Discussion Paper Series

# Estimation and Inference in Unstable Nonlinear Least Squares Models

By

**Otilia Boldea and Alastair R. Hall**

Centre for Growth and Business Cycle Research, Economic Studies,
University of Manchester, Manchester, M13 9PL, UK

October 2009
Number 126

# Estimation and Inference in Unstable Nonlinear Least Squares Models[*]

Otilia Boldea[†] and Alastair R. Hall[‡]

October 5, 2009

[*] We are grateful to Mehmet Caner, Manfred Deistler, John Einmahl, Atsushi Inoue and Denise Osborn for their comments, as well as for the comments of participants at the presentation of this paper at the *Conference on Breaks and Persistence in Econometrics*, London, UK, December 2006, *Inference and Tests in Econometrics*, Marseille, France, April 2008, *European Meetings of the Econometric Society*, Milan, Italy, August 2008, *NBER-NSF Time Series Conference*, Aarhus, Denmark, September 2008, *Fed St. Louis Applied Econometrics and Forecasting in Macroeconomics Workshop*, St. Louis, October 2008, and at the seminars in Tilburg University, University of Manchester, University of Exeter, University of Cambridge, University of Southampton, Tinbergen Institute, UC Davis and Institute for Advanced Studies, Vienna. The second author acknowledges the support of ESRC grant RES-062-23-1351.

[†] Corresponding author. Tilburg University, Dept. of Econometrics and Operations Research, Warandelaan 2, 5000 LE Tilburg, Netherlands, Email: O.Boldea@uvt.nl

[‡] University of Manchester, Email: Alastair.Hall@ manchester.ac.uk

## Abstract

In this paper, we extend Bai and Perron's (1998, Econometrica, pp. 47-78) method for detecting multiple breaks to nonlinear models. To that end, we consider a nonlinear model that can be estimated via nonlinear least squares (NLS) and features a limited number of parameter shifts occurring at unknown dates. In our framework, the break-dates are estimated simultaneously with the parameters via minimization of the residual sum of squares. Allowing for breaks in the marginal distribution of regressors and errors, and using new empirical process theory results, we derive the asymptotic distributions of both break-point and parameter estimates and propose several instability tests. Simulations show good finite sample properties of our procedure.

*JEL classification:* C12, C13, C22

*Keywords:* Multiple Change Points, Nonlinear Least Squares

# 1 Introduction

As pointed out by Lucas (1976), policy shifts and time-varying market conditions induce behavioral changes in the decisions of economic agents. Hence, over longer time spans, a stable model might not be the appropriate tool to capture the features of economic decisions. A popular way to capture instability in macroeconometric models is to impose sudden parameter shifts or break-points at unknown dates.

Both the econometric and statistical literature on break-point problems is extensive[1], and its main focus is on testing for breaks rather than estimation. For example, early work by Quandt (1960) suggests using a supremum (*sup*) type test for inference on a single unknown break-point. Whether in linear or nonlinear settings, most subsequent work - see *inter alia* Anderson and Mizon (1983), Andrews and Fair (1988), Ghysels and Hall (1990), Andrews (1993), Sowell (1996), Hall and Sen (1999) and Andrews (2003) - proposes tests that are designed against the alternative of a one-time parameter variation or of more general model misspecification. For parametric settings, Bai and Perron (1998) is among the few papers that proposes tests for identifying multiple breaks. Their tests are designed for linear models estimated via ordinary least-squares (OLS). While these tests are useful, the linear framework might be considered a limitation. In practice, researchers often argue that it can be difficult to discriminate between misspecification due to parameter instability or neglected nonlinearity. It is therefore desirable to develop a framework that allows both features. While tests such as the ones developed in Eitrheim and Teräsvirta (1996) can detect instability in some classes of nonlinear models, they are not particularly designed against an alternative with breaks nor offer an estimation framework that can allow for both smooth and sudden change. One of the aims of this paper is to provide change-point tests in the spirit of Bai

[1] For statistical literature surveys, see Zacks (1983), Krishnaiah and Miao (1988), Bhattacharya (1994), Csörgö and Horváth (1997); for recent developments in econometrics, see Dufour and Ghysels (1996) and Banerjee and Urga (2005).

and Perron's (1998) tests, but with a maintained nonlinearity assumption. These tests are valid for a large class of parametric nonlinear models, including *inter alia* smooth transition models, neural networks, partially linear models.

Compared to inference procedures, the issue of consistently estimating one or multiple change-points - when their location is unknown - has received considerably less attention in the literature. Within linear parametric models, there are a few methods that yield consistent estimates of the break-points, e.g. maximum likelihood - Quandt (1958), least-squares - Bai (1994), least absolute deviation - Bai (1995), minimum description length - Davis, Lee, and Rodriguez-Yam (2006). In Bai and Perron's (1998) paper, the break points are estimated simultaneously with the regression parameters via least-squares methods. Bai and Perron (1998) establish consistency and derive the convergence rate of the resulting break point fractions under fairly general assumptions. They also propose a sequential procedure for selecting the number of break points in the sample based on various tests for parameter constancy. This procedure is extended to models with cross-regime restrictions by Perron and Qu (2006), and to multivariate frameworks by Qu and Perron (2007). Hall, Han, and Boldea (2009) further extend Bai and Perron's framework to linear models with endogenous regressors. A slightly different approach is proposed by Davis, Lee, and Rodriguez-Yam (2006); they suggest estimating the number and location of breaks not separately, but simultaneously via minimization of the minimum description length (MDL) criterion of Rissanen (1989).

While useful, all analyses above are restricted to linear models. Nevertheless, several economic models exhibit asymmetries that are usually quantified as nonlinearities or breaks, but rarely both. For example, threshold models are largely viewed today as a special case of smooth transition models, when the smoothness parameter of the transition function approaches infinity. Similarly, change-points are viewed as a special case of smooth transition models with the state variable time and the smoothness parameter approaching infinity, see e.g. van Dijk, Teräsvirta, and Franses (2002). However, such a treatment is not desirable, since it is difficult to develop estimation and inference theory in the presence of parameters approaching infinity; even if these parameters are not the main object of inference, it is likely that their estimation will affect the estimation of other parameters of interest. While this discussion highlights the importance of distinguishing between breaks and time transitions with smoothness parameters close to infinity, it does not preclude the treatment of smooth and sudden change jointly. Since our assumptions about the form of the nonlinearity encompass smooth transition models, our results provide a framework for inference that allows modeling breaks and smooth transitions jointly, without treating them as mutually exclusive.

To that end, in this paper we consider a univariate nonlinear model that can be estimated via NLS - or under stronger assumptions, equivalent methods such as quasi-maximum likelihood - and exhibits multiple unknown breaks. Allowing for a limited number of breaks in the marginal distribution of the regressors and/or errors, we show that a minimization of the sum of squared residuals over all possible break dates and parameters yields consistent estimates of both the unknown break fractions and parameters. We further prove $T$-rate convergence of break fraction estimates, a key result because it implies that inference on parameters can be conducted as if the break-points were known *a priori*. To obtain this result, we make use of two new empirical process theory results: one in the context of piecewise strictly stationary processes, extending a result in Caner (2007), and another for more general empirical processes when the errors are independent. We also discuss sufficient but higher level assumptions for more general processes in the errors and the regressors.

Based on the above, we provide various structural stability tests - in the presence or absence of autocorrelation - that naturally generalize those proposed by Bai and Perron (1998). We consider global tests for zero versus a fixed number of breakpoints, known or unknown, as well as sequential tests for an additional break. These tests can be used to develop a sequential method for finding the number and locations of breaks, as suggested by Bai and Perron (1998) in linear settings. Moreover, the sequential Wald test we propose - similar to Hall, Han, and Boldea (2009) - allows for marginal breaks in the distributions of regressors, at the same time extending the strategy of identifying the number of breaks to settings where autocorrelation is present.

For forecasting purposes, it is still of interest to know with certain confidence when the last break occurred. As Bai (1994, 1995, 1997) shows, change-point distributions in linear models can be derived in two cases: when the magnitude of parameter shifts is constant and when it shrinks to zero at a certain rate. Because in the first case, the confidence intervals depend on the distribution of the data, the device of shrinking shifts is used to ensure that shifts disappear at a slow enough rate so that pivotal statistics can still be obtained. In practice, this framework can be viewed as one of moderate shifts, according to Bai and Perron (1998). A local analysis of small shifts is presented in Elliott and Müller (2007) for linear models, but providing a similar framework here is beyond the scope of our paper.

We consider each of the two cases above in turn. For the first case, we provide an asymptotic approximation to the exact change-point distribution, but this approximation is - as for linear cases its exact counterpart - dependent on the distribution of the data. For the second case, we obtain a similar asymptotic distribution as in Bai (1997). We validate the usefulness of our estimators, tests and confidence intervals via simulations.

The paper is organized as follows: Section 2 describes our model. Section 3 reveals the assumptions needed for our estimation method. We outline the consistency and limiting distributions results in Section 4. Section 5 rederives - in a nonlinear context - two classes of stability tests. Section 6 shows good finite properties of our break-point estimators, tests and number of break-points. Section 7 concludes. Sketch proofs are relegated to the Appendix, while the detailed proofs can be found in a Supplemental Appendix that is available from the authors upon request.

# 2 Model

In this section, we introduce a nonlinear model with breaks. Consider a univariate nonlinear data generation process with $m$ unknown change-points:

$$y_t = f(x_t, \theta^0_{i+1}) + u_t \qquad t \in I^0_i = [T^0_i + 1, T^0_{i+1}] \qquad i = 0, 1, \ldots m \tag{1}$$

where $T^0_0 = 0$ and $T^0_{m+1} = T$ by convention. Here $y_t$ is the dependent variable, $x_t$ $(s \times 1)$ are the regressors, $\theta^0_{i+1}$ $(p \times 1)$ are parameters that change at dates $T^0_i$, $f : \mathbb{R}^s \times \Theta \to \mathbb{R}$ is a known measurable function on $\mathbb{R}$ for each $\theta \in \Theta$, and $T$ is the sample size. To begin, we consider $m$ to be a known finite positive integer, but we allow for the break dates to be unknown to the researcher; we consider the question of how to estimate $m$ in Section 6. For simplicity, let $f_t(\theta) = f(x_t, \theta)$ and denote by $\bar{T}^m \equiv (T_0 = 1, T_1, \ldots, T_m, T_{m+1} = T)$ any $m$-partition of the interval $[1, T]$. To further simplify the notation, we will stack column vectors such as $\theta^0_{i+1}$ and $\theta_{i+1}$ into two corresponding $(m + 1)p \times 1$ vectors, $\theta^c_0$ and $\theta^c$. For a given sample partition and given parameter values $\theta^c$, denote by $S_T(\bar{T}^m, \theta^c)$ the sum of squares.[2]

One of our main goals is to provide a method for estimating the unknown parameters and change points. As in Bai and Perron (1998), the estimation method we propose is based on the least-squares principle[3] and follows in two steps. First, we obtain the sub-sample NLS estimators for each partition:

$$\hat{\theta}^c_T(\bar{T}^m) = \underset{\theta^c(\bar{T}^m)}{\operatorname{argmin}} \; S_T(\, \bar{T}^m, \theta^c(\bar{T}^m)\,) \tag{2}$$

Second, we search over all possible partitions to obtain the break-point estimates.

[2] We use superscript $c$ to distinguish between $(m + 1)p \times 1$ parameter vectors and the $p \times 1$ parameter vectors at which $f_t(\cdot)$ is evaluated.

[3] Note that an extension to more general settings such as generalized method of moments (GMM) is non-trivial because minimizing a GMM criterion over all possible partitions does not yield consistent estimates of the break-fractions indexing the break-points even for linear models and one break under reasonable conditions, see Hall, Han, and Boldea (2009).

The estimates $\hat{T} = (1, \hat{T}_1, \ldots, \hat{T}_m, T)$ for change-points and $\hat{\theta}^c_T = (\hat{\theta}_1, \ldots, \hat{\theta}_{m+1})$ for parameters are obtained as follows:

$$\hat{T} = \underset{\bar{T}^m}{\operatorname{argmin}}\, S_T\ (\ \bar{T}^m, \hat{\theta}^c_T(\bar{T}^m)\ )\ \text{ and } \hat{\theta}^c_T = \hat{\theta}^c_T\ (\hat{T}) \tag{3}$$

The above is an NLS estimation with an appropriate modification to allow for multiple break-points, and can be legitimately performed provided that $E[u_t f_t(\theta^0_{i+1})] = 0$ for each $t = T^0_i + 1, \ldots, T^0_{i+1}$ $(i = 0, 1, \ldots m)$.

# 3 Assumptions

To derive the statistical properties of our estimators, we establish a framework that combines elements of asymptotic theory in stable nonlinear models and unstable linear models. As pointed out by Hansen (2000), the marginal distributions of regressors and/or errors may change, possibly at different locations in the sample than the population parameters of the equation of interest. Our framework is designed to achieve as much generality as possible with respect to changes in marginal distributions.[4] In dealing with nonlinear asymptotics, we impose usual smoothness and boundedness assumptions that closely follow e.g. White and Domowitz (1984) and Gallant and White (1988), as well as a new assumption that originates from arguments of uniform convergence in $\theta \times r$ we need to consider. To deal with instability, we adapt some assumptions from Bai and Perron (1998).

**Assumption 1.** *Let $v_t = (x'_t, u_t)'$. Then either (i) $\{v_t\}$ is a piece-wise strictly stationary process on intervals $[T^*_{j-1}+1, T^*_j], j = (1, \ldots, m^*+1), T_0 = 0; T_{m^*+1} = T$ for a fixed $m^*$; $T^*_j = [T\lambda^*_j]$, where $0 < \lambda^*_1 < \ldots < \lambda^*_{m^*} < 1$; $\{v_t\}$ is also an $\alpha$-mixing process of size $-2s/(s-2)$, where $s > 2$, and the errors are uncorrelated*

[4] Allowing for these types of changes is important in many settings. For example, when estimating a possibly asymmetric (nonlinear) interest rate reaction function, regressors such as output gap or inflation gap may exhibit changes in variance, due to a period of Great Moderation - see e.g. Stock and Watson (2002) - and these changes may occur at different locations than those in the parameters of the equation of interest.

*with the regression function, i.e.* $E\left[u_t f_t(\theta)\right] = 0$ *for all* $\theta, t$ *or (ii) The errors* $\{u_t\}$ *are independent of each other and of the regressors, i.e.* $u_t \perp x_k, u_s$ *for all* $k$ *and* $t \neq s$*;* $\{x_t\}$ *is a near-epoch dependent (n.e.d.) sequence of size* $-1$ *on an underlying* $\alpha$*-mixing process of size* $-2s/(s-2)$*, with* $s > 2$*.*

**Assumption 2.** *The function* $f_t(\cdot)$ *is a known measurable function, twice continuously differentiable in* $\theta$ *for each* $t$*.*

**Assumption 3.** *Let* $F_t(\theta) = \partial f_t(\theta)/\partial\theta$*,* $p \times 1$ *vector and* $f_t^{(2)}(\theta)$*, a* $p \times p$ *matrix with* $(i,j)^{th}$ *element* $f_{t,i,j}^{(2)} = \partial^2 f_t(\theta)/(\partial\theta_i \partial\theta_j')$*. Also denote by* $\|\cdot\|$ *the Euclidean norm. Then (i) the common parameter space* $\Theta$ *is a compact subset of* $\mathbb{R}^p$*; (ii)* $E[\sup_{t \leq n_0, \theta} |u_t f_t(\theta)|] < \infty$*, for some fixed* $n_0$*, and* $\sup_{t,\theta} E|u_t f_t(\theta)|^{2s} < \infty$*; (iii)* $\sup_{t,\theta} E\|u_t F_t(\theta)\|^{2s} < \infty$*; (iv) For* $i, j = 1, \ldots p$*,* $\sup_{t,\theta} E\|u_t f_{t,i,j}^{(2)}(\theta)\|^s < \infty$ *for all* $t$*.*

**Assumption 4.** *(i)* $S(\theta^c) = plim\ T^{-1} S_T(\theta^c)$ *has a unique global minimum at* $\theta_0^c$*; (ii) Let* $A_{i,T}(\theta_i^0) = Var\ T^{-1/2} \sum_{t \in I_{i-1}^0} u_t F_t(\theta_i^0)$*, for* $i = 1, \ldots, m+1$*, and* $A_T(\theta, r) = Var\ T^{-1/2} \sum_{t=1}^{[Tr]} u_t F_t(\theta)$*. Then* $A_{i,T}(\theta_i^0) \xrightarrow{p} A_i(\theta_i^0)$*, and* $A_T(\theta, r) \xrightarrow{p} A(\theta, r)$ *, where the two limits are finite positive definite matrices not depending on* $T$*, and the last convergence is uniform in* $\theta \times r$*. (iii) Let* $D_{i,T}(\theta_i^0) = T^{-1} \sum_{t \in I_{i-1}^0} F_t(\theta_i^0) F_t(\theta_i^0)'$ *and* $D_T(\theta, r) = T^{-1} \sum_{t=1}^{[Tr]} F_t(\theta)\ F_t(\theta)'$*. Then* $D_{i,T}(\theta_i^0) \xrightarrow{p} D_i(\theta_i^0)$ *and* $D_T(\theta, r) \xrightarrow{p} D(\theta, r)$*, where the two limits are finite positive definite matrices not depending on* $T$*, and the last convergence is uniform in* $\theta \times r$*; (iv)* $E[f_t(\theta_i^0)] \neq E[f_t(\theta_{i+1}^0)]$*, for each* $i = 1, 2, \ldots, m$*.*

**Assumption 5.** $T_i^0 = [T\lambda_i^0]$*, where* $0 < \lambda_1^0 < \ldots < \lambda_m^0 < 1$*.*

Assumption 1(i) allows for a fixed number of shifts in the distribution of errors and/or regressors, $m^*$, possibly at different locations than the parameter breaks in (1), and ensures that the change-point fractions indexing the change-points in the distribution of $v_t$ are asymptotically distinct.[5] We alternatively derive our results

[5] Note that $m^*$ as well as $\lambda_j^*$ are taken as given and are not objects of inference here, unless

under Assumption 1(ii). A further discussion about why these assumptions arise in the context of nonlinear models with breaks, along with ways of relaxing them, can be found after Lemma 1.

Assumption 1 also ensures that the model can be estimated via NLS, since the errors are uncorrelated with the regression function. Assumption 2 and 3 are overall typical smoothness and boundedness assumptions encountered in nonlinear models.[6] Assumption 4 (i) is the usual NLS identification assumption, while (ii)-(iv) are required because of the unstable structure of the model. Part (ii) and (iii) are similar to those in Bai and Perron (1998) and refer to the existence of limiting variances within regimes, while (iv) ensures that the parameter shifts across regimes can be identified. Note that (ii)-(iii) require uniform convergence in $\theta \times r$, as compared to uniform convergence in $r$ in linear models; this renders the asymptotic theory we use non-standard. Assumption 5 is a typical assumption for unstable models, allowing the break-fractions to be fixed and hence the break-points to be asymptotically distinct.

# 4 Asymptotic Behavior of Estimates

## 4.1 Consistency of Break-Fraction Estimates

In Section 2, we described a least-squares based method similar to its linear counterpart in Bai and Perron (1998). To elucidate the connection between linear and nonlinear settings, we will provide a heuristic discussion first. As Gallant (1987) shows, NLS estimators have the same form as OLS estimators (in stable

$m = m^*$ and all breaks in $\{v_t\}$ and the parameters of (1) coincide. Knowledge of $m^*$ and $\lambda_j^*$ is irrelevant as far as asymptotic distribution results are concerned, but may be of course crucial for both getting consistent estimates of certain asymptotic variances, as well as obtaining the null distribution of stability tests if the break-points in parameters do not coincide with the ones in the distribution of $\{v_t\}$ - see Hansen (2000) and Section 5.

[6] Note that Assumption 3(ii) is slightly stronger, implying that the expectation of the first $n_0$ realizations of the process $\{\sup_\theta |u_t f_t(\theta)|\}$ exists. This assumption is only needed to obtain the result in Lemma 1 under Assumption 1(ii) and is not imposed in other parts of the analysis.

models) up to a first-order approximation. To see that, denote by $X$ the $T \times s$ and $f(X, \theta)$ the $T \times 1$ regressors in stable OLS, respectively NLS models, and let $F = \partial f(X, \theta^0)/\partial\, \theta$, where $\theta^0$ is the true parameter value. The similarity between OLS and NLS can be seen from the equation below:

$$OLS = (X'X)^{-1}X'y; \qquad NLS = (F'F)^{-1}F'y + o_p(T^{-1/2}) \tag{4}$$

Given this similarity, extending Bai and Perron's (1998) methodology to nonlinear settings may seem straightforward. However, consistency of parameters estimates, and related to this, the Taylor expansion needed to obtain a similar formula as in (4) for unstable NLS estimates cannot be legitimately obtained prior to deriving the consistency and convergence rate of break-fraction estimates. For the latter we require different proof strategies, but the results are similar to Bai and Perron (1998) and are summarized in Theorems 1 and 2.

**Theorem 1.** *For each $i = 1, \ldots, m$, let $\hat{\lambda}_i$ be the smallest number such that $\hat{T}_i = [T\hat{\lambda}_i]$. Then, under Assumptions 1-5, $\hat{\lambda}_i \xrightarrow{p} \lambda_i^0$.*

For intuition and because they are informative for Assumption 1, we outline the main steps of the proof here, the details being relegated to the Appendix.

Define $\hat{u}_t = y_t - f_t(\hat{\theta}_{k+1})$, for $t \in \hat{I}_k$ and $d_t = \hat{u}_t - u_t = f_t(\theta^0_{j+1}) - f_t(\hat{\theta}_{k+1})$, for $t \in I^0_j \cap \hat{I}_k$, with $I^0_j = [T^0_j + 1, T^0_{j+1}]$ and $\hat{I}_k = [\hat{T}_k + 1, \hat{T}_{k+1}]$ and $k, j = 0, 1, \ldots, m$. Also, denote $\psi_t(\theta) = u_t f_t(\theta)$, a mean zero process governed by Assumption 1. Then:

$$T^{-1}\sum_{t=1}^{T} u_t d_t = T^{-1}\sum_{i=0}^{m}\sum_{I^0_i}\psi_t(\theta^0_i) - T^{-1}\sum_{i=0}^{m}\sum_{\hat{I}_i}\psi_t(\hat{\theta}_i) = I + II.$$

The proof of consistency rests on showing that $I + II$ is $o_p(1)$. While $I = o_p(1)$ by a simple law of large numbers, the analysis of $II$ is more complicated as this term contains not only sums with random endpoints but summands that depend on the parameter estimators, which in turn depend on the random endpoints. In showing $II$, we appeal to the following result:

**Lemma 1.** *Under Assumptions 1-2 and 3(i)-(ii), $T^{-1}\sum_{t=1}^{[Tr]}\psi_t(\theta) = O_p(T^{-1/2})$ uniformly in $\theta \times r \in \Theta \times [0,1]$.*

Lemma 1 was shown by Caner (2007) under the assumption that $\{v_t\}$ is a strictly stationary process. To address issues raised in Hansen (2000), we first relax strict stationarity over the whole sample to piece-wise strict stationarity. However, there may be cases of interest where even piece-wise strict stationarity is not desired, so we also establish Lemma 1 in the Appendix under Assumption 1(ii), which restricts $x_t$ only to be n.e.d. but at the price of requiring independence of $\{x_t\}$ and $\{u_t\}$ and the serial independence of $\{u_t\}$. The proof uses Ottaviani's inequality and empirical process theory.

*Remark* The conditions of Lemma 1 allow for serial correlation in $u_t$, but rule out models with lagged dependent variables in $x_t$. The extension of the lemma to such processes is problematic due to the type of non-stationarity induced by breaks in the parameters of lagged dependent variables.[7] These issues seem to be mitigated under some assumptions, as stated in Andrews and Fair (1988), pp. 620, who describe without proof some classes of primitive conditions one would need to employ for a weak law of large numbers (but not a central limit theorem) uniform in $\theta \times r$. However, note that if Lemma 1 holds, all the other results in the paper hold with appropriate modifications, thus lagged dependent variables can be considered at the expense of imposing Lemma 1.

With Lemma 1 in mind and using the definition of the sum of squared residuals,

[7] Note that by dropping strict stationarity, the limit of the process $\Psi_T(\theta, r) = T^{-1/2}\sum_{t=1}^{[Tr]}\psi_t(\theta)$ in Lemma 1 becomes a non-standard Gaussian process, more general than the Kiefer process in Caner (2007), Lemma 1, since its variance-covariance function may depend nonlinearly on the fractions of the partial sums. In other words, $\Psi_T(\theta, r)$ may not have a unique limit for all $r$, because its terms undergo a number of shifts; however, that does not in general preclude uniform boundedness in $\theta \times r$, and the latter is all we need for Lemma 1 to go through.

one can show that:

$$T^{-1}\sum_{t=1}^{T} d_t^2 + 2T^{-1}\sum_{t=1}^{T} d_t u_t \leq 0 \tag{5}$$

Consistency follows from the following lemma:

**Lemma 2.** *Let Assumption 1-5 hold. Then: (i)* $T^{-1}\sum_{t=1}^{T} u_t d_t = o_p(1)$*; (ii) If* $\hat{\lambda}_j \overset{p}{\nrightarrow} \lambda_j^0$ *for some* $j$*, then* $\limsup P\left[T^{-1}\sum_{t=1}^{T} d_t^2 > C\right] > \epsilon$*, for some* $C > 0, \epsilon > 0$*.*

Given part (i) of Lemma 2 and inequality (5), it follows that $T^{-1}\sum_{t=1}^{T} d_t^2 = o_p(1)$. The latter is in contradiction with part (ii) of Lemma 2, establishing consistency of break-fraction estimates.

## 4.2 Rates of Convergence

A necessary next step involves determining the convergence rates of the break-fraction estimates. The results are summarized below:

**Theorem 2.** *Under Assumptions 1-5, for every* $\eta > 0$*, there exists a finite* $C > 0$ *such that for all large* $T$*,* $P(\mid T(\hat{\lambda}_k - \lambda_k^0) \mid > C) < \eta$*,* $(k = 1, \ldots, m)$*.*

Theorem 2 is useful since the consistency of $\hat{\theta}_T^c$ can be established provided that the difference between the estimated and the true objective function is no more than $o_p(1)$. This is the case here because Theorem 2 implies that the difference involves a bounded number of $o_p(1)$ terms. Given the $T$-rate convergence of break-fraction estimates, the limiting distributions of parameter estimates follow from standard NLS asymptotics:

**Theorem 3.** *Under Assumptions 1-5,* $\hat{\theta}_i$ *and* $\hat{\theta}_j$ *are asymptotically independent and* $T^{1/2}(\hat{\theta}_i - \theta_i^0) \xrightarrow{d} \mathcal{N}(0, \Phi_i(\theta_i^0))$*, where* $\Phi_i(\theta_i^0) = [D_i(\theta_i^0)]^{-1} A_i(\theta_i^0)[D_i(\theta_i^0)]^{-1}$ *for* $i, j = 1, \ldots, m+1, i \neq j$*.*

Theorems 1-3 allow us to estimate the covariance matrices $\Phi_i(\theta_i^0)$ by replacing $D_i(\theta_i^0)$ with $\hat{D}_i(\hat{\theta}_i) = T^{-1}\sum_{t=\hat{T}_{i-1}+1}^{\hat{T}_i} F_t(\hat{\theta}_i)F_t(\hat{\theta}_i)'$ and $A_i(\theta_i^0)$ with a heteroskedasticity and autocorrelation (HAC) robust covariance matrix estimator, $\hat{A}_i(\hat{\theta}_i)$. If we consider the special case:

**Assumption 6.** *(i) Assumption 1 holds with $m = m^*, T_i^* = T_i, i = 1, \ldots, m$, $E[u_t|x_t] = 0$ and $E[u_t u_s|x_k x_l] = 0$ for all $t \neq s$ and all $k, l$; (ii) The errors are homoskedastic within regimes: $E[u_t^2 \mid x_t] = \sum_{t=1}^{T} \sigma_i^2 \mathbf{1}\{t \in I_i^0\}$ for all $t$; (iii) Let $D_T(\theta, r) = T^{-1} \sum_{t=T_{i-1}^0+1}^{T_{i-1}^0+[Tr]} F_t(\theta)F_t(\theta)'$. Then $D_T(\theta, r) \xrightarrow{p} rD_i(\theta)$, $(T^{-1} < r \leq \lambda_i^0 - \lambda_{i-1}^0)$, where the latter is a positive definite matrix not depending on $T$, and the convergence is uniform in $\theta \times r$, with $D_i(\theta)$ not necessarily the same for all $i$; (iv) Let $A_T(\theta, r) =$ plim Var $T^{-1} \sum_{t=T_{i-1}^0+1}^{T_{i-1}^0+[Tr]} u_t(\theta)F_t(\theta)$. Then $A_T(\theta, r) \xrightarrow{p} rA_i(\theta)$, $(T^{-1} < r \leq \lambda_i^0 - \lambda_{i-1}^0)$, where the latter is a positive definite matrix not depending on $T$, and the convergence is uniform in $\theta \times r$, with $A_i(\theta)$ not necessarily the same for all $i$.*[8]

Then the covariance matrix in Theorem 3 simplifies to $\sigma_i^2[D_i(\theta_i^0)]^{-1}$, which can be estimated, for example, via $\hat{\sigma}_i^2[\hat{D}_i(\theta_i^0)]^{-1}$, where $\hat{\sigma}_i^2 = (\hat{T}_i - \hat{T}_{i-1})^{-1} \sum_{t=\hat{T}_{i-1}+1}^{\hat{T}_i} \hat{u}_t^2$, for $i = 1, \ldots, m + 1$.

Note that Assumption 6 allows for breaks in marginal distributions of regressors, as well as breaks in the error variance that occur at the same time as the true breaks in model (1).

## 4.3 Limiting Distribution of Break Dates

Similar work by Bai (1994, 1995, 1997) for linear models derives the non-standard distributions of change-point estimates. Hall, Han, and Boldea (2008) extend this method to models that can be estimated via two stage least squares. These papers find the distribution of the break-point estimators in two cases, fixed and shrinking magnitude of shifts. In the first case, in general, the distributions in linear models depend on the underlying distribution of the regressors and errors. The second case allows for magnitudes of shifts that shrink to zero as the sample size increases. We consider both cases in turn.

[8] Part (iv) is implicit from (ii)-(iii) given (i), but is used explicitly without (ii) for Theorems 8,9.

### 4.3.1 Fixed Magnitude of Shifts

Consider the following data generation process, with one break[9]:

$$y_t = \begin{cases} f(x_t, \theta_1^0) + u_t & t = 1, \ldots, k_0 \\ f(x_t, \theta_2^0) + u_t & t = k_0 + 1, \ldots, T. \end{cases}$$

An implicit assumption so far was that the parameter shifts are constant:

**Assumption 7.** *$\delta = \theta_2^0 - \theta_1^0$, a fixed number.*

Denote by $S_T(k, \theta_1, \theta_2)$ the sum of squared residuals evaluated at a potential break-point $1 \leq k \leq T$. Also, let $S_T(k) = \min_{\theta_1,\theta_2} S_T(k, \theta_1, \theta_2)$. Then we can write:

$$\hat{k} = \underset{1 \leq k \leq T}{\text{argmin}} \; \underset{\theta_1, \theta_2}{\text{argmin}} \; V(k, \theta_1, \theta_2) \tag{6}$$

where: $V(k, \theta_1, \theta_2) = S_T(k, \theta_1, \theta_2) - S_T(k_0, \theta_1^0, \theta_2^0)$. We obtain a large sample approximation to this finite distribution, given below:

**Theorem 4.** *Under Assumptions 1-5 and 7, for $m = 1$,*

$$\left[\hat{k} - k_0\right] - \underset{v \in \mathbf{R}}{\text{argmax}} \, J^*(v) \xrightarrow{p} 0,$$

*where $J^*(v)$ is a double-sided stochastic process with $J^*(0) = 0$, $J(v) = J_1^*(v), v < 0$, $J(v) = J_2^*(v), v > 0$, and*

$$J_1^*(v) = \sum_{t=k_0+v+1}^{k_0} \left[f_t(\theta_2^0) - f_t(\theta_1^0)\right]^2 - 2 \sum_{t=k_0+v+1}^{k_0} u_t \left[f_t(\theta_2^0) - f_t(\theta_1^0)\right]$$

$$J_2^*(v) = -\sum_{t=k_0+1}^{k_0+v} \left[f_t(\theta_2^0) - f_t(\theta_1^0)\right]^2 - 2 \sum_{t=k_0+1}^{k_0+v} u_t \left[f_t(\theta_2^0) - f_t(\theta_1^0)\right]$$

[9] The extension to multiple breaks is immediate because the $m + 1$ errors in the sub-samples are asymptotically independent given Assumption 1.

The result above is comparable to linear models. Under Assumption 1(ii), $J^*(v)$ becomes a two-sides random walk with stochastic drifts. If we also impose the strict stationarity Assumption 1(i) with $m^* = 0$, the limit is a two-sided Gaussian stochastic process with negative drift, and it is the same as the limit for shrinking shifts (see next section).

### 4.3.2 Shrinking Magnitude of Shifts

Instead of Assumption 7, consider Assumption 8, which imposes parameter shifts that are shrinking at a certain rate $w_T$:

**Assumption 8.** *For $i = 1, \dots, m$, $= \theta^0_{i+1,T} - \theta^0_{i,T} = \delta_i \, w_T$, where $\delta_i$ are fixed $p \times 1$ vectors and $\{w_T\}$ is a scalar series such that $w_T \to 0$ and $T^{1/2-\gamma} w_T^2 \to \infty$ as $T \to \infty$, for some $\gamma \in \left[0, \frac{1}{2}\right)$.*

This assumption ensures that the asymptotic distributions of the change-point estimates do not depend on the underlying distributions of $\{u_t, f_t(\theta)\}$. Similar assumptions are *inter alia* $T^{1/2-\gamma} w_T \to \infty$, for $\gamma \in \left(0, \frac{1}{2}\right)$ in Bai and Perron (1998) and $T^{1/2} w_T / (log T)^2 \to \infty$ in Qu and Perron (2007). Our assumption allows only shifts of order $T^{-1/4}$ or larger, but the simulation section discusses that, despite this, the coverage probability for the confidence intervals is good. Note that under shrinking magnitudes of shift, the asymptotic properties of parameter and break-fraction estimates need to be re-derived (see Appendix), with the break-fraction distribution presented below.

**Theorem 5.** *Let $\phi = \delta_1' A_2(\theta_1^0)\delta_1 / [\delta_1' A_1(\theta_1^0)\delta_1]$ and $\xi = \delta_1' D_2(\theta_1^0)\delta_1 / [\delta_1' D_1(\theta_1^0) \; \delta_1]$. Under Assumptions 1-5, 6(iii)-(iv), and 8, for $m = 1$,*

$$\frac{[\delta_1' D_1(\theta_1^0)\delta_1]^2}{\delta_1' A_1(\theta_1^0)\delta_1} w_T^2 [\hat{k} - k_0] \Rightarrow \underset{v}{\text{argmax}} \, Z(v)$$

*where $Z(v) = J_1(-v) - 0.5|v|, v \leq 0$, $Z(v) = \sqrt{\phi} J_2(v) - 0.5\xi|v|, v > 0$, $J_1(v)$, $J_2(v)$ are two independent standard scalar Gaussian processes defined on $[0, \infty]$, and '$\Rightarrow$'*

*denotes weak convergence in Skorohod metric.*

Details regarding this process can be found in Bai (1997). The density of $\text{argmax}_v\, Z(v)$ is characterized by Bai (1997) and he notes that it is not symmetric if $\phi \neq 1$ or $\xi \neq 1$. A confidence interval can be constructed as follows. Let $\hat{\omega}_{1,i} = (\hat{\theta}_2 - \hat{\theta}_1)'\hat{A}_i(\hat{\theta}_1)(\hat{\theta}_2 - \hat{\theta}_1)'$, $\hat{\omega}_{2,i} = (\hat{\theta}_2 - \hat{\theta}_1)'\hat{D}_i(\hat{\theta}_1)(\hat{\theta}_2 - \hat{\theta}_1)'$, $\hat{D}_i(\theta) = (\hat{T}_i - \hat{T}_{i-1})^{-1}\sum_{t=\hat{T}_{i-1}+1}^{\hat{T}_i} F_t(\theta)F_t(\theta)'$; $\hat{A}_i(\theta)$ a HAC estimator of the long-run variance $A_i(\theta)$, and $\hat{H} = \hat{\omega}_{2,1}^2/\hat{\omega}_{1,1}$. Also, let $\hat{\xi} = \hat{\omega}_{2,2}/\hat{\omega}_{2,1}$ and $\hat{\phi} = \hat{\omega}_{1,2}/\hat{\omega}_{1,1}$. Then, a $100(1-\alpha)\%$ confidence interval for $\hat{k}$ is:

$$(\,\hat{k} - [c_1/\hat{H}] - 1, \hat{k} + [c_2/\hat{H}] + 1\,) \tag{7}$$

where $c_1$ and $c_2$ are respectively the $(\alpha/2)^{th}$ and $(1-\alpha/2)^{th}$ quantiles for $\text{argmax}_v\, Z(v)$ which can be calculated using equations (B.2) and (B.3) in Bai (1997).

Theorem 5 can be extended to yield confidence intervals for the multiple break model, because given Assumption 1, the sample segments are asymptotically independent, allowing for the analysis of the limiting distribution to be carried out as in the one break case:

**Corollary to Theorem 5.** *Define* $\phi_i = \delta_i' A_{i+1}(\theta_i^0)\delta_i/[\delta_i' A_i(\theta_i^0)\delta_i]$ *and* $\xi_i = \delta_i' D_{i+1}(\theta_i^0)\delta_i$ $/[\delta_i' D_i(\theta_i^0)\delta_i]$. *Under Assumptions 1-5, 6(iii)-(iv) and 8,*

$$\frac{[\delta_i' D_i(\theta_i^0)\delta_i]^2}{\delta_i' A_i(\theta_i^0)\delta_i} w_T^2[\hat{k} - k_0] \Rightarrow \underset{v}{\text{argmax}}\, Z_i(v)$$

*where* $Z_i(v) = W_{i,1}(-v) - 0.5|v|, v \leq 0$, $W_i(v) = \sqrt{\phi_i}W_{i,2}(v) - 0.5\xi_i|v|, v > 0$ *and* $W_{i,1}(v)$, $W_{i,2}(v)$ *are independent standard scalar Gaussian processes defined on* $[0,\infty]$, *for* $i = 1, \ldots, m$.

Confidence intervals can thus be obtained by redefining the appropriate quantities in (7) for each break-point estimator.

# 5 Tests for Multiple Breaks

This section is concerned with finding the number of breaks $m$, so far treated as known. To that end, we consider similar tests in Bai and Perron (1998), as well as equivalent *sup* Wald tests that are useful when autocorrelation is present. Given the results in the previous sections, we are able to show that their distribution carry over from linear settings. The critical values are already tabulated in Bai and Perron (1998) and Bai and Perron (2003a).

## 5.1 Sup F-Tests

The $F$-tests based on differences in sum of squared residuals can be carried out as long as long as Assumption 6 holds. Extensions to serially correlated errors can be found in Section 5.2.

### 5.1.1 An F Test of No Breaks Versus a Fixed Number of Breaks

Consider the following hypothesis:

$$H_0 : m = 0 \qquad vs. \qquad H_A : m = k. \tag{8}$$

where $k$ is a fixed finite positive integer. For this purpose, consider a partition $(T_1, \ldots, T_k)$ of the $[1, T]$ interval such that $T_i = [T\lambda_i]$. We also need to restrict each change point to be asymptotically distinct and bounded away from the endpoints of the sample. To this end, define $\Lambda_\epsilon = \{\bar{\lambda}_k \equiv (\lambda_1, \ldots, \lambda_k) : |\lambda_{i+1} - \lambda_i| \geq \epsilon, \lambda_1 \geq \epsilon, \lambda_k \leq 1 - \epsilon\}$, where $\epsilon$ is a small number, in practice ranging from 0.05 to 0.15. As in Bai and Perron (1998), consider a generalized version of the *sup* $F$-type tests proposed in Andrews (1993):

$$\sup_{\bar{\lambda}_k \in \Lambda_\epsilon} F_T(k;p) = \sup_{\bar{\lambda}_k \in \Lambda_\epsilon} \frac{(SSR_0 - SSR_k)/kp}{SSR_k/[T-(k+1)p]} \tag{9}$$

where $SSR_0$ and $SSR_k$ are the sums of squared residuals under the null, respectively under the alternative hypothesis. Let $B_p(\cdot)$ be a $p$-vector of independent Brownian motions. The following theorem describes the distribution of the test under $H_0$:

**Theorem 6.** *Under Assumptions 2-6 and $H_0$ in* (8)*,*

$$\sup_{\bar{\lambda}_k \in \Lambda_\epsilon} F_T(k;p) \Rightarrow \frac{1}{kp} \sup_{\bar{\lambda}_m \in \Lambda_\epsilon} \sum_{i=1}^{k} \frac{\|\lambda_i B_p(\lambda_{i+1}) - \lambda_{i+1} B_p(\lambda_i)\|^2}{\lambda_i \lambda_{i+1}(\lambda_{i+1} - \lambda_i)}$$

It is worth noting that the distribution of the *sup-F* test under $H_0$ above does not depend on any nuisance parameters. As Bai and Perron (1998) show, the test above is consistent for its alternative. Of course, if autocorrelation is present, this F-test should be replaced with a Wald-type test of equality of parameters across regimes, and we describe such a test in the next section.

#### 5.1.2 A Double Maximum F Test

Next, one can consider testing against an unknown number of breaks $m < M$, $M$ being an upper bound on the number of change-points. To that end, consider the hypothesis:

$$H_0 : m = 0 \qquad vs. \qquad H_A : \text{ m unknown}, m < M, M \text{ fixed}. \tag{10}$$

As Bai and Perron (1998) point out, to test this hypothesis it suffices to take the maximum over weighted versions of the test statistics described in the previous section, where the weights are $(a_1, \ldots, a_M)$:

$$D \max F_T(M, a_1, \ldots, a_M) = \max_{1 \le m \le M} a_m \sup_{\bar{\lambda}_m \in \Lambda_\epsilon} F_T(m;p) \tag{11}$$

The distribution of the test statistic above is:

**Corollary to Theorem 6.** *Under Assumptions 2-6 and $H_0$ in* (10),

$$D\max F_T(M, a_1, \ldots, a_M) \Rightarrow \max_{1\leq m\leq M} \frac{a_m}{mp} \sup_{\bar{\lambda}_m\in\Lambda_\epsilon} \sum_{i=1}^{m} \frac{\|\lambda_i B_p(\lambda_{i+1}) - \lambda_{i+1}B_p(\lambda_i)\|^2}{\lambda_i\lambda_{i+1}(\lambda_{i+1}-\lambda_i)}$$

As Bai and Perron (1998) mention, the choice of weights remains an open question. It may reflect the imposition of some priors on the likelihood of various number of breaks. One possibility is to set all weights equal to unity. We denote this test as:

$$UD\max F_T(M,p) = \max_{1\leq m\leq M} \sup_{\bar{\lambda}_m\in\Lambda_\epsilon} F_T(m;p) \tag{12}$$

Note that, for fixed $m$ and break locations, $F_T(m;p)$ is the sum of $m$ dependent $\chi_p^2$ variables, each divided by $m$. This scaling by $m$ can be viewed as a prior that, as $m$ increases, a fixed sample becomes less informative about the hypotheses that it is confronted with. Since for any fixed $p$, the critical values of $\sup_{(\bar{\lambda}_k)\in\Lambda_\epsilon} F_T(m;p)$ decrease as $m$ increases, this implies that if we have a large number of breaks, we may get a test with low power, because the marginal p-values decrease with $m$. One way to keep marginal p-values of the tests equal across $m$ is to use weights that depend on $p$ and the significance level of the test, say $\alpha$. More precisely, let $c(p,\alpha,m)$ be the asymptotic critical value of the test $\sup_{\bar{\lambda}_m\in\Lambda_\epsilon} F_T(m;p)$. Define, as in Bai and Perron (1998), $a_1 = 1$ and $a_m = c(p,\alpha,1)/c(p,\alpha,m)$ for $1 < m \leq M$. The test obtained this way is:

$$WD\max F_T(M,p) = \max_{1\leq m\leq M} \frac{c(p,\alpha,1)}{c(p,\alpha,m)} \times \sup_{\bar{\lambda}_m\in\Lambda_\epsilon} F_T(m;p) \tag{13}$$

For consistency of Dmax tests and critical values of both its versions, UDmax and WDmax, see Bai and Perron (1998).

#### 5.1.3 An F Test of $\ell$ Versus $\ell+1$ Breaks

Consider the following hypothesis of interest:

$$H_0 : m = \ell \qquad vs. \qquad H_A : m = \ell + 1. \tag{14}$$

One would ideally construct such a test based on the difference between the sum of squared residuals for $\ell$ breaks and $(\ell + 1)$ breaks. Considering the different mismatches in end-points of partial sums obtained this way, it would be hard to describe the limiting behavior of such tests. An easier strategy involves imposing $\ell$ breaks and testing each segment for an additional break. The test statistic is:

$$F_T(\ell+1|\ell) = \max_{1\le i\le \ell+1} \frac{1}{\hat{\sigma}_i^2} \left\{ S_T(\hat{T}_1, \ldots, \hat{T}_\ell) - \inf_{\tau\in\Delta_{i,\ell}} S_T(\hat{T}_1, \ldots, \hat{T}_{i-1}, \tau, \hat{T}_i, \ldots, \hat{T}_\ell) \right\}$$

where under $H_0$:

$$\Delta_{i,\,\ell} = \{\tau : \hat{T}_{i-1} + (\hat{T}_i - \hat{T}_{i-1})\eta \le \tau \le \hat{T}_i - (\hat{T}_i - \hat{T}_{i-1})\eta\}, \text{ and } \hat{\sigma}_i^2 \xrightarrow{p} \sigma_i^2$$

The following result is proved in the Appendix:

**Theorem 7.** *Under Assumptions 2-6 and $H_0$ in* (14), $\lim P(F_T(\ell+1|\ell) \le x) = G_{p,\eta}^{\ell+1}$, *where $G_{p,\eta}$ is the distribution function of* $\displaystyle\sup_{\eta\le\mu\le 1-\eta} \frac{\|B_p(\mu) - \mu B_p(1)\|^2}{\mu(1-\mu)}$.

Note that this test allows for heterogeneity in regressors and errors across regimes, including breaks in the distribution of errors and/or regressors occurring simultaneously with the coefficient breaks.

If there are more than $\ell$ breaks, but we estimated a model with just $\ell$ breaks, then there must be at least one additional break not estimated. Hence, at least one of the $(\ell+1)$ segments obtained contains a nontrivial breakpoint, in the sense that both boundaries of this segment are separated from the true break-point by a positive fraction of the total number of observations. For this segment, the $sup\,F(1,p)$ test statistic diverges to infinity as the sample size increases, since this test is consistent. Then so does $F_T(\ell+1|\ell)$, hence this test is consistent too.

## 5.2 Tests in the Presence of Autocorrelation

In this section, we provide tests that are robust to types of autocorrelation allowed by Assumption 1. In particular, we extend the tests in Sections 5.1.1-5.1.3; the first two tests were developed for linear models in Bai and Perron (1998), while the last test is proposed for linear models in Hall, Han, and Boldea (2009).

### 5.2.1 A Wald Test of Zero Versus a Fixed Number of Breaks

The hypothesis in (8) can be re-written as: $H_0 : R_k\, \theta_0^c = 0$, where $R_k$ is the conventional matrix such that $(R_k\, \theta_0^c)' = (\theta_1^{0'} - \theta_2^{0'}, \ldots, \theta_k^{0'} - \theta_{k+1}^{0'})$. The corresponding sup Wald test statistic is:

$$\sup_{(\lambda_1,\ldots,\lambda_k)\in\Lambda_\epsilon} W_T(k;p) \; = \sup_{\bar{\lambda}_k\in\Lambda_\epsilon} \hat{\theta}^{c'}(\bar{T}_k) R_k' \, (R_k\, \hat{\Upsilon}(\bar{T}_k)\, R_k')^{-1}\, R_k\, \hat{\theta}^c(\bar{T}_k)$$

where $\hat{\theta}^{c'}(\bar{T}_k) = [\hat{\theta}_1'(\bar{T}_k), \ldots, \hat{\theta}_{k+1}'(\bar{T}_k)]$, $\hat{\Upsilon}(\bar{T}_k) = \text{diag}\,[\hat{\Upsilon}_1(\bar{T}_k), \ldots, \hat{\Upsilon}_{k+1}(\bar{T}_k)]$, and $\hat{\Upsilon}_i(\bar{T}_k) = T^{-1}[\hat{D}_i^{-1}(\hat{\theta}_i(\bar{T}_k))]\; [\hat{A}_i(\hat{\theta}_i(\bar{T}_k))]\; [\hat{D}_i^{-1}(\hat{\theta}_i(\bar{T}_k))]$, recalling that $\bar{T}_k$ was a certain $k$-partition of the sample interval.

To facilitate the presentation of an intuitive form for the distribution of the *sup* Wald tests, rewrite $R_k = \tilde{R}_k \otimes I_p$, with $\tilde{R}_k$ being the conventional $k \times (k+1)$ matrix such that $(\tilde{R}_k\beta)' = (\beta_1 - \beta_2, \ldots, \beta_k - \beta_{k+1})$, where $\beta_i$ the $i^{th}$ element of some $(k+1)\times 1$ vector $\beta$, and $I_p$ is the $p \times p$ identity matrix. From the Appendix, it follows that:

**Theorem 8.** *Under Assumptions 1-5, 6(iii)-(iv) and $H_0$ in (8),*

$$\sup_{\bar{\lambda}_k\in\Lambda_\epsilon} W_T(k;p) \;\Rightarrow\; \sup_{\bar{\lambda}_k\in\Lambda_\epsilon} \tilde{B}_k(\bar{\lambda}_k),$$

*where:* $\tilde{B}_k(\bar{\lambda}_k) = B_{p(k+1)}'\,\{\,[C_k^{-1}\tilde{R}_k'(\tilde{R}_k C_k^{-1}\tilde{R}_k')^{-1}\tilde{R}_k C_k^{-1}]\otimes I_p\,\}\, B_{p(k+1)}$, *with* $B_{p(k+1)} =$

$[B_p'(\lambda_1), B_p'(\lambda_2) - B_p'(\lambda_1), \ldots, B_p'(\lambda_{k+1}) - B_p'(\lambda_k)]'$, *a* $p(k+1) \times 1$ *vector of pairwise independent vector Brownian motions of dimensions* $p$, $C_k =$ *diag* $(\lambda_1, \lambda_2 - \lambda_1, \ldots, \lambda_{k+1} - \lambda_k)$ *and* $\lambda_{k+1} = 1$ *by convention.*

It can be shown that the $H_0$ distribution of the *sup* $W_T(k;p)$ is a scaled version of the corresponding distribution of the *sup* $F_T(k;p)$, with scaling factor $kp$.

### 5.2.2 Double Maximum Wald Tests

Given the result in Theorem 8, the $D\max F_T(M, a_1, \ldots, a_M)$ test has its corresponding autocorrelation-robust version:

$$D\max W_T(M, a_1, \ldots, a_M) = \max_{1 \leq m \leq M} \frac{a_m}{mp} \sup_{\bar{\lambda}_m \in \Lambda_\epsilon} W_T(m;p) \qquad (15)$$

whose distribution is:

**Corollary to Theorem 8.** *Under Assumptions 1-5, 6(iii)-(iv) and* $H_0$ *in* (10)*,*

$$D\max W_T(M, a_1, \ldots, a_M) \Rightarrow \max_{1 \leq m \leq M} \frac{a_m}{mp} \sup_{\bar{\lambda}_m \in \Lambda_\epsilon} \tilde{B}_m(\bar{\lambda}_m)$$

The scaling $mp$ is used not only to obtain the same asymptotic distributions as for the corresponding F-tests, but because, in the absence of scaling and equal weights $a_i$, this test will be equivalent to testing zero against $M$ breaks, since $\sup_{\bar{\lambda}_m \in \Lambda_\epsilon} \tilde{B}_m(\bar{\lambda}_m)$ is increasing in $m$ for a fixed $p$. Given the scaling, the discussion in Section 5.1.2. about picking $a_m$ is still valid. Thus, as in Section 5.1.2, we can use the unweighted version of the test, with $a_m = 1$, or the weighted version of the test, with $a_m = c(p, \alpha, 1)/\, c(p, \alpha, m)$ in (15).

### 5.2.3 A Wald Test of $\ell$ Versus $\ell + 1$ Breaks

For purposes of sequentially estimating the breaks in the presence of autocorrelation, it is desirable to develop a Wald-type test that is designed for testing $\ell$ versus

$\ell+1$ breaks; under $\ell+1$ breaks, this is equivalent to testing whether, there exists exactly one $i$ such that $\theta_i^0 \neq \theta_{i+1}^0$, where $i \in \{1, \ldots, \ell+1\}$.

Under $H_0$ in (14), for each index $q \in \{1, \ldots, \ell+1\}$ define the corresponding hypothesis: $R^*\,[\theta_q^{0'}, \theta_{q+1}^{0'}]' = 0$, where $R^* = \tilde{R}^* \otimes I_p$ and $\tilde{R}^* = [1, -1]$. For simplicity, let $\vartheta_q^0 = [\theta_q^{0'}, \theta_{q+1}^{0'}]'$ and $\hat{\vartheta}_q(\mu) = [\hat{\theta}_q(\mu)', \hat{\theta}_{q+1}(\mu)']'$, where we first estimated the model with $\ell$ breaks, imposed them as if they were the true ones, and then defined, for each feasible break $[T\mu] \in \Delta_{q,\ell}$ - with $\Delta_{q,\ell}$ defined in Section 5.1.3 - parameter estimates $\hat{\theta}_q(\mu), \hat{\theta}_{q+1}(\mu)$, for before and after the break.

The appropriate Wald test is:

$$W_T(\ell+1|\ell) = \max_{1 \leq q \leq \ell+1} \sup_{\tau \in \Delta_{q,\ell}} W_{T,\ell}(\tau, q)$$

where $W_{T,\ell}(\tau, q) \equiv W_{T,\ell}(\mu, q) = \hat{\vartheta}_q(\mu)' R^{*'} [R^* \hat{\Upsilon}_q^*(\mu) R^{*'}]^{-1} R^* \hat{\vartheta}_q(\mu)$, with $\hat{\Upsilon}_q^*(\mu) =$ diag $[\hat{\Upsilon}_{q,1}^*, \hat{\Upsilon}_{q,2}^*]$ with $\Upsilon_{q,j}^* = T[\hat{D}_{q,j}^*(\mu)]^{-1}\ \hat{A}_{q,j}^*(\mu)\ [\hat{D}_{q,j}^*(\mu)]^{-1}$, $(j = 1, 2)$, and $\hat{D}_{q,1}^*(\mu) = T^{-1} \sum_{t=\hat{T}_{q-1}+1}^{\tau} F_{t,q}(\mu) F_{t,q}(\mu)'$, $\hat{D}_{q,2}^*(\mu) = T^{-1} \sum_{t=\tau+1}^{\hat{T}_q} F_{t,q+1}(\mu) F_{t,q+1}(\mu)'$, while $\hat{A}_{q,1}^*(\mu)$ and $\hat{A}_{q,2}^*(\mu)$ are HAC estimators of the limiting variances of respectively $T^{-1/2} \sum_{t=\hat{T}_{q-1}+1}^{\tau} u_{t,q}(\mu) F_{t,q}(\mu)$, $T^{-1/2} \sum_{t=\tau+1}^{\hat{T}_q} u_{t,q+1}(\mu) F_{t,q+1}(\mu)$, with $F_{t,s}(\mu) = F_t(\hat{\theta}_s(\mu))$ and $u_{t,s}(\mu) = u_t(\hat{\theta}_s(\mu))$, $(s = q, q+1)$. Even though there are easier to compute estimates of the limiting variance of $\hat{\Upsilon}_q^*(\mu)$, for increasing the power of the test, we consider those that would be more relevant if the alternative is true.

Note that this test is useful for performing sequential estimation of breaks in the presence of autocorrelation. Not surprisingly, we find that the distribution of the above Wald test is the same as that of the corresponding F-test, but holds under more general assumptions:

**Theorem 9.** *Under Assumptions 1-5, 6(iii)-(iv) and* $H_0$ *in* (14)*, one can write* $\lim P(W_T(\ell+1|\ell) \leq x) = G_{p,\eta}^{\ell+1}$*, where* $G_{p,\eta}$ *is the cdf of* $\sup_{\eta \leq \mu \leq 1-\eta} \dfrac{\|B_p(\mu) - \mu B_p(1)\|^2}{\mu(1-\mu)}$.

### 5.3 Sequential Estimation of the Number of Breaks

Using the test statistics presented above, we can suggest a simple sequential method for obtaining an estimator, $\hat{m}_T$ say, of the number of breaks.

On the first step of the sequential estimation, use either $\sup F_T(1;p)$, $\sup W_T(1;p)$ or $Dmax\ F_T\ (M,p)$, $Dmax\ W_T(M,p)$, to test the null hypothesis that there are no breaks. If this null is not rejected then set $\hat{m}_T = 0$; else proceed to the next step. On the second step, use $F_T(2|1$ or $W_T(2|1)$ to test the null hypothesis of one against two breaks. If $F_T(2|1)$ or $W_T(2|1)$ does not reject, then $\hat{m}_T = 1$; else proceed to the next step. On the $\ell^{th}$ step, by means of $F_T\ (\ell+1|\ell)$ or $W_T\ (\ell+1|\ell)$, test the null hypothesis of $\ell$ breaks against $\ell+1$ breaks, and if the hypothesis is not rejected, then $\hat{m}_T = \ell$; else proceed to the next step. This sequential procedure stops when $M$, the ceiling on the number of breaks, is reached. If all statistics in the sequence are significant then the conclusion is that there are at least $M$ breaks. Note that this is not a proper sequential procedure, because with each sequential test, the breaks are re-estimated under the null with a global procedure.

## 6 Simulation Results

There are some clear computational advantages of the Bai and Perron (2003b) method for detecting breaks. As Bai and Perron (2003b) show, even when the number of change-points is large, we need not search over all possible partitions to find the true break. Imposing a minimum length on the segments in each partition, one need not perform more than $T(T+1)/2$ operations to find the estimated partition.

Here, we implement an algorithm for finding breaks similar to Bai and Perron (2003b). Along with nonlinearity additional issues arise, related to having no closed form for updating the sum of squares and parameter estimates when one more observation is present. Although approximate updating procedures such as an unscented Kalman filter can be useful, for simplicity we recalculate in each segment of the $T(T+1)/2$ new NLS estimates and sum of squares through global minimization of the concentrated sum of squares by a quasi Gauss-Newton algorithm.[10] As starting values for the nonlinear parameters, we use grid searches, Taylor expansions of up to $7^{th}$ order, as well as interpolations suggested in Gallant (1987) and Bates and Watts (1988).

We pick data generation processes (DGPs) with $m = 1, 2$, and a nonlinear function used in Gallant (1987) and Bates and Watts (1988):

$$f(x_t, \theta) = \theta_i^1 + \theta_i^2 \ \exp\left(-x_t\, \theta_i^3\right), \text{ with } t \in I_i^0, \text{ for } i = 1, \ldots, m+1$$

The true data was generated such that $x_t \sim N(0,1)$, $u_t \sim N(0,1)$ and $X \perp U$.[11]

Tables 1-3 are reported for 1000 simulations, $\epsilon = 0.15$ and 6 DGPs, with $m = 0, 1, 2$. Let $\iota_j$ be a $j$-vector of ones. We pick $DGP\,1 : m = 0, \theta_0^c = \iota_3$; $DGP\,2, 3, 4 : m = 1, \theta_0^{c'} = (1, 2) \otimes \iota_3'$, $(1, 1.5) \otimes \iota_3'$ and $(\iota_3'; (2, 1, 1))$; $DGP\,5, 6 : m = 2, \theta_0^{c'} = (1, 2, 1) \otimes \iota_3'$, $(1, 1.5, 1) \otimes \iota_3'$. The empirical coverage of the break-point $99\%, 95\%, 90\%$ confidence intervals are almost $100\%$ in each case. This is consistent with break-point estimates coinciding with the true break-points or being just one observation away. Table 1 shows very good size and power properties of $sup\, F$ tests; they improve as the sample size increases, for both $m = 1, m = 2$, and so do the properties of the estimate for number of breaks $\hat{m}_T$ in Table 2.

Parameter confidence interval coverages reported in Table 3 and are in all cases close to the nominal size. Overall, our methodology seems to work well in finite samples.

[10] The Levenberg-Marquardt algorithm provides similar results.

[11] To check whether our method works when parameters change in opposite directions, we ran simulations with $x_t \sim N(1,1)$ as well. The results are similar and are available upon request.

Table 1: Relative rejection frequencies of F-statistics

| *DGP* | *T* | *sup F* | | *seq F* | | *UDmax F* |
|---|---|---|---|---|---|---|
| | | 0:1 | 0:2 | 2:1 | 3:2 | |
| 1 | 100 | .05 | .05 | .01 | 0 | .05 |
| | 200 | .05 | .05 | .01 | 0 | .05 |
| 2 | 100 | 1.00 | 1.00 | .05 | 0 | 1.00 |
| | 200 | 1.00 | 1.00 | .03 | 0 | 1.00 |
| 3 | 100 | 1.00 | 1.00 | .04 | 0 | 1.00 |
| | 200 | 1.00 | 1.00 | .03 | 0 | 1.00 |
| 4 | 100 | .96 | .92 | .04 | 0 | .96 |
| | 200 | 1.00 | 1.00 | .04 | 0 | 1.00 |
| 5 | 100 | .97 | 1.00 | 1.00 | .02 | 1.00 |
| | 200 | 1.00 | 1.00 | 1.00 | .01 | 1.00 |
| 6 | 100 | .94 | 1.00 | .99 | .02 | 1.00 |
| | 200 | 1.00 | 1.00 | 1.00 | .01 | 1.00 |

*Notes:* *sup F* denotes the statistic $Sup\, F_T(k;1)$ and the second tier column heading under it denotes $k$; *seq F* denotes the statistic $F_T(\ell+1|\ell)$ and the second tier column beneath it denotes $\ell+1:\ell$; *UDmax F* denotes the statistic $UDmax\, F_T(5,1)$.

Table 2: Empirical distribution of the estimated number of breaks

| *DGP* | *T* | *sup F* | | | | *UDmax F* | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3,4,5 | 0 | 1 | 2 | 3,4,5 |
| 1 | 100 | .95 | .05 | 0 | 0 | .95 | .05 | 0 | 0 |
| | 200 | .95 | .05 | 0 | 0 | .95 | .05 | 0 | 0 |
| 2 | 100 | 0 | .95 | .05 | 0 | 0 | .95 | .05 | 0 |
| | 200 | 0 | .97 | .03 | 0 | 0 | .97 | .03 | 0 |
| 3 | 100 | 0 | .96 | .04 | 0 | 0 | .96 | .04 | 0 |
| | 200 | 0 | .96 | .04 | 0 | 0 | .96 | .04 | 0 |
| 4 | 100 | .04 | .93 | .03 | 0 | .04 | .93 | .03 | 0 |
| | 200 | 0 | .96 | .04 | 0 | 0 | .94 | .04 | 0 |
| 5 | 100 | .03 | 0 | .95 | .02 | 0 | 0 | .98 | .02 |
| | 200 | 0 | 0 | .99 | .01 | 0 | 0 | .99 | .01 |
| 6 | 100 | 0 | .96 | .04 | 0 | 0 | .96 | .04 | 0 |
| | 200 | 0 | 0 | .99 | .01 | 0 | 0 | .99 | .01 |

*Notes:* The blocks headed *sup F* or *UDmax F* give the empirical distribution of $\hat{m}_T$, obtained via the sequential strategy using $Sup\, F_T(1;1)$ or $UDmax\, F_T(5,1)$ on the first step with the maximum number of breaks set to five.

Table 3: Empirical coverage of parameter confidence intervals

| | | Confidence Intervals | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *DGP* | *T* | Regime | $\theta_1^0$ | | | $\theta_2^0$ | | | $\theta_3^0$ | | |
| | | | 99% | 95 % | 90 % | 99% | 95 % | 90 % | 99% | 95 % | 90 % |
| 2 | 100 | $1^{st}$ regime | .99 | .95 | .90 | .97 | .93 | .89 | .98 | .94 | .89 |
| | | $2^{nd}$ regime | .99 | .95 | .89 | .98 | .95 | .89 | .98 | .95 | .89 |
| | 200 | $1^{st}$ regime | .98 | .94 | .89 | .99 | .93 | .88 | .99 | .94 | .88 |
| | | $2^{nd}$ regime | .98 | .94 | .89 | .99 | .93 | .88 | .99 | .94 | .88 |
| 3 | 100 | $1^{st}$ regime | .99 | 95 | .90 | .97 | .93 | .89 | .99 | .94 | .89 |
| | | $2^{nd}$ regime | .99 | .95 | .89 | .98 | .95 | .89 | .98 | .95 | .89 |
| | 200 | $1^{st}$ regime | .98 | .94 | .89 | .98 | .94 | .89 | .99 | .94 | .90 |
| | | $2^{nd}$ regime | .99 | .94 | .89 | .98 | .94 | .89 | .98 | .94 | .90 |
| 4 | 100 | $1^{st}$ regime | .98 | .94 | .87 | .96 | .91 | .86 | .98 | .93 | .86 |
| | | $2^{nd}$ regime | .98 | .92 | .86 | .96 | .92 | .87 | .96 | .91 | .85 |
| | 200 | $1^{st}$ regime | .98 | .94 | .88 | .97 | .93 | .88 | .99 | .94 | .88 |
| | | $2^{nd}$ regime | .98 | .93 | .89 | .98 | .94 | .89 | .98 | .93 | .88 |
| 5 | 100 | $1^{st}$ regime | .96 | .91 | .87 | .94 | .89 | .85 | .97 | .92 | .87 |
| | | $2^{nd}$ regime | .96 | .91 | .87 | .98 | .93 | .86 | .97 | .92 | .86 |
| | | $3^{rd}$ regime | .99 | .94 | .90 | .97 | .93 | .89 | .98 | .93 | .88 |
| | 200 | $1^{st}$ regime | .98 | .94 | .89 | .99 | .93 | .88 | .99 | .94 | .88 |
| | | $2^{nd}$ regime | .98 | .94 | .89 | .98 | .93 | .89 | .98 | .93 | .89 |
| | | $3^{rd}$ regime | .98 | .94 | .90 | .98 | .94 | .90 | .98 | .94 | .89 |
| 6 | 100 | $1^{st}$ regime | .96 | .91 | .86 | .93 | .89 | .84 | .97 | .91 | .87 |
| | | $2^{nd}$ regime | .95 | .88 | .82 | .96 | .90 | .84 | .96 | .90 | .84 |
| | | $3^{rd}$ regime | .98 | .94 | .89 | .97 | .93 | .89 | .98 | .93 | .88 |
| | 200 | $1^{st}$ regime | .97 | .94 | .89 | .96 | .92 | .88 | .98 | .93 | .86 |
| | | $2^{nd}$ regime | .98 | .93 | .87 | .97 | .93 | .89 | .98 | .93 | .89 |
| | | $3^{rd}$ regime | .98 | .94 | .90 | .98 | .94 | .89 | .98 | .94 | .89 |

*Notes:* The column headed $100a\%$ gives the percentage of times the $100a\%$ confidence intervals for each parameter contains its true value.

# 7 Conclusions

In this paper, a nonlinear method for estimating and testing in NLS models with multiple breaks is developed. In our framework, the break-dates are estimated simultaneously with the parameters via minimization of the residual sum of squares. Using nonlinear asymptotic theory, we derive the asymptotic distributions of both break-point and parameter estimates and propose several instability tests. Our estimation procedure is very similar to that of Bai and Perron (1998), but the proofs are different since they require some empirical process theory results developed in this paper, results that may be useful in other settings as well. By construction, our method nests nonlinearities and breaks, and is useful in practice both for testing for breaks in the presence of nonlinearity, and for jointly modeling breaks and nonlinearity, should evidence for both be present.

Many other issues can be important for modeling nonlinearity jointly with breaks. Important macroeconomic applications that use structural equation models with endogeneity can be dealt with by extending the methodology in the current paper to multivariate, more general nonlinear models, as well as to partial structural change. On the other hand, developing primitive conditions along with new uniform convergence results for more general nonlinear time series processes is certainly of interest, and we leave this to future research.

## References

Anderson, G. J., and Mizon, G. E. (1983). 'Parameter Constancy Tests: Old and New', Discussion Paper 8325, Economics Department, University of Southampton.

Andrews, D. W. K. (1993). 'Tests for Parameter Instability and Structural Change with Unknown Change Point', *Econometrica*, 61: 821–856.

——— (2003). 'End-of-Sample Instability Tests', *Econometrica*, 71: 1661–1694.

Andrews, D. W. K., and Fair, R. C. (1988). 'Inference in Nonlinear Econometric Models with Structural Change', *The Review of Economic Studies*, 55: 615–639.

Bai, J. (1994). 'Least Squares Estimation of a Shift in Linear Processes', *Journal of Time Series Analysis*, 15: 453–472.

——— (1995). 'Least Absolute Deviation Estimation of a Shift', *Econometric Theory*, 11: 403–436.

——— (1997). ‘Estimation of a Change Point in Multiple Regression Models’, *Review of Economics and Statistics*, 79: 551–563.

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

——— (2003a). ‘Critical Values for Multiple Structural Change Tests’, *The Econometrics Journal*, 6: 72–78.

——— (2003b). ‘Multiple Structural Change Models: A Simulation Analysis’, in *Econometric Theory and Practice*, pp. 212–237. Cambridge Univ. Press, Cambridge.

Banerjee, A., and Urga, G. (2005). ‘Modeling Structural Breaks, Long Memory and Stock Market Volatility: An Overview’, *Journal of Econometrics*, 129: 1–34.

Bates, D. M., and Watts, D. (1988). *Nonlinear Regression Analysis and Its Applications.* John Wiley and Sons, New York.

Bhattacharya, P. K. (1994). ‘Some Aspects of Change-Point Analysis’, in E. Carlstein, H.-G. Müller, and D. Siegmund (eds.), *Change-Point Problems*, vol. 23 of *IMS Lecture Notes Monograph Series*, pp. 28–56. Institute of Mathematical Statististics.

Caner, M. (2007). ‘Boundedly Pivotal Structural Change Tests in Continuous Updating GMM with Strong, Weak identification and Completely Unidentified Cases’, *Journal of Econometrics*, 137: 28–67.

Csörgö, M., and Horváth, L. (1997). *Limit Theorems for Change Point Analysis.* Chichester-Wiley, Chichester.

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006). ‘Structural Break Estimation for Nonstationary Time Series Models’, *Journal of the American Statistical Association*, 101: 223–239.

Dufour, J.-M., and Ghysels, E. (1996). 'Editor's Introduction. Recent Developments in the Econometrics of Structural Change', *Journal of Econometrics*, 70: 1–8.

Eitrheim, Ø., and Teräsvirta, T. (1996). 'Testing the Adequacy of the Smooth Transition Autoregressive Models', *Journal of Econometrics*, 74: 59–75.

Elliott, G., and Müller, U. (2007). 'Confidence Sets for the Date of a Single Break in Linear Time Series Regressions', *Journal of Econometrics*, 141: 1196–1218.

Gallant, A. R. (1987). *Nonlinear Statistical Models*, Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York.

Gallant, A. R., and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Basil Blackwell, Oxford, UK.

Ghysels, E., and Hall, A. R. (1990). 'A Test for Structural Stability of Euler Conditions Parameters Estimated via the Generalized Method of Moments Estimator', *International Economic Review*, 31: 355–364.

Hall, A. R., Han, S., and Boldea, O. (2008). 'Asymptotic Distribution Theory for Break Point Estimators in Models Estimated via 2SLS', *Discussion Paper, Department of Economics, North Carolina State University.*

——— (2009). 'Inference Regarding Multiple Structural Changes in Linear Models with Endogenous Regressors', *Discussion Paper, Department of Economics, North Carolina State University.*

Hall, A. R., and Sen, A. (1999). 'Structural Stability Testing in Models Estimated by Generalized Method of Moments', *Journal of Business & Economic Statistics*, 17: 335–348.

Hansen, B. E. (2000). 'Testing for Structural Change in Conditional Models', *Journal of Econometrics*, 97: 93–115.

Krishnaiah, P. R., and Miao, B. Q. (1988). 'Review about Estimation of Change Points', in P. R. Krishnaiah and C. P. Rao (eds.), *Handbook of Statistics*, vol. 7, pp. 375–402. New York: Elsevier.

Lucas, R. (1976). 'Econometric Policy Evaluation: A Critique', in K. Brunner and A. Melzer (eds.), *The Phillips Curve and Labor Markets*, vol. 1 of *Carnegie-Rochester Conference Series on Public Policy*, pp. 19–46.

Perron, P., and Qu, Z. (2006). 'Estimating Restricted Structural Change Models', *Journal of Econometrics*, 134: 373–399.

Qu, Z., and Perron, P. (2007). 'Estimating and Testing Multiple Structural Changes in Multivariate Regressions', *Econometrica*, 75: 459–502.

Quandt, R. E. (1958). 'The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes', *Journal of the American Statistical Association*, 53: 873–880.

——— (1960). 'Tests of the Hypothesis that a Linear Regression System Obeys Two Separate Regimes', *Journal of the American Statistical Association*, 55: 324–330.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*, vol. 15 of *World Scientific Series in Computer Science.* World Scientific Publishing, Teaneck, NJ.

Sowell, F. (1996). 'Optimal Tests for Parameter Instability in the Generalized Method of Moments Framework', *Econometrica*, 64: 1085–1107.

Stock, J. H., and Watson, M. W. (2002). 'Has the Business Cycle Changed and Why?', *NBER Macroeconomics Annual*, pp. 159–218.

Vaart, v. d. A., and Wellner, J. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

van Dijk, D., Teräsvirta, T., and Franses, P. (2002). 'Smooth Transition Autoregressive Models: A Survey of Recent Developments', *Econometric Reviews*, 21: 1–47.

White, H., and Domowitz, I. (1984). 'Nonlinear Regression with Dependent Observations', *Econometrica*, 52: 143–161.

Zacks, S. (1983). 'Survey of Classical and Bayesian Approaches to the Change-Point Problem: Fixed Sample and Sequential Procedures of Testing and Estimation', in *Recent Advances in Statistics*, pp. 245–269. Academic Press, New York.

# 8 Appendix

This Appendix only contains a complete proof of Lemma 1. For the rest, an outline of the proofs is given; for complete proofs, see Supplemental Appendix. As a matter of notation, we will use $\| \cdot \|$ not only to denote the Euclidean vector norm, but also the matrix norm $\|A\| = [\text{tr}(A'A)]^{1/2}$, and $\psi_t(\theta) = u_t f_t(\theta)$, respectively $\Psi_t(\theta) = u_t F_t(\theta)$.

***Proof of Lemma 1.***

Lemma 1 translates into showing that under Assumptions 1-3, if $\Psi_T(\theta, r) = T^{-1/2} \sum_{t=1}^{[Tr]} \psi_t(\theta)$ is $O_p(1)$ uniform in $\theta$, it is also $O_p(1)$ uniform in $\theta \times r$. For simplicity, consider the case $m^* = 1$; the extension to $m^* > 1$ is immediate and omitted for simplicity.

From Assumption 1(i)-3 and Caner (2007), Lemma 1, it follows that for $r \leq \lambda_1^*$, the limit process of $T^{-1/2} \sum_{t=1}^{[Tr]} \psi_t(\theta)$ is a Kiefer process, say $K_1(\theta, r)$, restricted to $r \leq \lambda_1^*$; that is, $K_1(\theta, r)$ is the restriction to $r \leq \lambda_1^*$ of a Gaussian zero-mean process with variance covariance function $E\Psi_T(\theta_r, r)\ \Psi_T(\theta_s, s) = \min(r, s)\,\Omega(\theta_r, \theta_s)$ and

$\Omega(\theta_r, \theta_s) = \lim_{T\to\infty} E\Psi_T(\theta_r, 1)\Psi_T(\theta_s, 1)$. Hence, we have:

$$\sup_{\theta, 0<r\leq\lambda_1^*} |\Psi_T(\theta, r)| \leq \sup_{\theta, r\leq\lambda_1^*} |K_1(\theta, r)| + \xi_1 \tag{16}$$

where $\xi_1$ is an $o_p(1)$ term uniform in $\theta \times r$. If $r > \lambda_1^*$, then by Caner (2007) and Assumption 1(i)-3, the limit process of $T^{-1/2}\sum_{t=[T\lambda_1^*]+1}^{[Tr]} \psi_t(\theta)$ is a Kiefer process, say $K_2(\theta, r - \lambda_1^*)$, such that $\lambda_1^* < r \leq 1$, if $T \to \infty$. Hence:

$$\sup_{\theta, \lambda_1^*<r\leq 1} |\Psi_T(\theta, r)| \leq \sup_{\theta} |K_1(\theta, \lambda_1^*)| + \sup_{\theta, \lambda_1^*<r\leq 1} |K_2(\theta, r)| + \xi_2 \tag{17}$$

where $\xi_2$ is an $o_p(1)$ term uniform in $\theta \times r$. Thus, from (16) and (17), the desired result follows, since:

$$\sup_{\theta, 0<r\leq 1} |\Psi_T(\theta, r)| \leq \max\left\{ \sup_{\theta, r\leq\lambda_1^*} |K_1(\theta, r)|, \sup_{\theta} |K_1(\theta, \lambda_1^*)| + \sup_{\theta, \lambda_1^*<r\leq 1} |K_2(\theta, r)| \right\} + o_p(1)$$

where the $o_p(1)$ term is uniformly bounded in $\theta \times r$. It remains to show that this result also follows from Assumptions 1(ii), 2-3.

Proving Lemma 1 for $\psi_t(\theta)$ is equivalent to showing that for any $\epsilon > 0$, there exists real $\eta_\epsilon > 0, T_\epsilon > 0$ such that for $k = [Tr]$, $P[\max_{1\leq k\leq T} \sup_\theta |\Psi_T(\theta, r)| > \eta_\epsilon]$ $\leq \epsilon$. First, note that Ottaviani's inequality for independent sequences $\{\psi_t(\theta)\}$ - see van der Vaart and Wellner (1996), pp. 430 - can be rewritten as described below when letting $\lambda = \mu = \eta\sqrt{T}/2$ and provided all quantities below are measurable[12]:

$$\begin{aligned} P\left[\max_{1\leq k\leq T} \sup_{\theta\in\Theta} \left|\Psi_T(\theta, r)\right| > \eta\right] &\leq \frac{P\left[\sup_{\theta\in\Theta} \left|\Psi_T(\theta)\right| > \eta/2\right]}{1 - 2\max_{1\leq k\leq T} P\left[\sup_{\theta\in\Theta} \left|\Psi_k(\theta)\right| > \eta/4\right]} \\ &\leq \frac{\max_{1\leq k\leq T} L_k}{1 - 2\max_{1\leq k\leq T} L_k} \end{aligned} \tag{18}$$

[12] Measurability is satisfied by Assumption 2, thus we can replace outer probabilities $P^*$ in van der Vaart and Wellner (1996) with probabilities $P$.

where $\Psi_k(\theta) = k^{-1/2} \sum_{t=1}^{k} \psi_t(\theta)$ and $L_k = P[\sup_{\theta\in\Theta} |\Psi_k(\theta)| > \eta/4]$. Now consider $k > n_0$ for large enough but fixed $n_0$, the same as in Assumption 2. Then if $\Psi(\theta)$ is the mean-zero normal distribution with variance $A(\theta, 1)$ as defined in Assumption 4, then:

$$\max_{k>n_0} L_k \leq \max_{k>n_0} P \left[\sup_{\theta\in\Theta} |\Psi_k(\theta) - \Psi(\theta)| > \eta/8\right] + P \left[\sup_{\theta\in\Theta} |\Psi(\theta)| > \eta/8\right] \tag{19}$$

Since our Assumptions 1(ii), 2-3 involve a special case of the assumptions needed for the uniform CLT in $\theta$ in Gallant and White (1988), pp. 34, we have that, for any $\epsilon^* > 0$ and $\eta > 0$, $\max_{k>n_0} P\left[\sup_{\theta\in\Theta} |\Psi_k(\theta) - \Psi(\theta)| > \eta/8\right] < \epsilon^*/2$. On the other hand, since $\sup_{\theta\in\Theta} |\Psi(\theta)|$ has moments of all orders by van der Vaart and Wellner (1996), Proposition A.2.3, there is an $\eta_{\epsilon^*} > 0$ such that $P\left[\sup_{\theta\in\Theta} |\Psi(\theta)| > \eta_{\epsilon^*}/8\right] < \epsilon^*/2$. Using these and (19) into (18), we obtain, for $k^* \equiv \operatorname{argmax}_k L_k > n_0$,

$$P \left[\max_{1\leq k\leq T} \sup_{\theta\in\Theta} \left|\Psi_T(\theta, r)\right| > \eta^*\right] \leq \frac{\epsilon^*}{1 - 2\epsilon^*} \tag{20}$$

Upon defining $\epsilon^* = \epsilon/(1 + 2\epsilon)$ for any $\epsilon > 0$, noting that this definition implies $\epsilon^* < 1/2$ without restricting $\epsilon > 0$, and letting $\eta_{\epsilon^*} \equiv \eta_\epsilon$, Lemma 1 follows if $k^* > n_0$. If $k^* \leq n_0$, then, by Assumption 2(ii), $\max_{1\leq k\leq n_0} L_k \leq P(\sup_{\theta,t\leq n_0} |\psi_t(\theta)| > \eta/4) \leq 4n_0^{1/2}\eta^{-1}E[\sup_{\theta,t\leq n_0} |\psi_t(\theta)|] < M\eta^{-1}$, say. Then, for any $\epsilon^* > 0$ and $\eta = M/\epsilon^*$, (20) is satisfied. By redefining $\epsilon^*$ as in the case $k^* > n_0$, and since the two cases are mutually exclusive, the proof of Lemma 1 is complete.

□

Let $\hat{I}_i \equiv [\hat{T}_{i-1} + 1, \hat{T}_i]$ and $I_i^0 \equiv [T_{i-1}^0 + 1, T_i^0], (i = 1, \ldots, m + 1)$. To prove Lemma 2, we use the uniform law of large numbers (ULLN) in Gallant and White (1988), pp. 34. Note that their assumption of n.e.d. encompasses Assumption 1(i)-(ii):

***Proof of Lemma 2.***

*Part (i).* This part follows directly from Lemma 1.

*Part (ii).* Consider $\eta > 0$ such that $[T\eta]$ is an integer. Let $1^*$ and $2^*$ denote summing over the sets $I_1(\eta) = \{ \ [T\lambda_j^0] - T\eta + 1, \dots, [T\lambda_j^0] \ \}$, respectively $I_2(\eta) = \{ \ [T\lambda_j^0] + 1, \dots, [T\lambda_j^0] + T\eta \ \}$. If $\hat{\lambda}_j \overset{p}{\nrightarrow} \lambda_j^0$ for at least one $j$, then there is an $\eta$ such that with positive probability, $\hat{\theta}_k$ will be estimating $\theta_j^0$ on $I_1(\eta) \in \hat{I}_k$, but $\theta_{j+1}^0$ on $I_2(\eta) \in \hat{I}_k$. Hence, with positive probability greater than $\epsilon > 0$,

$$T^{-1} \sum_{t=1}^{T} d_t^2 \geq T^{-1} \sum_{1^*} d_t^2(\hat{\theta}_k, \theta_j^0) + T^{-1} \sum_{2^*} d_t^2(\hat{\theta}_k, \theta_{j+1}^0) \geq \inf_\theta H_T(\theta) \qquad (21)$$

where $d_t(\theta_A, \theta_B) = f_t(\theta_A) - f_t(\theta_B)$, with $\theta_A, \theta_B \in \Theta$, and for $i = 1, 2$, $H_{T,i}(\theta) = T^{-1} \sum_{i^*} d_t^2(\theta, \theta_{j-1+i}^0)$, and $H_T(\theta) = \sum_{i=1,2} H_{T,i}(\theta)$.

To prove $T^{-1} \sum_{t=1}^T d_t^2 > C$ with probability $> \epsilon$ and establish Lemma 2(ii), it is sufficient to prove uniform convergence in $\theta$ of $H_T(\theta)$ to a positive quantity $H(\theta)$. Uniform convergence can be established using the ULLN mentioned above, under Assumptions 1-4. It remains to show that $\inf_\theta H(\theta) > 0$. This can be establish by showing - see Supplemental Appendix:

$$E[H_T(\theta)] \geq \|\theta_j^0 - \theta_{j+1}^0\|^2 tr \left\{ \inf_t \inf_\theta E[F_t(\theta) F_t'(\theta)] \right\} > C$$

where the last inequality follows from Assumption 4(iii). □

***Proof of Theorem 2.***

The proof follows in three steps. The first step redefines the proof objective and introduces some notation. In the second step two distinct terms are analyzed and compared to finalize the proof.

***Step 1.*** As in Bai and Perron (1998), without loss of generality, we assume only three breaks. We will focus on proving Theorem 2 for $\hat{\lambda}_2$; the analyses for $\hat{\lambda}_1$ and $\hat{\lambda}_3$ are similar. For any $\epsilon > 0$, define $V_\epsilon = \{(T_1, T_2, T_3) : \ | T_i - T_i^0 \ |\leq \epsilon T \ (i = 1, 2, 3)\}$. Since $\hat{\lambda}_i \xrightarrow{p} \lambda_i^0$, $\lim P\{(\hat{T}_1, \hat{T}_2, \hat{T}_3) \in V_\epsilon\} = 1$. Hence, we need only examine the behavior of breakpoints contained in $V_\epsilon$. Consider, without loss of generality, the case $\hat{T}_2 < T_2^0$; the case $\hat{T}_2 \geq T_2^0$ can be handled by a symmetric argument. For

$C > 0$, define: $V_\epsilon(C) = \{(T_1, T_2, T_3) : \mid T_i - T_i^0 \mid\leq \epsilon T \ (i = 1, 2, 3);\ T_2^0 - T_2 > C\}$. Note that $V_\epsilon(C) \subset V_\epsilon$. We will show that the probability that the break-points are contained in $V_\epsilon(C)$ is very small. Hence, with large probability, $|\hat{T}_i - T_i^0| < C$, for $i = 1, 2, 3$, confirming the content of Theorem 2. So, for proving the latter, it suffices to show that the break-points will not be contained in $V_\epsilon(C)$ with large probability.

To that end, denote by $S_T(T_1, T_2, T_3)$ the minimized sum of squared residuals for a given 3-break-partition $(1, T_1, T_2, T_3, T)$ of the sample interval. By definition of minimized sum of squared residuals, $S_T(\hat{T}_1, \hat{T}_2, \hat{T}_3) \leq S_T(\hat{T}_1, T_2^0, \hat{T}_3)$. Let $\Delta_2 = T_2 - T_2^0$. We will show that for any $\eta > 0$, we can pick $\epsilon$ and $C$ such that:

$$P\left\{\min_{V_\epsilon(C)} (\Delta_2)^{-1}[S_T(T_1, T_2, T_3) - S_T(T_1, T_2^0, T_3)] < 0\right\} < \eta, \text{ for } T \geq T(\eta). \quad (22)$$

Equation (22) implies that for large $T$, with probability $\geq 1 - \eta$, $S_T(\hat{T}_1, \hat{T}_2, \hat{T}_3)$ $> S_T(\hat{T}_1, T_2^0, \hat{T}_3)$, contradicting the sum of squares minimization definition; thus, $\hat{T}_2 \notin V_\epsilon(C)$, completing the proof.

Define $SSR_1 = S_T(T_1, T_2, T_3)$, $SSR_2 = S_T(T_1, T_2^0, T_3)$ and introduce $SSR_3 = S_T(T_1, T_2, T_2^0, T_3)$. Then $S_T(T_1, T_2, T_3) - S_T(T_1, T_2^0, T_3) = (SSR_1 - SSR_3) - (SSR_2 - SSR_3)$. This approach helps carry out the analysis in terms of two problems involving a single structural change: the first imposing an additional break at $T_2^0$ between $T_2$ and $T_3$, and the second introducing an additional break at $T_2$ between $T_1$ and $T_2^0$. Let $(\theta_1^*, \theta_2^*, \theta_3^{**}, \theta_4^*)$, $(\theta_1^*, \theta_2^{**}, \theta_3^*, \theta_4^*)$ and $(\theta_1^*, \theta_2^*, \theta_2^\delta, \theta_3^*, \theta_4^*)$ be the NLS parameter estimates based on partitions $(1, T_1, T_2, T_3, T)$, respectively $(1, T_1, T_2^0, T_3, T)$ and $(1, T_1, T_2, T_2^0, T_3, T)$. Note that $\theta_2^*, \theta_2^\delta, \theta_2^{**}$ are all estimating $\theta_2^0$, while $\theta_3^*, \theta_3^{**}$ are both estimators of $\theta_3^0$.

In the light of proving (22), we need to locate the dominating terms in $(SSR_1 - SSR_2)$ and show that we can pick $\epsilon$ and $C$ such they are positive with large probability for large $T$. To that end, let $V_\epsilon(C)$ be the domain on which some quantity $q_T(\cdot)$ is defined. We will denote $q_T \sim O_p(T^b)$ $P(|q_T| > T^b) < \bar{\eta}$ for

$T \geq T(\bar{\eta})$ for some $b \in \mathbb{R}$ and any $\bar{\eta} > 0$, where $T$ as defined here is large. Note that the statement above depends on the choice of $C$ and $\epsilon$. We will write $q_T \sim O_p^+(T^b)$ if plim $q_T$ is positive (or positive definite for matrices). Similarly, let $q_T \sim O_p(T^b) + a_T$, if $q_T - a_T \sim O_p(T^b)$ for some $a_T$, and $q_T \sim O_p^+(T^b) + a_T$, if $q_T - a_T \sim O_p^+(T^b)$. Under this notation, equation (22) is equivalent to:

$$\Delta_2^{-1}(SSR_1 - SSR_2) \sim O_p^+(1) \tag{23}$$

because then the probability that $(SSR_1 - SSR_2)$ is negative is small. So, for proving Theorem 2, a proof of (23) suffices.

***Step 2:*** To further simplify the notation, let $I_1 = [1, T_1]$, $I_2 = [T_1 + 1, T_2]$, $I_2^\Delta = [T_2 + 1, T_2^0]$, $I_3 = [T_2^0 + 1, T_3]$, $I_4 = [T_3 + 1, T]$. Recall that $\Delta_2 = T_2^0 - T_2 > C$, and denote $e_t^2(\theta_A, \theta_B) \equiv u_t^2(\theta_A) - u_t^2(\theta_B)$. Consider $SSR_1 - SSR_3$ first:

$$\Delta_2^{-1}(SSR_1 - SSR_3) = \Delta_2^{-1} \sum_{I_2^\Delta} e_t^2(\theta_3^{**}, \theta_2^\delta) + \Delta_2^{-1} \sum_{I_3} e_t^2(\theta_3^{**}, \theta_3^*) = D_1 + D_2.$$

Heuristically speaking, $D_1$ involves a "mismatch" in estimators, because $\theta_3^{**}$ is estimating $\theta_3^0$, while $\theta_2^\delta$ is estimating $\theta_2^0$. This "mismatch" is not present in $D_2$, because $\theta_3^{**}$ and $\theta_3^*$ are both estimating $\theta_3^0$. Hence, $D_1$ should be dominating $D_2$ for a large enough $\Delta_2 > C$. To see this, note that,for $i = 1, \ldots, 4$, in an interval where $\theta_i^0$ is the true parameter value, and $\theta \in \Theta$, it can be shown that: $u_t^2(\theta) - u_t^2 = d_t^2(\theta, \theta_i^0) - 2u_t\, d_t(\theta, \theta_i^0)$. Also, the true parameter value on $I_2^\Delta$ is $\theta_2^0$. Then for any $\theta_A, \theta_B \in \Theta$ and $t \in I_2^\Delta$, $e_t^2(\theta_A, \theta_B) = d_t^2(\theta_A, \theta_2^0) - d_t^2(\theta_B, \theta_2^0) - 2u_t\, d_t(\theta_A, \theta_B)$. According to the above, we have:

$$D_1 = \Delta_2^{-1} \sum_{I_2^\Delta} d_t^2(\theta_3^{**}, \theta_2^0) - \Delta_2^{-1} \sum_{I_2^\Delta} d_t^2(\theta_2^\delta, \theta_2^0) + 2\Delta_2^{-1} \sum_{I_2^\Delta} u_t d_t(\theta_2^\delta, \theta_3^{**}) = \sum_{j=1}^{3} D_{1,j}$$

We will find the order of each of the terms above. In the proof of Lemma 2, we have shown that processes such as $\{d_t^2(\theta, \theta_2^0)\}$ satisfy the ULLN. In other words, if we pick $C$ large enough, $D_{1,1} \sim$ plim$\Delta_2^{-1} \sum_{I_2^\Delta} [d_t^2(\theta_3^{**}, \theta_2^0)] + o_p(1)$. To find this limit, note - from the supplemental Appendix - that $\theta_3^{**} - \theta_3^0 \sim O_p(T^{-1/2})$. So, by similar arguments as in the proof of Lemma 2(ii), we obtain:

$$D_{1,1} = \Delta_2^{-1} \sum_{I_2^\Delta} d_t^2(\theta_3^{**}, \theta_2^0) \ \sim O_p^+(1).$$

This will be the only positive dominating term in $SSR_1 - SSR_2$. For analyzing $D_{1,2}$, if we pick $C$ big enough, $\theta_2^\delta - \theta_2^0 \sim o_p(1)$. Hence, $D_{1,2} \sim o_p(1)$. Also, $D_{1,3} \sim o_p(1)$ by Lemma 1. It follows that for large $C$ and small $\epsilon$, $D_1 \sim O_p^+(1) - o_p(1) + o_p(1) = O_p^+(1)$.

Note that $D_2$ is different than $D_1$ given that we are summing over a different interval. For deriving the order of $D_2$, we have to consider two cases, $T_3 < T_3^0$ and $T_3 \geq T_3^0$ - see supplemental Appendix. For both cases, $D_2 \sim C^{-1}O_p(1)$. Since $D_1$ and $D_2$ determine the order of $SSR_1 - SSR_3$, for small $\epsilon$ and large $C$,$\Delta_2^{-1}(SSR_1 - SSR_3) = D_1 + D_2 \sim O_p^+(1) + C^{-1}O_p(1) = O_p^+(1)$. By similar arguments as for $D_2$, it can be shown that $\Delta_2^{-1}(SSR_2 - SSR_3) = C^{-1}O_p(1)$, if we pick $C$ large enough and $\epsilon$ small enough. Hence, $\Delta_2^{-1}(SSR_1 - SSR_2) \sim O_p^+(1) - C^{-1}O_p(1) = O_p^+(1)$, provided that $C$ is large enough and $\epsilon$ small enough, for large $T$. This is in fact (23), completing the proof. □

***Proof of Theorem 3***.

As usual for nonlinear consistency proofs, we need to show uniform convergence of the minimand, and then use uniqueness to establish consistency of parameter estimates. As a matter of notation, consider some partition of the interval $[1, T]$, denoted $(1, T_1, \ldots, T_m, T)$. Let $S_{T,I_i}(\theta) = T^{-1} \sum_{t=T_{i-1}}^{T_i} u_t^2(\theta)$ be the partial sum of squares in interval $I_i = [T_{i-1}+1, T_i]$, for $i = 1, \ldots, m+1$, and let $I_i^0 = [T_{i-1}+1, T_i]$, respectively $\hat{I}_i = [\hat{T}_{i-1} + 1, \hat{T}_i]$. Moreover, let $\hat{I}_i \ \nabla \ I_i^0 = (\hat{I}_i \setminus I_i^0) \cup (I_i^0 \setminus \hat{I}_i)$, and define as indicator function $\iota_i : \hat{I}_i \ \nabla \ I_i^0 \rightarrow \{-1, 1\}$, where $\iota_i(t) = \iota_{i,t} = 1$, if $t \in \hat{I}_i \setminus I_i^0$, and $\iota_{i,t} = -1$, if $t \in I_i^0 \setminus \hat{I}_i$. Then $S_{T,\hat{I}_i}(\theta) - S_{T,I_i^0}(\theta)$ is =

$\sum_{\hat{I}_i \,\nabla\, I_i^0} \iota_{i,t}\,[T^{-1}u_t^2] + \sum_{\hat{I}_i \,\nabla\, I_i^0} \iota_{i,t}\,[T^{-1}d_t^2(\theta,\theta_i^0)] + \sum_{\hat{I}_i \,\nabla\, I_i^0} \iota_{i,t}\,[T^{-1}2u_t\, d_t(\theta,\theta_i^0)]$. By Theorem 2, there can be no more than $2C$ integer values contained in $\hat{I}_i \,\nabla\, I_i^0$. By ULLN, $S_{T,\hat{I}_i}(\theta) - S_{T,I_i^0}(\theta) = o_p(1)$. Since we replaced the estimated break-points with the true breaks, standard nonlinear analysis tells us that under Assumptions 1-4, $\hat{\theta}_i \xrightarrow{p} \theta_i^0$, for $i = 1,\ldots,m$. One can also show - see Supplemental Appendix - that mean value expansions $T^{1/2}\,\partial S_{T,\hat{I}_i}/\partial\theta$ around $\theta_i^0$ are uniformly within $o_p(1)$ of the mean-value expansions using the true break-point estimates. Hence, standard nonlinear asymptotics shows that $\hat{\theta}_i$ have indeed the distribution given in Theorem 3. Asymptotic independence of $\hat{\theta}_i$ and $\hat{\theta}_j$ for $i \neq j$ follows from Assumption 1, completing the proof. □

***Proof of Theorem 4.***

The distribution of $\hat{k}$ depends on the distribution of $\operatorname{argmin}_{\theta_1,\theta_2}\, V_T(k,\theta_1,\theta_2)$. Assume $k < k_0$; the case $k \geq k_0$ can be handled similarly.

$$\begin{aligned} V_T(k,\hat{\theta}_1(k),\hat{\theta}_2(k)) \;&=\; \sum_{t=1}^{k}[u_t^2(\hat{\theta}_1(k)) - u_t^2(\theta_1^0)] \;+\; \sum_{t=k+1}^{k_0}[u_t^2(\hat{\theta}_2(k)) - u_t^2(\theta_1^0)] + \\ &+\; \sum_{t=k_0+1}^{T}[u_t^2(\hat{\theta}_2(k)) - u_t^2(\theta_2^0)] = \Sigma_1 + \Sigma_2 + \Sigma_3. \end{aligned} \tag{24}$$

Since we know the convergence rates of $\hat{k}$ and $\hat{\theta}_i(k)$, the minimization problem is defined over a neighborhood of $(k,\theta_1,\theta_2)$. Note that the asymptotic distributions of $\Sigma_1$ and $\Sigma_3$ do not depend on $v$, since the difference between the summations involving the true breaks and the estimated breaks is asymptotically negligible, uniformly in $v$. Hence, we can write $V_T(k,\hat{\theta}_1(k),\hat{\theta}_2(k)) = \mathfrak{D} + \Sigma_2 + o_p(1)$, where the $o_p(1)$ term is uniform in $v$. On the other hand, it can be shown that:

$$\Sigma_2 = \sum_{t=k+1}^{k_0} d_t^2(\theta_2^0,\theta_1^0) + \sum_{t=k+1}^{k_0} u_t\, d_t(\theta_2^0,\theta_1^0) + o_p(1)$$

with the $o_p(1)$ term uniform in $v$. Continuity of $f_t(\theta)$ guarantees that the maximum of $J^*(v)$ is unique almost surely, and we can use CMT to express the distribution of $\hat{k}$ as stated in Theorem 4. □

To prove Theorem 5, we need to show consistency of the break-fractions at a certain rate, as well as asymptotic normality of parameter estimates. Consistency is summarized by the following theorem.

**Theorem A 1.** *Under Assumptions 1-5 and 8, $\hat{\lambda}_i \xrightarrow{p} \lambda_i^0$, for $i = 1, \ldots, m$.*

***Proof of Theorem A1.***

The proof of Theorem A1 is similar to that of Theorem 1, but modifications are required to avoid the possibility that $T^{-1}\sum_{t=1}^{T} d_t^2 \xrightarrow{p} 0$ even if a break-fraction is not consistently estimated. Under Assumptions 1-5 or Lemma 1 and Assumptions 2-5, we have: $\sum_{t=1}^{T} u_t d_t \leq O_p(T^{1/2+\nu})$, uniformly over the space of all partitions and parameters $(T_1, \ldots, T_m) \times \theta$, with $\nu \geq 0$. On the other hand, by arguments similar to before, if at least one break-fraction is not consistently estimated, $\sum_{t=1}^{T} d_t^2 \geq \|\theta_j^0 - \theta_{j+1}^0\|^2 O_P(T) > CTw_T^2$. By Assumption 8, this term dominates $2T^{-1}\sum_{t=1}^{T} d_t u_t$, and $T^{-1}\sum_{t=1}^{T} d_t^2 + 2T^{-1}\sum_{t=1}^{T} d_t u_t \leq 0 \xrightarrow{p} \infty$. The latter contradicts equation (5), thus the break-points are consistent. □

Next, we state the rate of convergence for the break-fractions:

**Theorem A 2.** *Under Assumptions 1-5 and 8, for any $\eta > 0$, there is a $C > 0$ such that, for large $T$, $P(Tw_T^2|\hat{\lambda}_k - \lambda_k^0| > C) < \eta$, for any $k = 1, \ldots, m$.*

***Proof of Theorem A2.***

The proof of Theorem A2 proceeds in the same fashion as the proof of Theorem 2, except for convergence rates which are different given shrinking shifts; see Supplemental Appendix for proof. □

**Theorem A 3.** *Under Assumptions 1-5 and 8, $T^{1/2}(\hat{\theta} - \theta^0) \xrightarrow{d} \mathcal{N}(0, \Phi_i(\theta_i^0))$.*

***Proof of Theorem A3.*** The Proof of Theorem A3 is similar to that of Theorem 3 and can be found in the Supplemental Appendix. □

***Proof of Theorem 5***. Let $k < k_0$, the proof for $k \geq k_0$ is similar. Also let $v = k_0 - k$, $0 < v \leq C/v_T^2$; by similar arguments as for fixed shifts, using Theorems A1-A3, $V_T(k, \hat{\theta}_1(k), \hat{\theta}_2(k)) = \mathfrak{D} + o_p(1) + \Sigma_2$, where the $o_p(1)$ term is uniform in $v$ and $\mathfrak{D}$ is a distribution that does not depend on $v$. So, even in this case, $\Sigma_2$ will govern the distribution of the minimand for shrinking shifts. It can be shown that, uniformly in $v$, $\Sigma_2 = |v|\varpi_{2,1} - 2\varpi_{1,1}^{1/2} W_1(-v) + o_p(1)$, for $v \leq 0$. Since $C/v_T^2 \to \infty$, it follows that:

$$\hat{k} - k_0 = \underset{v \leq 0}{\operatorname{argmax}} \left[ \varpi_{1,1}^{1/2} W_1(-v) - 0.5|v|\varpi_{2,1} \right] + o_p(1) \tag{25}$$

Note that the limiting Brownian motions can only be obtained under Assumption 6(iii)-(iv), that is, when $\{u_t F_t(\theta)\}$ is second-order stationary within regimes, and $F_t(\theta)$ as well. Breaks in the variance of regressors are excluded, unless they coincide with the true value. By a change in variable in (25) - see Supplemental Appendix, we obtain the desired result. □

To prove Theorem 6, we need two additional Theorems. Denote by $\hat{\theta}_i$ and $\hat{\theta}_{1,i}$ the $[T_{i-1}+1, T_i]$, respectively the $[1, T_i]$- sub-sample estimators of $\theta^0$ where $T_i$ is the $i-th$ break belonging to a certain partition $\bar{T}^k$ on which $\hat{\theta}_i$ were defined as well. Then:

**Theorem A 4.** *Under Assumptions 2-6 and* $H_0 : m = 0$,
$T^{1/2}(\hat{\theta}_{1,i} - \theta^0) \Rightarrow \sigma \lambda_i^{-1} D^{-1/2}(\theta^0) B_p(\lambda_i)$, *where* $D(\theta)$ *is the common value of* $D_i(\theta)$ *in Assumption 4(iii), under* $H_0$.

**Theorem A 5.** *Under Assumptions 2-6 and* $H_0 : m = 0$,
$T^{1/2}(\hat{\theta}_i - \theta^0) \Rightarrow \sigma[\lambda_i - \lambda_{i-1}]^{-1} D^{-1/2}(\theta^0) \left[B_p(\lambda_i) - B_p(\lambda_{i-1})\right]$.

***Proof of Theorem A4.***
First, $\hat{\theta}_{1,i} \xrightarrow{p} \theta^0$ because it is just a sub-sample NLS estimator of $\theta^0$ in stable models. Using the mean value theorem, the desired result follows from Assumptions 2,3,4 and 6. The latter is essential for the limit to be a Brownian motion; thus, no

breaks in the variance of regressors and errors are allowed. The proof of Theorem A5 follows the same steps and is omitted for simplicity. □

***Proof of Theorem 6***.

First, under Assumptions 2-6 and $H_0$, $SSR_k/(T-(k+1)p) \xrightarrow{p} \sigma^2$, an immediate consequence of Lemma 2. On the other hand, it can be shown:

$$SSR_0 - SSR_k = \sum_{i=1}^{k} F^*_{T,i}, \text{with } F^*_{T,i} = D^R(1,i+1) - D^R(1,i) - D^U(i+1,i+1)$$

where the sum subscript $1,i$ indicates summing over interval $[1,T_i]$, while $i$ indicates, as before, summing over $[T_{i-1}+1,T_i]$, and $D^R(1,i) = \sum_{1,i}[u_t^2(\hat{\theta}_{1,i}) - u_t^2]$ and $D^U(i,i) = \sum_i[u_t^2(\hat{\theta}_i) - u_t^2]$. Using the last two theorems, it can be shown - see Supplemental Appendix - that under Assumptions 2-6, $D^R(1,i) \Rightarrow -\sigma^2\|B_p(\lambda_i)\|^2/\lambda_i$, $D^R(1,i+1) \Rightarrow -\sigma^2\|B_p(\lambda_{i+1})\|^2/\lambda_{i+1}$ and $D^U(i+1,i+1) \Rightarrow -\sigma^2\|B_p(\lambda_{i+1}) - B_p(\lambda_i)\|^2 \ /[\lambda_{i+1} - \lambda_i]$, yielding:

$$F^*_{T,i} \Rightarrow \sigma^2 \frac{\|\lambda_i B_p(\lambda_{i+1}) - \lambda_{i+1} B_p(\lambda_i)\|^2}{\lambda_i \lambda_{i+1}[\lambda_{i+1} - \lambda_i]}$$

□

***Proof of Theorem 7***.

Under $H_0 : m = \ell$, compute the estimated break-points, and let $SSR(\hat{T}_i, \hat{T}_j)$ be the minimized sum of squared residuals for the segment containing observations in the interval $[\hat{T}_i + 1, \hat{T}_j], i < j$. We can write:

$$F_T(\ell+1|\ell) = \max_{1\leq i\leq \ell} \sup_{\tau\in\Delta_{i,\eta}} F^*_{T,i}(\ell+1|\ell)/\hat{\sigma}_i^2, \tag{26}$$

$$\text{where } F^*_{T,i}(\ell+1|\ell) = SSR(\hat{T}_{i-1}, \hat{T}_i) - SSR(\hat{T}_{i-1}, \tau) - SSR(\tau, \hat{T}_i).$$

Using similar arguments to the previous theorem - see Supplemental Appendix:

$$\frac{F^*_{T,i}(\ell+1|\ell)}{\sigma_i^2} \Rightarrow \left[ sup_{\eta \le \mu \le 1-\eta} \frac{\|B_p(\mu) - \mu B_p(1)\|^2}{\mu(1-\mu)} \right]. \tag{27}$$

Since the regimes considered in $SSR(\cdot,\cdot)$ are non-overlapping, $F^*_{T,i}(\ell+1|\ell)$ are asymptotically independent for different $i$ by Assumption 6. Hence, the result in Theorem 7. □

***Proof of Theorem 8***. Recall that $H_0 : R_k\,\theta_0^c = 0$, implying that $\theta_1^0 = \ldots = \theta_{k+1}^0 = \theta^0$. Let $\Delta\lambda_i = \lambda_i - \lambda_{i-1}$, for $i = 1, \ldots, k+1$. By the uniform convergence statements in Assumption 6(iii) and (iv), it follows that $\hat{D}_i(\hat{\theta}_i(\bar{T}_k)) \xrightarrow{p} \Delta\lambda_i\, D(\theta^0)$ and $\hat{A}_i(\hat{\theta}_i(\bar{T}_k)) \xrightarrow{p} \Delta\lambda_i\, A(\theta^0)$, where $D(\cdot), A(\cdot)$ are the common value of $D_i(\cdot)$, respectively $A_i(\cdot)$ under $H_0$. For simplicity, let $A(\theta^0) \equiv A_0$ and $D(\theta^0) \equiv D_0$. Then:

$$T\hat{\Upsilon}(\bar{T}_k) \xrightarrow{p} [C_k^{-1} \otimes D_0^{-1}] \times [C_k \otimes A_0] \times [C_k^{-1} \otimes D_0^{-1}]$$
$$T^{1/2}(\hat{\theta}_i(\bar{T}_k) - \theta^0) \Rightarrow (\Delta\lambda_i)^{-1} D_0^{-1} A_0^{1/2}\, [B_p(\lambda_i) - B_p(\lambda_{i-1})]$$

Putting the last two equations together - see Supplemental Appendix - completes the proof of Theorem 8. The proof of ***Theorem 9*** follows similar steps and can be found in the Supplemental Appendix. □