymbol{\Omega}\mathbf{T}^{\top} \right] - \left[\boldsymbol{\Psi} + [N_1 - 1]\boldsymbol{\Psi}\mathbf{T}^{\top} \right. \\
&\quad \left. + [N_1 - 1]\mathbf{T}\boldsymbol{\Psi}^{\top} + [N_1 - 1][N_1 - 1]\mathbf{T}\boldsymbol{\Psi}\mathbf{T}^{\top} \right],
\end{aligned} \tag{A.13}$$

for $\mathbf{T}$, using $\hat{\boldsymbol{\Delta}}$, $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\Psi}}$. Either, this equation system is solved with respect to all $(MP_{\beta})^2$ components, or $\mathbf{T}$ is assumed to be symmetric, which reduces the number of unknowns to $(MP_{\beta}(MP_{\beta} + 1))/2$. The symmetry assumption is natural because $\mathbf{T}$ is used for a decorrelation transformation. Note that, because of the sum of scores remains zero after the transformation,

$$\sum_{i=1}^{N} \sum_{t=1}^{N_i} \hat{\mathbf{u}}_{it}^{*} = \sum_{i=1}^{N} \sum_{t=1}^{N_1} \hat{\mathbf{u}}_{it} + N_1(N_1 - 1)\mathbf{T} \sum_{i=1}^{N} \sum_{t=1}^{N_1} \hat{\mathbf{u}}_{it} = \mathbf{0}, \tag{A.14}$$

and the variance of the pre-decorrelated scores remains constant,

$$\begin{aligned}\mathrm{Var}(\hat{\mathbf{u}}^*_{it}) &= \mathrm{Var}\left(\hat{\mathbf{u}}_{it} + \mathbf{T}\sum_{t'=1,t'\neq t}^{N_1}\hat{\mathbf{u}}_{it'}\right)\\ &= \cdots\\ &= \boldsymbol{\Delta} + [N_1-1]\mathbf{T}\boldsymbol{\Omega}\mathbf{T}^\top + [N_1-1]\boldsymbol{\Omega}\mathbf{T}^\top + \mathbf{T}\left[[N_1-1]\boldsymbol{\Delta} + \frac{(N_1-1)(N_1-2)}{2}\boldsymbol{\Omega}\right]\mathbf{T}^\top \quad \forall\,(i,t),\end{aligned} \tag{A.15}$$

the equality with the third term in (Eq. (A.12)) holds automatically if the equality between the first two terms holds.

A.3. Imputation procedure for unbalanced data (Section 2.3.2)

The imputation for a missing observation t of individual i in model $\mathcal{M}$ (Eq. (1)) requires values for the design matrices $\mathbf{X}_{it}$ and $\mathbf{W}_{it}$ and the moderator $\mathbf{z}_{it}$. We propose to randomly draw these data from the N_i sets of observed predictor vectors of individual i. Next, $\mathbf{y}_{it}$ is randomly drawn from the conditional distribution $f(y_{it}|\hat{\mathbf{b}}_i; \mathbf{X}_{it}, \mathbf{W}_{it}, \mathbf{z}_{it})$ of the estimated model under H_0, in order to control the type I error of the test. To estimate the random coefficients $\mathbf{b}_i$ we use the posterior mean estimate, see Tutz (2012, Chap. 14.3.2).

A.4. Nodewise tests (Section 2.3.3)

This appendix specifies the properties of the nodewise, mean centered scores. Let $\hat{\mathbf{U}}^*_m$ be the $N_m \times MP_\beta$ matrix of pre-decorrelated scores (Section 2.3.2) corresponding to observations $\mathbf{z}_{it} \in \mathcal{B}_m$. Let $\hat{\mathbf{U}}^{**}_m$ be the $\hat{\mathbf{U}}^*_m$ minus its column means. In Section 2.3.2, we established that $\mathrm{Cov}(\hat{\mathbf{u}}^*_{it}, \hat{\mathbf{u}}^*_{i't'}) = -\frac{1}{N_T-1}\mathrm{Var}(\hat{\mathbf{u}}^*_{11})\ \forall\,(i,t) \neq (i',t')$. It follows that the covariance between the rows of $\hat{\mathbf{U}}^{**}_m$,

$$\begin{aligned}\mathrm{Cov}(\hat{\mathbf{u}}^{**}_{mit}, \hat{\mathbf{u}}^{**}_{mi't'}) &= \mathrm{Cov}\left(\hat{\mathbf{u}}^*_{mit} - \frac{1}{N_m}\sum_{i'',t''}\hat{\mathbf{u}}^*_{mi''t''},\ \hat{\mathbf{u}}^*_{mi't'} - \frac{1}{N_m}\sum_{i'',t''}\hat{\mathbf{u}}^*_{mi''t''}\right)\\ &= \cdots\\ &= \mathrm{Cov}(\hat{\mathbf{u}}^*_{m11}, \hat{\mathbf{u}}^*_{m21})\left[1 - 2\frac{N_m-1}{N_m} + \frac{(N_m-1)^2}{N_m^2}\right] + \mathrm{Var}(\hat{\mathbf{u}}^*_{m11})\left[-\frac{2}{N_m} + \frac{N_m}{N_m^2}\right]\\ &= -\frac{1}{N_m}\mathrm{Var}\left(\hat{\mathbf{u}}^*_{m11}\right) \quad \forall\,(i,t) \neq (i',t'),\end{aligned} \tag{A.16}$$

takes the required covariance structure (cf. Section 2.3.2). Further, the covariance between mean-centered scores of different nodes, say, $\mathcal{B}_m$ and $\mathcal{B}_{m'}$,

$$\begin{aligned}\mathrm{Cov}(\hat{\mathbf{u}}^{**}_{mit}, \hat{\mathbf{u}}^{**}_{m'i't'}) &= \mathrm{Cov}\left(\hat{\mathbf{u}}^*_{mit} - \frac{1}{N_m}\sum_{i'',t''}\hat{\mathbf{u}}^*_{mi''t''},\ \hat{\mathbf{u}}^*_{m'i't'} - \frac{1}{N_{m'}}\sum_{i'',t''}\hat{\mathbf{u}}^*_{m'i''t''}\right)\\ &= \cdots = \mathrm{Cov}(\hat{\mathbf{u}}^*_{11}, \hat{\mathbf{u}}^*_{21})\left[1 - \frac{N_m}{N_m} - \frac{N_{m'}}{N_{m'}} + \frac{N_m N_{m'}}{N_m N_{m'}}\right]\\ &= \mathbf{0},\end{aligned} \tag{A.17}$$

which implies that the tests on $\mathcal{B}_m$ are independent from tests on $\mathcal{B}_{m'}$.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2015.01.003.

References

Abdolell, M., LeBlanc, M., Stephens, D., Harrison, R.V., 2002. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. Stat. Med. 21, 3395–3409.
Alexander, W.P., Grimshaw, S.D., William, P., 1996. Treed regression. J. Comput. Graph. Statist. 5, 156–175.
Andrews, D.W.K., 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 61, 821–856.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, New York, USA.
Bürgin, R., 2015. vcrpart: tree-based varying coefficient regression for generalized linear and ordinal mixed models. URL http://cran.r-project.org/web/packages/vcrpart/. R package version 0.3-1.
Eo, S.H., Cho, H.J., 2014. Tree-structured mixed-effects regression modeling for longitudinal data. J. Comput. Graph. Statist. 23, 740–760.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models, second ed. In: Springer Series in Statistics, Springer-Verlag, New York, USA.
Field, C.A., Welsh, A.H., 2007. Bootstrapping clustered data. J. R. Stat. Soc. Ser. B 69, 369–390.
Hajjem, A., 2010. Mixed Effect Trees and Forests for Clustered Data (Ph.D. thesis), HEC Montréal.
Hajjem, A., Bellavance, F., Larocque, D., 2011. Mixed effects regression trees for clustered data. Statist. Probab. Lett. 81, 451–459.
Hastie, T., Tibshirani, R., 1993. Varying-coefficient models. J. R. Stat. Soc. Ser. B 55, 757–796.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning, second ed. In: Springer Series in Statistics, Springer-Verlag, New York, USA.
Hedeker, D., Gibbons, R.D., 1994. A random-effects ordinal regression model for multilevel analysis. Biometrics 50, 933–944.
Hjort, N.L., Koning, A., 2002. Tests for constancy of model parameters over time. J. Nonparametr. Stat. 14, 113–132.
Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Statist. 15, 651–674.
Kauermann, G., 2000. Modeling longitudinal data with ordinal response by varying coefficients. Biometrics 56, 692–698.
Kosmidis, I., 2014. Improved estimation in cumulative link models. J. R. Stat. Soc. Ser. B 76, 169–196.
Loh, W.Y., 2002. Regression trees with unbiased variable selection and interaction detection. Statist. Sinica 12, 361–386.
McCullagh, P., 1980. Regression models for ordinal data. J. R. Stat. Soc. Ser. B 42, 109–142.
Nyblom, J., 1989. Testing for the constancy of parameters over time. J. Amer. Statist. Assoc. 84, 223–230.
Oesch, D., Lipps, O., 2013. Does unemployment hurt less if there is more of it around? a panel analysis of life satisfaction in Germany and Switzerland. Eur. Sociol. Rev. 29, 955–967.
Quinlan, J.R., 1992. Learning with continuous classes. In: 5th Australian Joint Conference on Artificial Intelligence. World Scientific, Singapore, pp. 343–348.
R Core Team 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org.
Rusch, T., Zeileis, A., 2012. Gaining insight with recursive partitioning of generalized linear models. J. Stat. Comput. Simul. 83, 1–15.
Russell, S.J., Norvig, P., 2003. Artificial Intelligence: A Modern Approach, third ed. Pearson Education Inc., New Jersey, USA.
Sela, R., Simonoff, J.S., 2012. RE-EM trees: a data mining approach for longitudinal and clustered data. Mach. Learn. 86, 169–207.
Siddall, P.J., McClelland, J.M., Rutkowski, S.B., Cousins, M.J., 2003. A longitudinal study of the prevalence and characteristics of pain in the first 5 years following spinal cord injury. Pain 103, 249–257.
Strobl, C., Kopf, J., Zeileis, A., 2013. Rasch trees: a new method for detecting differential item functioning in the rasch model. Psychometrika 1–28, (forthcoming).
Su, X., Meneses, K., McNees, P., Johnson, W., 2011. Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. J. R. Stat. Soc. Ser. C 60, 457–474.
Taylor, M.F., John Brice, N.B., Prentice-Lane, E., 2010. British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices. University of Essex, Colchester, UK.
Tutz, G., 2012. Regression for Categorical Data. Cambridge Series in Statistical and Probabilistic Mathematics, In: Cambridge Series in Statistical and Probablilistic Mathematics, New York, USA.
Tutz, G., Hennevogl, W., 1996. Random effects in ordinal regression models. Comput. Statist. Data Anal. 22, 537–557.
Tutz, G., Kauermann, G., 2003. Generalized linear random effects models with varying coefficients. Comput. Statist. Data Anal. 43, 13–28.
Wang, J.C., Hastie, T., 2014. Boosted varying-coefficient regression models for product demand prediction. J. Comput. Graph. Statist. 23, 361–382.
Zeileis, A., Hornik, K., 2007. Generalized M-fluctuation tests for parameter instability. Stat. Neerl. 61, 488–508.
Zeileis, A., Hothorn, T., Hornik, K., 2008. Model-based recursive partitioning. J. Comput. Graph. Statist. 17, 492–514.
Zhang, D., 2004. Generalized linear mixed models with varying coefficients for longitudinal data. Biometrics 60, 8–15.