PSYCHOMETRIKA—VOL. 87, NO. 3, 1064–1080
SEPTEMBER 2022
https://doi.org/10.1007/s11336-021-09834-6

A NOTE ON THE STRUCTURAL CHANGE TEST IN HIGHLY PARAMETERIZED PSYCHOMETRIC MODELS

K. B. S. HUTH

UNIVERSITY OF AMSTERDAM

AMSTERDAM UNIVERSITY MEDICAL CENTER

CENTRE FOR URBAN MENTAL HEALTH

L. J. WALDORP

UNIVERSITY OF AMSTERDAM

J. LUIGJES

AMSTERDAM UNIVERSITY MEDICAL CENTER

A. E. GOUDRIAAN

AMSTERDAM UNIVERSITY MEDICAL CENTER

ARKIN MENTAL HEALTH INSTITUTE

R. J. VAN HOLST

AMSTERDAM UNIVERSITY MEDICAL CENTER

CENTRE FOR URBAN MENTAL HEALTH

M. MARSMAN

UNIVERSITY OF AMSTERDAM

Equal parameter estimates across subgroups is a substantial requirement of statistical tests. Ignoring subgroup differences poses a threat to study replicability, model specification, and theory development. Structural change tests are a powerful statistical technique to assess parameter invariance. A core element of those tests is the empirical fluctuation process. In the case of parameter invariance, the fluctuation process asymptotically follows a Brownian bridge. This asymptotic assumption further provides the basis for inference. However, the empirical fluctuation process does not follow a Brownian bridge in small samples, and this situation is amplified in large psychometric models. Therefore, common methods of obtaining the sampling distribution are invalid and the structural change test becomes conservative. We discuss an alternative solution to obtaining the sampling distribution—permutation approaches. Permutation approaches estimate the sampling distribution through resampling of the dataset, avoiding distributional assumptions. Hereby, the tests power are improved. We conclude that the permutation alternative is superior to standard asymptotic approximations of the sampling distribution.

Key words: parameter invariance, parameter stability, structural change test, finite sample behavior, permutation test.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-021-09834-6.

© 2022 The Author(s)

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

The assumption of parameter invariance across subgroups—assessed as measurement invariance, differential item functioning, heterogeneity, or parameter stability in other contexts—underlies virtually all statistical tests (Bechger and Maris 2015; Hjort and Koning 2002; Hansen 1997; Mellenbergh 1989). Formally, we can define parameter invariance as,

$$f(y \mid v, \theta) = f(y \mid \theta),$$

where $f(\cdot)$ is a parametric distribution that is indexed by a parameter θ, used to model an observed variable y, and v is an auxiliary variable against which we are testing parameter invariance. Thus, parameter invariance implies that an identical model holds for different subgroups (e.g., males and females, older and younger persons, and persons with different ethnic backgrounds) or across measurement occasions (Putnick and Bornstein 2016; van de Schoot et al. 2015). Violations of this assumption can lead to misspecified models, spurious parameter estimates and test results, therefore, concealing differences key for theory development, diagnostic procedures, and treatment design (e.g., Kapur et al. 2012; Breslau et al. 2008). Unfortunately, researchers often neglect parameter invariance, which poses a threat to research development (Borsboom 2006).

Structural change tests (SCTs) allow us to assess parameter invariance across subgroups (Brown et al. 1975). These tests were initially proposed by Andrews (1993) for parameter stability assessment in econometric time-series models, but since then have been adapted to assess models across the statistical sciences (e.g., Chang and Su 2014; Mulaudzi 2016; O'Connell et al. 2018; Strobl et al. 2015; Zeileis et al. 2008; Merkle et al. 2014). SCTs have become a popular method for assessing parameter invariance because they can be straightforwardly implemented, even for complicated psychometric models: SCTs do not require explicit specification of which parameter diverges or which subgroups behave differently (Wang et al. 2018). Different type of SCTs exist that differ in the particular type of influence measure used [e.g., recursive residuals (Brown et al. 1975) or single-shift test statistics (Andrews 1993)], but depicting closely related asymptotic properties. In this paper, we focus on SCTs using *scores* as influence measures (Merkle and Zeileis 2013; Zeileis 2006; Hjort and Koning 2002). Scores are partial derivatives of the log-likelihood function with respect to a particular parameter and can be considered similar to asymptotic influence functions which are used to determine the effect of single observations on the estimate (Hampel et al. 2005). SCTs assume that if parameter invariance holds, aggregated scores randomly fluctuate about zero and converge to a Brownian bridge (Hjort and Koning 2002); a process that starts and ends at zero and randomly fluctuates about zero in between with its individual elements being normally distributed. However, if the aggregated scores systematically change in line with changes of an auxiliary variable v, parameter invariance is violated (Zeileis 2006). This result is used to determine the sampling distribution of the SCT's test statistic.

The test statistic's sampling distribution is well-determined for large sample sizes (Hansen 1997; Estrella 2003). However, in finite samples, concerns arise. For a simple linear model, the test shows both sub-optimal power and a type 1 error rate below the expected nominal α-level (Zeileis and Hothorn 2013). The asymptotic sampling distribution does not exploit the significance level and is of poor quality compared to the exact conditional and conditional asymptotic distribution (Hothorn and Zeileis 2008; Zeileis and Hothorn 2013). In short, the SCT becomes increasingly conservative in finite samples (Jones et al. 2020; Merkle and Zeileis 2013; Strobl et al. 2015). In this paper, we delve into this issue, focusing on large psychometric models. Our goals are twofold. Our first goal is to assess the SCT's behavior in finite samples. In particular, we investigate the sampling distribution and the distribution of the p value, which should be uniformly distributed under the null hypothesis. We show that for finite samples, the sampling distribution is misspecified and p values are not uniformly distributed. This problem becomes more pronounced the larger the model (i.e., the more parameters to be estimated). Our second goal is to show a solution

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

to this misspecification—permutation approaches—allowing for the SCTs application in large psychometric models. Permutation approaches are useful for estimating the sampling distribution when distributional assumptions do not hold or are analytically intractable (Mooney and Duval 1993). We show that permutation approaches provide a correct type I error and tend to increase the test's statistical power in finite samples (Zeileis and Hothorn 2013). In this way, permutation methods are superior to standard asymptotic approaches.

The remainder of this paper is organized as follows. First, we introduce the SCT in detail. Second, we investigate the SCT's finite sample behavior, and in particular, the distribution of p values under the null hypothesis. Here, we establish that the asymptotically derived sampling distribution is incorrect for finite sample sizes. Third, we elaborate on an alternative approach to obtaining the sampling distribution—permutation approaches. To illustrate the issues and our solution, we will use a linear regression model and a Gaussian Graphical model throughout this paper.

1. Structural Change Tests

The SCT assesses the equivalence of all k model parameters (i.e., $j = 1, \ldots, k$) across subgroups defined by an auxiliary variable v (Andrews 1993). Under the null hypothesis, the SCT assumes that a parameter θ_j is the same for all subgroups v_g, $g = 1, \ldots, m$, of the auxiliary variable. That is,

$$\mathcal{H}_0 : \ \theta_{j\, v_g} = \theta_j; \forall\, 1 \leq g \leq m, 1 \leq j \leq k,$$

where $\theta_{j\, v_g}$ denotes the parameter value of subgroup v_g for parameter θ_j. The SCT comprises three steps: First, one estimates the model of interest and determines its parameter influence measure. Here, we focus on score-based SCTs (Hjort and Koning 2002; Zeileis and Hornik 2007). Secondly, so-called empirical fluctuation processes are derived from the scores. Thirdly, the fluctuation processes are aggregated into a test statistic and compared against the sampling distribution to compute the p value. We outline each of these steps below.

The first step consists of estimating the k parameters of a model of interest. This paper will focus on estimating the model parameters through maximum likelihood estimation (MLE; for other approaches, see, for example, Kuan and Hornik 1995). With the MLEs, the score for every particular parameter and observation can be calculated. The score is the gradient of the log-likelihood function and for a parameter θ_j and observation y_i, it is denoted by

$$\mathrm{s}(\theta_j, y_i) = \frac{\partial \log \mathrm{L}(\theta; y_i)}{\partial \theta_j},$$

where L is the likelihood function of the model, θ_j the focal parameter and y_i the data for an observation i. Since the MLEs maximize the log-likelihood function, the sum of the scores for a parameter j across all n observations will sum to zero:

$$\sum_{i=1}^{n} \mathrm{s}(\hat{\theta}_j, y_i) = 0, \tag{1}$$

which holds for all parameters in the model.

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

In the second step, the accumulations of scores across observations are interpreted as empirical fluctuation processes. These fluctuation processes are analyzed separately for every parameter of the model. To obtain the fluctuations, the scores are first ordered along the auxiliary variable v and then aggregated across observations:

$$\Psi(t; \hat{\theta}_j) = n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \mathrm{s}(\hat{\theta}_j, y_i),$$

where $\lfloor nt \rfloor$ is the floor function of $n \times t$. Note, t is a real number and in practice with discrete measurements commonly a proportion, such as here a fraction of the n participants (i.e., $t = i/n$ for $i = 1, \ldots, n$). $\sum_{i=1}^{\lfloor nt \rfloor}$ therefore describes the sum of all scores up until the $(n \times t)-$th term, which is referred to as the cumulated score. To ensure that the cumulative scores are independent across parameters, $\Psi(t; \hat{\theta}_j)$ is decorrelated (Merkle and Zeileis 2013):

$$\mathrm{B}(t; \hat{\theta}) = \hat{I}^{-1/2} \Psi(t; \hat{\theta}),$$

where $\hat{I}$ is the asymptotic covariance matrix of the scores, thus, an estimate of the Fisher information matrix (Zeileis 2006). Observe that the cumulated scores $\mathrm{B}(t; \hat{\theta}_j)$ are zero for $t = 0$ and also at $t = 1$. At $t = 1$, the scores of all observations have been summed up, which by definition of the MLE is zero, e.g., Eq. (1).

Under $\mathcal{H}_0$, the fluctuation processes asymptotically converge to a Brownian bridge (Hjort and Koning 2002; Andrews 1993) and for a model with k-parameters to k-independent Brownian bridges,

$$\mathrm{B}(\cdot; \hat{\theta}) \xrightarrow{d} \mathrm{B}^0(\cdot),$$

where $\xrightarrow{d}$ denotes weak convergence of $\mathrm{B}(\cdot; \hat{\theta})$ to a k-dimensional Brownian bridge $\mathrm{B}^0(\cdot)$. Parameter stability can now be visually assessed by plotting the fluctuation process. The fluctuation process randomly varies about zero if $\mathcal{H}_0$ were true and parameter invariance holds. However, in the case of parameter non-invariance, the process systematically deviates from zero. Figure 1 provides an illustration.

In the third and final step, the cumulative scores are combined into a test statistic which can be conducted in various ways (Merkle and Zeileis 2013; Hjort and Koning 2002). We will introduce the three test statistics commonly used in literature: The double maximum statistic (DM), the Cramér–von Mises statistic (CvM), and the maximum Lagrange Multiplier statistic (maxLM).

$$\mathrm{DM} = \max_{i=1,\ldots,n} \max_{j=1,\ldots,k} \left| \mathrm{B}_{ij} \right| \tag{2}$$

$$\mathrm{CvM} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} \mathrm{B}_{ij}^2 \tag{3}$$

$$\mathrm{maxLM} = \max_{i=\underline{i},\ldots,\bar{i}} \left(\left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1}^{k} \mathrm{B}_{ij}^2 \right), \tag{4}$$

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

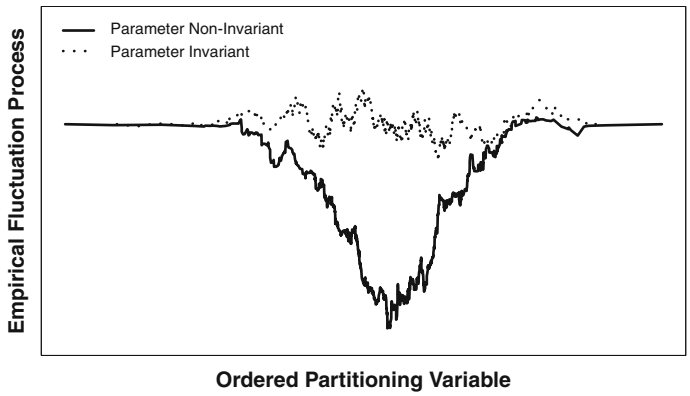


FIGURE 1.
Visualization of empirical fluctuation processes for two exemplary parameters. The dotted line represents the cumulative scores for a parameter with a random fluctuation around zero; thus, the fit for that parameter does not depend on the auxiliary variable. The solid line represents a systematic fluctuation coinciding with the auxiliary variable; parameter invariance is violated.

where B_{ij} denotes the fluctuation process at an observation i for a parameter $\hat{\theta}_j$, i.e., $B_{ij} = B(t = \frac{i}{n}; \hat{\theta}_j)$. The DM statistic takes the maximum of the cumulated scores across observations and parameters, and is used to test if any fluctuation process deviates too strongly from zero at any time. The CvM captures fluctuations that change across a variety of observations and parameters. Lastly, the maxLM statistic is suited if all k fluctuation processes change along the same observation i. To circumvent precision issues, the fluctuation process's tails are not considered when computing the maxLM statistic.

In null hypothesis significance testing, the test statistic computed from observed data is compared against the sampling distribution to obtain a p value. Critical values can be obtained by simulating observations from a Brownian bridge and applying the relevant statistic to the generated data (Andrews 1993; Zeileis 2006). Also, closed form solutions exist for specific situations. For example, Ploberger and Kramer (1992) derive the sampling distribution for DM-type statistics and Hjort and Koning (2002) show that CvM-type statistics follow an approximate χ^2-distribution that depends on the amount of parameters and focal change point assessed. Furthermore, Hansen (1997) established for the maxLM statistic that if the focal parameter changes close to the half-point of the auxiliary variable, the sampling distribution converges to a χ^2-distribution. Hansen presented critical statistics and p values for specific combinations of model parameters k and change point location. Estrella (2003) extends those results for high-dimensional scores. We will use the simulation method for determining the sampling distribution (Zeileis 2006).

2. Small Sample Behavior of the Structural Change Test

Simulations of the SCT's behavior concluded that the SCT shows suboptimal behavior in finite samples (Hothorn and Zeileis 2008; Merkle and Zeileis 2013; Strobl et al. 2015; Jones et al. 2020). Its power dwindles in finite samples and surprisingly, the type 1 error decreases with increasing model complexity, staying below the respective significance threshold. To provide a better understanding of the nature and severity of this issue as well as potential ways to mitigate it, we will delve into the problem. We will analyze it in two ways, first through simulation and second through mathematical derivation of the convergence rate.

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

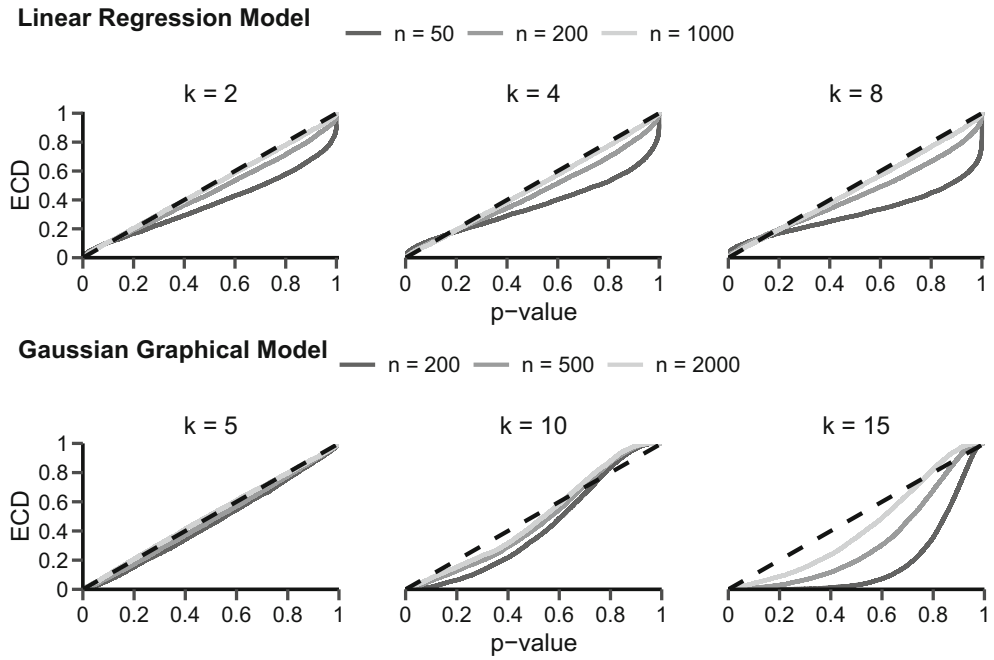


FIGURE 2.
Empirical cumulative distributions (ECDs) for the p value under the null hypothesis for different models and simulation settings. The top row shows the linear regression model results and the bottom row shows the GGM results. Here, n denotes the sample size and k the number of covariates for the linear regression model and the number of nodes for the GGM. In each plot, the black, dashed line shows the expected uniform distribution.

2.1. *Simulation of Small Sample Behavior*

We will analyze the SCT's behavior through simulation for two models: A simple linear regression model and a more complex Gaussian graphical model (GGM). Our simulations vary the sample size n and the covariates/nodes in the models. For the linear regression, we simulated models with two, four, and eight covariates (i.e., regression coefficients) for 50, 200, and 1000 observations each. For the GGM, we simulated networks with five, ten, and fifteen nodes for 200, 500, and 2000 observations. Each combination was run 5000 times. Datasets were simulated as a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ with a sparse interaction matrix Σ (i.e., probability of interaction was 0.2) without any dependency on an auxiliary variable. Thus, data were generated under the null hypothesis of equal parameters across subgroups. All simulations were run in the software R (R-Core-Team 2020); the SCT was conducted using the `strucchange` function for the linear regression model (Zeileis et al. 2002) and the `partykit::mob` function for the GGM (Zeileis et al. 2008; Hothorn and Zeileis 2015).

We will focus on the results for the maxLM statistic with a continuous auxiliary variable. In the online appendix, we report the results for a binary auxiliary variable as well as the results for CvM and DM statistics.[1] The simulated p value distributions are shown in Fig. 2. The p value is expected to follow a uniform distribution under the null hypothesis, which is indicated with the dashed, black line in each of the plots in Fig. 2. Observe that the p values do not follow this uniform distribution for the linear regression model in the smaller sample sizes but approximate a uniform distribution if the sample size increases. For the GGM, the p value is nearly uniformly distributed

[1]The code, simulation results and online appendix can be found on the project repository https://github.com/KarolineHuth/sctpermutation.

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

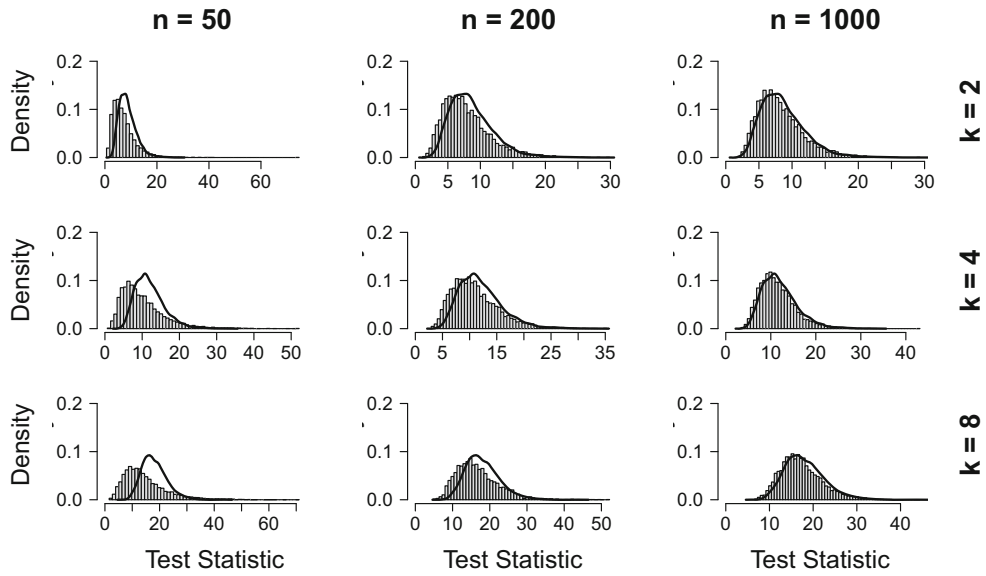


FIGURE 3.
Distributions of the maxLM statistic under the null hypothesis for the linear regression model. The expected sampling distribution is depicted as a black line and was obtained by simulating observations from a Brownian bridge and applying the maxLM statistic to them (e.g., see Zeileis 2006).

in small networks for all sample sizes. However, for larger networks, the p value distribution deviates. A result that appears to be independent of the sample size used in our simulations. The deviation between the simulated p value distribution and the correct uniform distribution is largest for networks with 15 nodes and 200 observations; here, even with 2000 observations, the p value does not follow a uniform distribution.

The simulated sampling distributions are shown in Fig. 3 for the linear regression model and in Fig. 4 for the GGM. The asymptotic sampling distributions are indicated with a black solid line in these graphs. They were generated by repeatedly simulating values from a Brownian bridge and then computing the maxLM statistic on the generated data (e.g., see Andrews 1993; Zeileis 2006). In computing the maxLM statistic, a choice is made to cut off the empirical fluctuation process's tails to avoid precision issues. The choice of cut-off points can, in principle, improve the fit of the estimated sampling distribution and lead to cherry-picking cut-offs that improve the fit of the sampling distribution. For the GGM, we chose to cut off the process's tails before n_p and after $n - n_p$ observations, respectively, where n_p denotes the number of free parameters in the model (i.e., similar to Jones et al. 2020). For the linear model, we chose to cut-off the bottom and upper 10%. Results in Figs. 3 and 4 show clearly that for the linear regression model, the sampling distribution is specified correctly for larger sample sizes, independent of model complexity, but not for smaller sample sizes. For the GGM, the sampling distribution is properly specified for small networks, but large discrepancies are found for larger networks.

From the assessment thus far, the true nature of the finite sample misspecification is unknown. The k-independent Brownian bridges approximation and thereby the correct specification of sampling distribution depends on two assumptions: the normal approximation to the score distribution and the accuracy of the estimated information matrix. In order to assess these two fundamental assumptions, we conducted some small scale simulations. Results indicate that the scores of a threshold parameter do follow a normal distribution fairly well for all number of nodes and observations (see Figure S5 in the online appendix). However, for an interaction parameter the score

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

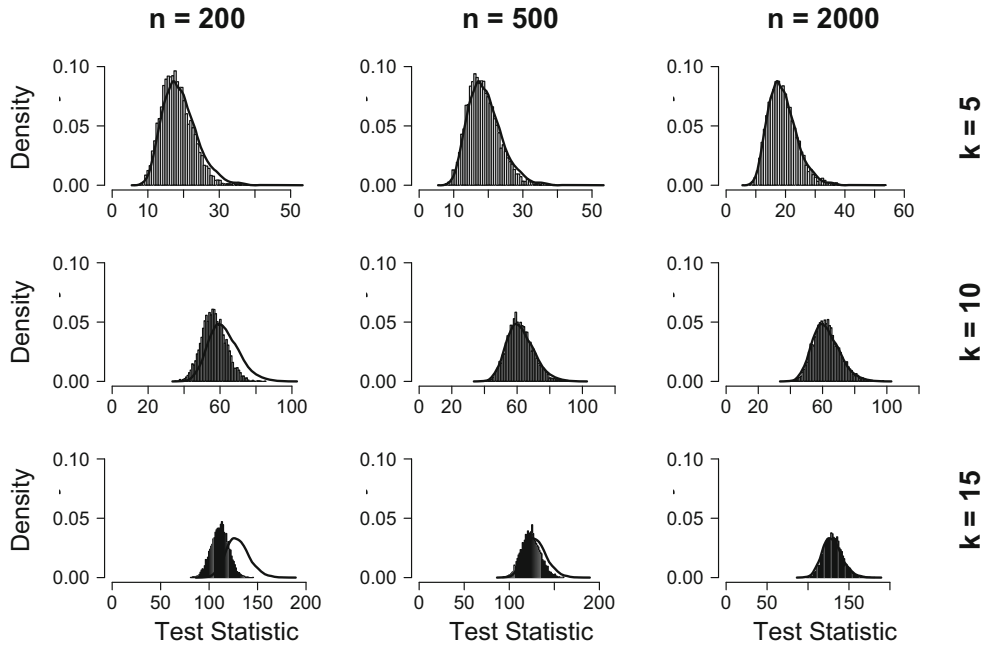


FIGURE 4.
Distributions of the maxLM statistic under the null hypothesis for the GGM. The expected sampling distribution is depicted as a black line and was obtained by simulating observations from a Brownian bridge and applying the maxLM statistic to them (e.g., see Zeileis 2006).

distribution deviates highly from the expected one; the distributions are skewed, multi-modal, and peaked (see Figure S6 in the online appendix). Furthermore, the estimated Fisher information matrix is biased. However, the bias reduces with increased sample size, especially for larger models (see Figure S7 in the online appendix). A detailed description of the results can be found in the online appendix.

In sum, the *p* value and test statistic do not follow the expected distributions in finite samples for the simple linear regression model and the more complex GGM. This problem was especially pronounced for small sample sizes in combination with complex models. Both fundamental requirements—the normal distribution of the scores and the unbiased estimation of the Fisher information matrix—are not met, however, even in setups where the sampling distribution seems properly specified.

2.2. *Formal Analysis of the Approximation Error and Convergence Rate*

Now, we turn to a formal analysis of this error and derive the convergence rate of the normal approximation for the fluctuation process. The full analysis can be found in Appendix.

Hjort and Koning (2002) used a linear (i.e., second-order) Taylor expansion to derive the normal approximation of the fluctuation process for $\hat{\theta}$ near θ_0, which tends to be accurate in large samples. They furthermore show that the aggregated fluctuation process (i.e., the canonical monitoring process) approximates several independent Brownian bridges under the null hypothesis (see Hjort and Koning 2002 Eqs. (2.3) and (2.4), p. 116]). This approximation provides the basis to derive the SCT's sampling distribution. Hence, the sampling distribution will be valid (i.e., have correct type I error rate), if the error of the normal approximation of the fluctuation process goes to zero sufficiently fast. The error of the approximation can be assessed by looking at the Lagrange

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

remainder. Whereas Hjort and Koning ignored this Lagrange remainder in their derivations, we will assess the rate at which it convergences to zero.

The Lagrange remainder of a linear (i.e., second-order) Taylor approximation is defined as

$$E_2(\theta) = \frac{f''(\theta_s)}{2}(\hat{\theta} - \theta_0)^2,$$

for θ_s between θ_0 and $\hat{\theta}$. Therefore, the full Taylor expansion for the fluctuation process is:

$$\Psi(t;\hat{\theta}) = \overbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt \rfloor} s(y_i,\theta_0) + \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt \rfloor} i(y_i,\theta_0)(\hat{\theta}-\theta_0)}^{\text{Linear Taylor Approximation}} + \overbrace{\frac{1}{2\sqrt{n}}\sum_{i=1}^{\lfloor nt \rfloor} j(y_i,\theta_s)(\hat{\theta}-\theta_0)^2}^{\text{Lagrange Remainder}},$$

where $j(y_i,\theta)$ denotes the third-order derivative of the log-likelihood function (i.e., the second-order derivative of the score function). The Lagrange remainder consists of two parts $(\hat{\theta}-\theta_0)^2$ and $j(y_i,\theta_s)$. Observe that $(\hat{\theta}-\theta_0)^2$ is the standard error of the estimator, which tends to zero at rate $\frac{1}{\sqrt{n}}$ for an unbiased or asymptotically unbiased estimator. In Appendix, we show that the third-order derivative $j(y_i,\theta_s)$ for exponential family models, such as normal linear regression model and the GGM, is constant, if their moments are bounded. As a result, the Lagrange remainder is bounded by $\frac{1}{\sqrt{n}}$, thus tends to zero for sufficiently large sample sizes. Unfortunately, this convergence rate is not very fast, which means that the approximation error could be significant; thus, the fluctuation process cannot be accurately described using the Brownian bridge. In this case, the sampling distribution is misspecified (Zeileis 2006; Estrella 2003; Hansen 1997; Hjort and Koning 2002) and the reported p value is wrong.

3. A Monte Carlo Permutation Approach to the Structural Change Test

Permutation testing is a popular nonparametric method for statistical testing if distributional assumptions are not met. Zeileis and Hothorn (2013) used a permutation test approach to increase the power of the SCT in small samples for linear regression models. Even though their results were positive, the permutation test alternative to the asymptotic version of the SCT has found limited application. Here, we want to assess if a permutation test approach can estimate the correct sampling distribution in finite samples, even for large psychometric models, and consequently control the type I error rate. In permutation tests, first introduced by Fisher (1951), sampling distributions are obtained by calculating the test statistic values under all possible rearrangements of the observed data points. Applied to the SCT, it would thus consider all $n!$ rearrangements of the auxiliary variable v, and then compute a test statistic for every possible arrangement. Since the labels are exchangeable under the SCT's null hypothesis, the permutation test approach provides exact significance levels (Good 1993). Compared to parametric tests (e.g., the t-test, or F-test), permutation tests are equally powerful in large samples (Bickel and van Zwet 2012); however, permutation approaches are more powerful, if assumptions of the parametric tests are not met. The permutation approach's major drawback is that recomputing the statistic for all possible rearrangements can become unwieldy. A Monte Carlo approach has been proposed in which possible rearrangements are randomly sampled (Kaiser 2007; Frank and Witten 1998). This alternative overcomes the exact permutation tests' computational burden and provides an approximate permutation test. We will use the approximate permutation test approach and illustrate

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

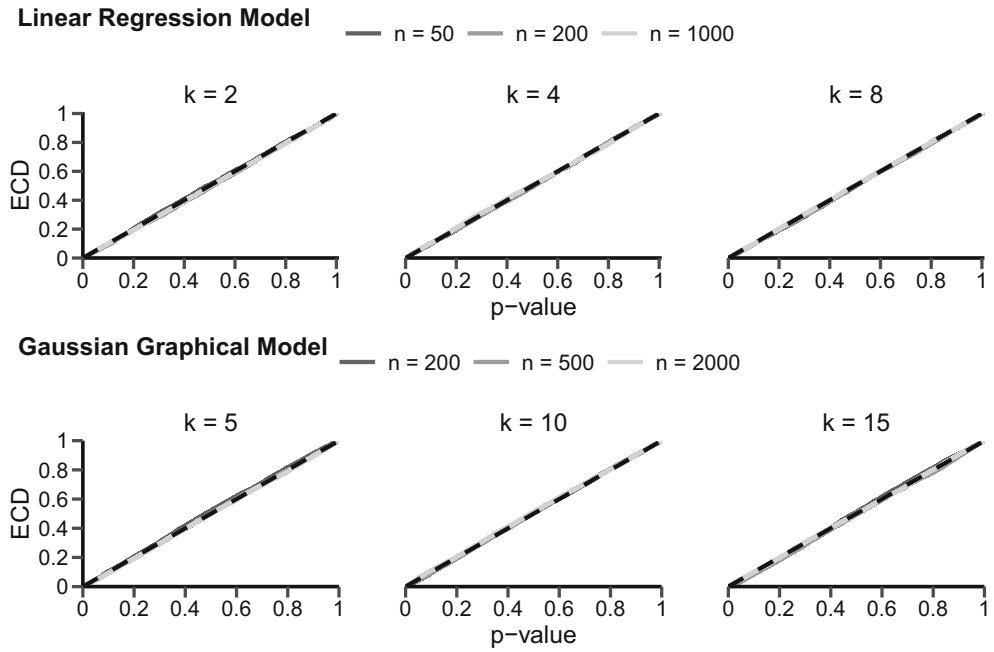


FIGURE 5.
Empirical cumulative distributions (ECDs) for the *p* value under the null hypothesis using the permutation approach. The top row shows the linear regression model results and the bottom row the results for the GGM. Here, *n* represents the sample size and *k* the number of covariates for the linear regression model and number of nodes for the GGM. In each plot, the black, dashed line shows the expected uniform distribution.

that it provides accurate sampling distributions, even for small sample sizes and large psychometric models.

The Monte Carlo permutation approach to the SCT comprises three steps. The test statistic for the original dataset is computed in the first step. We will consider the maxLM test statistic in Eq. (4) here. In the second step, we randomly rearrange the values of the grouping variable v. For example, say we have an original dataset with six observations belonging to two subgroups (i.e., Group A: 1, 2, and 3; Group B: 4, 5, and 6). After rearranging, observations three, four, and six might now belong to Group A and observations one, two, and five to Group B (i.e., Group A: 3, 4, and 6; Group B: 1, 2, 5). The maxLM test statistic is computed for every random rearrangement. We have used 5,000 random rearrangements in our simulations. It gave a good trade-off between accuracy and computation speed; however, the more samples are obtained, the more accurate the determined *p* value. In the final step, we estimate the *p* value by calculating how many resampled test statistics were larger than the original statistic.

We revisit the previous section's simulations to illustrate the SCT's behavior when combined with the Monte Carlo permutation test approach. The results are shown in Fig. 5. It is evident that the *p* values now nicely follow a uniform distribution in all simulation setups. No differences can be found depending on sample size or model complexity. Thus, the permutation test approach has solved the underlying issue of the misspecified sampling distribution. Furthermore, we have investigated the power of the SCT under the permutation alternative for GGMs. The effect of the permutation approach on the power of the SCT has been studied before for linear models (e.g., Hothorn and Zeileis 2008; Zeileis and Hothorn 2013). Previous results indicate that permutation testing leads to small power improvements in large samples but leads to huge power improvements in small samples. We show the results of our analyses for GGM in Table 1. In line with previous

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

TABLE 1.
Power of the SCT using the common asymptotic approach and the permutation alternative

		k = 5			k = 10			k = 15		
n	Distribution / $\Delta\theta$ =	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
200	Asymptotic Approach	0.57	0.85	0.96	0.01	0.02	0.21	0.00	0.00	0.00
	Permutation Alternative	0.60	0.85	0.97	0.05	0.15	0.52	0.03	0.10	0.19
500	Asymptotic Approach	0.70	0.98	1.00	0.03	0.30	0.94	0.01	0.02	0.31
	Permutation Alternative	0.69	0.98	1.00	0.07	0.45	0.98	0.07	0.22	0.65
2,000	Asymptotic Approach	0.90	1.00	1.00	0.13	0.99	1.00	0.06	0.86	1.00
	Permutation Alternative	0.92	1.00	1.00	0.19	1.00	1.00	0.09	0.91	1.00

We altered sample size (i.e., n = 200, 500, and 2,000), amount of nodes (i.e., k = 5, 10, and 15) as well as the size of the parameter invariance (i.e., $\Delta\theta$ = 0.1, 0.3, and 0.5). Datasets were simulated using a similar simulation setup as Jones et al. (2020).

results for linear models, we find that the power of the permutation test approach is similar to the power based on the asymptotic sampling distribution for large samples. In small samples, however, we see considerable improvements.

Based on these results, we conclude that the Monte Carlo permutation approach is a valuable method to perform the SCT, particularly for large psychometric models and small sample sizes. It leads to correct specification of the sampling distribution under the null hypothesis and tends to improve power.

Unfortunately, the permutation alternative comes with high additional computational efforts, which might hamper the use of the permutation approach in large models. However, the biggest misspecification is established for larger models and thus the permutation approach has the greatest benefit here. Therefore, alternatives should be considered that allow for the assessment of parameter invariance but employ sound inference also in finite samples. As a concrete example, we could feed the SCT's scores into the conditional inference tree (CTree; Hothorn et al. 2006). CTree is a recursive partitioning algorithm assessing parameter invariance using permutation approaches to obtain the sampling distribution (Schlosser et al. 2019). The algorithm evaluates the association between the transformed responses $h(Y)$ and each of the transformed splitting variables $g(Z_j)$. CTree requires the specification of an influence function $h(\cdot)$ and the transformed split variable function $g(\cdot)$. In case a parametric model is fitted to the observed data, one can obtain a model-based transformation function $h(\cdot)$, for example, a score-function. Here, CTree closely resembles the SCT, however, leveraging a conditional inference framework. Through simulations, we could confirm that CTree circumvents the small sample issues in finite samples also for large models; the p value distribution is uniform for all setups of k and n (see Figure S8 in the online appendix). Thus, CTree serves as an additional solution to circumvent a misspecified p value in small samples, which is readily implemented in the R-package `networktree` for GGMs (Jones et al. 2020).

4. Discussion

This paper has shown that the score-based structural change test's (SCT's) small sample behavior can be problematic, especially for large psychometric models. The SCT assumes that the accumulation of scores for a parameter across observations resembles a Brownian bridge. This property holds for large samples but not for small samples. As a result, standard methods that rely on this asymptotic property cannot determine the SCT's correct sampling distribution.

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

However, using a Monte Carlo permutation test approach, the proper sampling distribution can be obtained. As a result, the correct p values can be determined even for small samples and large psychometric models. This permutation approach also improves the power of the test.

Previous research has observed finite sample problems for the SCT, concluding that the SCT constitutes a conservative test with decreasing sample size (Strobl et al. 2015; Merkle and Zeileis 2013; Hothorn and Zeileis 2008). The SCT's sampling distribution is misspecified in small samples, which leads to problems with controlling the type I error, thus, incorrect inference. These findings are concerning. In particular, since the SCT has been adapted to larger psychometric models where small sample issues are amplified (e.g., Strobl et al. 2015; Jones et al. 2020). These adaptations additionally combine large models with the model-based recursive partitioning (MOB). The MOB is a recursive algorithm that uses the SCT to detect parameter invariance and splits data into smaller subsets for which it recalculates the SCT (Merkle and Zeileis 2013). When using subsets of the data in recursive applications of the SCT, its small sample properties can become more pronounced. In sum, the small sample properties of the SCT are a timely matter.

The Monte Carlo permutation approach offers a straightforward alternative to obtain the sampling distribution. We showed that the alternative approach leads to a correct specification of the sampling distribution, and consequently, a correct specification of the p value distribution. Also, we were able to corroborate previous demonstrations of the increased power of the permutation approach over asymptotic approximations in linear models (Hothorn and Zeileis 2008; Zeileis and Hothorn 2013), by extending these results to the GGM. We found higher power for the permutation approach, especially for large psychometric models combined with small sample sizes. Therefore, the permutation approach solves the issue of the misspecified sampling distribution and increases the power of the test, leading to an optimal result in finite samples.

The permutation test algorithm comes with some drawbacks. First, the algorithm has a highly increased computational effort. Our current implementation is in base R (R-Core-Team 2020) and could be improved through computationally more efficient programming languages like the Rccp plugin for R (Eddelbuettel and François 2011). Second, researchers should be wary that the permutation results depend on the number of permutations and random seeds of the sampling algorithm. It is therefore advised to use a sufficiently large number of permutations and test different random seeds. Nonetheless, by solving the misspecified sampling distribution and increasing the power, we conclude—similar to previous researchers—that the additional computational power needed for the permutation approach is justified and necessary in finite samples (Hothorn and Zeileis 2008; Zeileis and Hothorn 2013).

Still, researchers should consider alternative parameter invariance tests. In the paper, we discussed the readily implemented algorithm CTree (Schlosser et al. 2019; Hothorn et al. 2006). CTree provides a correct p value distribution also in finite samples, however, has even less power than the asymptotic SCT algorithm. Thus, if researchers are mainly concerned with a high power, it would still be advisable to apply the permutation approach to the SCT. Other alternatives could still be derived. Here, we need to understand the nature of the problem; in particular, the two fundamental requirements: the normal distribution of the scores and the unbiased estimation of the information matrix. Our preliminary assessment showed that the scores of the interaction parameters do not follow a normal distribution and the bias of the Fisher information matrix is inflated. Astoundingly, the assumptions are not met, even in situations where the sampling distribution is properly specified. More research is needed to understand the nature of this issue and perhaps offer parametric alternatives for the sampling distributions.

To conclude, finite sample misspecification of the structural change test needs to be acknowledged. Here, permutation approaches are a superior method to standard asymptotic approximations of the sampling distribution. Especially in large psychometric models, a wider adaptation of permutation approaches for SCTs is advisable.

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix: Formal Analysis of the Approximation Error and Rate of Convergence

We derive the error associated with the Brownian bridge approximation of the fluctuation process. First, we introduce the derivation of this approximation as shown in Hjort and Koning (2002). Second, we derive the error associated with the approximation, the Lagrange remainder in the Taylor approximation and conclude its convergence rate is bounded by $1/\sqrt{n}$.

The Cumulative Score Process

Let $s(y_i, \theta)$ denote the first-order derivative of the log-likelihood function g with respect to θ—the score—and $i(y_i, \theta)$ the second-order derivative. To determine whether scores fluctuate along a third variable of interest (e.g., gender, time), we compute the cumulative sum of the score:

$$\Psi(t; \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} s(y_i, \theta_0),$$

where $\lfloor nt \rfloor$ is the floor function of $n \times t$—index n the sample size and index t a fraction of all n participants (i.e., $t = i/n$ for $i = 1, \ldots, n$). Here, θ_0 describes the parameter estimate under the null-hypothesis. The mean of the cumulative score process is zero and the variance the information matrix $J = -E(i(y_i, \theta_0))$. Given the Donsker and Cramér-Wold Theorem, one can derive

$$\Psi(t; \theta_0) \xrightarrow{d} Z_0(t) \text{ in } D_p[0, 1],$$

where $Z_0(t)$ is a zero-mean Gaussian, which is a linear transformation of independent Brownian motions (Hjort and Koning 2002). This convergence takes place in the the space $D_p[0, 1]$ thus for t being in the range zero to one (i.e., $t \in [0, 1]$).

The Estimated Cumulative Score Process

Given that θ_0 is commonly unknown, we use the maximum likelihood estimator (MLE)—$\hat{\theta}$—and calculate the cumulative score process as

$$\Psi(t; \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} s(y_i, \hat{\theta}).$$

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

For MLE estimators the cumulative score process is bounded at zero, both at $t = 0$ and $t = 1$. Using a Taylor expansion up to the first-order derivative, e.g., which for a function f would be,

$$f(\hat{\theta}) = f(\theta_0) + f'(\theta_0)(\hat{\theta} - \theta_0),$$

Hjort and Koning approximate the cumulative score process for $\hat{\theta}$ near θ_0 as

$$\Psi(t; \hat{\theta}) \doteq \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} s(y_i, \theta_0) + \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} i(y_i, \theta_0)(\hat{\theta} - \theta_0)$$

where $\doteq$ denotes an approximate equation. The linear approximation using the first- and second-order derivative at θ_0 tends to approximate the cumulative score process of $\hat{\theta}$ in probability. Hjort and Koning use this Taylor expansion to derive a canonical monitoring process that approximates several independent Brownian bridges under the null hypothesis (see Hjort and Koning 2002, Eqs. (2.3) and (2.4), p. 116).

The Approximation Error

Hjort and Koning ignore the Lagrange remainder of the Taylor expansion. The Lagrange remainder characterizes the error associated with the approximation of $\hat{\theta}$. The full Taylor expansion for a function f is commonly written as:

$$f(\hat{\theta}) = f(\theta_0) + f'(\theta_0)(\hat{\theta} - \theta_0) + E_2(\theta),$$

where $E_2(\theta)$ denotes the Lagrange remainder. The largest term in $E_2(\theta)$ can be described by:

$$E_2(\theta) = \frac{f''(\theta_s)}{2}(\hat{\theta} - \theta_0)^2,$$

for θ_s between θ_0 and $\hat{\theta}$. More specifically, the full Taylor expansion for the cumulative score process is:

$$\Psi(t; \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} s(y_i, \theta_0) + \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} i(y_i, \theta_0)(\hat{\theta} - \theta_0) + \frac{1}{2\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} j(y_i, \theta_s)(\hat{\theta} - \theta_0)^2,$$

where $j(y_i, \theta)$ denotes the third-order derivative of the log-likelihood function g.
We will next assess the Lagrange remainder to determine the magnitude of the error associated with the linear Taylor approximation dependent on the sample size. We will discuss both facets composing the error: $(\hat{\theta} - \theta_0)^2$ and $f''(\theta_s)$. First, we evaluate $(\hat{\theta} - \theta_0)^2$, where θ_0 denotes the parameter estimate under $\mathcal{H}_0 : \theta_1 = \ldots = \theta_n$. Since θ_0 is unknown, we use its MLE—$\hat{\theta}$. Observe that in this case $(\hat{\theta} - \theta_0)^2$ is the standard error of the approximation and for an unbiased or asymptotically unbiased estimator it holds that $(\hat{\theta} - \theta_0)^2 = \mathcal{O}_p(1/\sqrt{n})$. Here, $\mathcal{O}_p$ is the Big-O in probability notation for random variables X_n and set of constants m_n. The notation $X_n = \mathcal{O}_p(m_n)$ states that there is a finite, positive M and an n_0 such that $P(|X_n/m_n| > M) < \epsilon$ for all $n > n_0$ and any positive ϵ. Thus, $(\hat{\theta} - \theta_0)^2$ is bounded by $1/\sqrt{n}$.

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

Second, we evaluate $f''(\theta_s)$ which comprises the individual third-order derivatives $j(y_i, \theta_s)$ of the log-likelihood function g. To illustrate this derivative, we assume g is part of the exponential family:

$$p(x \mid \eta) = h(x)e^{\eta^\mathsf{T} t(x) - A(\eta)},$$

where $h(x)$ denotes the base function, η the natural parameter of the model, $t(x)$ denotes the sufficient statistic, and $A(\eta)$ the log-normalizing constant—$\int_x h(x)\exp(\eta^\mathsf{T} t(x))\mathrm{d}x$—that ensures that the density integrates to one. The first-, second-, and third-order derivatives of exponential family distributions w.r.t. the natural parameter are

$$\begin{aligned}
\frac{\partial p(x \mid \eta)}{\partial \eta_i} &= h(x)e^{\eta^\mathsf{T} t(x) - A(\eta)} \left(t(x)_i - \frac{\partial}{\partial \eta_i} A(\eta) \right), \\
\frac{\partial^2 p(x \mid \eta)}{\partial^2 \eta_i} &= h(x)e^{\eta^\mathsf{T} t(x) - A(\eta)} \left(\left(t(x)_i - \frac{\partial}{\partial \eta_i} A(\eta) \right)^2 + \frac{\partial^2}{\partial^2 \eta_i} A(\eta) \right), \\
\frac{\partial^3 p(x \mid \eta)}{\partial^3 \eta_i} &= h(x)e^{\eta^\mathsf{T} t(x) - A(\eta)} \\
&\quad \times \left(\left(t(x)_i - \frac{\partial}{\partial \eta_i} A(\eta) \right) \left(\left(t(x)_i - \frac{\partial}{\partial \eta_i} A(\eta) \right)^2 - 3\frac{\partial^2}{\partial^2 \eta_i} A(\eta) \right) - \frac{\partial^3}{\partial^3 \eta_i} A(\eta) \right).
\end{aligned}$$

The third-order derivative consists of two parts $h(x)e^{\eta^\mathsf{T} t(x) - A(\eta)}$ and everything inside the bracket. Note that the first part is the distribution itself and is bounded to lie between zero and one. Therefore, we need to take a closer look at the second part, which mainly depends on the derivatives of $A(\eta)$. It is a convenient feature of the exponential family distributions that the moments of the sufficient statistics can be derived from the derivatives of $A(\eta)$. We will show this for the first moment, but it can be shown for all other moments.

$$\begin{aligned}
\frac{\partial}{\partial \eta_i} A(\eta) &= \frac{\partial}{\partial \eta_i} \left\{ \log \int h(x)e^{\eta^\mathsf{T} t(x)} \mathrm{d}x \right\} \\
&= \frac{\int t(x)_i h(x)e^{\eta^\mathsf{T} t(x)} \mathrm{d}x}{\int h(x)e^{\eta^\mathsf{T} t(x)} \mathrm{d}x} \\
&= \int t(x)_i h(x)e^{\eta^\mathsf{T} t(x) - A(\eta)} \mathrm{d}x \\
&= \mathbb{E}[t(x)_i]
\end{aligned}$$

Thus, if the moments of the specific exponential family distribution are bounded, the third-order derivative is bounded and $f'''(\hat{\theta}) = \mathcal{O}(1)$.
Taking everything together—$(\hat{\theta} - \theta_0)^2 = \mathcal{O}_p(1/\sqrt{n})$ and $f'''(\hat{\theta}) = \mathcal{O}(1)$—we obtain:

$$\Psi(t; \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} s(y_i, \theta_0) + \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} i(y_i, \theta_0)(\hat{\theta} - \theta_0) + \mathcal{O}_p\left(\frac{1}{\sqrt{n}} \right).$$

This shows that the approximation error depends on the sample size, and the error will be larger for smaller samples. The approximation error tends to zero as the sample size grows. If the sample size is sufficiently large, the calculations of Hjort and Koning hold.

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

References

Andrews, D. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica, 61,* 821–856.

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika, 80,* 317–340. https://doi.org/10.1007/s11336-014-9408-y.

Bickel, P. J., & van Zwet, W. R. (2012). Asymptotic expansions for the power of distribution-free tests in the two-sample problem. In S. van de Geer & M. Wegkamp (Eds.), *Selected Works of Willem van Zwet* (pp. 117–184). New York, NY: Springer.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44,* 176–181. https://doi.org/10.1097/01.mlr.0000245143.08679.cc.

Breslau, J., Javaras, K. N., Blacker, D., Murphy, J. M., & Normand. (2008). Differential item functioning between ethnic groups in the epidemiological assessment of depression. *The Journal of Nervous and Mental Disease, 196,* 297–306. https://doi.org/10.1097/NMD.0b013e31816a490e.

Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society: Series B (Methodological), 37,* 149–163. https://doi.org/10.1111/j.2517-6161.1975.tb01532.x.

Chang, M. - J., & Su, C. - Y. (2014). The dynamic relationship between exchange rates and macroeconomic fundamentals: Evidence from Pacific Rim countries. *Journal of International Financial Markets, Institutions and Money, 30,* 220–246. https://doi.org/10.1016/j.intfin.2014.03.002.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v040.i08.

Estrella, A. (2003). Critical values and p values of bessel process distributions: Computation and application to structural break tests. *Econometric Theory*. https://doi.org/10.1017/S0266466603196107.

Fisher, R. A. (1951). *The design of experiments* (6th ed.). New York, NY: Hafner.

Frank, E., & Witten, I. H. (1998). Using a permutation test for attribute selection in decision trees. In *Proceedings of the fifteenth international conference on machine learning* (pp. 152–160). Morgan Kaufmann.

Good, P. (1993). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. Springer.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P., & Stahel, W. (Eds.). (2005). *Robust statistics: The approach based on influence functions (digital print ed.)*. Wiley.

Hansen, B. E. (1997). Approximate asymptotic P values for structural-change tests. *Journal of Business & Economic Statistics, 15,* 60. https://doi.org/10.2307/1392074.

Hjort, N. L., & Koning, A. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics, 14,* 113–132. https://doi.org/10.1080/10485250211394.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15,* 651–674. https://doi.org/10.1198/106186006X133933.

Hothorn, T., & Zeileis, A. (2008). Generalized Maximally Selected Statistics Generalized Maximally Selected Statistics. *Biometrics, 64,* 1263–1269. https://doi.org/10.1111/j.1541-0420.2008.00995.x.

Hothorn, T., & Zeileis, A. (2015). Partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research, 16,* 3905–3909.

Jones, P. J., Mair, P., Simon, T., & Zeileis, A. (2020). Network trees: A method for recursively partitioning covariance structures. *Psychometrika, 85,* 926–945. https://doi.org/10.31219/osf.io/ha4cw.

Kaiser, J. (2007). An exact and a Monte Carlo proposal to the Fisher–Pitman permutation tests for paired replicates and for independent samples. *The Stata Journal: Promoting Communications on Statistics and Stata, 7,* 402–412. https://doi.org/10.1177/1536867X0700700307.

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry, 17,* 1174–1179. https://doi.org/10.1038/mp.2012.105.

Kuan, C.-M., & Hornik, K. (1995). The generalized fluctuation test: A unifying view. *Econometric Reviews, 14,* 135–161. https://doi.org/10.1080/07474939508800311.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127–143. https://doi.org/10.1016/0883-0355(89)90002-5.

Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika, 79,* 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,* 59–82. https://doi.org/10.1007/s11336-012-9302-4.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference* (Vol. 95). Sage.

Mulaudzi, M. C. (2016). Testing measurement invariance of the learning programme management and evaluation scale across academic achievement. *SA Journal of Human Resource Management, 15,* 1–8. https://doi.org/10.4102/sajhrm.v14i1.730.

O'Connell, C. S., Ruan, L., & Silver, W. L. (2018). Drought drives rapid shifts in tropical rainforest soil biogeochemistry and greenhouse gas emissions. *Nature Communications, 9,* 1–9. https://doi.org/10.1038/s41467-018-03352-3.

Ploberger, W., & Kramer, W. (1992). The CUSUM test with OLS residuals. *Econometrica, 60,* 271. https://doi.org/10.2307/2951597.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41,* 71–90. https://doi.org/10.1016/j.dr.2016.06.004.

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press

R-Core-Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
Schlosser, L., Hothorn, T., & Zeileis, A. (2019). The power of unbiased recursive partitioning: A unifying view of CTree, MOB, and GUIDE. arXiv:1906.10179.
Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika, 80,* 289–316. https://doi.org/10.1007/s11336-013-9388-3.
van de Schoot, R., Schmidt, P., de Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology, 6,* 1064. https://doi.org/10.3389/fpsyg.2015.01064.
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. https://doi.org/10.1007/s11336-017-9591-8.
Zeileis, A. (2006). Implementing a class of structural change tests: An econometric computing approach. *Computational Statistics & Data Analysis, 50,* 2987–3008. https://doi.org/10.1016/j.csda.2005.07.001.
Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica, 61,* 488–508. https://doi.org/10.1111/j.1467-9574.2007.00371.x.
Zeileis, A., & Hothorn, T. (2013). A toolbox of permutation tests for structural change. *Statistical Papers, 54,* 931–954. https://doi.org/10.1007/s00362-013-0503-4.
Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics, 17,* 492–514. https://doi.org/10.1198/106186008X319331.
Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2002). Strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software, 7,* 1–38. https://doi.org/10.18637/jss.v007.i02.

Manuscript Received: 10 DEC 2020
Final Version Received: 13 DEC 2021
Published Online Date: 1 FEB 2022

https://doi.org/10.1007/s11336-021-09834-6 Published online by Cambridge University Press