Achim Zeileis and Torsten Hothorn
Permutation Tests for Structural Change

Original Citation:

Zeileis, Achim and Hothorn, Torsten
(2006)
Permutation Tests for Structural Change.
This version is available at: https://epub.wu.ac.at/1182/
Available in ePubWU: December 2006
ePubWU, the institutional repository of the WU Vienna University of Economics and Business, is provided by the University Library and the IT-Services. The aim is to enable open access to the scholarly output of the WU.
Permutation Tests for Structural Change

Achim Zeileis
Wirtschaftsuniversität Wien

Torsten Hothorn
Friedrich-Alexander-Universität Erlangen-Nürnberg

Abstract

The supLM test for structural change is embedded into a permutation test framework for a simple location model. The resulting conditional permutation distribution is compared to the usual (unconditional) asymptotic distribution, showing that the power of the test can be clearly improved in small samples. Furthermore, generalizations are discussed for binary and multivariate dependent variables as well as model-based permutation testing for structural change. The procedures suggested are illustrated using both artificial and real-world data (number of youth homicides, employment discrimination data, structural-change publications, and stock returns).

Keywords: conditional inference, asymptotic distribution, exact distribution, maximally-selected statistics.

1. Introduction

Methods for detecting structural changes in series of observations have been receiving increased interest in the theoretical and applied literature, both in econometrics and statistics—see, e.g., Stock and Watson (1996) for a discussion of their relevance to econometric practice. Since the suggestion of the Quandt test (supremum of Chow statistics, see Chow 1960; Quandt 1960), several ideas for capturing structural instabilities in tests statistics have emerged. However, tracking the distribution of such test statistics turned out to be difficult so that the (asymptotical) distribution of the Quandt test remained unknown for a long time. The breakthrough in deriving an asymptotic approximation for structural change test statistics came with the discovery of suitable functional central limit theorems, first for CUSUM statistics (Brown, Durbin, and Evans 1975; Ploberger and Krämer 1992), then for supF statistics (Andrews 1993)—a unifying view on both types of tests is given in Zeileis (2005). Test procedures based on these asymptotic distributions are predominantly used in econometric practice, although some approaches for finite samples also employ other approximations, e.g., based on simulation or bootstrap sampling.

In this paper, we consider a different approach, namely conditional inference methods, also known as permutation tests. This powerful general principle for deriving a suitable reference distribution for a test statistic was described more than 70 years ago by Fisher (1935); its asymptotical properties have been investigated early, e.g., by Pitman (1938). The approach has gained much popularity in the statistics literature in recent years (Ludbrook and Dudley 1998; Strasser and Weber 1999; Pesarin 2001; Ernst 2004). Permutation tests have been found particularly useful because of their flexibility, distribution-free nature and intuitive formulation, which makes it easy to communicate the general principles of such test procedures to practitioners.

In econometrics, permutations tests are still less popular compared to the statistics community—however, several applications exist, see e.g., Kennedy (1995) for an overview of how to employ permutation tests in econometrics. In the following, we discuss how a wide class of permutation tests for structural change can be established, pointing out their strengths and weaknesses. The tests are derived within the framework of Strasser and Weber (1999) as discussed by Hothorn, Hornik, van de Wiel, and Zeileis (2006a), and using ideas of Kennedy (1995). In Section 2, the
permutation distribution of the supLM test of Andrews (1993) is derived for the location-shift model and compared to the established (unconditional) asymptotic distribution, both on artificial and real-world data. In Section 3, a general class of permutation tests for structural change is suggested and specific tests for binary and multivariate observations are derived as well as model-based permutation tests. The procedures are illustrated using real-world data on the number of youth homicides in Boston, a case of employment discrimination, structural-change publications in econometrics and statistics, and Dow Jones industrial average stock returns. Section 4 concludes the paper with a brief discussion.

2. Structural changes in the mean

For comparing unconditional and conditional inference techniques in a structural change context, we focus initially on the simple, yet important, special case of location shifts in a univariate series of observations. First, we establish some general notation as well as the general testing problem which is subsequently specialized to location shifts for which test statistics and sampling distributions are derived. Further important special cases of the general testing problem are considered in Section 3.

2.1. Test problem and statistics

Consider a sequence of \( n \) observations \( Y_i (i = 1, \ldots, n) \), possibly vector-valued, which is ordered with respect to \( t_i \), which usually corresponds to time (but could also be some other ordering variable such as income etc.). In the following, we assume that time \( t \) has been scaled to the unit interval such that it gives the fraction of observations up to the current time (without loss of generality). In the simplest case of \( n \) totally ordered observations along \( i = 1, \ldots, n \), this is simply \( t_i = i/n \)—in the more general case, there could be ties in \( t \) (i.e., several observations were made at the same time).

The structure of the sequence \( Y \) is stable if the distribution of the observations \( Y_i \sim F_t \) does not depend on the time \( t_i \). Thus, structural change tests are concerned with testing the hypothesis

\[
H_0 : \quad F_t = F \quad (t \in [0,1])
\]

against the alternative that the distribution \( F_t \) does depend on \( t \) in some way. As it is not possible (nor desired, typically) to derive tests that have good power properties under arbitrary alternatives (because there are infinitely many different ways how \( F_t \) can depend on \( t \)), specific test statistics are typically derived for certain patterns of deviation from the null hypothesis. The alternative most commonly of interest in this context is the single shift alternative, where the distribution remains constant up to an unknown breakpoint \( t^* \) and shifts to a different distribution afterwards. Test statistics derived for this particular alternative will, of course, also be able to pick up other structural changes albeit with less power. However, the loss in power is usually small if the true alternative can be described sufficiently well by a single shift.

Even a single shift alternative, however, is still too general if it is not specified which aspects of \( F \) are subject to change at \( t^* \). To illustrate the basic approach and focus on the derivation of the conditional distribution of the test statistic, we consider in the remainder of this section the simplest case: only the first moment of \( F \) changes at \( t^* \)—more general types of changes in \( F \) are discussed in Section 3. In the case of location shifts, the model can be formulated more conveniently as

\[
Y_i = \mu_{t_i} + \varepsilon_i, \tag{2}
\]

where \( \varepsilon_i \) is a zero mean disturbance term. The null hypothesis and alternative can then be written as:

\[
H_0 : \quad \mu_t = \mu_0 \quad (t \in [0,1])
\]

\[
H_A : \quad \mu_t = \mu_0 \quad (t \leq t^*)
\]

\[
\mu_t = \mu_0 + \delta \quad (t > t^*) \tag{3}
\]
To test for single shift alternatives, the supF tests of Andrews (1993) are probably the tests employed most often in practice. In a mean shift model and using Lagrange multiplier (LM) statistics, the supF test is based on the statistics

\[ F_\pi = n \cdot R^2_\pi = n \cdot \left( 1 - \frac{\text{RSS}_\pi}{\text{RSS}_0} \right), \]

where \( \text{RSS}_0 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \) is the residual sum of squares (RSS) under the null hypothesis and \( \text{RSS}_\pi \) is the RSS if the mean of the observations up to \( \pi \) is estimated by \( \bar{Y}_{1,\pi} \) and the mean of the observations afterwards by \( \bar{Y}_{2,\pi} \). A sequence of LM statistics \( F_\pi (\pi \in \Pi) \) is computed for each conceivable breakpoint \( \pi \in \Pi = [\bar{\pi}, \bar{\Pi}] \subset [0, 1] \) and the overall null hypothesis is rejected if their supremum \( \sup_{\pi \in \Pi} F_\pi \) is too large. The interval \( \Pi \) is typically derived using some trimming, e.g., \( \Pi = [0.1, 0.9] \), which we use in our simulations and applications below.

For deriving the asymptotic conditional distribution in the following sub-section, it will be useful to transform the LM statistics from their usual “F type” to a “t type”, essentially by taking the square root. The statistic \( F_\pi \) can be rewritten as follows

\[ F_\pi = \frac{\text{RSS}_0 - \text{RSS}_\pi}{\text{RSS}_0/n} = \frac{n_{1,\pi}n_{2,\pi} (\bar{Y}_{1,\pi} - \bar{Y}_{2,\pi})^2}{\text{RSS}_0/n}, \]

where \( n_{1,\pi} \) and \( n_{2,\pi} \) are the observations up to \( \pi \) and after \( \pi \), respectively. Therefore, the supLM test can also be carried out by rejecting the null hypothesis if \( \sup_{\pi \in \Pi} \left| Z_\pi \right| \) is too large, where

\[ Z_\pi = \sqrt{\frac{n_{1,\pi}n_{2,\pi}}{n}} \frac{\bar{Y}_{1,\pi} - \bar{Y}_{2,\pi}}{\sqrt{\text{RSS}_0/(n - 1)}}. \]

We have slightly rescaled \( Z_\pi \) by using \( n - 1 \) instead of \( n \) for standardization. The reason for this is the derivation of the asymptotic conditional distribution and will be explained in more detail below. For notational convenience, we will sometimes replace the supremum by the maximum in the formulation of the test statistic \( \max_{\pi \in \Pi} \left| Z_\pi \right| \)—in these cases, \( \Pi \) is taken to be the elements of \( [\bar{\pi}, \bar{\Pi}] \) observed in the sample, i.e., \( \Pi = \{t_i \mid \bar{\pi} \leq t_i \leq \bar{\Pi}\} = \{\pi_1, \ldots, \pi_m\} \).

### 2.2. Distribution of the test statistic

In the previous sub-section, two equivalent formulations of the supLM test have been established: reject the null hypothesis if \( \max_{\pi \in \Pi} F_\pi \) or \( \max_{\pi \in \Pi} \left| Z_\pi \right| \) becomes “too large”. To render this test useful, the distribution \( D \) (or at least an approximation thereof) of the test statistic under the null hypothesis is required to compute critical values or equivalently \( p \) values. In general, unfortunately, the distribution \( D = D_F \) depends on the unknown distribution \( F \) and is therefore unknown as well. However, there are several strategies to dispose of this dependency by using a suitable approximation of \( D \). The most popular strategy in classical statistics and econometrics is to use the (unconditional) asymptotical distribution \( D_\infty \), i.e., to derive the limit of \( D_F \) for \( n \to \infty \) analytically under some (typically mild) regularity conditions.

In the case of the supLM test, this problem was solved in the seminal paper of Andrews (1993) who showed that a functional central limit theorem holds for the sequence of LM statistics \( F_\pi \) which converge to a squared standardized tied-down Bessel process under fairly general regularity conditions. Thus, the unconditional limiting distribution \( D_\infty \) is given by \( \sup_{\pi \in \Pi} (\pi(1 - \pi))^{-1/2} B^2(\pi) \), where \( B(t) \) (\( t \in [0, 1] \)) is a standard Brownian bridge. This distribution is non-standard but efficient numerical algorithms for computing approximate \( p \) values from this distribution have been derived by Hansen (1997).

A fundamentally different strategy is to replace the unknown null distribution by the conditional null distribution, i.e., the distribution of the test statistic given the observed data. This approach
leads to permutation tests which—albeit already having been established by Fisher (1935)—received increasing interest in the statistics literature only recently (see Ludbrook and Dudley 1998; Strasser and Weber 1999; Pesarin 2001; Ernst 2004, among others) because modern computers render this approach computationally feasible. In econometrics, the interest in permutation or randomization tests also increased but less compared to the statistics community. An introduction to permutation tests in econometrics—highlighting both advantages and problems—is given by Kennedy (1995). Conceptually, carrying out a permutation test for structural change is extremely easy: If the distribution of the $Y_t$ does not depend on the time $t$, the $Y_t$ can be permuted on the $t_i$, breaking up the original ordering. The exact conditional distribution $D_{\pi|Y}$ of the test statistic can then be derived by computing the test statistic for each permutation $\pi \in S$ of the observations $Y_t$. As the size of $S$ is $n!$, it is only feasible for very small $n$ to actually compute all permutations. Otherwise, either specialized algorithms are required for computing the exact distribution (which are only available in certain special cases) or it can be always approximated arbitrarily precisely by drawing a sufficiently large number of permutations $P$ from $S$. In the following, we always draw $P = 10,000$ permutations to approximate the exact conditional distribution $D_{\pi|Y}$ (except in one application where $n = 7$ and the computation of the exact distribution is feasible). See the appendix in Kennedy (1995) for a discussion of some practical considerations concerning the number of permutations.

Instead of drawing a large number of permutations $P$, there also exists another approximation to the conditional distribution: its limiting counterpart. Thus, we can employ the conditional asymptotical distribution $D_{\infty|Y}$ which is obtained from $D_{\pi|Y}$ for $n \to \infty$. For the supLM test, the joint asymptotical conditional distribution of the vector of standardized statistics $Z = (Z_{\pi_1}, \ldots, Z_{\pi_n})^\top$ is multivariate normal. Therefore, it is relatively easy to compute $D_{\infty|Y}$ because efficient numerical algorithms are available for computing $p$-values for the maximum of a multivariate normal statistic $Z$ (Genz 1992). Thus, it is computationally cheap (for small to moderate $n$) to compute the asymptotical conditional distribution $D_{\infty|Y}$ while the advantage of the somewhat more costly computation of $D_{\pi|Y}$ is that the quality of this approximation can be controlled by choosing a sufficiently large $P$.

The asymptotical normality of $Z$ stated in the previous paragraph still needs to be stated more precisely and, of course, proved. It can be shown that expectation, variance and covariance of $Z$ under $H_0$ and given all permutations $\pi \in S$ is:

$$E_{\pi}[Z_\pi] = 0 \quad (\pi \in \Pi)$$

$$\text{VAR}_{\pi}[Z_\pi] = 1 \quad (\pi \in \Pi)$$

$$\text{COV}_{\pi}[Z_\tau, Z_\pi] = \frac{n_{1,\pi}n_{2,\pi}}{\sqrt{n_{1,\pi}n_{2,\pi}n_{1,\tau}n_{2,\tau}}} \quad (\pi < \tau)$$

Collecting the variances and covariances in the matrix $\Sigma$, the multivariate normality of $Z$ can be compactly stated as $Z \sim \Sigma$. A formal proof is given in the appendix which is obtained by embedding the test statistics $Z_\pi$ and $\max_{\pi \in \Pi} |Z_\pi|$ into the framework of Strasser and Weber (1999) who establish asymptotic normality for a general class of permutation tests. This is also the reason for using $n - 1$ rather than $n$ in the standardization of $Z_\pi$. Here, we follow the formulation of Strasser and Weber (1999) whereas for $F_\pi$ we use the standard $n$ in the LM statistic. Note that this only influences the $p$ values computed from the two asymptotical distributions $D_{\infty}$ and $D_{\infty|Y}$ whereas the $p$ values from $D_{\pi|Y}$ remain unaffected. Furthermore, the difference in standardization only has an influence for small $n$ and will lead to slightly smaller $p$ values for the unconditional asymptotical distribution $D_{\infty}$ (but as we will see below, this does not make any difference in practice).

The assumptions under which the unconditional and conditional distributions are valid reference distributions for the test statistic are as different as the underlying conceptual frameworks. The unconditional asymptotics can be established under different sets of assumptions such as those given in Andrews (1993), typically requiring some weak dependence of the series $Y_t$ ($i = 1, \ldots, n$) and certain regularity assumptions for the estimators employed. The assumptions for the conditional permutation tests, on the other hand, are simpler but in a time-series setup somewhat more
restrictive: they require exchangeability of the observations $Y_i$ (or the errors $\varepsilon_i$ in the model-based view, respectively). See Remark 2.4 of Strasser and Weber (1999) and the discussion in Kennedy (1995).

### 2.3. Finite sample performance

To illustrate the quality of the reference distributions for the test statistic $D$ in scenarios with small sample size $n$, a Monte Carlo study of a local alternative model is conducted:

$$Y_i = 0 + n^{-1/2} \cdot (t^* - t_i) + \varepsilon_i,$$

where $1_I$ is the indicator function for the interval $I$, $\delta$ controls the intensity and $t^*$ the timing of the shift. Thus, the mean of $Y_i$ jumps from 0 to $n^{-1/2}\delta$ after time $t^*$. The standardized time is simply $t_i = i/n$ and the disturbances $\varepsilon_i$ are standard normal.

To study the influence of the various parameters of the model, the number of observations $n$ is set to 10, 20 and 50, respectively, $t^* = 0.2, 0.35, 0.5$ and $\delta = 0.5, 10, 15$. The earliest shift is $t^* = 0.2$ so that for the smallest sample size $n = 10$ there are two observations in the first regime.

For comparing the performance of the distributions $D$, power curves (at significance level 5%) are estimated from 10,000 replications for each parameter combination. The values for the shift intensity $\delta$ also include 0 to analyze size as well as power of the tests. This setup corresponds to power/size “conditional on assignment” (in the terminology of Kennedy 1995) allowing for a fair comparison between the unconditional and conditional version of the supLM test.

The results from the Monte Carlo experiment are summarized both in Table 1 and Figure 1. These clearly indicate that the (approximated) exact conditional distribution $D_{\gamma|\Pi}$ performs best, both in terms of power and size, independent of the timing of the shift. Among the asymptotic distributions, the conditional asymptotic distribution $D_{\infty|\gamma}$ outperforms the unconditional asymptotic distribution $D_{\infty}$. However, the differences are only large for very small sample sizes $n$ and diminish with increasing $n$: for $n = 50$ the power curves are already almost indistinguishable. This justifies the usage of the unconditional limiting distribution $D_{\infty}$ (typically computed using the algorithm of Hansen 1997) in moderate to large samples. For small samples, however, the conditional inference approach using permutation tests for structural change proves to be a more powerful strategy.

### 2.4. An illustration

To illustrate the different versions of the reference distribution $D$ on a real-world data set, we re-analyze a time series giving the number of youth homicides in Boston, USA. To address the problem of high youth homicide rates in Boston a policy initiative called the “Boston Gun Project” was launched in early 1995, implementing in particular an intervention called “Operation Ceasefire” in the late spring of 1996. As a single shift alternative seems plausible but the precise start of the intervention cannot be determined, Piehl, Cooper, Braga, and Kennedy (2003) chose to model the number of youth homicides in Boston using modifications of the $F$ tests for structural change of Andrews (1993) based on monthly data ($n = 77$ observations) from 1992(1) to 1998(5) (see Figure 2) and assessing the significance via Monte Carlo results instead of the standard reference distribution $D_{\infty}$.

Here, we take a similar approach and test whether the number of homicides ($Y = \log(\text{homicides} + 0.5)$) changes over time using the supLM test of Andrews (1993) and compare the outcome of all three reference distributions $D$: The test statistic is $\max_{c \in \Pi} |Z_c| = 5.374$ (or equivalently $\max_{c \in \Pi} F_c = 29.261$) with the standard asymptotic unconditional distribution $D_{\infty}$ yielding a $p$ value of $2.55 \cdot 10^{-6}$, the asymptotic conditional distribution $D_{\infty|\gamma}$ a $p$ value of $1.01 \cdot 10^{-6}$, and the approximated conditional distribution $D_{\gamma|\Pi}$ a $p$ value of $1 \cdot 10^{-4}$ (i.e., not a single of the 10,000 permutations produced a greater test statistic). Thus, all three $p$ values are very similar and lead to practically equivalent solutions, providing firm evidence for a change in the number of homicides. The maximal LM statistic is assumed in 1996(7) (an estimate for the timing of the shift $t^*$) at about the time the Operation Ceasefire was implemented.
Table 1: Simulated power (in %) of the supLM test.

| n  | $D_{\infty}$ | $D_{\infty}|Y$ | $D_{\sigma}|Y$ |
|----|--------------|----------------|--------------|
| 10 | 0.0          | 0.6            | 4.0          |
|    | 0.2          | 3.8            | 14.3         |
|    | 2.6          | 26.2           | 49.6         |
|    | 15.4         | 70.2           | 85.2         |
|    | 0.0          | 0.5            | 3.6          |
|    | 0.3          | 5.4            | 26.3         |
|    | 4.6          | 40.2           | 79.6         |
|    | 26.6         | 85.3           | 97.6         |
|    | 7.1          | 50.6           | 97.9         |
|    | 0.0          | 3.6            | 3.9          |
|    | 0.4          | 7.6            | 34.8         |
|    | 7.1          | 50.6           | 99.8         |

Note: Table 1 shows simulated power (in %) at the 5% level for different break intensities $\delta$ and sample sizes $n$.

Figure 1: Simulated power of the supLM test using reference distributions: asymptotic unconditional $D_{\infty}$ (solid), asymptotic conditional $D_{\infty}|Y$ (dashed), approximated conditional $D_{\sigma}|Y$ (dotted).
Table 2: Number of youth homicides in Boston: monthly counts and annual averages.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Monthly</td>
<td>2 1 1</td>
<td>5 1 4</td>
<td>2 5 6</td>
<td>3 4 2</td>
<td>4 2 1</td>
<td>1 3 1</td>
<td>1 1 0</td>
</tr>
<tr>
<td></td>
<td>1 3 3</td>
<td>2 7 3</td>
<td>1 1 6</td>
<td>1 2 3</td>
<td>3 4 2</td>
<td>0 2 3</td>
<td>0 2 2</td>
</tr>
<tr>
<td></td>
<td>3 5 4</td>
<td>4 5 7</td>
<td>3 3 7</td>
<td>7 3 10</td>
<td>3 1 2</td>
<td>0 1 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7 2 5</td>
<td>2 4 4</td>
<td>4 3 1</td>
<td>5 3 3</td>
<td>0 1 2</td>
<td>1 0 3</td>
<td></td>
</tr>
<tr>
<td>Annual</td>
<td>3.083</td>
<td>4.000</td>
<td>3.167</td>
<td>3.833</td>
<td>2.083</td>
<td>1.250</td>
<td>0.800</td>
</tr>
</tbody>
</table>

Figure 2: Number of youth homicides in Boston: monthly counts (gray), annual averages (black).

So far, we essentially confirmed the findings of Piehl et al. (2003) and also the impression from our simulation study that already for moderately sized $n$ all three reference distributions $\mathcal{D}$ lead to virtually identical results. However, imagine that we would not have been provided with such detailed monthly observations but with annual averages instead (see Table 2 and Figure 2). Then, with only $n = 7$ observations, would we still be able to show that the policy intervention had an effect on the number of homicides? Using the raw means (instead of logs because we already averaged) leads to a test statistic of $\max_{\pi \in \Pi} |Z_\pi| = 2.246$ (or equivalently $\max_{\pi \in \Pi} F_\pi = 5.885$). This corresponds to a $p$ value of 20.25% computed from the standard asymptotic unconditional distribution $\mathcal{D}_{\infty}$, 10.62% for the asymptotic conditional distribution $\mathcal{D}_{\infty|Y}$, and 5.71% for the exact conditional distribution $\mathcal{D}_{\sigma|Y}$. Thus, we observe a similar phenomenon as in the simulation study: the standard $\mathcal{D}_{\infty}$ lacks power and results in a clearly non-significant $p$ value whereas the conditional $p$ values are considerably smaller. The exact $p$ value is on the verge of being significant (at 5% level) and in fact there was not a single permutation yielding a greater test statistic, all 5.71% permutations are ties with the observed maximal test statistic (which is assumed for the year 1995).

3. Extensions

In the previous section, we discussed how the conditional distributions $\mathcal{D}_{\sigma|Y}$ and $\mathcal{D}_{\infty|Y}$ can be established by embedding the supLM test of Andrews (1993) into the framework of Strasser and Weber (1999) for the location model (2). Here, we discuss how this general framework can be employed more generally in a structural change setup. Strasser and Weber (1999) provide a very general approach for assessing the dependence of a
sequence of possibly multivariate observations $Y_i$ on another variable $t_i$ by employing test statistics of the form

$$T = \text{vec} \left( \sum_{i=1}^{n} g(t_i) h(Y_i, (Y_1, \ldots, Y_n))^\top \right),$$

where $g(\cdot)$ and $h(\cdot)$ are possibly vector-valued transformations of $t_i$ and $Y_i$, respectively. They are also called regression function and influence function, respectively, where the latter may depend on the full sequence of observations $(Y_1, \ldots, Y_n)$, however, only in a permutation-symmetric way. Given exchangeability of the observations, the asymptotic conditional multivariate normality of $T$ under the null hypothesis of independence of $Y_i$ and $t_i$ is derived by Strasser and Weber (1999).

The choice of $g(\cdot)$ and $h(\cdot)$ determines against which types of dependence of $Y_i$ on $t_i$ tests based on $T$ have good power. For a single shift alternative with unknown breakpoint as in (3), it is straightforward to use a multivariate regression function constructed from indicator functions for all potential breakpoints $g(t) = (1_{[0, \pi_1]}(t), \ldots, 1_{[0, \pi_m]}(t))^\top$. While $g(\cdot)$ reflects the type of time dependence, the choice of $h$ determines what types of changes in the distribution $F_t$ can be captured (well): For shifts in location using the identity $h(Y) = Y$ is suitable. To aggregate the multivariate statistic $T$ to a single scalar test statistic, typically the maximum of the standardized $T$ is used

$$\max \left| \frac{T - \mathbf{E}_\sigma[T]}{\sqrt{\text{diag} \mathbf{VAR}_\sigma[T]}} \right|$$

which corresponds to taking the maximum over the components of $g(\cdot)$ (i.e., the various potential breakpoints) and of $h(\cdot)$ (if it is multivariate). For the location model, more details are given in the appendix.

In the following, other choices of $h(\cdot)$ are discussed which are suitable for assessing changes in binary observations $Y_i$, multivariate series, stratified data (including certain type of panel data) and parametric models, respectively. In all illustrations, the exact conditional distribution $\mathcal{D}_{\sigma|\gamma}$ is used for computing $p$ values and approximated by drawing $P = 10,000$ permutations.

**Structural changes in binary variables.** For binary observations $Y_i$, the distribution $F_{t_i}$ is binomial with a certain success probability $\mu_{t_i}$ which could depend on the time $t_i$. The null hypothesis of structural stability can again be written as in (3) corresponding to constancy of the success probability. A test statistic $T$ that compares empirical proportions from two subsamples defined by a set of potential break points $\pi_1, \ldots, \pi_m$ can simply be obtained by using a dummy coding for $Y_i$. This corresponds to using the influence function $h(Y) = 1_{\{\text{success}\}}(Y)$ while the remaining ingredients of the test remain the same and can be applied out of the box. As an illustration we use the data provided in Table 3 from an employment discrimination case described in Freidlin and Gastwirth (2000). The issue in the case was whether the hiring policy was gender neutral and a charge was filed in May 1994. Freidlin and Gastwirth (2000) supported the court’s decision that there was evidence that the employer switched from under-hiring of females (compared to a fraction of 3.43% in the qualified labor force in the labor market) to over-hiring after the charge was filed. Employers can use such strategies to obscure discriminations in data aggregated over time (using both pre- and post-charge periods). Here, we re-analyze the data set in a simple structural change setup, i.e., without employing the additional knowledge of the fraction of females in the qualified labor force. Using the test procedure described above, we show that the fraction of hired females changed significantly over the years. Although there $n = 988$ observations,

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Female hires</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Male hires</td>
<td>427</td>
<td>86</td>
<td>104</td>
<td>180</td>
<td>111</td>
<td>59</td>
</tr>
</tbody>
</table>

Table 3: Annual hiring data of employment discrimination case.
there are only \( m = 5 \) potential breakpoints (in 1991, \ldots, 1995) as the data is reported annually. Thus, the regression function is \( g(t) = (1_{[1991,1991]}(t), \ldots, 1_{[1991,1995]}(t))^\top \) and \( h(Y) = 1_{\text{female}}(Y) \) yielding a 5-dimensional statistic \( T \). The maximum of the standardized statistics is 10.49 assumed in 1995, corresponding to a \( p \) value of \( 10^{-4} \) (i.e., not a single permutation yielded a greater test statistic) conforming with the findings of Freidlin and Gastwirth (2000).

Structural changes in multivariate series. If the observations \( Y_i \) are vector-valued, several scenarios are conceivable: all components correspond to dependent variables, or some might also correspond to dependent variables, or there might be one stratifying variable. The latter two scenarios are dealt with in the next paragraphs, here we focus on the case of a multivariate dependent series of observations \( Y_i \). Typically, a multivariate influence function \( h(\cdot) \) is used which is obtained by applying a suitable univariate influence function to each component of \( Y_i \). Thus, a sequence of standardized statistics (over potential breakpoints) is computed for each component and the test rejects the null hypothesis of stability if there is evidence for structural change in any of the components. By using the joint distribution of all standardized statistics, this procedure corrects appropriately for multiple testing via incorporation of the full correlation structure over time and components. This allows not only for identification of the timing of the shift (as in the previous illustrations) but also of the component of \( Y_i \) affected by it. For illustration purposes, we investigate the bivariate series of the number of structural-change-related publications per year between 1986 and 2005 published in seven econometrics and eight statistics journals, respectively. The data have been obtained from the IS\textit{i} Web of Science (The Thomson Corporation 2006) and are given in Table 4, more details are provided in the appendix. Here, we aim to find out whether the number of structural-change publications has changed over the last 20 years in either the econometrics or statistics communities (or both). As the observations \( Y_i \) are counts, we employ a log transformation as the influence function \( h(Y) = \log(Y) \) (which is consequently bivariate here) and the regression function \( g(t) \) corresponds to \( m = 17 \) potential breakpoints (1987, \ldots, 2003, i.e., a trimming of 10%) yielding a statistic \( T \) of dimension 17 \( \times \) 2. The maximum of the standardized statistics is 3.83 corresponding to a \( p \) value of \( 2 \cdot 10^{-4} \). This maximal statistic is assumed in 1989 for the statistics series, but the maximal statistic for the econometrics series is almost as large with 3.76 in 1991. Both statistics clearly exceed the 95% critical value of 2.93, thus signalling a significant change in both communities (occurring somewhat earlier for the statistics series). An intuitive interpretation for this increase in the number of publications is that after the theoretical foundations of many structural change procedures had been established in the late 1980s and early 1990s, many more publications emerged that extended the results and applied them to various types of models. Also note that the three peaks in the econometrics series in 1992, 1996 and 2005 are associated with special issues of the Journal of Business \& Economic Statistics (“Breakpoints and Unit Roots” in 1992) and the Journal of Econometrics (“Recent Developments in the Econometrics of Structural Change” in 1996 and “Modelling Structural Breaks” in 2005), which also reflect the increased importance that structural changes gained during the previous decade.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Econometrics</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>19</td>
<td>8</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>Statistics</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>14</td>
<td>15</td>
<td>15</td>
<td>21</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Econometrics</td>
<td>28</td>
<td>16</td>
<td>22</td>
<td>16</td>
<td>18</td>
<td>21</td>
<td>22</td>
<td>27</td>
<td>16</td>
<td>31</td>
</tr>
<tr>
<td>Statistics</td>
<td>25</td>
<td>28</td>
<td>20</td>
<td>13</td>
<td>21</td>
<td>15</td>
<td>18</td>
<td>11</td>
<td>30</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 4: Annual number of structural-change publications in econometrics and statistics journals.

Achim Zeileis, Torsten Hothorn
Figure 3: Annual number of structural-change publications in econometrics and statistics journals (in logs).

Structural changes in stratified observations. If one of the components of a multivariate sequence $Y_i$ stratifies the observations into independent blocks, the following simple strategy can be used: Compute the statistic $T$ and its association expectation and covariance (given $\sigma \in S$) for each block and aggregate the blockwise statistics by taking their sum. Due to independence of the blocks the expectation and covariance of the aggregated statistic is also obtained by taking sums. In addition to independence the (hypothesized) block-wise breakpoints should be identical (or at least similar) for the test to have good power. In econometric applications, such situations occur less often compared to planned experiments in statistical applications—however, some situations are conceivable (e.g., panel data from independent companies). Another application would be to use a multivariate influence function $h(\cdot)$ and treat its components as blocks. This is useful for model-based tests (see below) when decorrelated score functions for different parameters are used for $h(\cdot)$.

Structural changes in parametric models. To assess changes in certain aspects of the distribution $F_t$ a parametric model could be useful, in particular if the observations can be split up into dependent and explanatory variables $Y_i = (y_i, x_i)^T$. As the influence function may depend on the full set of observations (in a permutation symmetric way), the model and its corresponding parameter estimate can be easily incorporated into $h(\cdot)$. As the most important special case, we first consider some options for the linear regression model

$$y_i = x_i^T \theta + \epsilon_i,$$

where the assumption of exchangeability now has to be fulfilled for the disturbances $\epsilon_i$ (under the null hypothesis) to render the permutation approach valid. Then, structural changes in the conditional mean of the $y_i$ can be easily assessed by using the usual ordinary least squares (OLS) residuals in the influence function $h(Y_i) = \hat{\epsilon}_i = y_i - \hat{y}_i$. For univariate observations $Y_i = y_i$, this is equivalent to the supLM test described in Section 2. Analogously, changes in the variance of the disturbances can be captured by basing the test on the squared residuals $h(Y_i) = \hat{\epsilon}_i^2$. Moreover, changes in any component of the vector of regression coefficients $\theta$ can be tested by using the full
OLS model scores $h(Y_i) = \hat{\varepsilon}_i x_i$. Similar to the vector-valued influence function for multivariate series, this leads again to a statistic for each combination of potential breakpoint and parameter component. The same ideas apply not only to linear regression models, but to more general parametric models as long as the exchangeability assumption can be assured. More precisely, we could consider a model for univariate or multivariate observations $Y \sim \mathcal{F}$ (under the null hypothesis) that are modelled by a parametric distribution $\mathcal{G}(\theta)$ (which might or might not include the true distribution $\mathcal{F}$). Then, some model scores or moment conditions derived from the model $\mathcal{G}(\theta)\psi_\theta(Y) = \text{const}$ say, could be used for testing the model stability via the influence function $h(Y) = \psi_\theta(Y)$. This uses the same ideas as in Nyblom-Hansen test (Nyblom 1989; Hansen 1992), see also Zeileis (2005) for a unified approach discussing different score functions $\psi_\theta(Y)$. Here, we use these ideas to investigate the stability of the distribution of Dow Jones industrial average stock returns based on weekly closing prices from 1971-07-02 to 1974-08-02. The series of prices is provided in Hsu (1979) and the corresponding log-difference returns ($\times 100$) are depicted in Figure 4. Following Hsu (1979), we model the returns $Y_i$ as approximately normally distributed with mean and variance $\theta = (\mu, \sigma^2)^\top$ leading to the maximum likelihood scores $\psi_\theta(Y) = (Y - \mu, (Y - \mu)^2 - \sigma^2)^\top$ corresponding to the usual moment conditions. For assessing the stability of both mean and variance of the returns, we employ the empirical model scores as the influence function $h(Y) = \psi_{\hat{\theta}}(Y)$ (note that using the simpler $h(Y) = (Y, (Y - \hat{\mu})^2)^\top$ would lead to identical results) and again a 10% trimming for deriving $g(t)$. This yields a maximal standardized statistic of 4.96 assumed on 1973-03-16 for the variance, corresponding to a $p$ value of $2 \cdot 10^{-4}$—the maximal statistic for the mean, on the other hand, is considerably smaller with 1.89, remaining clearly below the 95% critical value of 3.19. Thus, there is evidence for a clear shift in the variances in mid-March 1973 (matching the break found by Hsu 1979) while the mean remains constant throughout the sample period.

4. Conclusions
The supLM test for structural change of Andrews (1993) is embedded into a permutation test framework for the location-shift model. This yields the conditional permutation distribution of the test statistic (and its asymptotic counterpart) which can be used for inference instead of the usual unconditional asymptotic distribution. Comparing the size and power of the test procedures based on different versions of the reference distribution shows that (unconditional) asymptotics work well already for moderately large samples. In small samples, however, performance can be improved
significantly by employing the conditional approach, in particular by computing/approximating the exact conditional distribution. Permutation tests for structural change from the framework of Strasser and Weber (1999) can, in fact, not only be derived for the simple location model: The flexible class of tests considered includes both non-parametric and parametric (model-based) permutation tests. However, the results have to be taken with a grain of salt: Exchangeability of the errors might be a too strong assumption in time-series applications where the dependence structure of the observations cannot be fully captured within the model. Although there are time-series applications where the errors are not correlated (and exchangeability is fulfilled as in the illustrations presented above), this assumption impedes the application of permutation methods to many other models of interest.

Computational details

The results in this paper were obtained with R 2.3.1 (R Development Core Team 2006), in particular using the packages coin 0.5-2 (Hothorn, Hornik, van de Wiel, and Zeileis 2006b) and strucchange 1.3-1 (Zeileis, Leisch, Hornik, and Kleiber 2002). Both, R itself and the packages, are freely available at no cost under the terms of the GNU General Public Licence (GPL) from the Comprehensive R Archive Network at http://CRAN.R-project.org/.

Acknowledgements

We are indebted to James Bullard (University of California, Berkeley) for his assistance with searching ISI Web of Science. The work of T. Hothorn was supported by Deutsche Forschungsgemeinschaft (DFG) under grant HO 3242/1-3.

References


A. Proofs

The (asymptotic) distribution of the multivariate statistic \( Z = (Z_{\pi_1}, \ldots, Z_{\pi_m})^T \) is derived by embedding the statistic into the framework of Strasser and Weber (1999), as discussed in Hothorn et al. (2006a). More precisely, the test statistic \( \max_{\pi \in \Pi} Z_{\pi} \) considered above is the maximum-type statistic \( c_{\text{max}} \) of Hothorn et al. (2006a) if the influence function \( h(Y) = Y \) is used for the observations \( Y_i \) and the transformation \( g(t) = (1_{[0, \pi_i]}(t), \ldots, 1_{[0, \pi_m]}(t))^T \) is used for the associated timings \( t_i \). The transformation \( g \) used the indicator function \( 1_I \) of the interval \( I \) and thus corresponds to a vector of indicators for the time up to the timings \( \pi_j \) (\( j = 1, \ldots, k \)).

Using these transformations \( h(\cdot) \) and \( g(\cdot) \), the unstandardized test statistic \( T \) is in the notation of Hothorn et al. (2006a)

\[
T = \operatorname{vec}
\left( \sum_{i=1}^{n} g(t_i) h(Y_i)^T \right) = (n_{1, \pi_1} \bar{Y}_{1, \pi_1}, \ldots, n_{1, \pi_m} \bar{Y}_{1, \pi_m})^T. \tag{9}
\]

Under \( H_0 \), given all permutations \( \sigma \in S \) of the observations \( Y_1, \ldots, Y_n \), the unstandardized statistic has expectation

\[
E_{\sigma}[T] = \operatorname{vec}
\left( \sum_{i=1}^{n} g(t_i) n^{-1} \sum_{i=1}^{n} h(Y_i)^T \right) = (n_{1, \pi_1}, \ldots, n_{1, \pi_m})^T \bar{Y}. \tag{10}
\]

and each unstandardized statistic has variance

\[
\operatorname{VAR}_{\sigma}[T_{\pi}] = \left( n_{1, \pi} - \frac{n^2}{n} \right) \frac{\operatorname{RSS}_0}{n-1} = \frac{n_{1, \pi} n_{2, \pi}}{n} \frac{\operatorname{RSS}_0}{n-1}, \tag{11}
\]

where the residual sum of squares is \( \operatorname{RSS}_0 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \).

Standardizing the vector of raw statistics \( T = (T_{\pi_1}, \ldots, T_{\pi_m})^T \) by their respective mean and standard deviation yields the vector of statistics \( Z = (Z_{\pi_1}, \ldots, Z_{\pi_m})^T \):

\[
Z_{\pi} = \frac{T_{\pi} - E_{\sigma}[T_{\pi}]}{\sqrt{\operatorname{VAR}_{\sigma}[T_{\pi}]}} = \frac{n_{1, \pi} \bar{Y}_{1, \pi} - n_{1, \pi} \bar{Y}}{\sqrt{\frac{n_{1, \pi} n_{2, \pi}}{n} \frac{\operatorname{RSS}_0}{n-1}}} = \frac{n_{1, \pi} n_{2, \pi}}{n} \frac{\bar{Y}_{1, \pi} - \bar{Y}_{2, \pi}}{\sqrt{\operatorname{RSS}_0/(n-1)}},
\]

because of the following simple relationship between \( \bar{Y}_{1, \pi}, \bar{Y}_{2, \pi} \) and \( \bar{Y} \):

\[
\bar{Y} = \frac{n_{1, \pi} \bar{Y}_{1, \pi} + n_{2, \pi} \bar{Y}_{2, \pi}}{n}.
\]

Consequently, \( Z \) has zero mean and unit variance given all permutations \( \sigma \in S \). Similarly, the covariance between two elements of \( Z, Z_{\pi} \) and \( Z_{\tau} \), say, is

\[
\frac{\operatorname{RSS}_0}{n-1} \left( \sum_{i=1}^{n} 1_{[0, \pi]}(t_i) 1_{[0, \tau]}(t_i) \right) - \frac{1}{n} \frac{\operatorname{RSS}_0}{n-1} \left( \sum_{i=1}^{n} 1_{[0, \pi]}(t_i) \right) \left( \sum_{i=1}^{n} 1_{[0, \tau]}(t_i) \right) = \frac{n \min(n_{1, \pi}, n_{1, \tau}) - n_{1, \pi} n_{1, \tau}}{n} \frac{\operatorname{RSS}_0}{n-1}.
\]

Assuming that \( \pi < \tau \) and using the variance computed above, the correlation is thus

\[
\frac{n_{1, \pi} n_{2, \tau}}{\sqrt{n_{1, \pi} n_{2, \pi} n_{1, \tau} n_{2, \tau}}}
\]
Given that we derived the first two moments of $Z$ by embedding the statistic into the framework of Strasser and Weber (1999), the asymptotic normality of $Z$ follows by application of their Theorem 2.3.

**B. Publications data**

The bibliographic information about publications related to structural change was obtained from *ISI Web of Science* on 2006-12-12 (*The Thomson Corporation* 2006). The query was based on the Science Citation Index Expanded (SCI-Expanded) and Social Sciences Citation Index (SSCI) for the years 1986–2005. Seven econometrics journals (Econometric Theory, Econometrica, Economics Letters, Journal of Applied Econometrics, Journal of Business & Economic Statistics, Journal of Econometrics, Review of Economics and Statistics) and eight statistics journals (Biometrics, Biometrika, Communications in Statistics: Simulation and Computation, Communications in Statistics: Theory and Methods, Journal of Multivariate Analysis, Journal of Statistical Planning and Inference, Statistics & Probability Letters, The Annals of Statistics) were used. Title, abstract or keyword had to match one of the following topics: structural change, structural break, structural stability, structural instability, parameter instability, parameter stability, parameter constancy, change point, changepoint, change-point, breakpoint, break-point, break point, CUSUM, MOSUM.