Comput Stat (2013) 28:2585–2598
DOI 10.1007/s00180-013-0419-4

ORIGINAL PAPER

Non-monotonic penalizing for the number of structural breaks

Erhard Reschenhofer · David Preinerstorfer · Lukas Steinberger

Received: 8 July 2012 / Accepted: 27 April 2013 / Published online: 10 May 2013

Abstract This paper first reduces the problem of detecting structural breaks in a random walk to that of finding the best subset of explanatory variables in a regression model and then tailors various subset selection criteria to this specific problem. Of particular interest are those new criteria, which are obtained by means of simulation using the efficient algorithm of Bai and Perron (J Appl Econom 18:1–22, 2003). Unlike conventional variable selection methods, which penalize new variables entering a model either in the same way (e.g., AIC and BIC) or milder (e.g., MRIC and FPE_{sub}) than already included variables, they do not follow any monotonic penalizing scheme. In general, their non-monotonicity is more pronounced in the case of fat tails. The characteristics of the different criteria are illustrated using bootstrap samples from the Nile data set.

Keywords Breaks in the drift · Random walk · Subset selection · Variable selection

Mathematics Subject Classification (2000) 62M10 · 62M20

1 Introduction

Estimating the number of breaks in the drift of a random walk requires balancing model fit and model complexity. This can be done with the help of a model selection

E. Reschenhofer (✉) · D. Preinerstorfer · L. Steinberger
Department of Statistics and Operations Research, University of Vienna, Vienna, Austria
e-mail: erhard.reschenhofer@univie.ac.at

D. Preinerstorfer
e-mail: david.preinerstorfer@univie.ac.at

L. Steinberger
e-mail: lukas.steinberger@univie.ac.at

criterion like AIC (Akaike 1973) or BIC (Schwarz 1978). An alternative approach is to use the sequential testing procedure proposed by Bai and Perron (1998, 2003). A big advantage of this procedure is that it allows for general forms of serial correlation and heteroscedasticity. However, the results of simulation experiments (Bai and Perron 2006) show that in all cases except the base case, where there is neither serial correlation nor heterogeneity across segments, considerable care should be taken in choosing a required trimming parameter, which is inversely related to the minimal length of a segment.

In this paper, we focus on the first method. For the simplest case of an independent normal sequence with shifts in the mean, Yao (1988) established the consistency of estimating the number of breaks by minimizing the Bayesian information criterion

$$BIC(k) = n \log\left(\frac{RSS(k)}{n}\right) + p \log(n) \tag{1}$$

Schwarz (1978), where $RSS(k)$ is obtained by minimization of the residual sum of squares over all sets of k breakpoints and p is the total number of model parameters that have to be estimated. When all q regression parameters of a linear regression model are subject to change and the variance of the errors is constant, there are $p = q(k+1)+1$ conventional model parameters (excluding the k break dates). In the simplest case, where the model describes only shifts in the mean, we have $q = 1$. Yao and Au (1989) proved the consistency of a class of criteria satisfying certain conditions (for improved results see Kuehn 2001). Liu et al. (1997) compared a particular element of this class, namely

$$YA(k) = n \log\left(\frac{RSS(k)}{n}\right) + pc_1 n^{\alpha} \tag{2}$$

with $c_1 = 0.368$ and $\alpha = 0.7$, to BIC as well as to their modified BIC

$$\begin{aligned} LWZ(k) &= n \log\left(\frac{RSS(k)}{n-p}\right) + pc_0(\log(n))^{2+\delta_0} \\ &= n \log\left(\frac{RSS(k)}{n}\right) + n \log\left(\frac{n}{n-p}\right) + pc_0(\log(n))^{2+\delta_0} \end{aligned} \tag{3}$$

with $c_0 = 0.299$ and $\delta_0 = 0.1$. Their simulations suggest that both YA and LWZ outperform BIC in a simple framework with only two breaks. In a further simulation study (Perron 1997), BIC and LWZ performed reasonably well in the absence of serial correlation. In contrast, Akaike's information criterion

$$AIC(k) = n \log\left(\frac{RSS(k)}{n}\right) + 2p \tag{4}$$

performed very badly. Bai and Perron (2006) found that LWZ worked better than BIC under the null hypothesis of no break but performed much worse under the alternative hypothesis because of its higher penalty. They did not consider AIC any further

because of its bad performance in Perron's (1997) study. Unfortunately, penalizing the number of breaks in addition to the regular parameters (Ninomiya 2005) has only a slight toughening effect on AIC. We therefore take another approach. By rewriting the structural-break problem as a subset-selection problem we avoid to distinguish between regular and non-regular parameters and obtain penalties which are comparable to those of YA and LWZ. However, our penalties have the advantage that they do not depend on constants like c_1 and α or c_0 and δ_0, which are often chosen arbitrarily or based on simulation studies. Also the BIC in its standard form has been derived by ignoring a remainder term (Schwarz 1978), which depends on the specification of the prior distributions for the model parameters. Different priors imply different remainder terms and therefore also different versions of BIC (see, e.g., Kass and Wasserman 1995; Reschenhofer 1996).

The estimation of linear regression models with multiple breaks and unknown break dates poses not only a statistical but also a numerical problem. A simple grid search procedure to find the k break dates, which minimize the sum of squared residuals globally, requires least squares operations of order $O(n^k)$ and is therefore only feasible if k is small. In our simulations and our empirical analysis we have therefore used the efficient algorithm of Bai and Perron (2003), which is based on the principle of dynamic programming and requires at most least squares operations of order $O(n^2)$ for any k.

In the next section, we will extend the bias correction approach, on which the AIC is based, to the task of estimating the number of breaks. Section 3 presents the results of an empirical investigation. Section 4 concludes.

2 Tailoring criteria for the estimation of the number of breaks

Let

$$y_t = \mu_t + \varepsilon_t \tag{5}$$

be the increments of a random walk with drift, whose innovations ε_t are i.i.d. with mean 0 and variance σ^2. We assume that k breaks in the drift μ_t occur at times $1 < t_1 < \cdots < t_k < n$, i.e.,

$$\mu_t = \begin{cases} \mu^{(1)}, & 0 = t_0 < t \le t_1, \\ \vdots & \vdots \\ \mu^{(k)}, & t_{k-1} < t \le t_k, \\ \mu^{(k+1)}, & t_k < t. \end{cases} \tag{6}$$

The problem of estimating the number and the dates of the breaks can be reduced to the problem of selecting the optimal submodel

$$\mathbf{y} = \mathbf{X_S}\boldsymbol{\beta_S} + \boldsymbol{\varepsilon}$$

of the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{7}$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \tag{8}$$

and $\mathbf{X_S}$ is a submatrix of columns of $\mathbf{X}$. The total number of columns of the matrix $\mathbf{X}$ determines the number of candidate submatrices for each dimension. For example, for $n = 100$ there are 100 submatrices with 1 column, 4,950 submatrices with 2 columns, 161,700 submatrices with 3 columns, etc. A disadvantage of using criteria such as AIC and BIC for the selection of the best submodel (submatrix) is that they do not take these differences into account. They just penalize the best submodel of each dimension in the same way as in the case of nested models, where there is only one submodel for each dimension. In contrast, specially designed subset selection criteria like RIC (Foster and George 1994), MRIC (George and Foster 2000), FPE_{sub} (Reschenhofer 2004), and FPE_0 (Reschenhofer et al. 2012) take also the respective numbers of submodels into account when they compare different model dimensions. The last two criteria may be regarded as extensions of the bias correction approach, on which criteria like AIC and FPE (Rothman 1968; Akaike 1969) are based, to the case of non-nested models. In the following, we adapt these two criteria for the specific task at hand, namely the estimation of the number of breaks in the drift of a random walk.

If a fixed submodel of dimension K, which is represented by a subset S of K column indices, is correctly specified, i.e., if

$$E\,\hat{\mathbf{y}}_\mathbf{S} = E\,\mathbf{X_S}\hat{\boldsymbol{\beta}}_S = E\,\mathbf{X_S}(\mathbf{X}_\mathbf{S}'\mathbf{X_S})^{-1}\mathbf{X}_\mathbf{S}'\mathbf{y} = \boldsymbol{\mu},$$

then

$$\hat{\sigma}_S^2 = \frac{1}{n-K} RSS(S) = \frac{1}{n-K} \left\|\mathbf{y} - \hat{\mathbf{y}}_\mathbf{S}\right\|^2 \tag{9}$$

will be an unbiased estimator of σ^2 and

$$FPE(S) = RSS(S)\frac{n+K}{n-K} \tag{10}$$

will be an unbiased estimator of the mean squared prediction error

$$MSPE(S) = E\left\|\mathbf{z} - \hat{\mathbf{y}}_\mathbf{S}\right\|^2,$$

where $\mathbf{z}$ is an independent sample from the same distribution as $\mathbf{y}$. For large n, model selection by FPE is practically equivalent to model selection by AIC, because

$$\begin{aligned} n\log(FPE(S)) &= n\log\left(RSS(S)\left(1+\frac{2K}{n-K}\right)\right) \\ &= n\log(RSS(S)) + \log\left(\left(1+\frac{2K}{n-K}\right)^n\right) \\ &\sim n\log(RSS(S)) + 2K, \end{aligned}$$

which differs from

$$AIC(S) = n\log\left(\frac{RSS(S)}{n}\right) + 2(K+1) \tag{11}$$

only by the additive constant $-n\log(n)+2$. In general, additive penalties P^a can be obtained from multiplicative penalties P^m via the transformation

$$P^a(K) = n\log(P^m(K)). \tag{12}$$

Turning to the case of non-nested models, we will now compare different model dimensions rather than different fixed models. Accordingly, we will consider that submodel of dimension K, which minimizes the residual sum of squares, rather than a fixed submodel of dimension K. Let $\hat{\mathbf{y}}_\mathbf{K}$ denote that estimator of μ which is based on the best submodel of dimension K. Then

$$FPE_{sub}(K) = RSS(K)\frac{n+\zeta_1(K,N)}{n-\zeta_1(K,N)}, \tag{13}$$

where N is the number of potential regressors (in our case N equals n) and $\zeta_1(K,N)$ is the expected value of the sum of the K largest of N independent $\chi^2(1)$-variables, will be an unbiased estimator of

$$MSPE(K) = E\left\|\mathbf{z}-\hat{\mathbf{y}}_\mathbf{K}\right\|^2,$$

provided that the regressors are orthogonal and the theoretical regression coefficients vanish (see Reschenhofer 2004). In the next two subsections, we will try to relax these apparently very restrictive conditions.

2.1 Relaxing the assumption of vanishing regression coefficients

Under the assumption of vanishing regression coefficients, it seems okay to use $\zeta_1(K,N)$ as benchmark for assessing the effect of the apparently most important subset of size K. However, if there are $K-1$ dominant regressors, which are certain to be selected, there are only $N-K+1$ regressors remaining, which may be selected in addition to the certain regressors. For the assessment of the importance of the first

of the remaining regressors, it might be more appropriate to regard it as the most important in a random sample of size $N - K + 1$ rather than as the Kth most important in a random sample of size N. Unfortunately, the obvious alternative to replace the term $\zeta_1(K, N)$, which occurs both in the numerator and in the denominator of the multiplicative penalty of FPE_{sub}, by the sum $\zeta_1(1, N) + \cdots + \zeta_1(1, N - K + 1)$ (Reschenhofer et al. 2012) has the disadvantage that the denominator can become very small even if K is not very large. We therefore propose a new criterion, FPE_Δ, which is more stable and still avoids overfitting in the presence of some dominant regressors. Its additive penalty terms are given recursively by

$$P^a_\Delta(K) = P^a_\Delta(K-1) + n \log\left(\frac{n + (K-1) + \zeta_1(1, N-(K-1))}{n - (K-1) - \zeta_1(1, N-(K-1))}\right) - n \log\left(\frac{n + (K-1)}{n - (K-1)}\right). \tag{14}$$

In the practically more relevant case, where the first regressor is certain to be included, we have

$$P^a_\Delta(1) = n \log\left(\frac{n+1}{n-1}\right).$$

Otherwise, if all regressors are treated equally, the first penalty would be

$$n \log\left(\frac{n + \zeta_1(1, N)}{n - \zeta_1(1, N)}\right).$$

In contrast to conventional criteria like AIC and BIC, FPE_Δ penalizes the first regressors to be included more than the later ones. However, the decline is insignificant as long as K is small compared to n (see Fig. 1). Using the approximation

$$\hat{\zeta}_1(1, N) = 2\log(N) - \log(\log(N))$$

Reschenhofer (2004) we arrive at the simpler criterion FPE_δ, whose additive penalty terms are given by elementary functions, i.e.,

$$P^a_\delta(K) = P^a_\delta(K-1) + n \log\left(\frac{n + (K-1) + \hat{\zeta}_1(1, N-(K-1))}{n - (K-1) - \hat{\zeta}_1(1, N-(K-1))}\right) - n \log\left(\frac{n + (K-1)}{n - (K-1)}\right). \tag{15}$$

2.2 Relaxing the orthogonality assumption

Even in the case of a pure random sample, there is a good chance of observing a cluster of very small or very large values. A certain reduction in the residual sum of

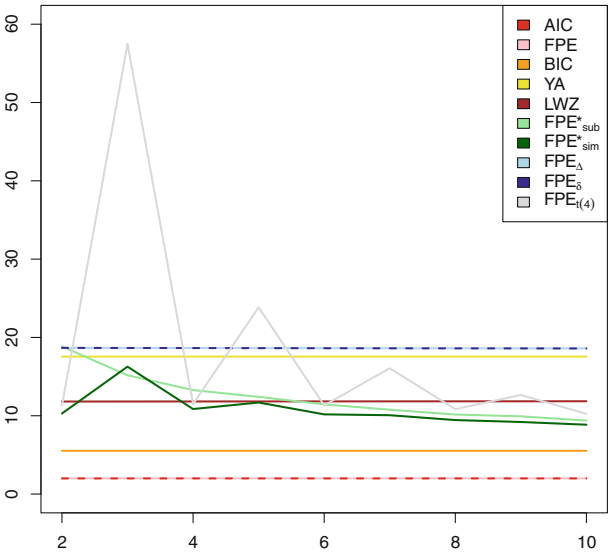

Fig. 1 Increments $P^a(K) - P^a(K-1)$, $K = 2, \ldots, 10$ of the additive penalties of the model selection criteria AIC (*red*), FPE (*pink*), BIC (*orange*), YA (*yellow*), LWZ (*brown*), FPE^*_{sub} (*lightgreen*), FPE^*_{sim} (*darkgreen*), FPE_Δ (*lightblue*), FPE_δ (*darkblue*), $FPE_{t(4)}$ (*lightgray*) (color figure online)

squares may then be achieved by introducing a structural break either at the start or at the end of this cluster. Clearly, the introduction of a second break (at the other end) will in general have a much larger effect. Thus, the second break should be penalized much harsher than the first. Similarly, the fourth breakpoint should be penalized more than the third (two clusters of extreme values), the sixth more than the fifth (three clusters), and so on. Following this line of reasoning, we end up with a completely new concept of penalizing, i.e., non-monotonic penalizing, which differs crucially from the conventional approach of using either constant increments in the penalties (e.g., AIC and BIC) or declining increments (e.g., MRIC, and FPE_{sub}). A precise description of this non-monotonicity has been obtained with the help of simulations. The details are given subsequently.

In the following, we will always include the first column of $\mathbf{X}$ in the submatrix $\mathbf{X_S}$. Accordingly, we must use a slightly modified version of FPE_{sub}, i.e.,

$$FPE^*_{sub}(K) = RSS(K)\frac{n+1+\zeta_1(K-1,N-1)}{n-1-\zeta_1(K-1,N-1)}, \tag{16}$$

as benchmark for our simulation study. In this study, $r = 100{,}000$ random samples $y(i)$ and $z(i)$ of size $n = 250$ from a standard normal distribution are generated and the ideal multiplicative penalties are approximated by the ratios

$$\frac{\sum_{i=1}^{r} \left\| z(i) - \hat{y}_K(i) \right\|^2}{\sum_{i=1}^{r} \left\| y(i) - \hat{y}_K(i) \right\|^2}$$

or, computationally more efficiently, just by

$$R(K) = \frac{\sum_{i=1}^{r} \left(n + \left\| \hat{y}_K(i) \right\|^2 \right)}{\sum_{i=1}^{r} \left\| y(i) - \hat{y}_K(i) \right\|^2}$$

(for all computations and graphics we use the free software environment R; see R Development Core Team 2011). The associated model selection criterion is given by

$$FPE^*_{sim}(K) = RSS(K)R(K). \tag{17}$$

Figure 1 compares the penalties of FPE^*_{sub} and FPE^*_{sim}. More precisely, the increments

$$P^a(K) - P^a(K-1)$$

of the corresponding additive penalties P^a are displayed.

In our case, the discrepancies between FPE^*_{sub} and FPE^*_{sim} are much larger than those observed for macroeconomic data by Reschenhofer et al. (2012). Most striking is the non-monotonicity of the FPE^*_{sim} increments. The peaks at $K = 3$ and $K = 5$ can be explained by the fact that we need in general three regimes for the description of one extreme observation (or one cluster of extreme observations), five regimes for two extreme observations, etc. This pattern becomes more apparent when the likelihood of extreme values is increased, e.g., by using the t-distribution with 4 degrees of freedom instead of the normal distribution to generate the random samples. The resulting criterion is called $FPE_{t(4)}$.

In contrast to conventional applications, where the set of candidate variables is different every time, our design matrix $\mathbf{X}$ always has the same form. It is therefore worthwhile to calculate our penalties for a large number of sample sizes and put the results into tables, which allow using our criteria without having to re-simulate. Tables 2 and 3 contain the non-monotonic increments in the penalties of the criteria FPE^*_{sim} and $FPE_{t(4)}$ for $n = 20, 30, \ldots, 250$. Overall penalties can be obtained by calculating cumulative sums starting at 2.0 ($K = 1$: one certain regressor).

3 Empirical results

To illustrate the performance of our new non-monotonic (FPE^*_{sim} and $FPE_{t(4)}$) and monotonic (FPE_δ) criteria we analyze the measurements of the annual flow of the river Nile at Aswan 1871–1970 (see, e.g., Cobb 1978; Zeileis et al. 2003). It is widely accepted that, due to the construction of the Aswan dam, there is a single structural break in 1898 (see Fig. 2). However, without further provisions AIC and BIC select a large number of additional breaks. Taking ad-hoc measures like imposing the restriction of a minimal regime length or treating the break dates as ordinary regression parameters is not a satisfactory solution. Firstly, there may well be applications where even individual observations must be modeled as separate regimes. Secondly, using asymptotic arguments it can be shown that break dates require a different treatment

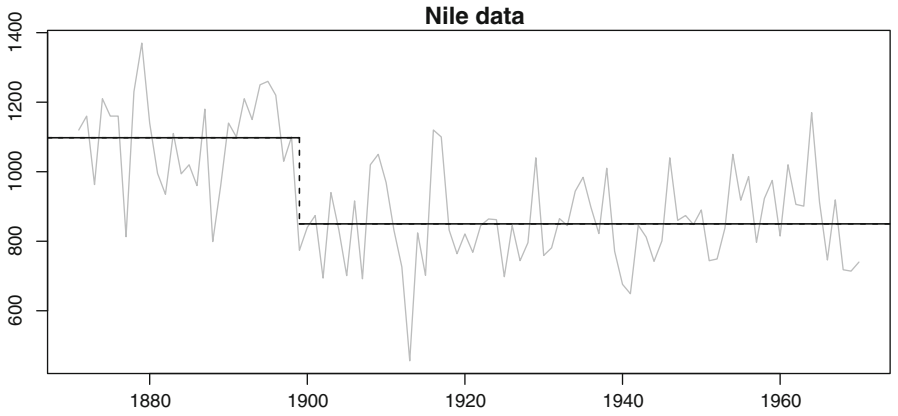

Fig. 2 Single break fit to the Nile data

Table 1 Proportion of single-break detection (Correct), mean number of selected breaks (Mean) and standard deviation (SD) of selected breaks for 100,000 samples drawn from the Nile data

	AIC	BIC	FPE^*_{sim}	$FPE_{t(4)}$	FPE_δ	YA	LWZ
Correct	0.000	0.000	0.983	1.000	0.998	0.892	0.841
Mean	9.000	7.672	1.031	1.000	1.002	1.133	1.229
SD	0.012	1.948	0.323	0.000	0.047	0.433	0.635

than ordinary regression parameters (see Ninomiya 2005). Instead of just mending improper methods it seems therefore much better to try different approaches. Indeed, YA, LWZ, and all of our new criteria select only one break. Of course, this comparison is not conclusive because it is based on only one sample. In order to get a more comprehensive assessment, we proceed as follows: (i) Calculate the means $\bar{x}_1$ and $\bar{x}_2$ for the two regimes 1871–1898 and 1899–1970 and the corresponding residuals $\hat{u}_1, \ldots, \hat{u}_{28}$ and $\hat{v}_1, \ldots, \hat{v}_{72}$. (ii) Draw $\hat{u}^*_1, \ldots, \hat{u}^*_{28}$ and $\hat{v}^*_1, \ldots, \hat{v}^*_{72}$ from the empirical distribution of the residuals in the first and second regime, respectively. (iii) Generate a sample

$$(\bar{x}_1 + \hat{u}^*_1, \ldots, \bar{x}_1 + \hat{u}^*_{28}, \bar{x}_2 + \hat{v}^*_1, \ldots, \bar{x}_2 + \hat{v}^*_{72}). \tag{18}$$

(iv) Estimate the number of structural breaks using AIC, BIC, FPE^*_{sim}, $FPE_{t(4)}$, FPE_δ, YA and LWZ with a maximum of nine possible breaks.

The results obtained by repeating steps (ii)–(iv) 100,000 times are summarized in Table 1. The conventional criteria AIC and BIC come off badly. They are clearly unsuitable for the detection of breaks. YA and LWZ perform much better but are clearly outperformed by our new criteria. Although we might, in view of the obvious non-normality of the data, have expected that the non-monotonic criterion $FPE_{t(4)}$, which has been designed for the case of fat-tailed distributions, is particularly accurate, its excellent performance is still striking. It selects the correct model in practically every single case.

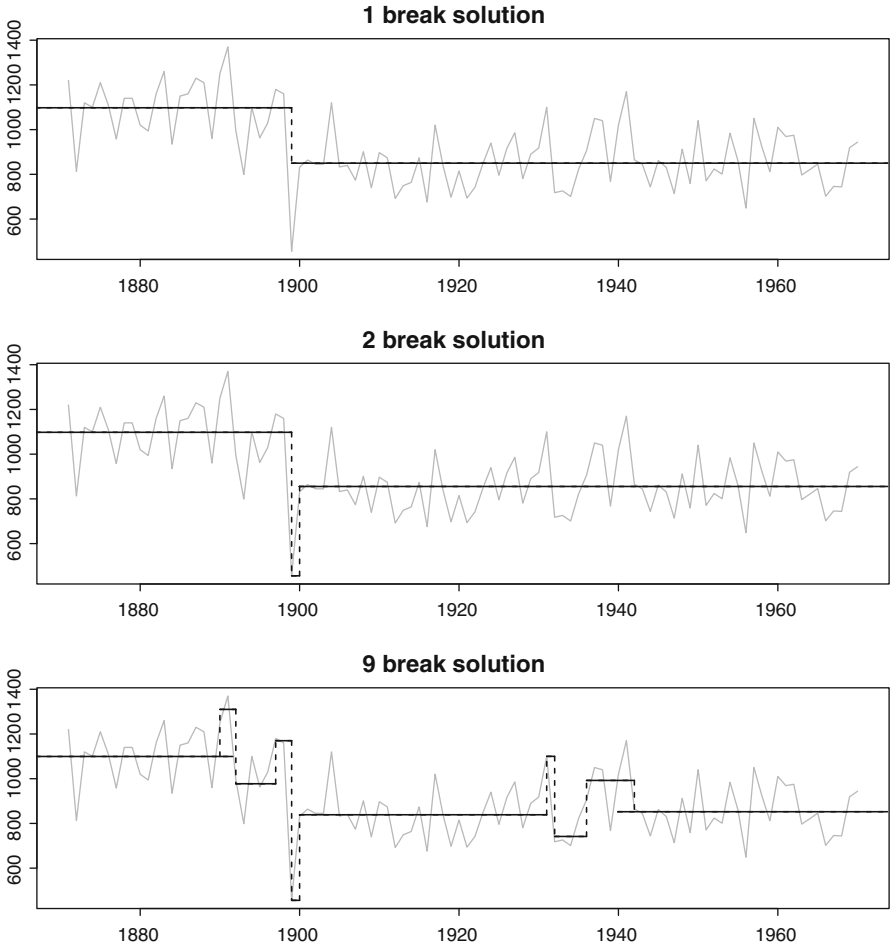

Fig. 3 One break (selected by FPE^*_{sim}, $FPE_{t\,(4)}$ and FPE_δ), two break (selected by YA and LWZ) and nine break (selected by AIC and BIC) solution for one specific sample drawn from the Nile data

To illustrate the possible effect of extreme observations, we present one specific synthetic sample generated as outlined above. In this specific sample, a very small observation occurs just after the break (see Fig. 3). In the best solution with two breaks, the second regime consists only of this single observation. This wrong solution is chosen by both YA and LWZ. As usual, AIC and BIC select the maximum number of breaks. The plot of the solution selected by AIC and BIC (see Fig. 3) shows that these criteria poorly distinguish signal from noise.

4 Discussion

In the light of Kempthorne's (1984) finding that all post-model-selection estimators are admissible, Kabaila's (2002) criticism that Shibata's (1980, 1981) asymptotic optimality results for the AIC hold only pointwise and may therefore be misleading,

Table 2 Increments $P^a(K) - P^a(K-1)$, $K = 2, \ldots, 10$, of the additive penalties of the criterion FPE^*_{sim} for different sample sizes n

n	K								
	2	3	4	5	6	7	8	9	10
20	7.2	8.0	6.2	6.0	5.7	5.7	5.7	5.9	6.2
30	7.8	9.1	6.8	6.6	6.1	5.9	5.7	5.7	5.7
40	8.2	10.0	7.3	7.1	6.4	6.2	6.0	5.8	5.7
50	8.5	10.7	7.7	7.5	6.8	6.5	6.2	6.1	5.9
60	8.7	11.3	8.0	7.9	7.1	6.9	6.5	6.3	6.1
70	8.9	11.8	8.3	8.3	7.4	7.1	6.8	6.5	6.3
80	9.0	12.3	8.6	8.6	7.7	7.4	7.0	6.8	6.5
90	9.2	12.7	8.9	8.9	8.0	7.7	7.2	7.0	6.7
100	9.3	13.0	9.1	9.2	8.2	7.9	7.4	7.2	6.9
110	9.4	13.4	9.3	9.4	8.4	8.1	7.6	7.4	7.1
120	9.5	13.7	9.4	9.7	8.5	8.3	7.8	7.6	7.3
130	9.6	13.9	9.6	9.9	8.7	8.5	8.0	7.7	7.4
140	9.7	14.2	9.7	10.1	8.9	8.6	8.2	7.9	7.6
150	9.8	14.5	9.9	10.3	9.0	8.8	8.3	8.0	7.7
160	9.8	14.7	10.0	10.4	9.2	9.0	8.4	8.2	7.9
170	9.9	14.9	10.1	10.6	9.3	9.1	8.6	8.3	8.0
180	10.0	15.1	10.2	10.8	9.4	9.3	8.7	8.4	8.1
190	10.0	15.3	10.3	10.9	9.6	9.4	8.8	8.6	8.2
200	10.1	15.5	10.4	11.1	9.7	9.5	8.9	8.7	8.4
210	10.1	15.7	10.5	11.2	9.8	9.6	9.1	8.8	8.5
220	10.2	15.8	10.6	11.3	9.9	9.7	9.2	8.9	8.6
230	10.2	16.0	10.7	11.5	10.0	9.9	9.3	9.0	8.7
240	10.3	16.1	10.8	11.6	10.1	10.0	9.4	9.1	8.8
250	10.3	16.3	10.9	11.7	10.2	10.1	9.4	9.2	8.9

Each value is based on 100,000 random samples of size n from a normal distribution

and Yang's (2005, 2006) conclusion that no model selection criterion can share the main strengths of AIC (pointwise asymptotic optimality in nonparametric scenarios) and BIC (consistency in parametric scenarios) simultaneously, we cannot hope to find anything like a universally best model selection criterion. However, we could try to find a method for assessing whether AIC or BIC is more appropriate for a specific dataset. Unfortunately, such a method will typically depend on tuning parameters (Liu and Yang 2011). Moreover, more than one assessment method might be proposed. So instead of just selecting the model selection criterion we would have to select the assessment method first which we could then use to select the model selection criterion. On top of that, it is not a priori clear why we should content ourselves with only AIC and BIC. These criteria have been designed for non-nested models. None of the two takes the total number of candidate models into account. In general, even the BIC-penalties are much too small to ensure consistency for selecting among non-nested

Table 3 Increments $P^a(K) - P^a(K-1)$, $K = 2, \ldots, 10$, of the additive penalties of the criterion $FPE_{t(4)}$ for different sample sizes n

n	K								
	2	3	4	5	6	7	8	9	10
20	7.7	13.0	6.3	6.7	5.8	5.9	5.8	6.0	6.3
30	8.4	16.4	6.9	7.7	6.3	6.2	5.9	5.8	5.8
40	8.9	19.8	7.5	8.8	6.7	6.8	6.2	6.1	5.9
50	9.2	22.1	8.0	9.8	7.2	7.3	6.5	6.4	6.1
60	9.5	24.7	8.3	10.7	7.6	7.8	6.9	6.8	6.4
70	9.7	27.0	8.7	11.6	7.9	8.4	7.2	7.1	6.7
80	9.9	29.5	9.0	12.4	8.3	8.9	7.5	7.5	7.0
90	10.0	31.6	9.2	13.3	8.5	9.4	7.8	7.9	7.2
100	10.2	33.7	9.5	14.1	8.8	9.9	8.1	8.2	7.5
110	10.3	35.2	9.7	14.8	9.1	10.4	8.4	8.5	7.7
120	10.5	37.3	9.9	15.6	9.3	10.8	8.6	8.9	7.9
130	10.5	39.3	10.0	16.3	9.5	11.3	8.8	9.2	8.2
140	10.7	41.2	10.2	17.0	9.7	11.7	9.0	9.5	8.4
150	10.7	42.7	10.4	17.7	9.9	12.1	9.3	9.8	8.6
160	10.9	44.5	10.5	18.4	10.1	12.6	9.4	10.1	8.8
170	10.9	46.3	10.6	19.1	10.3	13.0	9.6	10.4	9.0
180	10.9	47.2	10.8	19.7	10.4	13.4	9.8	10.7	9.2
190	11.0	48.6	10.9	20.4	10.5	13.8	10.0	11.0	9.3
200	11.0	50.8	11.0	20.9	10.7	14.2	10.2	11.3	9.5
210	11.1	53.1	11.1	21.5	10.8	14.6	10.3	11.6	9.7
220	11.2	53.4	11.2	22.2	11.0	15.0	10.4	11.9	9.8
230	11.3	54.9	11.3	22.8	11.1	15.3	10.6	12.1	10.0
240	11.3	55.8	11.4	23.4	11.2	15.7	10.7	12.4	10.1
250	11.3	57.5	11.5	23.9	11.3	16.1	10.8	12.7	10.2

Each value is based on 100,000 random samples of size n from a t distribution with 4 degrees of freedom

models (Sin and White 1996; Hong and Preston 2012). It might therefore make more sense to tailor different model selection criteria to the different types of applications rather than to look for some universal criterion. Our paper focuses on a very specific application, namely the detection of structural breaks in a simple regression model.

The variable selection criteria YA and LWZ have been put to use for the detection of structural breaks in financial time series (see, e.g., Zeileis et al. 2010) because of the inadequate performance of AIC (see, e.g., Perron 1997) and BIC (see, e.g., Liu et al. 1997; Zeileis et al. 2003). They differ from these standard criteria only in the size of their penalties. They just penalize harder but still treat all breaks in the same way. There is a simple linear relationship between the size of the penalty and the number of breaks. In contrast, all of the criteria proposed in this paper are genuine subset selection criteria and take therefore also the much larger number of potential breakpoints into account. Our new criteria are obtained from existing subset selection criteria by relaxing two simplifying assumptions, that of orthogonal regressors and

that of vanishing regression coefficients. The relaxation of the second assumption yields a criterion which outperforms YA and LWZ in our empirical study. Moreover, unlike YA and LWZ, it does not depend on the specification of tuning parameters. Finally, the relaxation of the first assumption yields criteria with penalties that exhibit a specific type of non-monotonicity. This non-monotonicity is positively related to the likelihood of extreme values and negatively to serial dependencies such as conditional heteroscedasticity. Such a pattern appears striking at first sight, but much less so on closer inspection. Clearly, we cannot expect that a cluster of unusual observations occurs just at the begin or at the end of the observation period. More likely, it will occur somewhere in the middle and therefore require two breaks (three regimes) for its description. Consequently, the penalty for the second break should be higher than that for the first. An analogous argument holds for the case of 2, 3, ... clusters of unusual observations, where 5, 7, ... regimes will be needed.

In summary, we recommend to use the non-monotonic criteria FPE^*_{sim} and $FPE_{t(4)}$ (see Tables 2 and 3) in situations where it is a priori not clear whether there are any structural breaks at all. FPE^*_{sim} is more appropriate in the Gaussian case and $FPE_{t(4)}$ in the case of fat-tailed distributions (e.g., in financial applications). In situations where it is clear that some major breaks exist and the only question is whether there are also some minor breaks, the criterion FPE_δ should be used.

Acknowledgments The authors very much appreciate the referee's comments, which helped to substantially improve this paper.

References

Akaike H (1969) Fitting autoregressive models for prediction. Ann Inst Stat Math 21:243–247

Akaike H (1973) Information theory and an extension of the maximum likeliihood principle. In: Petrov BN, Csaki F (eds) Second international symposium on information theory. Akademia Kiado, Budapest, pp 267–281

Bai J, Perron P (1998) Estimating and testing linear models with multiple structural changes. Econometrica 66:47–78

Bai J, Perron P (2003) Computation and analysis of multiple structural change models. J Appl Econom 18:1–22

Bai J, Perron P (2006) Multiple structural change models: a simulation analysis. In: Corbea D, Durlauf S, Hansen BE (eds) Econometric theory and practice: frontiers of analysis and applied research. Cambridge University Press, Cambridge, pp 212–237

Cobb GW (1978) The problem of the nile: conditional solution to a changepoint problem. Biometrika 65:243–251

Foster DP, George EI (1994) The risk inflation criterion for multiple regression. Ann Stat 22:1947–1975

George EI, Foster DP (2000) Calibration and empirical Bayes variable selection. Biometrika 87:731–747

Hong H, Preston B (2012) Bayesian averaging, prediction and nonnested model selection. J Econom 167:358–369

Kabaila P (2002) On variable selection in linear regression. Econom Theory 18:913–925

Kass RE, Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J Am Stat Assoc 90:928–934

Kempthorne PJ (1984) Admissible variable-selection procedures when fitting regression models by least squares for prediction. Biometrika 71:593–597

Kuehn C (2001) An estimator of the number of change points based on a weak invariance principle. Stat Probab Lett 51:189–196

Liu J, Wu S, Zidek JV (1997) On segmented multivariate regression. Statistica Sinica 7:497–526

Liu W, Yang Y (2011) Parametric or nonparametric? A parametericness index for model selection. Ann Stat 39:2074–2102

Ninomiya Y (2005) Information criterion for Gaussian change-point model. Stat Probab Lett 72:237–247
Perron P (1997) L'estimation de modeles avec changements structurels multiples. Actual Econ 73:457–505
R Development Core Team (2011) R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, ISBN 3-900051-07-0, http://www.R-project.org/
Reschenhofer E (1996) Approximating the Bayes factor. Stat Probab Lett 30:241–245
Reschenhofer E (2004) On subset selection and beyond. Adv Appl Stat 4:265–286
Reschenhofer E, Schilde M, Oberecker E, Payr E, Tandogan HT, Wakolbinger LM (2012) Identifying the determinants of foreign direct investment: a data-specic model selection approach. Stat Pap 53:739–752
Rothman D (1968) Letter to the editor. Technometrics 10:432
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Shibata R (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Ann Stat 8:147–164
Shibata R (1981) An optimal selection of regression variables. Biometrika 68:45–54
Sin C-Y, White H (1996) Information criteria for selecting possibly misspecified parametric models. J Econom 71:207–225
Yang Y (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. Biometrika 92:937–950
Yang Y (2006) Prediction/estimation with simple linear models: is it really that simple? Econom Theory 23:1–36
Yao Y-C (1988) Estimating the number of change-points via Schwarz' criterion. Stat Probab Lett 6:181–189
Yao Y-C, Au ST (1989) Least-squares estimation of a step function. Sankhya Indian J Stat Ser A 51:370–381
Zeileis A, Kleiber C, Krämer W, Hornik K (2003) Testing and dating of structural changes in practice. Comput Stat Data Anal 44:109–123
Zeileis A, Shahb A, Patnaik I (2010) Testing, monitoring, and dating structural changes in exchange rate rregimes. Comput Stat Data Anal 54:1696–1706