A Note on the Structural Change Test in Finite Samples: Using a Permutation Approach to Estimate the Sampling Distribution

Huth, K.[1,2,3], Waldorp, L. J.[1], Luigjes, J.[2], Goudriaan, A. E. [2,4], van Holst, R. J. [2,3], & Marsman, and M.[1]

1 Department of Psychology

University of Amsterdam

2 Department of Psychiatry

Amsterdam University Medical Center

3 Centre for Urban Mental Health

University of Amsterdam

4 Arkin Mental Health Institute

Author Note

Correspondence concerning this article should be addressed to Karoline Huth, University of Amsterdam, Psychological Methods, Nieuwe Achtergracht 129B, PO Box 15906, 1001 NK Amsterdam, The Netherlands. E-mail: k.huth@uva.nl.

Abstract

Equal parameter estimates across subgroups - measurement invariance - is a substantial requirement of statistical tests. Ignoring subgroup differences poses a major threat to study replicability, model specification, and theory development. One powerful statistical method that allows testing for measurement invariance is structural change tests. A core element of those tests is the empirical fluctuation process. In the case of measurement invariance, the fluctuation process asymptotically follows a Brownian bridge. This asymptotic assumption further provides the basis for inference. However, in this paper, we show that the empirical fluctuation process does not follow a Brownian bridge in small samples. Thus, methods of obtaining the sampling distribution are incorrect, and the p-value misspecified. Therefore, we propose and implement an alternative solution to obtaining the sampling distribution - permutation approaches. Permutation approaches obtain the sampling distribution through resampling of the dataset, avoiding unmet distributional assumptions. We show that the permutation approach serves as a viable alternative, permitting valid inferential conclusions.

Keywords: measurement invariance, parameter stability, structural change test, finite sample behavior, permutation test

A Note on the Structural Change Test in Finite Samples: Using a Permutation Approach to Estimate the Sampling Distribution

The assumption of measurement invariance —called 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). Formally, measurement invariance is defined as (Mellenbergh, 1989, p. 129),

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

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 measurement invariance. Thus, measurement 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 measurement occasions (Putnick and Bornstein, 2016; van de Schoot et al., 2015). Violations of measurement invariance 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 the measurement invariance assumption, which poses a major threat to research development (Borsboom, 2006).

Structural change tests (SCTs) allow us to test for 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 measurement invariance because they can be straightforwardly implemented, even for complicated statistical models: SCTs

do not require explicit specification of which parameter diverges or which subgroups behave differently (Wang et al., 2018). At their core, SCTs use scores (i.e., partial derivatives of the log-likelihood function with respect to a particular parameter) to determine whether parameters are invariant across subgroups. Scores are similar to asymptotic influence functions which are used to determine the effect of single observations on the estimate (Hampel et al., 2005). The basic premise is that if measurement 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. However, if the fluctuation of the aggregated scores systematically coincides with an auxiliary variable v, measurement 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), and is derived from the observation that aggregated scores behave like a Brownian bridge asymptotically. Unfortunately, the SCTs in finite samples are barely studied. Our concern is that the aggregated scores do not approximate a Brownian bridge in finite samples. An invalid asymptotic approximation has grave consequences for the SCT; the sampling distribution may be misspecified, and we cannot control for the type 1 error. In sum, the null-hypothesis statistical test would be wrong. In light of this concern, our goals are twofold. Our first goal is to assess the SCT's behavior in finite samples. In particular, we investigate the distribution of the p-value, which should be uniformly distributed under the null hypothesis (Hung et al., 1997). For finite samples, the p-values are not uniformly distributed under the null hypothesis, and this problem becomes more pronounced in complex models. Our second goal is to show that permutation approaches offer a simple and viable solution to the problem at hand. Permutation approaches allow for estimation of the sampling distribution when distributional assumptions do not hold or are analytically intractable (Mooney and Duval, 1993).

The remainder of this paper is organized as follows. First, we introduce the SCT in

detail. Then, 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. We then elaborate on an alternative approach to obtaining the sampling distribution - permutation approaches. To illustrate the issues and our solution, we will use a simple linear regression model and a more complex Gaussian Graphical model throughout this paper.

Structural Change Tests

The SCT assesses the equivalence of model parameters 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 scores. 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)). Once the MLEs of the model parameters are obtained, the score for every particular parameter and observation can be calculated. The score is defined as the gradient of the log-likelihood function. For a parameter θ_j the score of an observation y_i 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, we know that 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.

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 $n \times t$, t 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):

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

where $\hat{I}$ is the asymptotic covariance matrix of the scores —i.e., the Fisher information matrix (Zeileis, 2006). Observe that the cumulated scores $\mathrm{B}(t;\, \hat{\theta}_j)$ are zero for $t = 0$ and $t = 1$. 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). Looking at a model with k-parameters, the

fluctuation processes asymptotically approximate 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 measurement invariance holds. However, in the case of measurement non-invariance, the process systematically deviates from zero. Figure 1 provides an illustration.

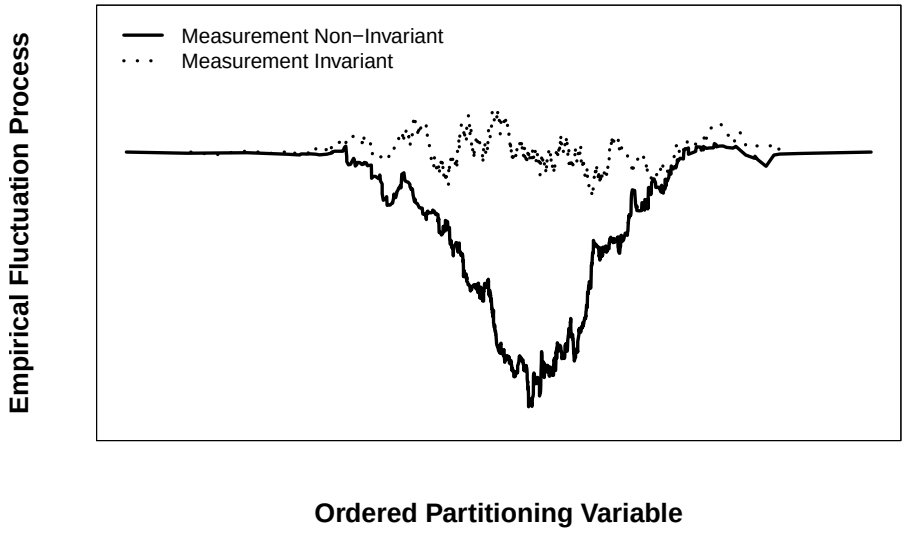

Figure 1

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; measurement invariance for this parameter is violated.

The cumulative scores are combined into a test statistic in the third and final step. The cumulative scores can be combined in various ways (Merkle and Zeileis, 2013; Hjort and Koning, 2002). We will denote the fluctuation process at an observation i for a parameter $\hat{\theta}_j$ with B_{ij}, i.e., $\mathrm{B}_{ij} = \mathrm{B}(t = \frac{i}{n};\, \hat{\theta}_j)$. Next, we introduce the three test statistics

that are commonly used in literature: The double maximum statistic (DM), the Cramér-von Mises statistic (CvM), and the maximum Lagrange Multiplier statistic (maxLM).

$$\text{DM} = \max_{i=1,\dots,n} \max_{j=1,\dots,k} |\text{B}_{ij}| \tag{2}$$

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

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

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 recent applications of the SCT, the maxLM statistic is the most popular of the three (Jones et al., 2019; Wang et al., 2018).

In null hypothesis significance testing, the test statistic computed from observed data is compared against the sampling distribution to obtain a p-value. The asymptotic sampling distributions of the three statistics above have been analyzed in several papers. For example, Hjort and Koning (2002) state that CvM-type statistics follow an approximate χ^2-distribution, and Zeileis (2006) show that this is also the case for DM-type statistics. Furthermore, Hansen (1997) and Estrella (2003) show that the sampling distribution of maxLM-type statistics also approximates a χ^2-distribution. In sum, the sampling distribution of all three statistics converges to a χ^2-distribution in the large sample limit. There have been different suggestions to set the degrees of freedom of these limiting distributions. It depends on the number of parameters and the point where the focal parameter changes value in a non-trivial way. For specific combinations of test

statistic, number of parameters, and change-point, tables with critical values can be found in Andrews (1993), Hansen (1997), and Estrella (2003). An alternative, more general procedure is to produce a sampling distribution by first simulating observations from a Brownian bridge and then computing the relevant statistic from the generated data (Andrews, 1993; Zeileis, 2006). This is currently the most popular method for determining the sampling distribution and the one we will use here.

Small Sample Behavior of the Structural Change Test

While the SCT's large sample behavior has been extensively studied, this is not the case for its behavior in small samples. To the best of our knowledge, only one study looked at the probability of a type 1 error of the SCT in small samples. Jones et al. (2019) simulated data for Gaussian Graphical models and looked at the type 1 error, altering sample size, network size, and invariance violation. Their simulations indicated that the type 1 error decreased with increasing model complexity; surprisingly, the type 1 error was always below their significance level. However, Jones et al. (2019) did not assess the assumptions that underly the SCT in their study, i.e., if the sampling distribution was properly specified. We turn to this analysis next.

We will analyze the SCT's behavior 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 linear regression, we simulated models with two, four, and eight covariates 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 5,000 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 measurement invariance. 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).

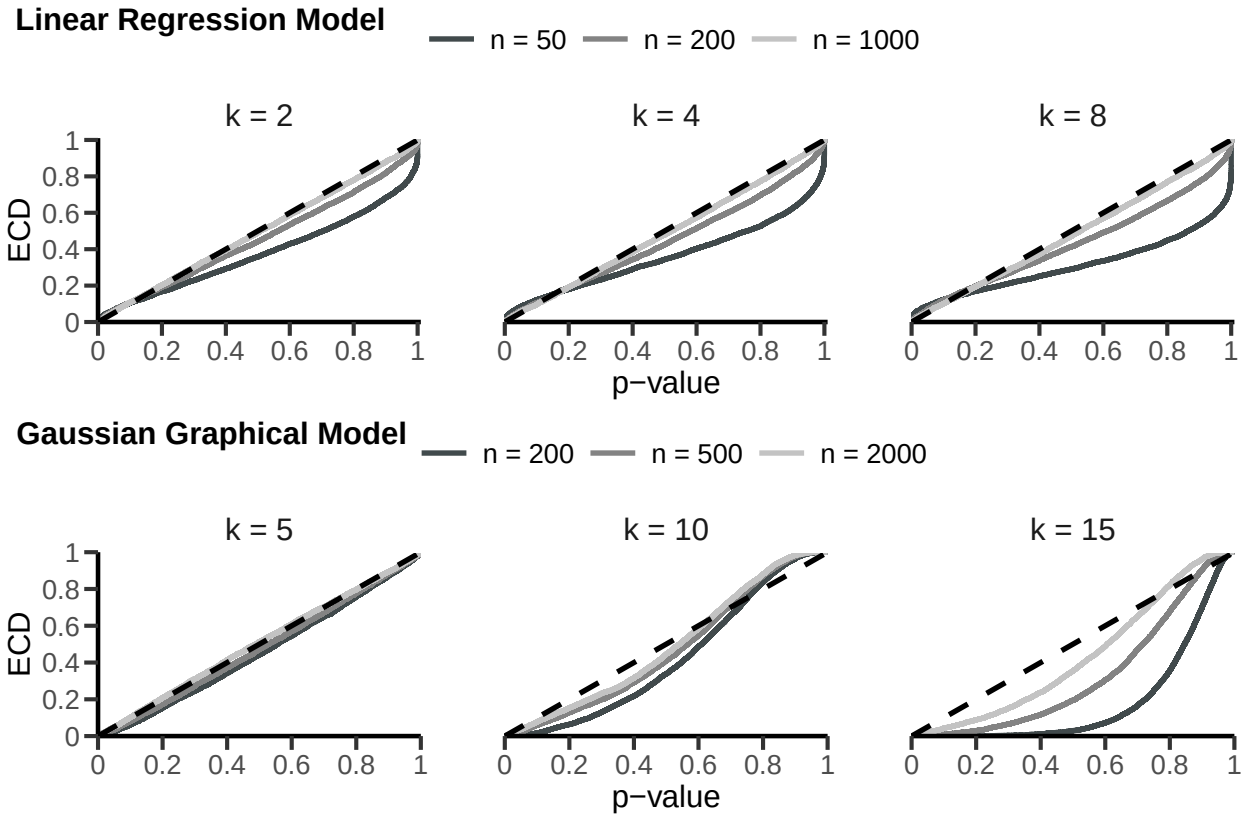

Figure 2

Empirical cumulative distributions 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 of the results for the GGM. In each plot, the black, dashed line represents the expected uniform distribution.

We will focus here on the results for the maxLM statistic and report the results for the CvM and DM statistics in the online appendix.[1] The simulated p-value distributions are shown in Figure 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 Figure 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

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

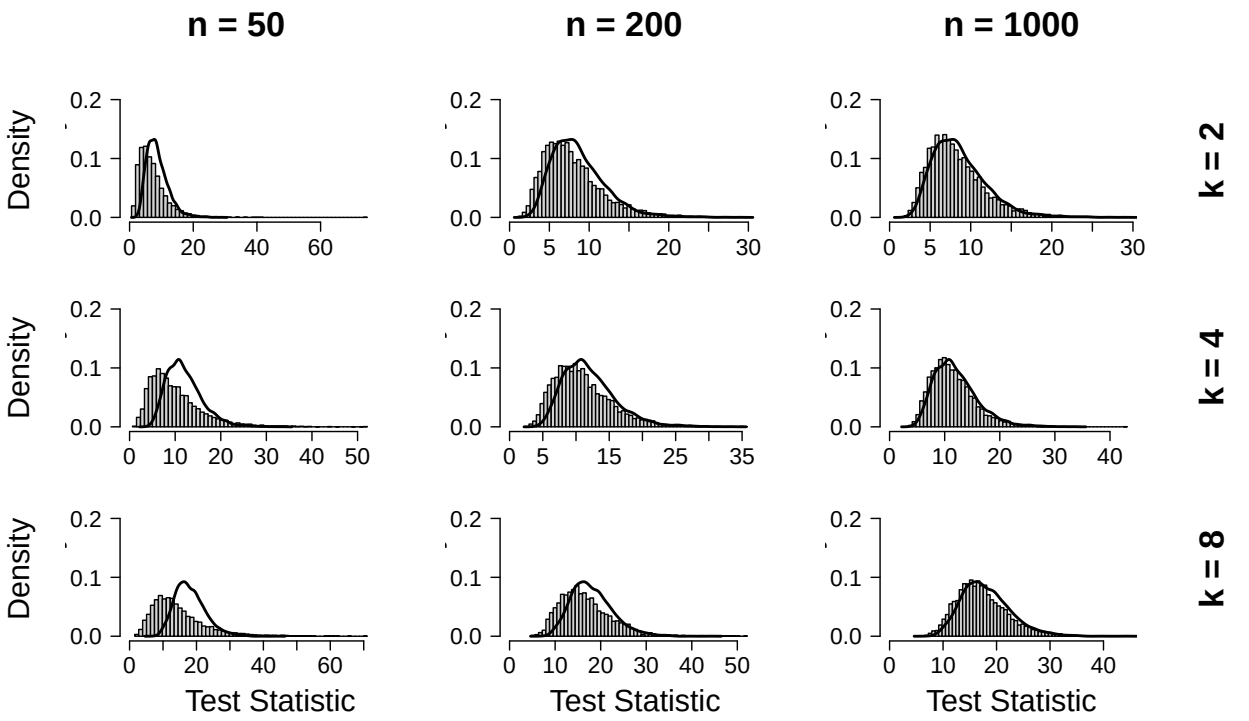

Figure 3

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)).

sample size increases. For the GGM, the p-value is nearly uniformly distributed 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. However, even with 2000 observations, the p-value does not follow a uniform distribution. In sum, the p-value is not necessarily uniformly distributed under the null hypothesis in finite samples, and the deviation between its distribution from the correct uniform distribution increases with model complexity.

The simulated sampling distribution are shown in Figure 3 for the linear regression model and in Figure 4 for the GGM. The asymptotic sampling distributions are indicated

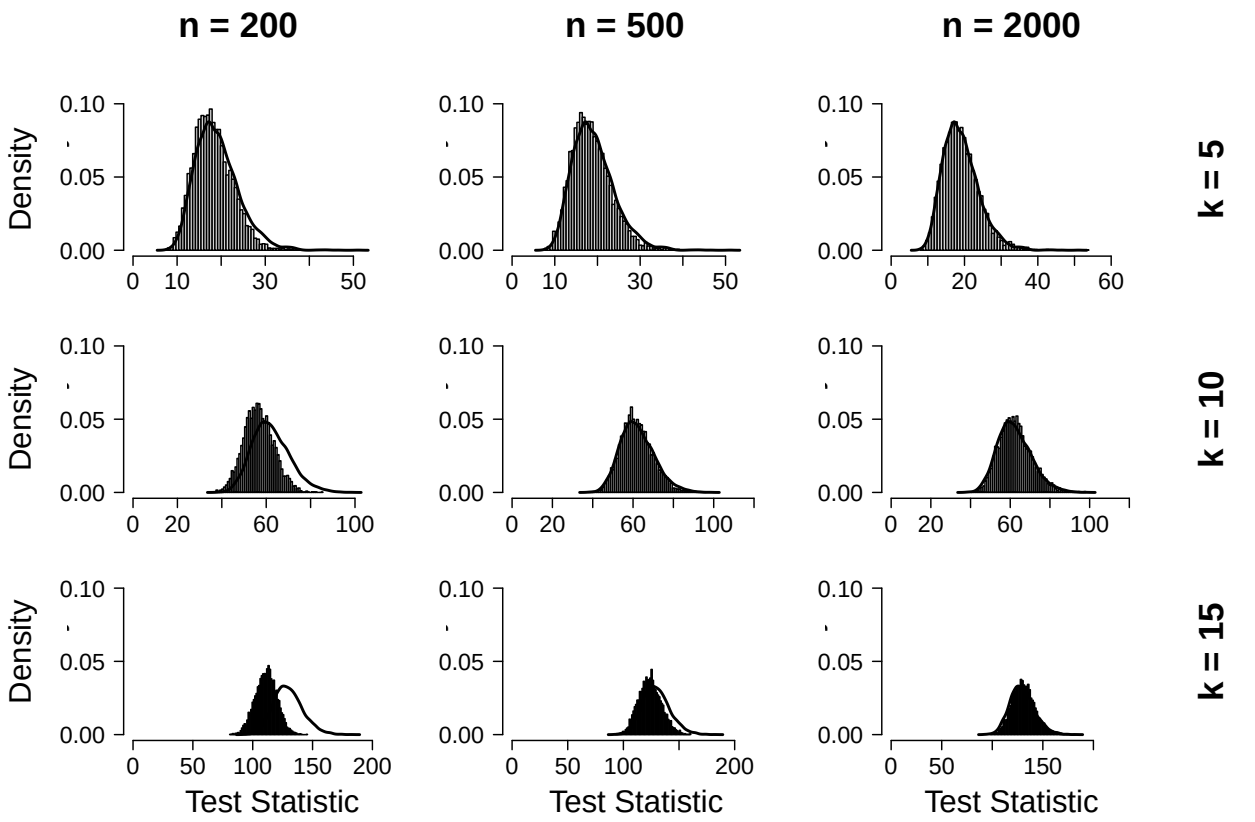

Figure 4

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)).

with a black solid line in these graphs. They were generated by repeatedly simulating values from a Brownian bridge and then computing the statistic on the generated data (e.g., see Andrews, 1993; Zeileis, 2006). It is clear 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.

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 is not arbitrary as it can have a large impact on the obtained sampling distributions. By cherrypicking the cut-off points, we could, in principle, improve the fit of the estimated

sampling distribution. Unfortunately, we do not know which cut-off values will lead to a good approximation a priori. 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. These cut-off points were also used by (Jones et al., 2019). For the linear model, which has few free parameters, we chose to cut-off the bottom and upper 10%.

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. The distribution of the p-value and the test statistic deviate from their expected distributions since the fluctuation process does not sufficiently approximate a Brownian bridge in small samples. The formal derivations of the error in this approximation is detailed in the Appendix. If the fluctuation process does not approximate a Brownian bridge, the sampling distribution is unjustified (Zeileis, 2006; Estrella, 2003; Hansen, 1997; Hjort and Koning, 2002) and the reported p-value is wrong. Conclusions drawn from the SCT in finite samples are therefore incorrect. To overcome this problem, we introduce an alternative approach to obtain the sampling distribution. Our approach will allow the use of the SCT in finite samples, even for complex models.

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. 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 (Kaiser, 2007). 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 might be more powerful if the assumptions of the parametric tests are not met. There are also relatively few assumptions that underlie permutation tests: The underlying distribution is symmetric and/or the alternative hypothesis states simple shifts in parameter values (Good, 1993). However, the permutation approach's major drawback is that recomputing the statistic for all possible rearrangements can become unwieldy. A Monte Carlo approach, in which possible rearrangements are randomly sampled, has been proposed to overcome the exact permutation tests' computational burden and provide an approximate permutation test (Kaiser, 2007). We will use the approximate permutation test approach and illustrate that it provides accurate sampling distributions, even for small sample sizes and complex 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 1,000 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 Figure 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

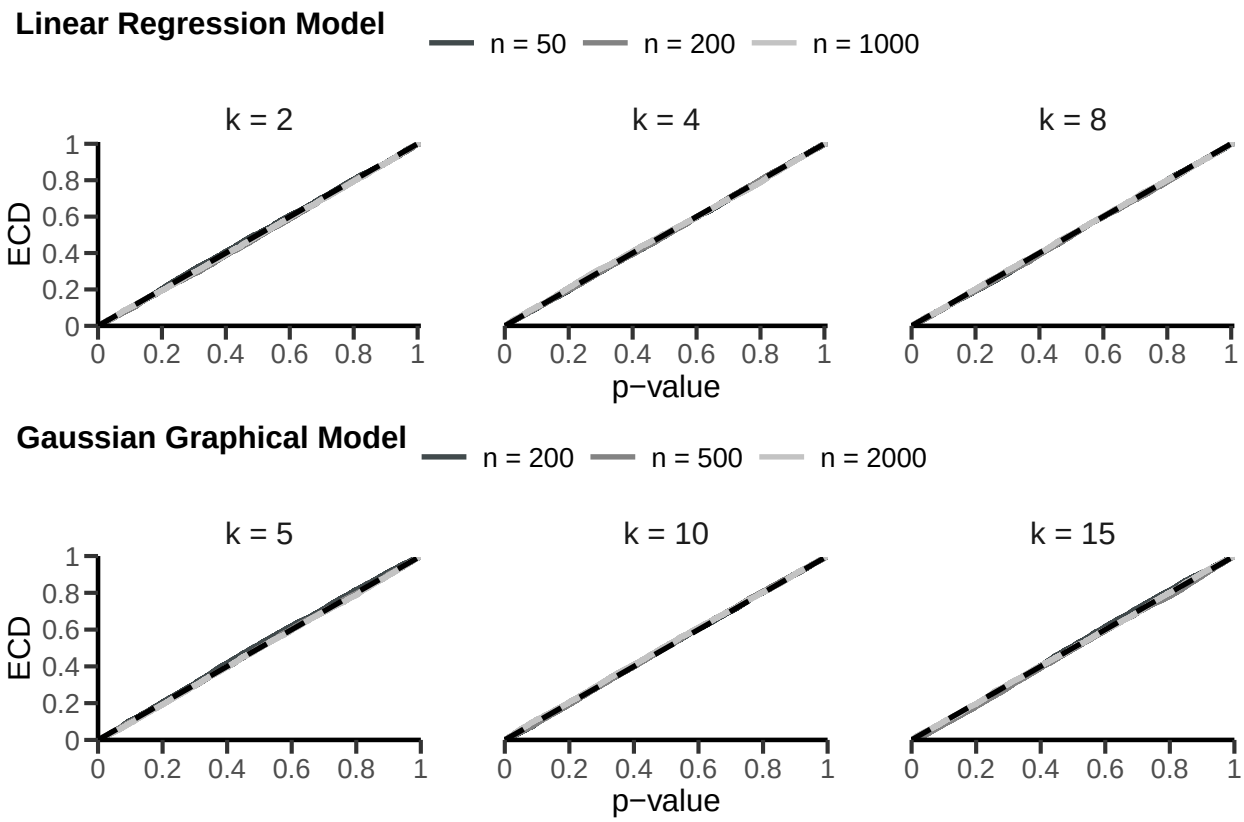

Figure 5

Empirical cumulative distributions 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. In each plot, the black, dashed line represents the expected uniform distribution.

complexity. Therefore, we conclude that the Monte Carlo permutation approach is a valuable method to obtain the SCT's p-value, particularly for complex models and small sample sizes.

Conclusion

This paper has shown that the Structural Change test's (SCT's) small sample behavior can be problematic, but this issue can be resolved using a Monte Carlo Permutation test approach. Specifically, we illustrated that the sampling distribution is misspecified in small samples and for complex models, which leads to an incorrect p-value in practice. The SCT assumes that the accumulation of scores for a parameter across

observations resembles a Brownian bridge, a property that holds for large samples but not for small samples (as shown in the Appendix). As a result, standard methods that rely on this asymptotic property cannot determine the SCT's correct sampling distribution. The permutation test approach is a simple but effective nonparametric method to calculate the sampling distribution when distributional assumptions cannot be met. We have successfully used a Monte Carlo permutation test approach to estimate the correct sampling distribution. As a result, correct p-values can be determined using this approximate permutation version of the SCT, even for small samples and complex models.

References

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

Bechger, T. M. and Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, 80:317–340.

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

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, 44:176–181.

Breslau, J., Javaras, K. N., Blacker, D., Murphy, J. M., and Normand, S.-L. T. (2008). Differential item functioning between ethnic groups in the epidemiological assessment of depression. *The Journal of nervous and mental disease*, 196:297–306.

Brown, R. L., Durbin, J., and 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.

Chang, M.-J. and 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.

Estrella, A. (2003). Critical values and p values of bessel process distributions: Computation and application to structural break tests. *Econometric Theory*, 19.

Fisher, R. A. (1951). *The Design of Experiments*, volume 6th Ed. Hafner, New York, NY.

Good, P. (1993). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.* Springer, New York.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P., and Stahel, W., editors (2005). *Robust Statistics: The Approach Based on Influence Functions.* Wiley Series in Probability and Mathematical Statistics. Wiley, New York, digital print edition.

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

Hjort, N. L. and Koning, A. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, 14:113–132.

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

Hung, H. M. J., O'Neill, R. T., Bauer, P., and Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, 53:11–22.

Jones, P. J., Mair, P., Simon, T., and Zeileis, A. (2019). Network Model Trees. Preprint, Open Science Framework.

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.

Kapur, S., Phillips, A. G., and 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.

Kuan, C.-M. and Hornik, K. (1995). The generalized fluctuation test: A unifying view. *Econometric Reviews*, 14:135–161.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13:127–143.

Merkle, E. C., Fan, J., and Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika*, 79:569–584.

Merkle, E. C. and Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika*, 78:59–82.

Mooney, C. Z. and Duval, R. D. (1993). *Bootstrapping: A Nonparametric Approach to Statistical Inference*, volume 95. Sage Publishing, Newbury Park.

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.

O'Connell, C. S., Ruan, L., and Silver, W. L. (2018). Drought drives rapid shifts in tropical rainforest soil biogeochemistry and greenhouse gas emissions. *Nature Communications*, 9:1–9.

Putnick, D. L. and 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.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Strobl, C., Kopf, J., and Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, 80:289–316.

van de Schoot, R., Schmidt, P., de Beuckelaer, A., Lek, K., and Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology*, 6:1064.

Wang, T., Strobl, C., Zeileis, A., and Merkle, E. C. (2018). Score-Based tests of differential item functioning via pairwise maximum likelihood estimation. *Psychometrika*, 83:132–155.

Zeileis, A. (2006). Implementing a class of structural change tests: An econometric computing approach. *Computational Statistics & Data Analysis*, 50:2987–3008.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, 17:492–514.

Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). Strucchange : An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7:1–38.

Appendix

The Error in the Brownian Bridge Approximation of Empirical Fluctuation Processes

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. This error 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

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

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

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

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. It 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 and discuss both factors composing the error (i.e., $(\hat{\theta} - \theta_0)^2$ and $f^{''}(\theta_s)$). First, we evaluate $(\hat{\theta} - \theta_0)^2$. Note that 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 thus 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 a 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 $1/\sqrt{n}$.

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:

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

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 —$\int_x h(x) \exp(\eta^{\intercal} t(x)) \mathrm{d}x$— 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

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

The third-order derivative consists of two parts $h(x)e^{\eta^\intercal t(x) - A(\eta)}$ 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.

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

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.