Generalized M-Fluctuation Tests for Parameter Instability

Achim Zeileis **Kurt Hornik**
Department of Statistics and Mathematics, Wirtschaftsuniversität Wien

Abstract

A general class of fluctuation tests for parameter instability in an M-estimation framework is suggested. Tests from this framework can be constructed by first choosing an appropriate estimation technique, deriving a partial sum process of the estimation scores which captures instabilities over time, and aggregating this process to a test statistic by using a suitable scalar functional. Inference for these tests is based on functional central limit theorems which are derived under the null hypothesis of parameter stability and local alternatives. For (generalized) linear regression models, concrete tests are derived which cover several known tests for (approximately) normal data but also allow for testing for parameter instability in regressions with binary or count data. The usefulness of the test procedures—complemented by powerful visualizations derived from these—is illustrated using Dow Jones industrial average stock returns, youth homicides in Boston, USA, and illegitimate births in Großarl, Austria.

Keywords: fluctuation test, functional central limit theorem, generalized linear model, M-estimation, parameter instability, structural change.

1. Introduction

Structural change is of central interest in many fields of research and data analysis: to learn if, when and how the structure of the data generating mechanism underlying a set of observations changes. In parametric models, structural change is typically described by parameter instability. If this instability is ignored, parameter estimates are generally not meaningful, inference is severely biased and forecasts lose accuracy. Therefore, structural changes and parameter instability have been receiving attention in many disciplines in particular in the econometrics and the statistics communities with varying emphasis on different aspects like diagnostic checking, data exploration or breakpoint estimation. Fields of application are manifold and include, for example, finance (see e.g., Andreou and Ghysels 2002), economics (see e.g., Stock and Watson 1996), political and social sciences (see e.g., Piehl, Cooper, Braga, and Kennedy 2003), or biostatistics and medical statistics (see e.g., Muggeo 2003).

Usually, in structural change problems it is known with respect to which quantity the instability occurs: e.g., in time-series regression it is natural to ask whether the relationship between dependent and explanatory variables changes over time. In other applications often problems arise where a regression relationship changes with respect to some other variable, e.g., firm size in an economic model, or some prognostic factor in biometric studies. The situations described have in common that the observations have some unique ordering with respect to which a structural change occurs but that the (potential) breakpoint is unknown. Starting from the Recursive CUSUM test of Brown, Durbin, and Evans (1975) a large variety of tests has been suggested in both the econometrics and the statistics literature, in particular for (approximately) normal data and the linear regression model. For the linear regression model, Kuan and Hornik (1995) introduce a generalized and unifying framework of fluctuation tests for structural changes without assuming a specific alternative. For models estimated by maximum likelihood, Hjort and Koning (2002) use a similar

This is a preprint of an article published in *Statistica Neerlandica*, **61**(4), 488–508.
 doi:10.1111/j.1467-9574.2007.00371.x

approach and suggest a general class of fluctuation tests based on maximum likelihood scores. In this paper, we substantially generalize these frameworks to fluctuation tests for general parametric models and a wide range of M-type estimation techniques. From the resulting class of generalized M-fluctuation tests a broad spectrum of tools from exploratory to significance procedures can be derived in the spirit of Brown *et al.* (1975) who point out that this framework "... includes formal significance tests but its philosophy is basically that of data analysis as expounded by Tukey (1962). Essentially, the techniques are designed to *bring out departures from constancy in a graphic way* instead of parametrizing particular types of departure in advance and then developing formal significance tests intended to have high power against these particular alternatives." (Brown *et al.* 1975, pp. 149–150)

The generalized M-fluctuation test framework can be thought of as a toolbox for the construction of parameter instability tests in various situations. A particular test can be constructed in the following steps: First, an appropriate estimation technique (or equivalently its score function) has to be chosen. Second, an empirical partial sum process that is governed by a functional central limit theorem is used to capture the instabilities in the estimated model. Third, the fluctuation within this process needs to be measured by a scalar functional yielding a test statistic that can be compared with the distribution of the functional as applied to the corresponding limiting process. Furthermore, strategies are outlined for enhancing traditional significance testing with visualization methods for detecting the timing of a potential shift and which parameter is affected by it. Based on this framework, we construct a class of parameter instability tests for generalized linear regression models, in particular for normal, poisson and binomial data. For the normal model the resulting tests include many well-established tests from the literature as special cases (including tests based on ML scores, F statistics and OLS residuals, see Zeileis 2005) but yield new tests for poisson and binomial models which are often much more suitable for modeling data in biostatistics, economics or the social siences.

The remainder of this paper is organized as follows: In Section 2 we first introduce a general parametric model and formulate the null hypothesis before we derive functional central limit theorems for partial sum processes of M-scores under the hypothesis of parameter stability and under local alternatives. These form the basic probabilistic ingredient for our parameter instability test toolbox. The remaining components are discussed in Section 3: choice of an estimation technique (or score functions, respectively), functionals for measuring fluctuation in empirical fluctuation processes and corresponding means of visualization. Section 4 applies the previously introduced general framework to generalized linear models and derives concrete tests based on ML scores. The usefulness of the proposed tests is illustrated in Section 5 based on time-series data for Dow Jones industrial average stock prices, the number of youth homicides in Boston, USA, and illegitimate births in Großarl, Austria, in a policy intervention framework. Section 6 offers some concluding remarks.

2. Generalized M-fluctuation processes

In this section, we suggest a general class of fluctuation processes that can capture instabilities in parametric models. After introducing a general parametric model in Section 2.1, the limiting behaviour of the associated fluctuation processes is derived in Section 2.2, both for the case where θ_0 is known and where it has to be estimated. In Section 2.3, the results are generalized to local alternatives.

2.1. Model frame

We assume n possibly vector-valued independent observations

$$Y_i \quad \sim \quad F(\theta_i) \qquad (i = 1, \dots, n). \tag{1}$$

distributed according to some distribution F with k-dimensional parameter θ_i. We also assume that the index $i = 1, \dots, n$ stems from the ordering with respect to some external variable, usually time. These assumptions are much more restrictive than required for establishing the results given in the following. They are employed to present the properties of the generalized M-fluctuation test without too much technical overhead and unnecessarily complicated notation focusing on the conceptual design of the framework. Weaker assumptions for regression models and dependent data can be used as in Krämer, Ploberger, and Alt (1988), Bai (1997) or Andrews (1993) and will be discussed in Sections 2.4 (dependent data) and 4 (regression models).

We are interested in testing the hypothesis

$$H_0: \theta_i \quad = \quad \theta_0 \qquad (i = 1, \dots, n) \tag{2}$$

against the alternative that (at least one component of) θ_i varies over "time". For this alternative to be sensible the ordering assumption is necessary: If a parameter instability with a single breakpoint, say, occurs with respect to a certain ordering of the variables this single breakpoint interpretation would be lost by re-ordering in such a way that observations from the two regimes are mixed.

2.2. Theoretical and empirical fluctuation processes

Consider some suitably smooth k-dimensional score function $\psi(\cdot)$, independent of n and i (see also Stefanski and Boos 2002, for extensions to nonsmooth score functions), with

$$\mathsf{E}[\psi(Y_i, \theta_i)] \quad = \quad 0 \tag{3}$$

and define the following matrices

$$\begin{aligned} A(\theta) &= \mathsf{E}[-\psi'(Y, \theta)], & (4)\\ B(\theta) &= \mathsf{VAR}[\psi(Y, \theta)], & (5)\\ C(\theta) &= \mathsf{E}[\psi(Y, \theta) u(Y, \theta)^\top] & (6) \end{aligned}$$

where $Y \sim F(\theta_0)$, $\psi'(\cdot)$ is the partial derivative of $\psi(\cdot)$ with respect to θ, and $u(\cdot, \theta)$ is

$$u(y, \theta) \quad = \quad \frac{\partial \log f(y, \theta)}{\partial \theta}, \tag{7}$$

and $f(\cdot, \theta)$ is the probability density function corresponding to F. Hence, u is the partial derivative of the log likelihood with respect to θ, also called Maximum Likelihood (ML) score. The first two matrices $A(\theta)$ and $B(\theta)$ are standard in M-estimation, $C(\theta)$ is only needed in Section 2.3. Note that given ψ the matrices $A(\theta)$ and $B(\theta)$ but not $C(\theta)$ can be estimated without further knowledge of F or f respectively.

Under mild regularity conditions, e.g., as given in White (1994, Theorem 6.10, p. 104), the following theorems can be established.

Theorem 1 *For the cumulative score process given by*

$$W_n(t, \theta) \quad = \quad n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \psi(Y_i, \theta) \tag{8}$$

and under the assumptions stated above and under H_0 the following functional central limit theorem (FCLT) holds:

$$W_n(\cdot, \theta_0) \quad \xrightarrow{\text{d}} \quad Z(\cdot),$$

where $Z(\cdot)$ is a Gaussian process with continuous paths, mean function $\mathsf{E}[Z(t)] = 0$ and covariance function $\mathsf{COV}[Z(t), Z(s)] = \min(t, s) \cdot B(\theta_0)$.

Proof: The proof follows by direct application of Donsker's theorem (Billingsley 1999).

Corollary 1 *If $B(\theta_0)$ is non-singular, the following FCLT holds for the decorrelated fluctuation process*

$$B(\theta_0)^{-1/2} W_n(\cdot, \theta_0) \quad \xrightarrow{\mathrm{d}} \quad W(\cdot),$$

where $W(\cdot)$ is a k-dimensional Wiener process or standard Brownian motion.

Usually, in applications the parameter θ_0 is not known but has to be estimated. A suitable estimator can be based on the function $\psi(\cdot)$: the full sample M-estimator $\hat{\theta}_n$ is defined by the equation

$$\sum_{i=1}^{n} \psi(Y_i, \hat{\theta}_n) \quad = \quad 0. \tag{9}$$

Theorem 2 *Under H_0 the following FCLT holds for the empirical cumulative score process with M-estimated parameters*

$$W_n(\cdot, \hat{\theta}_n) \quad \xrightarrow{\mathrm{d}} \quad Z^0(\cdot),$$

where $Z^0(t) = Z(t) - tZ(1)$.

Proof: As the asymptotic properties of M-estimators are very similar to those of ML estimators (see e.g., White 1994; Stefanski and Boos 2002), the arguments of (Hjort and Koning 2002, Section 2.2) for their ML-based cumulative score processes can be repeated yielding the proof for the empirical cumulative M-score processes above.

Corollary 2 *If $B(\theta_0)$ is non-singular, the following FCLT holds for the decorrelated empirical fluctuation process with M-estimated parameters*

$$\hat{B}_n^{-1/2} W_n(\cdot, \hat{\theta}_n) \quad \xrightarrow{\mathrm{d}} \quad W^0(\cdot),$$

where $W^0(t)$ is a standard k-dimensional Brownian bridge with $W^0(t) = W(t) - tW(1)$ and $\hat{B}_n$ is some consistent and non-singular covariance matrix estimate, e.g., $\hat{B}_n = n^{-1} \sum_{i=1}^{n} \psi(Y_i, \hat{\theta}_n) \psi(Y_i, \hat{\theta}_n)^\top$.

In the following, the empirical fluctuation process is also denoted

$$\mathit{efp}(t) \quad = \quad \hat{B}_n^{-1/2} W_n(t, \hat{\theta}_n).$$

2.3. Local alternatives

In parameter instability problems or structural change situations an alternative of interest is the local alternative

$$H_A: \ \theta_i \quad = \quad \theta_0 + n^{-1/2}\, g\left(\frac{i}{n}\right), \tag{10}$$

where $g(\cdot)$ is a function of bounded variation on $[0, 1]$ which describes the pattern of departure from stability of the parameter θ_0 (Kuan and Hornik 1995; Hjort and Koning 2002) including various types of abrupt and gradual changes (see e.g., Hušková 1999) or partial changes (Andrews 1993).

Then Y_i has the probability density function

$$f(y, \theta_i) \doteq f(y, \theta_0) \left\{1 + u(y, \theta_0)^\top \, n^{-1/2} \, g\left(\frac{i}{n}\right)\right\},$$

which can be easily derived from first order Taylor expansion of f; $a_n \doteq b_n$ denotes that $a_n - b_n$ tends to zero (in probability if a_n or b_n are stochastic).

Therefore, under a local alternative like (10) the components of the fluctuation process (8) no longer have zero mean in general but

$$\mathsf{E}[\psi(Y_i,\theta_0)] \doteq \int \psi(y,\theta_0) f(y,\theta_0)\ \mathrm{d}y + \int \psi(y,\theta_0) u(\theta_0)^\top f(y,\theta_0)\ n^{-1/2}\ g\left(\frac{i}{n}\right)\ \mathrm{d}y \quad (11)$$

$$= 0 + n^{-1/2}\ C(\theta_0)\ g\left(\frac{i}{n}\right). \quad (12)$$

In fact, based on the ideas of Kuan and Hornik (1995) and Hjort and Koning (2002) and with the same arguments as in Section 2.2, the whole fluctuation process can be split into one part which is governed by the FCLT from Theorem 1 and a second part which is determined by the function g from (10):

$$W_n(\cdot,\theta_0) \quad \xrightarrow{\mathrm{d}} \quad Z^A(\cdot), \quad (13)$$

where $Z^A(t) = Z(t) + C(\theta_0)G(t)$ and $G(\cdot)$ is the antiderivative of g with $G(t) = \int_0^t g(y)\ \mathrm{d}y$.

Finally, the following limiting process can be derived for the decorrelated empirical fluctuation process:

$$\hat{B}_n^{-1/2} W_n(t,\hat{\theta}_n) \doteq \hat{B}_n^{-1/2}\left\{W_n(t,\theta_0) - tW_n(1,\theta_0)\right\} \quad (14)$$

$$\doteq W^0(t) + B(\hat{\theta}_n)^{-1/2} C(\hat{\theta}_n) G^0(t), \quad (15)$$

with $G^0(t) = G(t) - tG(1)$, provided $B(\cdot)$ is consistent under H_A.

The results above include the results from Section 2.2 as special cases because under the null hypothesis of parameter stability (2) the function g is identical to zero $g \equiv 0$. But the results also imply that tests based on the empirical fluctuation processes will be consistent against suitable local alternatives of type (10).

2.4. Dependent data

As stated above, the assumption of independent observations is often (but not necessarily) violated, in particular when dealing with time-series data. Several approaches are conceivable when the methodology introduced above is to be applied to dependent data.

When using ML estimation techniques the parameters can be estimated from a fully specified likelihood or from a conditional likelihood and the fluctuation processes can be derived accordingly. However, in many situations this is not necessary as consistent estimates $\hat{\theta}$ can be obtained from the usual estimating equations (Godambe 1985; Liang and Zeger 1986). As Lumley and Heagerty (1999) point out, it is crucial for inference in such models to compute consistent estimates for the covariance matrix $\hat{B}$. Heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimators are available, e.g., the kernel HAC estimators with automatic bandwidth selection of Andrews (1991), the prewhitening extensions of Andrews and Monahan (1992) or the class of weighted empirical adaptive variance estimators of Lumley and Heagerty (1999). These can be plugged into the fluctuation processes described above which renders the asymptotic theory valid again.

3. Generalized M-fluctuation tests

We choose the M-estimation framework for estimation of the parameters θ as it contains many other estimation techniques as special cases by choosing a suitable score function ψ. Other classes of estimators are not strictly special cases but are strongly related to M-estimation and the principles introduced in Section 2 can be used to construct fluctuation processes with the same asymptotic properties. A few of these generalizations are outlined in the following.

3.1. Choice of the scores

One of the most common choices for ψ is the partial derivative of some objective function Ψ

$$\psi(y, \theta) \quad = \quad \frac{\partial \Psi(y, \theta)}{\partial \theta}, \tag{16}$$

where Ψ could be the residual sum of squares or the log-likelihood, yielding the OLS or ML estimators $\hat{\theta}$ respectively (the dependence of $\hat{\theta}$ on the number of observations n is suppressed in the following). In both cases the cumulative sums of the first order conditions ψ lead very naturally to fluctuation processes as described in the previous section.

In a misspecification context the score function ψ is typically directly chosen: e.g., robust M-estimation (Huber 1972) or Quasi-Maximum Likelihood (White 1994) which both account for violation of some of the standard model assumptions.

Another approach is to not fully specify a model via its likelihood but to use some estimating equations (Godambe 1960, 1985) which are satisfied by the true model. Similarly, some moment or orthogonality conditions can be exploited to derive estimating functions which again yield parameter estimates. This approach is used in estimation techniques like instrumental variables in linear models (IV, Sargan 1958), the generalized method of moments (GMM, Hansen 1982) for the estimation of economic models or the generalized estimating equations (GEE, Liang and Zeger 1986) for models for longitudinal or time-series data in biostatistics. Further discussion of related estimation approaches can be found in Bera and Bilias (2002). Usually, these are regression models which are not yet covered by the methodology introduced above. However, with modifications as described in Section 4 fluctuation processes with rather similar properties can be derived.

All these methods have in common that the estimation of θ is based on some score or estimating function ψ or a moment or orthogonality condition similar to (9) whose partial sums yield fluctuation processes satisfying the FCLT from Theorem 2. Based on this, tests for parameter instability are constructed in the following section.

3.2. Test statistics

We derived the empirical fluctuation processes because they can capture departures from the null hypothesis (2) of parameter stability. Therefore, simple visual inspection already conveys information about whether H_0 is violated or not. But this alone is, of course, not enough and we want to derive tests based on empirical fluctuation processes while maintaining their capabilities as exploratory tools. The straightforward strategy for constructing test statistics is to consider some scalar functional λ that can be applied to the fluctuation processes.

Given a finite sample of size n as in model (1) an empirical fluctuation process is an $n \times k$ array $(\mathit{efp}_j(i/n))_{i,j}$—where $i = 1, \dots, n$ and $j = 1, \dots, k$ and $\mathit{efp}_j(\cdot)$ is the jth component of $\mathit{efp}(\cdot)$—that converges to a k-dimensional Brownian bridge which is continuous in time. To aggregate this empirical process to a scalar test statistic several suitable functionals of the form $\lambda(\mathit{efp}_j(i/n))$ are conceivable. The limiting distribution for these test statistics can be determined fairly easily, it is just the corresponding (asymptotic) functional applied to the limiting process (Chu, Hornik, and Kuan 1995). Although closed form results for certain functionals of Brownian bridges exist, the critical values are typically best derived by simulation so there are no constraints for the choice of λ.

The functional λ can usually be split into two components: λ_{time} which aggregates over time and λ_{comp} which aggregates over the components of ψ. Common choices for λ_{time} are the absolute maximum, the mean or the range. Typical functionals λ_{comp} include the maximum norm (or L_∞ norm, denoted as $||\cdot||_\infty$) or the squared Euclidean norm (or L_2 norm, denoted as $||\cdot||_2^2$).

As the decorrelated processes are asymptotically independent it seems to be more intuitive to first aggregate over time and then have k independent univariate test statistics $\lambda_{\mathrm{time}}\left(\mathit{efp}_j(i/n)\right)$ $(j = 1, \dots, k)$, each associated with one component of the process which can usually be matched with one component of the parameter vector θ. Using the maximum for λ_{comp} gives

$$\max_{j=1,\ldots,k} \left|\lambda_{\mathrm{time}}\left(\mathit{efp}_j(i/n)\right)\right|$$

and yields a multiple testing procedure controlling the family-wise error rate. If the overall hypothesis is rejected, the component(s) of θ which caused the instability can be identified: All components associated with a significant statistic, i.e., crossing some critical value c, are in conflict with the null hypothesis.

On the other hand, when there is evidence for a structural change a very natural question is *when* it occured. To focus on this question it is obviously better to first aggregate over j and then inspect the resulting univariate process for excessive fluctuation. This can be also done visually, e.g., by checking whether this process crosses some boundary $b(t) = c \cdot d(t)$, where c determines the significance level and $d(\cdot)$ the shape of the boundary. The resulting test statistic is

$$\max_{i=1,\ldots,n} \left|\frac{\lambda_{\mathrm{comp}}\left(\mathit{efp}_j(i/n)\right)}{d(i/n)}\right|,$$

i.e., a weighted maximum of the absolute values of the process aggregated by λ_{comp}. Natural choices are to weigh all observations equally, i.e., $d(t) = 1$, or by the (asymptotic) standard deviation of the fluctuation process, i.e., $d(t) = \{t(1-t)\}^{1/2}$ for the cumulative score process. The visulizations of all these test statistics convey information about the timing of the change(s) (if any) and are powerful exploratory tools in the spirit of Brown *et al.* (1975) (see Section 5 for several examples). Furthermore, this exploratory analysis of the timing of structural changes can also be complemented by formal breakpoint estimation techniques in multiple shift models (Bai 1997; Bai and Perron 1998).

Let us point to three functionals or test types respectively that have certain desirable properties and that will also be used in the applications in Section 5:

Double max test: The only class of test statistics which allows for both identification of the component j as well as the timing i/n of a potential structural instability is when the maximum is used for aggregating over both time and components, i.e., when the L_∞ norm is applied to the (possibly weighted) empirical fluctuation process. In this paper, we use a double maximum test with a constant boundary $d(t) = 1$

$$\max_{i=1,\ldots,n} \max_{j=1,\ldots,k} \left|\mathit{efp}_j(i/n)\right| \tag{17}$$

where the $\mathit{efp}_j(i/n)$ which cross some absolute critical value c can be regarded as violating the hypothesis of stability (Mazanec and Strasser 2000). The limiting distribution is given by $\max_{j=1,\ldots,k} ||W_j^0(t)||_\infty$, i.e., the maximum of a Brownian bridge. A closed form solution for computing the distribution of the maximum of a Brownian bridge is available (see e.g., Ploberger and Krämer 1992).

Supremum of LM statistics test: A strategy which is both theoretically and practically appealing for testing for a single abrupt change of unknown timing is to compute some test statistic for all potential change points in an interval Π and reject if their maximum is too large (Andrews 1993). The sequence of single break LM statistics can be computed by taking the squared Euclidean norm of the empirical fluctuation process weighted by its variance, thus the $\sup LM$ statistic is

$$\sup_{t \in \Pi} \frac{||\mathit{efp}(t)||_2^2}{t(1-t)}. \tag{18}$$

The limiting distribution is given by $\sup_{t\in\Pi}(t(1-t))^{-1}||W^0(t)||_2^2$ and approximate asymptotic p values can be computed with the methods of Hansen (1997).

Cramér-von Mises type test: Nyblom (1989) derives a Cramér-von Mises type test for parameter instability based on ML scores and shows that this test is the locally most powerful test against the alternative that the parameters follow a random walk. For this test, the functional λ is the average of the squared Euclidean norm yielding the test statistic

$$\frac{1}{n}\sum_{i=1}^{n}\left|\left|efp\left(\frac{i}{n}\right)\right|\right|_2^2, \tag{19}$$

i.e., first the L_2 norm is used to aggregrate over the components and then the mean of the resulting aggregated process is used as the test statistic. The corresponding limiting distribution is given by $\int_0^1 ||W^0(t)||_2^2 \, dt$ and simulated critical values are published in various places in the literature (see e.g., Anderson and Darling 1952; Hansen 1992).

Many other test statistics known from the statistics and econometrics literature are contained as special cases in the rich class of generalized M-fluctuation tests. Specifically, the residuals-based tests of Kuan and Hornik (1995), the ML score-based framework of Hjort and Koning (2002) as well as other tests based on F statistics (Andrews 1993; Andrews and Ploberger 1994) are contained. A unifying view is given in Zeileis (2005).

3.3. Optimality

By combination of an empirical M-fluctuation process—based on a score function ψ—with some fluctuation capturing functional λ we derived a very rich and flexible class of M-fluctuation tests. In applications, this flexibility is often a curse rather than a blessing because a particular test has to be chosen from the general class of tests. Whereas the choice of the score function ψ is typically determined by the application and the model of interest respectively, the choice of the functional λ is less obvious. Ideally, one would like to use the functional λ which yields the optimal test (for the problem at hand). Unfortunately, finding some uniformly most powerful test for the general alternative of structural change or parameter instability is a futile task as this vast alternative encompasses arbitrarily many conceivable ways in which the parameters can vary. Therefore, there is a multitude of structural change tests (not necessarily from the M-fluctuation framework) which are not dominated by any other test over all conceivable alternatives.

However, if the alternative is restricted there are some optimality results for structural change tests available in certain special cases: Nyblom (1989) shows that under normality the Cramér-von Mises type test from Equation (19) is locally most powerful against the alternative that the parameters follow a random walk. For a single shift alternative (with unknown breakpoint) Andrews and Ploberger (1994) derive two tests—namely the ave$F$ and exp$F$ test—with certain optimality properties. They derive the FCLT for a sequence of LR, Wald or LM test statistics (where the latter also fall into the class of M-fluctuation tests, as pointed out above) and the average and exp functionals employed direct power against weak and strong single shifts (of unknown timing), respectively. Of course, these tests also have non-trivial power against other alternatives but are known to be outperformed by other functionals in various situations, e.g., for multiple break alternatives by fluctuation tests based on MOSUM functionals (Kuan and Hornik 1995).

To sum up, some advice can be given for guiding the choice of a (M-fluctuation) test in applications: If a significance test should be carried out in a situation where a single shift alternative (of unknown timing) is plausible the tests of Andrews and Ploberger (1994) are probably most useful. If, however, additionally to the timing of the shift the component affected by it should be revealed some (possibly weighted) double maximum type M-fluctuation test might be more useful (especially in an exploratory setting). For random walk alternatives a Cramér-von Mises type statistic and for multiple shifts MOSUM type statistics will be good choices. Further discussion of the choice of functionals for fluctuation tests—in particular of different optimal weighting schemes—can be found in Hjort and Koning (2002). The determination of good or even optimal functionals for other classes of alternatives is of great interest but beyond the scope of this paper.

4. Testing for parameter instability in GLMs

The generalized M-fluctuation test framework presented in Section 3 solves the conceptual problems—choice of scores/estimation technique, computation of fluction processes, construction/visualization of test statistics—in the general design of fluctuation tests in parametric models and provides the guidelines how to derive tools in specific situations. In this section, we apply the general framework to one such specific model: the generalized linear model. To construct fluctuation tests in a regression setup where the observations can be split into $Y_i = (y_i, x_i)^\top$ with a dependent variable y_i and additional regressors x_i, the usual approach is to model the conditional distribution $f(y_i \,|\, x_i, \theta_i)$ of y_i given the x_i. The standard assumption is that the x_i form a weakly dependent process without deterministic or stochastic trends, see Andrews (1993) for technical details or also Krämer *et al.* (1988), Hansen (1992) or Hjort and Koning (2002). Under such suitable assumptions the same asymptotic distribution can be derived for the processes based on estimates of the regression coefficients which can be obtained by various procedures as discussed in Section 3.1.

To make the dependence on the covariates obvious the score function ψ from (3) is now written as

$$\psi(Y_i, \theta_i) \quad = \quad \psi(y_i, x_i, \theta_i).$$

Of course, it is still required to have zero expectation (with respect to $f(y_i \,|\, x_i, \theta_i)$) and we assume that the variances stabilize:

$$\frac{1}{n} \sum_{i=1}^{n} \mathsf{COV}[\psi(y_i, x_i, \theta_0)] \;=\; J_n \quad \xrightarrow{\text{p}} \quad J,$$

where the matrix J in a regression context corresponds to $B(\theta)$ from (5) in the setup without regressors. This follows for example from the weak dependence assumption stated above.

It is easy to show that functional central limit theorems similar to Theorem 1 and 2 hold for the resulting theoretical and empirical fluctuation processes based on $\psi(y_i, x_i, \theta_i)$, in particular the limiting processes are the same with J instead of $B(\theta_0)$.

In the following, we derive explicit test statistics for the generalized linear model (GLM, McCullagh and Nelder 1989).

4.1. The generalized linear regression model

To fix notations, y_i is a response variable distributed according to a distribution $F(\theta, \phi)$ where θ is the canonical parameter and ϕ is the dispersion parameter common to all y_i. The probability density has the form

$$f(y_i \,|\, \theta, \phi) \quad = \quad \exp\left\{ \frac{y_i \theta - q(\theta)}{w(\phi)} + p(y_i, \phi) \right\}, \tag{20}$$

for some known functions $p(\cdot)$, $q(\cdot)$ and $w(\cdot)$, so that $\mathsf{E}(y_i) = \mu_i = q'(\theta)$ and $\mathsf{VAR}(y_i) = w(\phi) q''(\theta) = w(\phi) V(\mu_i)$. The following relationship is assumed for covariates and responses:

$$\mu_i \quad = \quad h(\eta_i) \quad = \quad h(x_i^\top \beta_i) \qquad (i = 1, \dots, n),$$

where $h(\cdot)$ is a known inverse link function, β is again the vector of regression coefficients and η_i is the linear predictor.

The regression coefficients β are usually estimated by ML and ϕ is treated as a nuisance parameter (or is known anyway). The resulting score function for β is

$$\psi(y_i, x_i, \beta) \quad = \quad x_i \; h'(x_i^\top \beta) \; V(\mu_i)^{-1} (y_i - \mu_i),$$

where $h'(\cdot)$ is the derivative of the inverse link function. The corresponding covariance matrix J_n is given by

$$J_n \quad = \quad \frac{1}{n} \sum_{i=1}^{n} h'(x_i^\top \beta)^2 w(\phi) \; V(\mu_i)^{-1} x_i x_i^\top.$$

An alternative approach (as pointed out in Section 2.4) to the estimation of β is to adopt a quasi-likelihood (White 1994) or GEE view of the model (Liang and Zeger 1986): The same score function is used (thus, leading to the same parameter estimate) but it is not assumed that it is derived from the full correctly specified likelihood. Consistent estimates of the parameters can typically still be derived (e.g., when the observations are correlated), but inference has to be based on different covariance matrix estimators. Therefore, instead of using a simple plug-in estimator for J_n (based on estimates for β and ϕ) different estimators are appropriate, e.g., HAC covariances.

In the following, we give explicit formulae for the empirical fluctuation processes in three important special cases of the GLM: the linear regression model with normal errors, the log-linear poisson model and the binomial (logistic) regression model.

4.2. Normal model

The normal GLM with canonical identity link is usually written as the linear regression model

$$y_i \quad = \quad x_i^\top \beta \, + \, u_i \qquad (i = 1, \dots, n),$$

where the disturbances u_i are normally distributed with zero mean and common variance $\phi = \sigma^2$. This is closely related to the general linear model which usually does not impose normality on the errors and in which the parameters are typically estimated by OLS. As this leads to almost identical processes and test statistics we do not treat the general linear model explicitly here.

Usually, the variance σ^2 is treated as a nuisance parameter and the regression coefficient β are estimated by ML (or equivalently OLS) which leads to the following scores

$$\psi_\beta(y_i, x_i, \beta) \quad = \quad x_i(y_i - x_i^\top \beta) \; = \; x_i(y_i - \mu_i).$$

These yield the usual estimate $\hat{\beta}$ whose stability can be assessed using the empirical fluctuation process

$$efp(t) \quad = \quad \hat{J}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \psi_\beta(y_i, x_i, \hat{\beta}), \tag{21}$$

where $\hat{J} = n^{-1} \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 x_i x_i^\top$ and $\hat{\mu}_i$ is the predicted mean of y_i.

Alternatively, the error variance can be treated as a full parameter of the model and an additional component of the empirical fluctuation process can be derived for the corresponding ML estimate $\hat{\sigma}^2$:

$$\psi_{\sigma^2}(y_i, x_i, \beta, \sigma^2) \quad = \quad (y_i - x_i^\top \beta)^2 - \sigma^2.$$

4.3. Poisson model

If the y_i are poisson distributed $\mathrm{Poi}(\mu_i)$ then $V(\mu) = \mu$ and $w(\phi) = 1$. Using the canonical log link yields $h'(x) = \exp(x)$ so that the covariance can be estimated by $\hat{J} = n^{-1} \sum_{i=1}^n \hat{\mu}_i x_i x_i^\top$.

The empirical fluctuation process simplifies to

$$efp(t) \quad = \quad \hat{J}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} (y_i - \hat{\mu}_i) x_i. \tag{22}$$

With a simple modification both processes can not only be used in poisson but also in quasi poisson models where overdispersion is allowed. Then, the variance is not required to be equal to the mean but can be $\mathsf{VAR}(y_i) = \phi\mu_i$. Note that in this case the density function is not given by (20) with $w(\phi) = \phi$. The dispersion parameter is a nuisance parameter and can be consistently estimated by $X^2/(n-k)$, where X^2 is the usual Pearson χ^2 statistic. To obtain properly standardized fluctuation processes, $efp(t)$ from (22) has to be multiplied by $\hat{\phi}^{-1/2}$ which can then be used as usual for testing the constancy of the regression coefficients β.

4.4. Binomial model

Let y_i be the proportion of successes from m trials such that my_i is binomially distributed $\mathrm{Bin}(\mu_i, m_i)$. Then the variance is determined by $w(\phi) = 1/m_i$ and $V(\mu) = \mu(1-\mu)$ and if the canonical logit link is used then $h'(x) = \exp(x)/(1+\exp(x))^2$. Given the ML estimates $\hat{\beta}$ and the corresponding fitted values $\hat{\mu}_i$, the covariance matrix for the M-fluctuation process can be estimated by

$$\hat{J} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \frac{h'(x_i^\top \hat{\beta})^2}{m_i \hat{\mu}_i (1-\hat{\mu}_i)} x_i x_i^\top.$$

The empirical fluctuation process is then given by

$$\mathit{efp}(t) \quad = \quad \hat{J}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} h'(x_i^\top \hat{\beta}) \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i (1-\hat{\mu}_i)} x_i. \tag{23}$$

This process together with the mean L_2 (or Cramér-von Mises) functional from (19) is applied to a historical binomial time series in Section 5.3.

5. Illustrations and applications

We illustrate a few of the tests discussed above by applying them to the following three models: a normal model for Dow Jones industrial average stock returns, a poisson GLM for the number of youth homicides in Boston and a binomial GLM (logistic regression) for the proportion of illegitimate births in pre-industrial Austria. For the latter model, an ML approach is employed, while for the first two models, a quasi-likelihood approach is adopted that is based on GLM scores and HAC covariances. All computations have been carried out in the R system for statistical computing (R Development Core Team 2007, `http://www.R-project.org/`), in particular using the R package `strucchange` (Zeileis, Leisch, Hornik, and Kleiber 2002) which includes all three data sets. Both, R itself and the package `strucchange`, 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/`.

5.1. Dow Jones industrial average stock returns

Hsu (1979) investigates the stability of the volatility of weekly Dow Jones industrial average stock returns based on weekly closing prices from 1971-07-02 to 1974-08-02, the corresponding log-difference returns ($\times 100$) are depicted in Figure 1. Hsu (1979) assumes the mean to be constant and (approximately) known and assesses only the stability of the variance of the series finding a clear break in mid-March 1973. We follow the approach of Hsu (1979) by modeling the data as a series of approximately normal observations, however, we assess the stability of both the mean and the variance of the series. The estimated parameters are $\hat{\mu} = \hat{\beta} = -0.104$ and $\hat{\sigma}^2 = 4.889$ which are tested using the M-fluctuation process defined in Equation (21) and the double max functional from Equation (17). The process from (21) is slightly modified by using the combined scores ψ_β and ψ_{σ^2} and estimating $\hat{J}$ by a quadratic spectral kernel HAC estimator with VAR(1) prewhitening and automatic bandwidth selection based on an AR(1) approximation to account for potential autocorrelations.

This test is performed graphically in Figure 2 by plotting the processes for both parameters with a horizontal boundary for $\pm$ the critical value at 5% level. As one of the processes crosses its boundary, there is evidence for an overall parameter instability in the model. As there is only moderate non-significant fluctuation in the process for the mean, but a clear boundary crossing for the variance process, it can be concluded that this instability is caused by a change in the variance while the mean remains constant. Furthermore, the clear peak in the variance process conveys the information that there is a single abrupt increase in the variance in March 1973, matching the break found by Hsu (1979) (highlighted by the dotted vertical line) very well. The corresponding p value is < 0.001.

5.2. Boston homicide data

To address the problem of continuing high homicide rates in Boston, in particular among young people, a policing initiative called the "Boston Gun Project" was launched in early 1995. This project implemented what became known as the "Operation Ceasefire" intervention in the late spring of 1996 which aimed at lowering homicide rates by a deterrence strategy. As a single shift alternative seems reasonable but the precise start of the intervention cannot be determined, Piehl *et al.* (2003) chose to model the number of youth homicides in Boston using modifications of the F tests for structural change of Andrews (1993) and Andrews and Ploberger (1994) based on monthly data from 1992(1) to 1998(5) (see Figure 3) and assessing the significance via Monte Carlo results. In their regression model A, they include control variables like the population or a factor coding the month, however both have no significant influence at 10% level. Hence, we use a simple poisson model for only the mean of the number of homicides which is estimated as $\hat{\mu} = \exp(\hat{\beta}) = 2.766$. We assess the stability of the mean using the M-fluctuation process defined in Equation (22) with the supremum of LM statistics functional from Equation (18) and a trimming of $\Pi = [0.1, 0.9]$, again using a kernel HAC estimator for $\hat{J}$ as in the previous section to account for potential autocorrelations.

The empirical fluctuation process of LM statistics can be seen in Figure 4. As it crosses its 5% level boundary, there is evidence for a decrease of the number of homicides. Furthermore, the peak in the process indicates that the change seems to have occurred around early 1996 when the Operation Ceasefire was implemented (dotted line). The corresponding p value is 0.008.

5.3. Illegitimate births in Großarl

In 18th century Salzburg, Austria, human reproductive behaviour was confined to marital unions due to sanctions by the catholic church and the legal system. Nevertheless, illegitimate births conceived outside wedlock happened although the catholic church tried to prevent them by moral regulations of increasing severity. Veichtlbauer, Zeileis, and Leisch (2006) discuss the impact of these and other policy interventions on the population system in Großarl, a small village in the Austrian Alps in the region of the archbishopric Salzburg.

Here, we consider a binomial regression for the number of illegitimate and legitimate births between 1700 and 1800 (see Figure 5). There were between 48 and 63 births (inter-quartile range) per year in Großarl during the 18th century, about seven of which were illegitimate yielding an estimated fraction of $\hat{\mu} = \text{logit}(\hat{\beta}) = 12.792\%$. During this time the close linkage between religiosity and morality and between church and state led to a policy of moral suasion and social disciplining, especially concerning forms of sexuality that were not wanted by the catholic church. Moral regulations aimed explicitly at avoiding such unwanted forms of sexuality, e.g., by punishing fornication by stigmatizing corrections, corporal punishment, compulsory labour or workhouse-prison. After secularization such regulations were abolished in the 19th century.

To assess whether these moral interventions had any effect on the mean proportion of illegitimate births, we employ the M-fluctuation test based on ML-scores from a binomial GLM as defined in Equation (23) with the Cramér-von Mises functional from Equation (19). Figure 6 shows the L_2 norm of the score-based fluctuation process and the dashed horizontal line represents the test statistic (the mean L_2 norm), which clearly exceeds its 5% critical value (solid horizontal line). Thus, the test provides evidence for a decrease in the fraction of illegitimate births suggesting that the moral regulations were effective. In addition to the information that the test finds evidence for structural change in the data, the peak in the process conveys that there has been at least some structural break at about 1750, but two minor peaks on the left and the right can also be seen in the process. These match the three major moral interventions of the catholic church in 1736, 1753 and 1771, respectively (indicated by dotted lines). The corresponding p value is < 0.001.

6. Conclusions

In this paper, we propose a general class of tests for parameter instability based on partial sum fluctuation processes of M-estimation scores. Based on functional central limit theorems for these fluctuation processes both under the null hypothesis and local alternatives it is shown how structural changes in parametric models, with a special emphasis on generalized linear regression models, can be discovered by test statistics that capture the fluctuation in the M-score processes.

The strategy derived can be summarized as follows: Given a data set, choose a model which should be tested for parameter stability or in which structural changes should be revealed, and choose an estimation technique suitable for the data under consideration. From the estimation technique the choice of the scores ψ follows naturally yielding an empirical fluctuation process. To check for excessive fluctuation in this process a functional has to be chosen which brings out either the timing of a structural change or the component incorporating the instability or both. The resulting significance test can be enhanced by a visualization method that not only displays the result of the test procedure but also conveys information about the type of structural instability and thus allows for better understanding of the structure of the data.

Acknowledgements

The authors are grateful for support by the Austrian Science Foundation (FWF) under grant SFB#010 ('Adaptive Information Systems and Modeling in Economics and Management Science').

References

Anderson TW, Darling DA (1952). "Asymptotic Theory of Certain Goodness-of-fit Criteria Based on Stochastic Processes." *Annals of Mathematical Statistics*, **23**, 193–212.

Andreou E, Ghysels E (2002). "Detecting Multiple Breaks in Financial Market Volatility Dynamics." *Journal of Applied Econometrics*, **17**, 579–600.

Andrews DWK (1991). "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica*, **59**, 817–858.

Andrews DWK (1993). "Tests for Parameter Instability and Structural Change with Unknown Change Point." *Econometrica*, **61**, 821–856.

Andrews DWK, Monahan JC (1992). "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator." *Econometrica*, **60**(4), 953–966.

Andrews DWK, Ploberger W (1994). "Optimal Tests When a Nuisance Parameter is Present Only Under the Alternative." *Econometrica*, **62**, 1383–1414.

Bai J (1997). "Estimation of a Change Point in Multiple Regression Models." *Review of Economics and Statistics*, **79**, 551–563.

Bai J, Perron P (1998). "Estimating and Testing Linear Models with Multiple Structural Changes." *Econometrica*, **66**, 47–78.

Bera AK, Bilias Y (2002). "The MM, ME, ML, EL, EF and GMM Approaches to Estimation: A Synthesis." *Journal of Econometrics*, **107**, 51–86.

Billingsley P (1999). *Convergence of Probability Measures*. Wiley, New York, 2nd edition.

Brown RL, Durbin J, Evans JM (1975). "Techniques for Testing the Constancy of Regression Relationships over Time." *Journal of the Royal Statistical Society B*, **37**, 149–163.

Chu CSJ, Hornik K, Kuan CM (1995). "MOSUM Tests for Parameter Constancy." *Biometrika*, **82**, 603–617.

Godambe VP (1960). "An Optimum Property of Regular Maximum Likelihood Estimation." *Annals of Mathematical Statistics*, **31**, 1208–1211.

Godambe VP (1985). "The Foundations of Finite-Sample Inference in Stochastic Processes." *Biometrika*, **48**, 419–428.

Hansen BE (1992). "Testing for Parameter Instability in Linear Models." *Journal of Policy Modeling*, **14**, 517–533.

Hansen BE (1997). "Approximate Asymptotic p Values for Structural-Change Tests." *Journal of Business & Economic Statistics*, **15**, 60–67.

Hansen L (1982). "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica*, **50**, 1029–1054.

Hjort NL, Koning A (2002). "Tests for Constancy of Model Parameters Over Time." *Nonparametric Statistics*, **14**, 113–132.

Hsu DA (1979). "Detecting Shifts of Parameter in Gamma Sequences with Applications to Stock Price and Air Traffic Flow Analysis." *Journal of the American Statistical Association*, **74**, 31–40.

Huber PJ (1972). "Robust Statistics: A Review." *Annals of Mathematical Statistics*, **43**, 1041–1067.

Huškova M (1999). "Gradual Changes Versus Abrupt Changes." *Journal of Statistical Planning and Inference*, **76**, 109–125.

Krämer W, Ploberger W, Alt R (1988). "Testing for Structural Change in Dynamic Models." *Econometrica*, **56**(6), 1355–1369.

Kuan CM, Hornik K (1995). "The Generalized Fluctuation Test: A Unifying View." *Econometric Reviews*, **14**, 135–161.

Liang KY, Zeger SL (1986). "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, **73**, 13–22.

Lumley T, Heagerty P (1999). "Weighted Empirical Adaptive Variance Estimators for Correlated Data Regression." *Journal of the Royal Statistical Society B*, **61**, 459–477.

Mazanec JA, Strasser H (2000). *A Nonparametric Approach to Perceptions-based Market Segmentation: Foundations*. Springer-Verlag, Berlin.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. Chapman & Hall, London, 2nd edition.

Muggeo VMR (2003). "Estimating Regression Models with Unknown Break-Points." *Statistics in Medicine*, **22**, 3055–3071.

Nyblom J (1989). "Testing for the Constancy of Parameters Over Time." *Journal of the American Statistical Association*, **84**, 223–230.

Piehl AM, Cooper SJ, Braga AA, Kennedy DM (2003). "Testing for Structural Breaks in the Evaluation of Programs." *Review of Economics and Statistics*, **85**(3), 550–558.

Ploberger W, Krämer W (1992). "The CUSUM Test with OLS Residuals." *Econometrica*, **60**(2), 271–285.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org/.

Sargan JD (1958). "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica*, **26**, 393–415.

Stefanski LA, Boos DD (2002). "The Calculus of M-Estimation." *The American Statistician*, **56**, 29–38.

Stock JH, Watson MW (1996). "Evidence on Structural Instability in Macroeconomic Time Series Relations." *Journal of Business & Economic Statistics*, **14**, 11–30.

Tukey JW (1962). "The Future of Data Analysis." *Annals of Mathematical Statistics*, **33**, 1–67.

Veichtlbauer O, Zeileis A, Leisch F (2006). "The Impact of Interventions on a Pre-Industrial Austrian Alpine Population." *Collegium Antropologicum*, **30**, 1–11.

White H (1994). *Estimation, Inference and Specification Analysis*. Cambridge University Press.

Zeileis A (2005). "A Unified Approach to Structural Change Tests Based on ML Scores, F Statistics, and OLS Residuals." *Econometric Reviews*, **24**(4), 445–466. doi:10.1080/07474930500406053.

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**(2), 1–38. URL http://www.jstatsoft.org/v07/i02/.

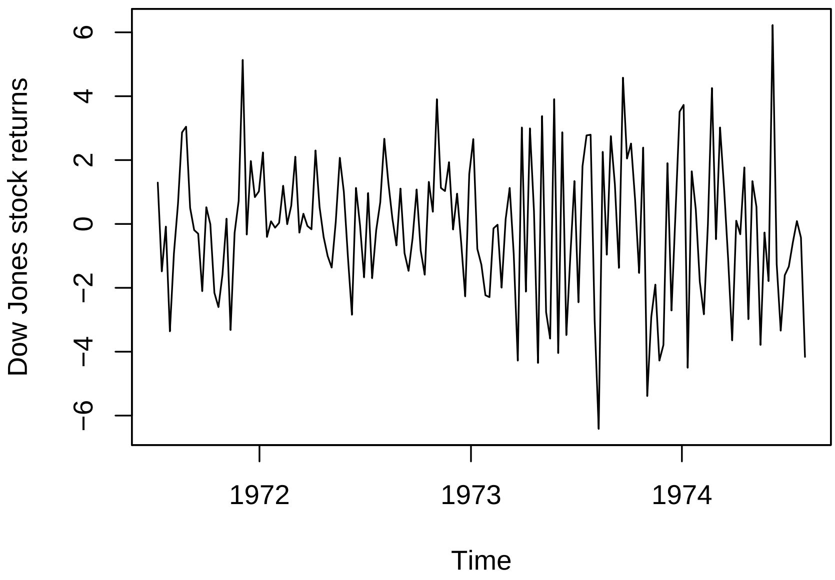

Figure 1: Dow Jones industrial average log-difference returns (in percent).

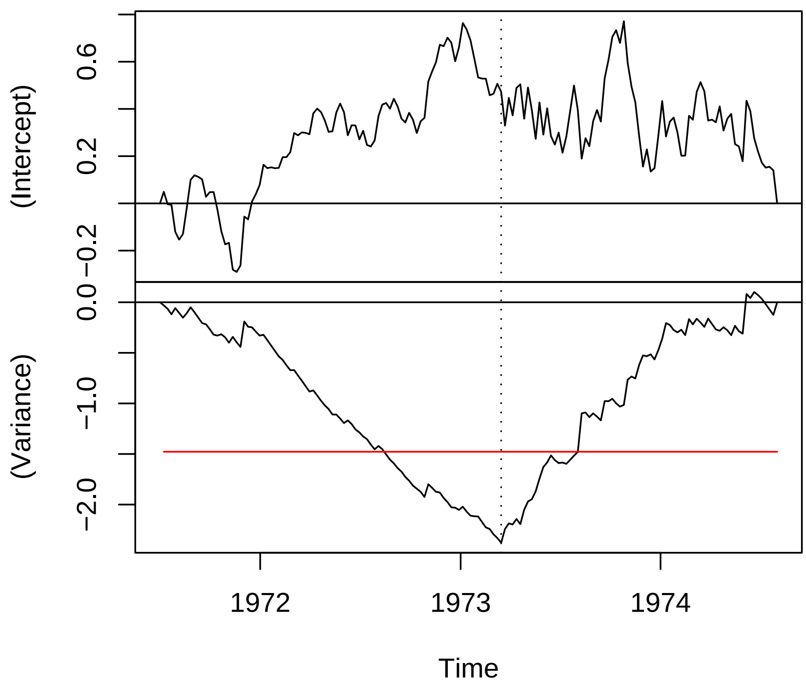

Figure 2: Bivariate normal M-fluctuation process for Dow Jones data.

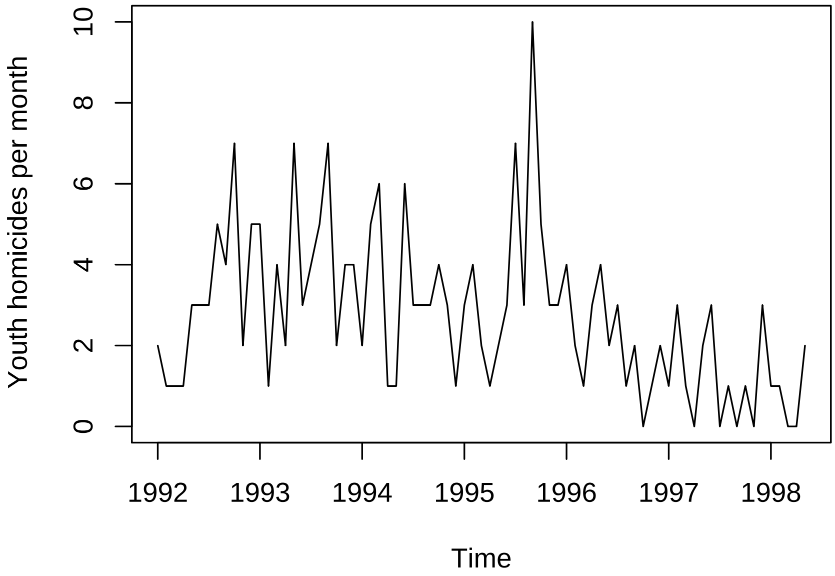

Figure 3: The Boston homicide data.

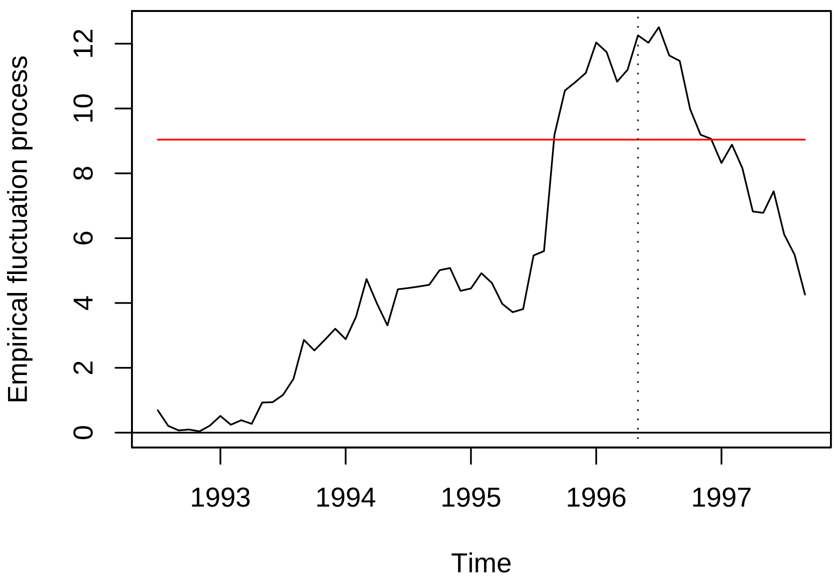

Figure 4: Poisson M-fluctuation process (LM statistics) for Boston homicide data.

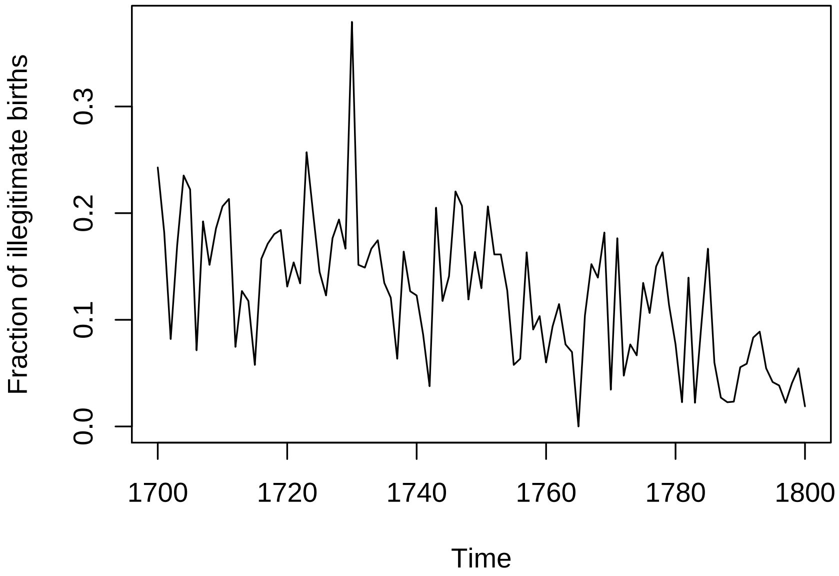

Figure 5: Illegitimate births in Großarl.

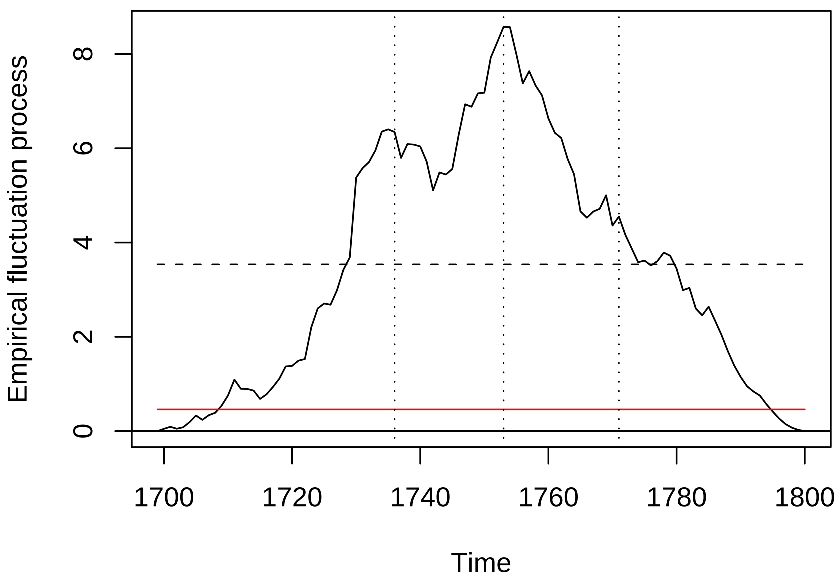

Figure 6: Binomial M-fluctuation process (L_2 norm) for Großarl data.