ISSN 1440-771X

MONASH University

Australia

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

STRUCTURAL-BREAK MODELS UNDER MIS-SPECIFICATION: IMPLICATIONS FOR FORECASTING

Bonsoo Koo and Myung Hwan Seo

April 2013

Working Paper 11/13
(Revised Version 08/13)

Structural-break models under mis-specification: implications for forecasting*

Bonsoo Koo† Monash University

Myung Hwan Seo‡ London School of Economics

April 2013

Abstract

This paper revisits the least squares estimator of the linear regression with a structural break. We view the model as an approximation to the true data generating process whose exact nature is unknown but perhaps changing over time either continuously or with some jumps. This view is widely held in the forecasting literature and under this view, the time series dependence property of all the observed variables is unstable as well. We establish that the rate of convergence of the estimator to a properly defined limit is much slower than the standard super consistent rate, even slower than the square root of the sample size T and as slow as the cube root of T. We also provide an asymptotic distribution of the estimator and that of the Gaussian quasi likelihood ratio statistic for a certain class of true data generating process. We relate our finding to current forecast combination methods and bagging and propose a new averaging scheme. The performance of various contemporary forecasting methods is compared to ours using a number of macroeconomic data.

Key words: structural break, forecasting, mis-specification, cube-root asymptotics, bagging.

Journal of Economic Literature Classification: C13, C22, C53

*We thank Andreas Pick and Jing Tian for sharing their codes with us. Their codes are of great help for conducting our applications. We thank Taya Dumrongrittikul for her excellent research assiatance.

†Department of Econometrics and Business Statistics, Monash University, PO Box 11E, Clayton Campus, VIC 3800, Australia; e-mail: bonsoo.koo@monash.edu

‡Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, United Kingdom; e-mail: M.Seo@lse.ac.uk

1 Introduction

Structural breaks have been observed in many economic time series and economic models (Stock and Watson, 1996). Documented examples include interest rates (Garcia and Perron, 1996), GDP (Ben-David and Pappel 1998, McConnell and Perez-Quiros, 2000) and labour productivity (Hansen, 2001). Consequently, various aspects of econometric analyses of structural break models have been investigated throughout the literature. For relevant surveys, see Bhattacharya (1994), Stock (1994), van Dijk *et al.* (2002) and Perron (2006).

A break, also known as a change-point, is often associated with a change in parameter values of the underlying regression model along an observable variable and the change involves a jump or discontinuity in the regression function. As observed by Bai and Perron (1998), Hansen (2000), and Perron and Qu (2006), among many others, this generates a convenient oracle property: an estimator for the location of the break converges to a true break point faster than estimators for other parameters converge, and they are asymptotically independent of each other. This means that distribution theory for structural break models can be established as if break dates were known *a priori*, once break dates are consistently estimated. This property extends to nonparametric regression models (Delgado and Hidalgo, 2000) and to cointegration models (Seo, 2011). Consequently, standard estimation and inference procedures involve estimation of break dates, followed by estimation of, and inference for, other model parameters conditional on these estimated change-points. To the best of our knowledge, all existing distributional theory pertaining to structural break models is based on the assumption of knowledge of the correct specification of the true data generating process (DGP).

In practice, however, we are not certain whether structural break models correctly specify the true underlying DGP or not. Therefore, economic and statistical models for structural breaks are subject to mis-specification. We show that the oracle property is unlikely to hold for structural-break models under possible model mis-specification. No asymptotic theory, such as the rate of convergence, and limiting distributions of estimators, is yet available under these conditions. Our establishment of this theory thus constitutes a significant contribution to the literature. In the presence of model uncertainty, estimation of break dates has an influence on the estimation and inference of all remaining unknown parameters in structural break models, meaning the break dates cannot be treated as known *a priori*. Instead, all parameters in structural break models (including break dates) should be evaluated jointly.

Indeed, many works on forecasting with structural break models have viewed the structural instability in economic time series and models as ongoing and its exact nature as unknown. For instance, see Clements and Hendry (1998) and a recent review by Rossi (2012), and other works cited in this paragraph. Under the premise of model uncertainty, it is often argued that forecasting based solely on the post-break data is not necessarily optimal and alternative methods could outperform the traditional post-break forecasting method. Furthermore, the jump point in the estimated break model does not need to be a discontinuity point of the constantly changing true data generating scheme. Then, the literature naturally considered certain forecast averaging methods as in Pesaran and Timmerman (2007), for example. In

particular, they include averaging over different estimation windows (Pesaran and Pick, 2011), the optimal and robust weighting forecasting approach (Pesaran *et al.* 2011), reverse ordered CUSUM weighting, and more weighting on the recent data (Tian and Anderson, 2012), and weighted averaging between the models with and without breaks based on a Mallows criterion (Hansen, 2009).

The objectives of this paper are the following. First, we establish asymptotic theories for the least squares estimator of the regression model with a structural break, under the assumption that we do not know the true regression function. Our setup allows for a time-varying system such as a time-varying coefficient model of Robinson (1989) and locally stationary models, although our conditions are not restricted to these cases only. In particular, we show that the estimator converges at a slower rate than $T^{-1/2}$, where T is the sample size, and can be as slow as $T^{-1/3}$, and that the estimator can have a well-defined asymptotic distribution under certain conditions. Second, we shed light on contemporary averaging forecasting methods and develop a simple averaging forecasting procedure incorporating our new distribution theory. In so doing, we provide a rationale behind why recent forecasting procedures based on averaging schemes could yield better results. While various averaging forecasting methods have emerged, very little asymptotic distribution theory has been established. Our finding explains why averaging schemes could perform better in the presence of possible model mis-specification. This paper also proposes a procedure for enhancing the forecasting performance of the structural-break model. Specifically, our distribution theory provides us with a range over which averaging is performed.

One way to think about our forecasting method is in the context of *bootstrap aggregating* (Bagging), which was introduced by Breiman (1996) . It demonstrated that the aggregation improves the forecast by reducing its sampling variation when the underlying model is *unstable*. This instability is reflected in our asymptotic result in terms of slower convergence rates of the estimates. This is in line with Bühlmann and Yu (2002) and extends their analysis to dependent data. On the other hand, the standard asymptotic experiment does not capture this feature well due to the super-consistency of the break estimate. However, we do not attempt the bootstrap as it cannot estimate the asymptotic distribution of the estimator under the cube-root type asymptotics as shown by e.g. Abrevaya and Huang (2005).

Yet another way to understand our approach is to consider the well-known trade-off between bias and variance. Averaging approaches adopted in most recent literature are based on the trade-off between bias and variance of a given model with respect to its forecasting performance. When the model accommodates various features of the underlying true process, the prediction bias could be smaller but the prediction variance may be larger. Conversely, if the model is too simple and therefore fails to capture core features of the true process, the prediction variance could be smaller but the prediction bias could be substantial. A balance between parsimony and flexibility is thus required in order to minimize a combination of the prediction bias and variance. This latter motivation for averaging methods is not inconsistent with motivating our strategy via the instability of the prediction method. Rather, averaging performs better in terms of the bias/variance trade-off, but this trade-off could result from instability in the

prediction method. Our approach is thus relevant to both of these empirical trends within the literature.

Lastly, we provide a comparison of studies utilizing various methods, including ours where possible. It is worth highlighting that our approach is similar to other averaging methods mentioned above as it also uses averaging. However, our approach is differentiated by the basing of our strategy on newly developed distribution theory. Moreover, this method is easy to apply.[1] We provide two empirical studies, one for autoregressive models and the other for a leading indicator regression model. Our empirical studies lend clear support to our method. We show this by utilizing macroeconomic datasets (from Stock and Watson (1996) and publicly available from G7 countries) widely used in the time series literature. We compare the strategy and empirical results found through our research with results produced by other methods. We show that our forecasting procedure leads to a reduction in forecasting error of approximately 20% in most cases, compared with the approach where the oracle property is used. In addition, our comparison study with other contemporary averaging methods confirms that our forecasting method performs better than other methods in most cases, which lends strong support to the validity of our forecasting method.

The remainder of this paper is organized as follows. Section 2 introduces the model and the estimation procedure for the unknown parameters. Section 3 develops distribution theories along with related assumptions. We then propose our forecasting method which incorporates our newly developed distribution theory in Section 4. Empirical application of our forecasting procedure to various time series data is provided in Section 4.2. Section 5 concludes. The mathematical proofs are relegated to the Appendix.

2 Structural Break Model under Mis-specification

This section revisits the classical linear regression model with a possible break. That is,

$$y_{t,T} = x'_{t,T}\beta_1 1(\tau \leq \gamma) + x'_{t,T}\beta_2 1(\tau > \gamma) + e_{t,T} \tag{1}$$

where $1(\cdot)$ is an indicator function, $\tau = t/T$ and unknown parameters $\theta = (\beta', \gamma)' \in \Theta$ with $\beta = (\beta'_1, \beta'_2)'$. In particular, $\gamma \in \Gamma$, which is a closed interval in $(0, 1)$. The regressors, $x_{t,T} \in \mathbb{R}^p$, may contain lags of the dependent variable and lagged explanatory variables. The array notation is used to allow for general types of processes. Prime denotes transpose. Some elements of β_2 could be zero. When there are no zeros in β_2, this is a pure structural break model. Let $\delta = \beta_2 - \beta_1$. If $\delta = 0$, the parameter γ is not identified.

The standard least squares estimator $\hat{\theta}$ of θ minimizes the sum of squared residuals

$$\begin{aligned} \mathbb{S}_T(\theta) &= \frac{1}{T}\sum_{t=1}^{T}[y_{t,T} - x'_{t,T}\beta_1 1(\tau \leq \gamma) - x'_{t,T}\beta_2 1(\tau > \gamma)]^2 \qquad (2) \\ &= \frac{1}{T}\sum_{t=1}^{\lfloor \gamma T \rfloor}(y_{t,T} - x'_{t,T}\beta_1)^2 + \frac{1}{T}\sum_{t=\lfloor \gamma T \rfloor + 1}^{T}(y_{t,T} - x'_{t,T}\beta_2)^2, \end{aligned}$$

[1] Programme codes (Matlab or GAUSS) are available upon request from the authors.

where $\lfloor t \rfloor$ is the biggest integer less than or equal to t. That is,

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \, \mathbb{S}_T(\theta).$$

For a given $\gamma \in \Gamma$, we can obtain the concentrated sum of squared residuals,

$$\begin{aligned} \mathbb{S}_T(\gamma) &\equiv \mathbb{S}_T(\hat{\beta}_1(\gamma), \hat{\beta}_2(\gamma), \gamma) \\ &= \frac{1}{T} \sum_{t=1}^{\lfloor \gamma T \rfloor} (y_{t,T} - x'_{t,T} \hat{\beta}_1(\gamma))^2 + \frac{1}{T} \sum_{t=\lfloor \gamma T \rfloor + 1}^{T} (y_{t,T} - x'_{t,T} \hat{\beta}_2(\gamma))^2, \end{aligned} \tag{3}$$

where $\hat{\beta}_1(\gamma)$ and $\hat{\beta}_2(\gamma)$ are the OLS estimates in the two subsamples. Then,

$$\hat{\gamma} = \underset{\gamma \in \Gamma}{\operatorname{argmin}} \, \mathbb{S}_T(\gamma). \tag{4}$$

In fact, $\hat{\gamma}$ is given as an interval as $\mathbb{S}_T(\gamma)$ is a step function and conventionally $\hat{\gamma}$ is defined as the minimum of the interval. Then,

$$\hat{\theta} = \left(\hat{\beta}_1(\hat{\gamma}), \hat{\beta}_2(\hat{\gamma}), \hat{\gamma} \right).$$

In our mis-specification analysis, we allow for instability of the regression function over time not to mention the non-linearity of the regression function, which is often the focus in the classical robust inference study for the OLS. There are several possibilities of mis-specification in the model (1), in which we are interested. First, the linearity of the regression function in each subsample can be an approximation as in the robust inference for the OLS estimator in the classical linear regression. Second, the one time break model may be approximating a more general time-varying coefficient model, whether or not there is a jump. For instance, $y_{t,T} = x'_{t,T} \beta_t + \varepsilon_{t,T}$, where $\beta_t = \beta(t/T)$ for some function β defined on $[0,1]^p$. Third, it might be that the true regression function is neither linear nor time-invariant, where we may write $y_{t,T} = f_t(x_{t,T}) + \varepsilon_{t,T}$. The interpretation of the quantity that $\hat{\theta}$ estimate would change each case, although we are more interested in the second and third cases by considering the break model. To be more precise, we need to look at the probability limit of $\hat{\theta}$, which is the pseudo true value θ_0 and exists under certain conditions as given in the next section.

That is, the pseudo true value θ_0 is defined as the minimizer of the limit of the following average mean squared error loss

$$\mathbf{S}_T(\theta) = \mathrm{E}\mathbb{S}_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \mathrm{E}[y_{t,T} - x'_{t,T} \beta_1 1(\tau \leq \gamma) - x'_{t,T} \beta_2 1(\tau > \gamma)]^2, \tag{5}$$

assuming that the integral and the limit are well-defined and the minimizer is unique. That is,

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmin}} \; \mathcal{S}(\theta) := \lim_{T \to \infty} \mathbf{S}_T(\theta). \tag{6}$$

The limit $\mathcal{S}$ exists under fairly general conditions on the distributions of $\{y_{t,T}, x_{t,T}\}_{t=1}^{T}, T = 1, 2, ...$ Assumption 1 in the following section gives a sufficient condition for this and does not

require stationarity of the regressors and the regression errors. In the subsequent section, we set out the conditions under which $\hat{\theta}$ converges to θ_0 in probability and those under which we establish the rate of convergence. The latter relies more on a certain local property of $\mathcal{S}$ at γ_0, while the former depends more on the global shape of $\mathcal{S}$. For instance, if the time-varying coefficient model were the true data generating process, the step function $\beta_1 1(\tau \leq \gamma) + \beta_2 1(\tau > \gamma)$ approximates the function β. It is worthwhile to note that the best approximation, in particular, the best split point γ_0 does not seek to find a discontinuity in the function β. This motivates Assumption 2 in the following section. This assumption is not necessary to derive the consistency of $\hat{\theta}$ to the pseudo true value θ_0 but is crucial to determine the convergence rate of the estimate.

Under our scenario the number of breaks is predetermined, perhaps by the practitioner's judgement, as there is no such thing as the true number of breaks. If $\beta_{10} = \beta_{20}$, however, γ_0 is not identified. Standard structural break tests (see e.g. Andrews 1993 and Hidalgo and Seo 2013) may be employed to test this null hypothesis, which have power against general types of parameter instability in the linear regression. Furthermore, we expect our subsequent findings can be extended to the model with more than one break but we focus on the one break model for the clarity of our exposition.

3 Asymptotics

This section establishes the asymptotic theories associated with the proposed estimator $\hat{\theta} = (\hat{\beta}_1', \hat{\beta}_2', \hat{\gamma})'$. Since the model we estimate is mis-specified, we need to make certain assumptions on the unknown true regression function to get some meaningful asymptotic distribution. Thus, we establish conditions for the convergence of $\hat{\theta}$ to θ_0 in probability, i.e., consistency, and those for convergence rate and asymptotic distribution. These conditions are fairly general but this section shows that under these assumptions the asymptotic property of $\hat{\theta}$ changes dramatically from the standard result, where $\hat{\gamma}$ is super-consistent and $\hat{\beta}$ can be estimated as if γ_0 were known *a priori*. Throughout the paper, $|\cdot|$ denotes the euclidean norm of a vector or matrix.

3.1 Consistency and Rate of Convergence

This section derives the consistency of $\hat{\theta}$ to θ_0 under the assumption that the data is near epoch dependent, which needs not be stationary. The definition is reiterated from Davidson (1994) for completeness. It is a rather minimal assumption to ensure the uniform convergence of the objective function $\mathbb{S}_T(\theta)$.

Definition 1 *For a stochastic array $\{\{V_{t,T}\}_{t=-\infty}^{+\infty}\}_{T=1}^{\infty}$, where $V_{t,T}$ is possibly vector-valued and mixing on a probability space $(\Omega, \mathcal{F}, P)$, let $\mathcal{F}_{t-m,T}^{t+m} = \sigma(V_{t-m,T}, ...V_{t+m,T})$. If an integrable array $\{\{X_{t,T}\}_{t=-\infty}^{+\infty}\}_{T=1}^{\infty}$, satisfies*

$$\left\| X_{t,T} - \mathrm{E}(X_{t,T} | \mathcal{F}_{t-m,T}^{t+m}) \right\|_p \leq d_{tT} \nu_m,$$

where $\nu_m \to 0$ and $\{d_{tT}\}$ is an array of positive constants, it is said to be triangular arrays of Near Epoch Dependent in L^p-norm (L_p-NED) on $\{V_{t,T}\}$. $\{d_{tT}\}$ and $\{\nu_m\}$ are called NED norms and coefficients, respectively. If $\nu_m = O(m^{-\kappa})$ for $\kappa > \kappa_0$, the array is said to be L_p-NED of size $-\kappa_0$.

Next, $\mathbf{S}_T(\theta)$ is a deterministic function and it needs to have a well-defined limit. To state this, define $\mu_{xx}^T(t/T) = \mathrm{E}\left(x_{t,T}x_{t,T}'\right), \mu_{xy}^T(t/T) = \mathrm{E}(x_{t,T}y_{t,T})$, and $\mu_{yy}^T(t/T) = \mathrm{E}y_{t,T}^2$. This notation allows for the nonstationarity and array structure for the data. It is convenient to define a vector-valued function μ^T, which is defined on $[0,1]$ and is the collection of the distinctive elements of μ_{xx}^T, which is a symmetric matrix valued function, μ_{xy}^T, and μ_{yy}^T. Then, we need a well-defined limit of μ^T, say μ, to define $\mathcal{S}$. That is, we assume the following.

Assumption 1 *(i) $\{y_{t,T}, x_{t,T}\}$ are triangular arrays of L_p-NED of size -1 for $p > 2$ on $\{V_{t,T}\}$ such that $\sup_{t,T}\mathrm{E}|y_{t,T}|^{p+d} < \infty$ and $\sup_{t,T}\mathrm{E}|x_{t,T}|^{p+d} < \infty$ for some $d > 0$. Here, $\{V_{t,T}\}$ is α-mixing of size $-q$ for some $q > p(p+d)/d$.*
(ii) There exists a vector-valued function μ defined on $[0,1]$ such that it is bounded and integrable and $\max_{u\in[0,1]}\left|\mu^T(u) - \mu(u)\right| = o\left(T^{-1/2}\right)$.
(iii) For any $\varepsilon > 0$, there exists $\eta > 0$ such that $\inf_{|\theta-\theta_0|>\varepsilon}\mathcal{S}(\theta) - \mathcal{S}(\theta_0) \geq \eta$.

The first condition ensures the uniform convergence of $\mathbb{S}_T - \mathbf{S}_T$. Since the data might be heterogeneous and come from arrays, we impose the second condition for the convergence of $\mathbf{S}_T$. However, the function μ does not need to be continuous. A special case of interest is the one where $\mu_{xx}(t/T) = \mathrm{E}\left(x_{t,T}x_{t,T}'\right)$ for all T. The last one is an asymptotic identification condition and demands that the unique minimizer of $\mathcal{S}$ be well separated. Then, we obtain the following consistency.

Theorem 1 *Under Assumption 1, $\hat{\theta} \xrightarrow{p} \theta_0$.*

Next, we derive the convergence rate of our estimator $\hat{\theta}$. Our estimation problem is non-standard. First, the estimating equation contains a discontinuity in γ. It often leads to non-standard rates such as $T^{-1/3}$ or T^{-1} as in e.g. the maximum score estimation or the estimation of the change point in the threshold regression, respectively. Second, our model is mis-specified. The nature of the true data generating process determines the asymptotic property of $\hat{\theta}$ and the following Assumption 2 is important in this regard. It contrasts with the standard case of estimating the correctly specified break model where $\mathcal{S}$ is not differentiable at θ_0. Note that $\mathcal{S}$ may not be differentiable at points other than θ_0 and thus the true relationship between $y_{t,T}$ and $x_{t,T}$ may change over time in many different ways. The function may not be continuous. However, the best split point does not need to be a jump point.

Assumption 2 *The function $\mathcal{S}$ is twice continuously differentiable at θ_0 with a positive definite second derivative matrix $\mathcal{S}_{\theta\theta}$.*

Assumption 2 requires among others that δ_0 should not be zero.

Theorem 2 *Suppose that Assumption 1 with $p \in (2,4]$ and Assumption 2 hold. Then,*

$$\hat{\theta} = \theta_0 + O_p(T^{\frac{-p}{4p-4}}).$$

Theorem 2 shows that the convergence rate for $\hat{\theta}$ cannot be as fast as $T^{-1/2}$ and can be as slow as $T^{-1/3}$. It states that under the current asymptotic experiment the least squares estimator for θ_0 converges at a rate dramatically different from the conventional rates, say $O_p(T^{-1})$ for the break point estimator $\hat{\gamma}$ and $O_p(T^{-1/2})$ for the slope coefficients estimator β, respectively. This implies that the model uncertainty results in much larger variance for $\hat{\theta}$ than predicted by the conventional theory under correct specification. In particular, the break point is not estimated as precisely as we expect from the conventional theory, which also affects the precision of the estimator for the slope coefficients. It also suggests that the limiting distribution of estimators for both regression slope coefficients and the break point should be evaluated simultaneously rather than the break point is treated to be known *a priori*. That is, we cannot expect the asymptotic independence between the slope estimator $\hat{\beta}$ and the break estimator $\hat{\gamma}$.

3.2 Asymptotic Distribution

To derive the limiting distribution of our estimators in a clear closed-form, we impose more structure on the nature of the process $\{y_{t,T}\}$ given $\{x_{t,T}\}$. Specifically, we assume that $\{x_{t,T}, \varepsilon_{t,T}\}$ is strictly stationary and henceforth write x_t and ε_t dropping the subscript T. Furthermore, assume that there exists a unique function β on $[0,1]$ such that

$$y_{t,T} = x_t'\beta\left(t/T\right) + \varepsilon_t, \tag{7}$$

where $\mathrm{E}\left(\varepsilon_t|\mathcal{F}_t\right) = 0$, and $x_t \in \mathcal{F}_t$. That is, there exists a linear time-varying structural model for y_t given x_t as in e.g. Robinson (1989).

Assumption 3 *(i) The sequence $\{x_t, \varepsilon_t\}$ is strictly stationary and L_4-NED of size $-1/2$ such that $\mathrm{E}|x_t|^{4+d} < \infty$, $\mathrm{E}|\varepsilon_t|^{4+d} < \infty$ for some $d > 0$, and $\mathrm{E}\left(\varepsilon_t|\mathcal{F}_t\right) = 0$ and $x_t \in \mathcal{F}_t$.*
(ii) For any $a \in (0, 1/2)$ and $m = aT$, $\frac{1}{m}\sum_{t=1}^{m} x_t x_t'$, $\frac{1}{m}\sum_{t=T-m+1}^{T} x_t x_t'$, $\frac{1}{m}\sum_{t=\lfloor\gamma_0 T\rfloor-m+1}^{\lfloor\gamma_0 T\rfloor} x_t x_t'$, and $\frac{1}{m}\sum_{t=\lfloor\gamma_0 T\rfloor+1}^{\lfloor\gamma_0 T\rfloor+m} x_t x_t'$ are invertible when $m \geq p$ and those four matrices have stochastically bounded norms uniformly in m.
(iii) $\sum_{s=0}^{\infty} s\left|\mathrm{cov}\left(\delta_0' x_t x_t, \delta_0' x_{t+s} x_{t+s}\right)\right| < \infty$.

Then, the first order condition for the minimization implies that

$$\left.\frac{\partial \mathcal{S}(\theta)}{\partial \gamma}\right|_{\theta=\theta_0} = \mathrm{E}[\delta_0' x_t x_t'(\beta(\gamma_0) - (\beta_{10} + \beta_{20})/2] = 0 \tag{8}$$

$$\left.\frac{\partial \mathcal{S}(\theta)}{\partial \beta}\right|_{\theta=\theta_0} = \begin{pmatrix} \mathrm{E}[x_t x_t'] \int_0^{\gamma_0} (\beta(u) - \beta_{10}) du \\ \mathrm{E}[x_t x_t'] \int_{\gamma_0}^{1} (\beta(u) - \beta_{20}) du \end{pmatrix} = 0. \tag{9}$$

In particular, the condition (9) is equivalent to

$$\beta_{10} = \gamma_0^{-1} \int_0^{\gamma_0} \beta(u)du \text{ and } \beta_{20} = (1-\gamma_0)^{-1} \int_{\gamma_0}^{1} \beta(u)du.$$

The best split point γ_0 is the point minimizing the approximation error to $\beta(u)$ in terms of L_2 loss. Note that we do not assume β is everywhere continuous in the domain of $[0,1]$. Furthermore, it is important to note that γ_0 is not intended to find a discontinuity point of $\beta(u)$. Thus, we assume the following, which is analogous to Assumption 2.

Assumption 4 *The function β is bounded, integrable and continuously differentiable at γ_0 and furthermore $S_{\theta\theta}$ is positive definite.*

Specifically, the second derivative matrix of $\mathcal{S}(\theta)$ at θ_0 is given by

$$\mathcal{S}_{\theta\theta} = \begin{bmatrix} \mathcal{S}_{\beta\beta} & \mathcal{S}_{\beta\gamma} \\ \mathcal{S}_{\gamma\beta} & \mathcal{S}_{\gamma\gamma} \end{bmatrix} \tag{10}$$

where

$$\mathcal{S}_{\beta\beta} = 2 \begin{pmatrix} \mathrm{E}[x_t x_t']\gamma_0 & 0 \\ 0 & \mathrm{E}[x_t x_t'](1-\gamma_0) \end{pmatrix} \tag{11}$$

$$\mathcal{S}_{\gamma\gamma} = 2\mathrm{E}[\delta_0' x_t x_t' \frac{\partial}{\partial u}\beta(\gamma_0)] \tag{12}$$

$$\mathcal{S}_{\beta\gamma} = 2 \begin{bmatrix} -(\beta(\gamma_0)-\beta_{10})'\mathrm{E}[x_t x_t'] \\ (\beta(\gamma_0)-\beta_{20})'\mathrm{E}[x_t x_t'] \end{bmatrix} \tag{13}$$

and $\mathcal{S}_{\gamma\beta} = \mathcal{S}_{\beta\gamma}'$ with $\frac{\partial}{\partial u}\beta(\gamma_0)$, the first derivative of $\beta(u)$ evaluated at γ_0.

Assumption 4 requires that $\delta_0 \neq 0$ and $\frac{\partial}{\partial u}\beta(\gamma_0) \neq 0$ as well as $\mathrm{E}(x_t x_t')$ being positive definite. This is related to the estimated regression function being discontinuous at γ_0 with a positive probability. Assumption 3.(ii) is standard in structural-break literature. See Bai (1997) and Bai and Perron (1998). Assumption 3.(iii) is a regularity condition to apply a Functional Central Limit Theorem (FCLT).

Theorem 3 *Let $\mathcal{B}$ and ω^2 denote the two-sided standard Brownian motion and the long-run variance of $2\delta_0' x_t(\varepsilon_t + x_t'(\beta(\gamma_0) - (\beta_{10}+\beta_{20})/2))$, respectively. Then, under Assumptions 1-4,*

$$\sqrt[3]{T}(\hat{\theta}-\theta_0) \xrightarrow{d} \underset{h}{\operatorname{argmin}} \left[\omega\mathcal{B}(g) + \frac{1}{2}h'\mathcal{S}_{\theta\theta}h\right]$$

where $h = (b', g)'$, of the same dimension as θ, and $\mathcal{S}_{\theta\theta}$ is given in (10).

A natural way to estimate ω^2 is the heteroskedasticity-autocorrrelation consistent variance estimation. That is, for some $K > 0$,

$$\hat{\omega}^2 = \frac{1}{2Kr_T^2} \sum_{j=0}^{m} w_{mj} \sum_{t=\lfloor T\hat{\gamma}-Kr_T^2 \rfloor + j}^{\lfloor T\hat{\gamma}+Kr_T^2 \rfloor} [\hat{u}_t\hat{u}_{t-j} - \bar{u}_T^2]$$

where $r_T = T^{1/3}$, $\hat{u}_t = \hat{\delta}' x_t \left(2y_t - x_t'(\hat{\beta}_1 + \hat{\beta}_2)\right)$ with $\hat{\delta} = (\hat{\beta}_2 - \hat{\beta}_1)$, $\bar{u}_T = (2K)^{-1} r_T^{-2} \sum_{t=\lfloor T\hat{\gamma} - Kr_T^2 \rfloor + 1}^{\lfloor T\hat{\gamma} + Kr_T^2 \rfloor} \hat{u}_t$, and w_{mj} is a weight function. Note that it is estimated only from observations that are in the neighborhood r_T^2 of the break point $T\hat{\gamma}$. Regarding the choice of w_{mj} see e.g. Davidson and de Jong (2000).

Another parameter unknown here is that $\mathcal{S}_{\theta\theta}$. For $S_{\beta\beta}$, we can plug in the sample moment of $x_t x_t'$ and $\hat{\gamma}$. For $S_{\beta\gamma}$, let

$$\hat{S}_{\beta\gamma} = 2 \begin{bmatrix} \frac{1}{Kr_T^2} \sum_{\lfloor T\hat{\gamma} - Kr_T^2 \rfloor + 1}^{\lfloor T\hat{\gamma} \rfloor} [-x_t(y_t - x_t'\hat{\beta}_1)] \\ \frac{1}{Kr_T^2} \sum_{\lfloor T\hat{\gamma} \rfloor + 1}^{\lfloor T\hat{\gamma} + Kr_T^2 \rfloor} [x_t(y_t - x_t'\hat{\beta}_2)] \end{bmatrix},$$

and $\hat{S}_{\gamma\beta} = \hat{S}_{\beta\gamma}'$.

However, since $S_{\gamma\gamma}$ involves $\frac{\partial}{\partial u}\beta(\gamma_0)$, we estimate $\beta(u)$ for $u \in (0,1)$ via a nonparametric estimation method and then obtain the first derivative of $\hat{\beta}(u)$ with respect to u evaluated at $\hat{\gamma}$. More specifically,

$$\hat{S}_{\gamma\gamma} = \frac{1}{Kr_T^2} \sum_{\lfloor T\hat{\gamma} - Kr_T^2 \rfloor + 1}^{\lfloor T\hat{\gamma} + Kr_T^2 \rfloor} \left\{ \hat{\delta}' x_t x_t' \frac{\partial}{\partial u} \hat{\beta}(u) \Big|_{u=\hat{\gamma}} \right\}$$

where $\hat{\delta} = \hat{\beta}_2 - \hat{\beta}_1$ and $\hat{\beta}(u) = \left[\sum_{t=1}^T K_h(u - t/T) x_t x_t'\right] \left[\sum_{t=1}^T K_h(u - t/T) x_t y_t\right]$ with a kernel function $K_h(\cdot) = K(\cdot/h)/h$ and a bandwidth h. This type of kernel estimator for $\beta(u)$ can be found in various works including Robinson (1989) and Koo and Linton (2012).

Remark 1 *The intuition behind the slower rate of convergence in Theorem 2 and 3 is that the limit criterion function $\mathcal{S}$ does not distinguish the pseudo true value γ_0 as well as the limit criterion function in the correctly specified break case. This is reflected in the twice differentiability in $\mathcal{S}$ compared to non-differentiability in case of correct specification. Furthermore, the sample criterion function $\mathbb{S}_T$ exhibits more sampling variation in finite samples than the standard $T^{1/2}$ estimation problems. The combination of the two yields the slower convergence for $\hat{\gamma}$, which in turn slows down the convergence of the other estimates. The latter does not have much to do with mis-specification while the former has. Indeed, if the function β had a big jump at γ_0, we may have ended up with the conventional T rate.*

3.2.1 Quasi-Likelihood Ratio

Since the estimated model is the linear projection for a fixed γ, the objective function $\mathbb{S}_T(\theta)$ from (2) can be analyzed via $\mathbb{S}_T(\gamma)$ from (3). It is possible to conduct inference on γ_0 using the sampling distribution of the break point estimate discussed in Theorem 3. However, another common approach is to use the quasi-likelihood ratio statistic (QLR) based on (3) as advocated in Hansen (2000) . We can tabulate the limiting distribution of the properly rescaled QLR statistic. Let

$$\mathcal{LR}_T(\gamma) = T^{2/3}(\ln \mathbb{S}_T(\gamma) - \ln \mathbb{S}_T(\hat{\gamma})).$$

Following our cube-root rate of convergence, $\mathcal{LR}_T(\gamma)$ is scaled by $T^{2/3}$ instead of the usual T.

Corollary 1 *Under Assumptions 1-4,*

$$\xi \mathcal{LR}_T(\gamma_0) \xrightarrow{d} \max_{g \in \mathbb{R}} \left(\mathcal{B}(g) - \frac{1}{2} g^2 \right), \tag{14}$$

where $\xi = \frac{1}{\mathbb{S}_T(\theta_0)} \left[(\mathcal{S}_{\gamma\gamma} - \mathcal{S}_{\gamma\beta}\mathcal{S}_{\beta\beta}^{-1}\mathcal{S}_{\beta\gamma})\omega^{-4} \right]^{1/3}$.

For the inference with respect to γ_0, we tabulate the limiting distribution of the rescaled QLR statistic in Table 1.

*****TABLE 1 ABOUT HERE*****

We also provide the plots of the limiting distribution of the rescaled quasi-likelihood ratio statistic, i.e. $\max_{g \in \mathbb{R}} \left(\mathcal{B}(g) - \frac{1}{2} g^2 \right)$ with the usual χ^2 distribution for the comparison purpose in Figure 1. The limiting distribution of the rescaled QLR statistic has a much thinner tail than that of the χ^2 distribution.

*****Figure 1 ABOUT HERE*****

4 Forecasting

After briefly discussing the implications of our theoretical findings for forecasting, this section proposes a new forecasting method for models subject to structural instability. It is in line with many forecasts averaging methods in the literature like Pesaran and Pick (2011) and others mentioned in the introduction. Our new forecasting method is termed as AveLR (Averaging based on a LR statistic) because the averaging is made over the interval by inverting the QLR statistic Corollary 1 suggests.

4.1 AveLR

Our results in Section 3 entail several implications for forecasting. Recall that, in the presence of model uncertainty, the estimators for the break point and regression coefficients in structural-break models have a much larger variance reflecting our cube-root asymptotic results. This leads to wider confidence intervals. Consequently, the forecast based on the estimator of structural-break models will have a much larger variance, which significantly reduces the advantage these models have over the linear model without a structural break. Although structural–break models result in a smaller bias, this comes at a large cost in terms of precision. These implications lead to our forecasting method. Our forecasting method involves simple averaging of forecasts over all possible break points. Possible candidates of the break point parameter are obtained by constructing a confidence band which incorporates the rate of convergence, $T^{1/3}$. Based on the $\mathcal{LR}$ statistic, due to Corollary 1, we can construct the confidence set as

$$\begin{aligned} \Phi &= \{\gamma : \mathcal{LR}_m(\gamma) \leq c\} \\ &= \{\gamma : [\ln \mathbb{S}_m(\gamma) - \ln \mathbb{S}_m(\hat{\gamma})] \leq c \cdot m^{-2/3}\} \end{aligned}$$

where c and m denote a critical value and the size of sample used for estimation respectively. Once the confidence set is constructed, we obtain a forecast from each possible candidate of the break location parameter belonging to the above confidence band. That is, for every possible candidate $\gamma \in \Phi$, we calculate $\hat{\beta}_2(\gamma)$ and use it to come up with a forecast given γ. Then, we average those forecasts to yield our final forecast.

We can see clearly the relevance of our cube-root asymptotics to our forecasting method since it sheds some light on the range over which averaging takes place. Under the correct specification of the true DGP, the confidence set in which the possible break is located is small since the rate of convergence is T^{-1}, hence averaging does not make much difference. On the other hand, the confidence set in which the possible break is located is relatively large in our setting of model uncertainty, since the rate of convergence for our estimator is $T^{-1/3}$ and hence, averaging can greatly reduce the forecasting variation.

Remark 2 *As seen from Corollary 1, c incorporates the scaling factor ξ. The choice of c is arbitrary even if ξ were known and remains open for a future research topic. Furthermore, the estimation of ξ can introduce too much sampling variation into the forecasts. In the following, we experiment with several choices of c for our empirical studies. Firstly, we choose several values of c arbitrarily. Secondly, we choose c such that the chosen c implies the $4m^{-1/3}$ percentile of the set, $\{\mathcal{LR}_m(\gamma) : \gamma \in \Gamma\}$ where Γ is a parameter set for the break point. The latter yields confidence bands less sensitive to the change in ξ or c.*

4.2 Empirical Implementation

In this section, we apply our AveLR, building on our cube-root asymptotics to a range of time series in many different set-ups. We investigate whether AveLR produces forecasts that improve upon the standard method, whereby forecasts are obtained using only post-break data once the break point is estimated (henceforth, Post break). We also compare the performance of our forecasting methods with other contemporary forecasting methods in the presence of structural breaks.[2]

In the first empirical study, we use autoregressive models for many macroeconomic time series data from Stock and Watson (1998) and Hansen (2001). In the second empirical study, we take an example of a leading indicator model whereby the yield curve is used as a leading indicator of GDP growth in G7 Countries and Australia along the line of Pesaran *et al.* (2011). Note that leading indicator models have been popular in macroeconomics literature, for instance, Estrella and Mishkin (1997). Detailed description of those models can be found in Anderson *et al.* (2007) and Pesaran *et al.* (2011). We use the lagged dependent variables as regressors for the first application whereas we use the yield curve as a regressor for the second one. For both applications, we trim the parameter space, Γ for γ by considering only the interval between 0.05 and 0.95.

[2]For this experiment, we modify MATLAB codes used by Pesaran *et al.* (2011) and GAUSS codes used by Tian and Anderson (2012) in addition to our GAUSS code, which is available upon request from the authors.

Before we proceed to the results of various forecasting methods (including ours), we briefly discuss a range of competing forecasting methods for the comparative study. There are a large number of methods that have been used empirically in the presence of structural breaks. However, for our analysis, we restrict ourselves to a sample of methods in which averaging approaches play an important role. For detailed explanation, see Pesaran *et al.* (2011) and Hansen (2009). We start with averaging forecasts from different sub-windows (AveW) studied in Pesaran and Pick (2011). In this method, once a minimum sub-window size is chosen, forecasts can be obtained by averaging over sub-windows within the given expanding window. This is quite similar to our AveLR method. However, AveW does not rely on estimates of break dates and sizes and hence there is no estimation of a break point and its size, let alone no need of a break testing. In our comparative study, we choose $m = T(1-v_{\min})+1$ windows with $v_{\min} = 0.05$. The second method we discuss is the optimal weight method (Optwgt) studied in Pesaran *et al.* (2011). This method gives past observations weights which minimize prediction mean squared error of the one-step ahead forecast. They consider two types of breaks, discrete and continuous breaks. This is a model-specific approach, however. The third method is the robust weight method (Rwgt) studied in Pesaran *et al.* (2011). This method integrates the optimal weights from Optwgt with respect to a uniformly distributed break dates over a possible range, say $[\underline{b}, \overline{b}]$, where $\underline{b}$ and $\overline{b}$ are the lower and upper range in which the break resides. In our study, we consider two pairs of $[\underline{b}, \overline{b}]$ to check the sensitivity of the PMSE to the choice of $\underline{b}$ and $\overline{b}$. That is, we conduct Rwgt with $\underline{b} = 0.5$ and $\overline{b} = 0.98$ (Rwgt1) and with $\underline{b} = 0$ and $\overline{b} = 1$ (Rwgt2), following Pesaran *et al.* (2011). We also conduct the reverse ordered CUSUM method (ROC) studied in Tian and Anderson (2012). In this method, averaging takes place over different estimation windows with different weights based on reverse ordered CUSUM. Finally, we include the averaging method based on a Mallows criterion (AveMal) studied in Hansen (2009). This method involves averaging between the structural break estimates and the no-break estimates in a way that the weight is selected to minimize the Mallows criterion.

4.2.1 Empirical Study I: Autoregressive Models

For the first empirical study, our data set consists of 40 major monthly US macroeconomic time series, which have been studied extensively throughout the literature in the context of the structural break models.[3] These series can be grouped into several categories: labour productivity (11 series), unemployment (2 series), wages (7 series), money (4 series), stock price indices (2 series), interest rates (7 series), bond yield (1 series), exchange rates (3 series), producer price indices (2 series), and consumer price index (1 series). The sample period of

[3] The data was sourced predominantly from the Stock and Watson (1998) paper, except for labour productivity, which is procured from Hansen (2001). In particular, Stock and Watson (1998) data codes are LHUR, LHU680, LEH, LEHCC, LEHM, LEHTU, LEHTT, LEHFR, LEHS, FM1, FM2, FM3, FMBASE, FSNCOM, FSPCOM, FYFF, FYCP, FYGM3, FYGM6, FYGT1, FYGT5, FYGT10, FYAAAC, EXRUS, EXRJAN, EXRUK, PWFSA, PWFCSA, PUNEW. To see a more detailed description of these series, refer to Stock and Watson (1998). Labour productivity data is from Hansen (2001) and hence, detailed description of those data can be found in Hansen (2001).

labour productivity is from January 1947 to April 2001. Labour productivity is classified into 11 sub-categories, and labour productivity growth is constructed as the first difference of the ratio of industrial production index to total work hours in logarithm form. For the other series, the sample runs from January 1959 to December 1996.[4] Following Stock and Watson (1998), we use seasonally adjusted data for series that have seasonal patterns. Also, some series such as wages, money, stock price indices, exchange rates, producer and consumer price indices were analyzed in logarithm form. In contrast, no preliminary transformations were used for interest rates, unemployment, and bond yield.

We conduct an experiment to examine pseudo out-of-sample performance of various forecasting methods. The experimental design is kept simple for the comparison. That is, for labour productivity growth, we use AR(1), following Hansen (2001), and for the other series, we use AR(4), following Stock and Watson (1998). We also consider AR(4) for labour productivity and AR(1) for the other series for the robustness check. Forecast errors are recursively computed beginning at $t = \lfloor T/2 \rfloor + 1$ and $\lfloor 2T/3 \rfloor + 1$ where T is the entire sample size. Only one break is allowed for each series.

Although standard break tests can be employed for our forecasting approach, our setting does not require testing of a break point. Therefore, we assess forecasting performance of an array of forecasting methods in both testing and non-testing cases. For the testing case, we use the asymptotic p-value for Quandt-Andrews structural break test (Hansen, 1997). If a break is not detected, we use an AR model. Alternatively, if a break is detected, based on the estimated break date $\widehat{\gamma}$, we construct a confidence set Φ for the unknown true break date γ_0. For the first application, our choice of c is an arbitrarily chosen fixed value. We choose three values, $c = 5, 10, 20$, to check sensitivity associated with choice of c. We also consider the average over different values of c (Ave c) for the robustness check.[5]

For evaluation, we focus on the one step ahead prediction mean squared error (PMSE) to lend support to the validity and performance of our forecasting method.

$$PMSE = \mathrm{E}[y_{t+1} - \hat{y}_{t+1}]^2$$

where $\hat{y}_{t+1} = x_t' \hat{\beta}_2$.

Out-of-Sample Forecasting Results and Analysis Our question is primarily concerned with whether our method improves significantly upon the Post break method. To answer this, we compute the percentage improvement of each forecasting method including ours relative to the Post break method. The percentage improvement of forecast method i over the Post break is computed as

$$\text{percentage improvement} = \frac{PMSE_{Pb} - PMSE_i}{PMSE_{Pb}} \times 100.$$

[4] The Japanese exchange rate starts from January 1973 because this series is flat in the period of fixed exchange rate regime. Also, due to data limitations, some series of wages (LEH, LEHTU, LEHTT, LEHFR, and LEHS) start from January 1964.

[5] We also used the quantile-based confidence set for verifying the forecasting performance of our method and obtained similar results.

This exercise tells us how much each method can improve one-step out-of-sample forecasting performance in terms of PMSE, compared to the Post break method, in different situations.

Table 2 About Here

Table 2 reports a simple average of percentage improvement of each method over the Post break in the case of no testing with an array of arbitrary c. The variance of percentage improvement is in parentheses. We find that there are significant gains from averaging upon our confidence sets. Careful examination of AveLR, the averaging based on the inversion of the LR statistic, can improve upon the Post break by approximately 20 percent. The results also suggest that the other forecasting methods outperform the Post break in all cases. Comparing AveLR with the other methods, we can clearly see the performance of AveLR is at least as good as the others in the case of no testing with some arbitrary c. As we expected, the variance of the percentage improvements for Ave c is smaller than those of $c = 10$ and $c = 20$. Table 3 shows the results in the case of testing in every recursion, with an array of arbitrary c. The results indicate similar dynamics to Table 2. We also compare AveLR with no break models for the purpose of sensitivity and robustness checks. No break models include AR(4) and AR(1), as well as a random walk process. Table 4 shows our method with arbitrary c works better than no break models.

Table 3 About Here
Table 4 About Here

4.2.2 Empirical Study II: Leading Indicator Model

For the second empirical study, we collect the spread between short-term and long-term interest rates and real GDP growth rates for G-7 countries and Australia. A detailed data description is provided in Appendix C. For the second empirical study, (thanks to the reasonable number of series), we plot the out-of-sample one-period ahead forecasting performance to compare our method (AveLR) with other methods, including a linear regression without a break (FullW), Post break, and Averaging window (AveW). For this study, we only report the non-testing case since the results for the testing case are similar. Unlike the first empirical study where c is chosen arbitrarily, we choose a value such that c determines the $4m^{-1/3}$ percentile of $\{\mathcal{LR}_T(\gamma) : \gamma \in \Gamma\}$. We provide the graphical representation of the forecasting performance of each forecasting method in the following way. We fix $n_0 = \lfloor 0.5T \rfloor$ where T is the entire sample size and define $m = n_0, n_0 + 1, ..., \lfloor 0.95T \rfloor$. Then, for each m, we estimate $\hat{y}_{m+1}$ based on $\{y_t\}_{t=1}^{m}$ and compute the prediction mean squared errors as $(T-m)^{-1} \sum_{t=1+m}^{T} (y_t - \hat{y}_t)^2$. Figure 2 describes the movements of real GDP growth and the spread for each country. Figure 2 shows that these two time series move similarly, confirming the relationship between the two. However, it also clearly shows the relationship varies over different time periods. Figure 3 plots the residual sum of squares when the corresponding time is chosen for a split point based on the whole sample. Figure 4 plots one period ahead PMSEs for each method (FullW, Post

break, AveW, and AveLR) along the x axis in order for us to evaluate different forecasting methods over time.[6]

Figure 2 About Here
Figure 3 About Here
Figure 4 About Here

Out-of-Sample Forecasting Results and Analysis As seen from Figure 3, in most European countries, there is a likely break towards the end of the sample reflecting the recent financial distress after the collapse of Lehman brothers. Also, Australia has a likely break early in the sample period. The Sup Wald test strongly supports the presence of those breaks although the results reported here are based on the non-testing method and hence we do not test the presence of a break but rather assume there is a break when various forecasting methods are implemented. We can relate this to the forecasting performance of different methods reported in Figure 4. Figure 4 confirms that our method performs quite well and produces the least PMSEs compared with the other methods. Averaging methods (AveLR, AveW) work well and the forecasting performance of AveLR is slightly better than AveW although those are quite similar. For the UK and Germany, it is not surprising that the forecasting performance of Post break is disappointing because they seem to have only one break towards the end of the sample. For Australia, it has one obvious break at the early stage and that's why the performance of FullW and Post break methods is similar. They are also better than averaging methods. For the other countries, averaging methods work well and are quite robust to many different scenarios.

5 Conclusion

This paper develops a new asymptotic theory for the standard least squares estimator of the linear regression with one-time structural break. It reveals that the sampling variation in the estimate is much larger than the standard theory has suggested. As a consequence, it also explains the reason why various forecast combination methods are more successful (especially the combinations from rather big windows) than the forecast based on only after-break data. An interesting future research will be developing some analogue of Bagging. The bagging is not straightforward in the current setup since the cube-root type asymptotics fails the naive bootstrap.

[6] It is worth noting that we also checked the percentage improvement of our method over the Post break and obtained similar results reported in the first empirical study. Moreover, we plotted the PMSEs for the first empirical study and confirmed that the results are almost the same as reported in the second empirical study.

References

Abrevaya, J and J. Huang (2005) On the Bootstrap of the Maximum Score Estimator. Econometrica 73, 1175-1204.

Anderson, H.M., G. Athanasopolous and F. Vahid (2007) Nonlinear Autoregressive Leading Indicator Models of Output in G7 Countries. Journal of Applied Econometrics 22, 63 – 87.

Andrews, T.W. (1988) Laws of Large Numbers for Dependent Non-identically Distributed Random Variables. Econometric Theory 4, 458-467.

Andrews, T.W. (1993) Tests for parameter instability and structural change with unknown change point. Econometrica 61, 821-856.

Bai, J. (1997) Estimation of a change point in multiple regression models. Review of Economic and Statistics 79, 551-563.

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

Ben-David, D. and D. Papell (1998) Slowdowns and Meltdowns: Postwar Growth Evidence from 74 Countries. Review of Economics and Statistics 80, 271-287.

Bhattacharya, P.K. (1994) Some aspects of change-point analysis. In Carlstein, E., Muller, H. G. Siegmund, D. (eds) Change Point Problems, IMS Lecture Notes – Monograph Series, vol 23, 28-56.

Breiman, L. (1996) Bagging Predictors. Machine Learning 24, 123-140.

Bühlmann, P. and B. Yu (2002) Analyzing Bagging. Annals of Statistics 30, 927-961.

Clements, M. P., and D. F. Hendry (1998a) Forecasting economic time series. CUP, Cambridge

Davidson, J. (1994) Stochastic Limit Theory, Oxford University Press.

Davidson, J. and R.M. de Jong (2000) Consistency of kernel estimators of heteroskedastic and autocorrelated covariance matrices. Econometrica 68, 407-424.

Delgado, M.A. and J. Hidalgo (2000) Nonparametric inference on Structural Breaks. Journal of Econometrics 96, 113-144.

Estrella, A. and F.S. Mishkin (1998) Predicting U.S. Recessions: Financial Variables as Leading Indicators. Review of Economics and Statistics 80, 45-61.

Garcia, R. and P. Perron (1996) An analysis of the real interest rate under regime shifts. Review of economics and statistics 78, 111-125.

Hansen, B.E. (1997) Approximate Asymptotic P Values for Structural-Change Tests. Journal of Business and Economic Statistics 15, 60-67.

Hansen, B.E. (2000) Sample Splitting and Threshold Estimation. Econometrica 68, 575-603.

Hansen, B.E. (2001) The new econometrics of structural change: Dating breaks in U.S. Labor productivity. Journal of Economic Perspectives 15, 117-128.

Hansen, B.E. (2009) Averaging estimators for regressions with a possible structural break. Econometric Theory 25, 1498-1514.

Hidalgo, F.J. and M.H. Seo (2013) Testing for structural stability in the whold sample. Journal of Econometrics, forthcoming.

Koo, B. and O. Linton (2012) Estimation of Semiparametric Locally Stationary Diffusion Models. Journal of Econometrics 170, 210-233.

McConnell, M.M. and G. Perez-Quiros (2000) Output fluctuations in the United States:What has changed since the early 1980's? American Economic Review 90, 1464-1476.

McLeish, D.L. (1975) A maximal inequality and dependent strong laws. The Annals of Probability 3, 829-839.

Perron, P. (2006) Palgrave Handbook of Econometrics: Volume 1 Econometric Theory

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

Pesaran, M.H. and A. Timmermann (2007) Selection of Estimation Window in the Presence of Breaks. Journal of Econometrics 137, 134-161.

Pesaran, M.H. and A. Pick (2011) Forecast Combination across Estimation Windows. Journal of Business and Economic Statistics 29, 307-318.

Pesaran, M.H., A. Pick and M. Pranovich (2011) Optimal Forecasts in the Presence of Structural Breaks. preprint.

Robinson, P. (1989) Nonparametric Estimation of Time-Varying Parameters. Statistical Analysis and Forecasting of Economic Structural Change. Springer-Verlag, 253-264.

Rossi, B. (2012) Advances in Forecasting Under Instability. in Handbook of Economic Forecasting, ed. by G. Elliott, and A. Timmermann. Elsevier-North Holland.

Seo, M.H. (2011) Estimation of Nonlinear Error Correction Models. Econometric Theory 27, 201-234.

Seo, M.H. (2012) Forecasting with a Regime-Switching Model, manuscript.

Stock, J.H. (1994) Unit roots, structural breaks and trends. In Handbook of Econometrics, vol. 4 (Engle, R. F., McFadden, D., eds), Elsevier, 2740-2841.

Stock, J.H. and M. Watson (1996) Evidence of structural instability in macroeconomic time series relations. Journal of Business and Economic Statistics 14, 11-30.

Tian, J. and H.M. Anderson (2012) Forecast Combinations Under Structural Break Uncertainty. Preprint.

van der Vaart, A.W. and J.A. Wellner (1996) Weak convergence and empirical process. Springer, New York.

van Dijk, D., T. Terasvirta and P.H. Franses (2002) Smooth transition autoregressive models – a survey of recent developments, Econometric Reviews 21, 1-47.

Wooldridge, J.M. and H. White (1988) Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes. Econometric Theory 4, 210-230.

A Proof of Theorems

Proof of Theorem 1. Given Assumption 1 (iii), Corollary 3.2.3 in van der Vaart and Wellner (1996) requires the uniform convergence of $\mathbb{S}_T(\theta)$ to $\mathcal{S}(\theta)$ in probability. We establish it by applying the maximal inequality of mixingale triangular arrays, i.e. Lemma 2 in this paper. Recall that $\mathbb{S}_T(\theta)$, $\mathbf{S}_T(\theta)$ and $\mathcal{S}(\theta)$ be defined as in (2), (5) and (6) respectively. Note that

$$|\mathbb{S}_T(\theta) - \mathcal{S}(\theta)| \leq |\mathbb{S}_T(\theta) - \mathbf{S}_T(\theta)| + |\mathbf{S}_T(\theta) - \mathcal{S}(\theta)|. \tag{A.1}$$

First, since Θ is bounded, for some $C < \infty$ and as $T \to \infty$

$$\begin{aligned} & \sup_{\theta\in\Theta} |\mathbf{S}_T(\theta) - \mathcal{S}(\theta)| \qquad (A.2) \\ \leq\ & \frac{1}{T}\sum_{t=1}^{T} \left|\mu_{yy}^T(t/T) - \mu_{yy}(t/T)\right| + \frac{2C^2}{T}\sum_{t=1}^{T} \left|\mu_{xx}^T(t/T) - \mu_{xx}(t/T)\right|_\infty \\ & + \frac{4C}{T}\sum_{t=1}^{T} \left|\mu_{xy}^T(t/T) - \mu_{xy}(t/T)\right|_\infty + \left|\frac{1}{T}\sum_{t=1}^{T}\mu_{yy}(t/T) - \int_0^1 \mu_{yy}(u)\,du\right| \\ & + 2C^2\left|\frac{1}{T}\sum_{t=1}^{T}\mu_{xx}(t/T) - \int_0^1 \mu_{xx}(u)\,du\right|_\infty + 4C\left|\frac{1}{T}\sum_{t=1}^{T}\mu_{xy}(t/T) - \int_0^1 \mu_{xy}(u)\,du\right|_\infty \\ \to\ & 0, \end{aligned}$$

where $|A|_\infty$ denotes the supremum norm of a matrix A and the convergence in the last line follows from Assumption 1 (ii).

Second, we turn to the uniform convergence of $|\mathbb{S}_T(\theta) - \mathbf{S}_T(\theta)|$ in $(A.1)$. We begin with establishing the stochastic equicontinuity of the process. Let us consider $\theta^* = (\beta_1^{*\prime}, \beta_2^{*\prime}, \gamma^*)'$ and $\theta = (\beta_1', \beta_2', \gamma)'$ with $\theta^* \in B(\theta, \epsilon)$ where $B(\theta, \epsilon)$ is a closed ball in Θ of radius $\epsilon \geq 0$ centered at θ. It is enough to show that for all T,

$$\sup_{\theta\in\Theta}\sup_{\theta^*\in B(\theta,\epsilon)} |\mathbb{S}_T(\theta) - \mathbb{S}_T(\theta^*)| \xrightarrow{p} 0, \tag{A.3}$$

as $\epsilon \to 0$. To this end, we define

$$\mathbb{S}_T(\theta) - \mathbb{S}_T(\theta^*) = \mathbb{H}_1 + \mathbb{H}_2,$$

where

$$\mathbb{H}_1 = \frac{1}{T}\left[\sum_{t=1}^{\lfloor T\gamma\rfloor}\left(y_{t,T} - x_{t,T}'\beta_1\right)^2 - \sum_{t=1}^{\lfloor T\gamma^*\rfloor}\left(y_{t,T} - x_{t,T}'\beta_1^*\right)^2\right]$$

and $\mathbb{H}_2$ is defined analogously for the summation after $\lfloor T\gamma\rfloor$ and $\lfloor T\gamma^*\rfloor$. Define $b_1 = (\beta_1 - \beta_1^*)$ and $g = (\gamma^* - \gamma)$. We focus on the case of $g > 0$ since the argument is analogous for the case of $g < 0$.

Note that, for $g > 0$,

$$\mathbb{H}_1 = A_1 + B_1 - 2C_1,$$

with

$$\begin{aligned} A_1 &= \frac{1}{T}\sum_{t=1+\lfloor T\gamma\rfloor}^{\lfloor T\gamma+Tg\rfloor} e_{1t,T}^{*2} \\ B_1 &= \frac{1}{T}\sum_{t=1}^{\lfloor T\gamma\rfloor} b_1'(x_{t,T}x_{t,T}' - \mathrm{E}x_{t,T}x_{t,T}')b_1 + \frac{1}{T}\sum_{t=1}^{\lfloor T\gamma\rfloor} b_1'\mathrm{E}x_{t,T}x_{t,T}'b_1 \\ C_1 &= \frac{1}{T}\sum_{t=1}^{\lfloor T\gamma\rfloor} (e_{1t,T}^{*}x_{t,T}' - \mathrm{E}\left(e_{1t,T}^{*}x_{t,T}'\right))b_1 + \frac{1}{T}\sum_{t=1}^{\lfloor T\gamma\rfloor} \mathrm{E}\left(e_{1t,T}^{*}x_{t,T}'\right)b_1. \end{aligned}$$

where $e_{1t,T}^{*} = y_{t,T} - x_{t,T}'\beta_1^{*}$. Since $e_{1t,T}^{*2} \geq 0$,

$$A_1 \leq A_1^{*} = \frac{1}{T}\sum_{t=1+\lfloor T\gamma\rfloor}^{\lfloor T\gamma+T\epsilon\rfloor} e_{1t,T}^{*2} = \frac{1}{T}\sum_{t=1+\lfloor T\gamma\rfloor}^{\lfloor T\gamma+Tg\rfloor} \left(e_{1t,T}^{*2} - \mathrm{E}e_{1t,T}^{*2}\right) + \frac{1}{T}\sum_{t=1+\lfloor T\gamma\rfloor}^{\lfloor T\gamma+Tg\rfloor} \mathrm{E}e_{1t,T}^{*2}.$$

Due to the moment conditions in Assumption 1 (i), it is straightforward to see that the second terms in $A_1^{*}, B_1,$ and C_1 are $O(\epsilon)$. For the first terms therein, we apply Lemma 2. We ensure the conditions for the lemma are satisfied. Note that the sum and product of pairs of NED triangular arrays are also NED. In particular, the sum of L_p-NED triangular arrays of size $-b$ is L_p-NED of size $-b$ whereas the product is $L_{p/2}$-NED of size $-b$. And from Corollary 17.6.(i) in Davidson (1994) regarding the relationship between NED and L_p-mixingale arrays, demeaned terms in A_1, B_1 and C_1 are $L_{p/2}$-mixingales of size -1 for $p > 2$ due to Assumption 1 (i). For the first term in A_1^{*}, we need a bit more elaboration. That is, for any $d > 0$ as $\epsilon \to 0$

$$\begin{aligned} \Pr\left\{\max_{\gamma\in\Gamma}\frac{1}{T}\sum_{t=1+\lfloor T\gamma\rfloor}^{\lfloor T\gamma+T\epsilon\rfloor}\left(e_{1t,T}^{*2} - \mathrm{E}e_{1t,T}^{*2}\right) > d\right\} &= \Pr\left\{\max_{\gamma\in\Gamma}\frac{1}{\sqrt{T}}\left|\xi_{\lfloor T\gamma\rfloor}(\epsilon)\right| > d\right\} \\ &\leq \mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\left|\xi_{\lfloor T\gamma\rfloor}(\epsilon)\right|^2\right)/d^2 \\ &\to 0, \end{aligned}$$

where $\xi_s(\epsilon) = \frac{1}{\sqrt{T}}\sum_{t=1+s}^{\lfloor s+T\epsilon\rfloor}\left(e_{1t,T}^{*2} - \mathrm{E}e_{1t,T}^{*2}\right)$ and $\mathrm{E}\left|\xi_s(\epsilon)\right|^2 \leq O(\epsilon)$ uniformly in s due to Lemma 2 and Assumption 1 (i).

In addition, from Lemma 1, for $\forall\theta \in \Theta$,

$$\mathbb{S}_T(\theta) - \mathbf{S}_T(\theta) \xrightarrow{p} 0. \tag{A.4}$$

Due to Theorem 21.9. in Davidson (1994), combining (A.3) and (A.4) concludes

$$\sup_{\theta\in\Theta}|\mathbb{S}_T(\theta) - \mathbf{S}_T(\theta)| \xrightarrow{p} 0. \tag{A.5}$$

■

Proof of Theorem 2. From Theorem 3.4.1 in van der Vaart and Wellner (1996), two conditions that we need to show are as follows. For some $d > 0$, and for any $0 < \epsilon < d$ and $\theta = (\beta_1', \beta_2', \gamma)'$,

$$\sup_{||\theta-\theta_0||<\epsilon}\left[-\mathcal{S}(\theta) + \mathcal{S}(\theta_0)\right] \leq -\epsilon^2 \tag{A.6}$$

$$\mathrm{E}\sup_{||\theta-\theta_0||<\epsilon}|\sqrt{T}[(-\mathbb{S}_T+\mathcal{S})(\theta)-(-\mathbb{S}_T+\mathcal{S})(\theta_0)]|\leq C\phi_T(\epsilon) \tag{A.7}$$

for functions ϕ_T such that $\epsilon\to\phi_T(\epsilon)/\epsilon^{\zeta}$ is decreasing for some $\zeta<2$. Note that (A.6) is satisfied due to Assumption 2 and the positive definiteness of $\mathcal{S}_{\theta\theta}$. Hence, we focus on (A.7). Note that we proceed similarly to the proof of Theorem 1.

Write as in $(A.1)$ and proceed as in $(A.2)$ to get

$$\sup_{||\theta-\theta_0||<\epsilon}\sqrt{T}[(\mathcal{S}-\mathbf{S}_T)(\theta)-(\mathcal{S}-\mathbf{S}_T)(\theta_0)]=o(1)+o(\epsilon).$$

Next decompose

$$\sqrt{T}[(\mathbb{S}_T-\mathbf{S}_T)(\theta)-(\mathbb{S}_T-\mathbf{S}_T)(\theta_0)]=\mathbb{I}_1+\mathbb{I}_2,$$

where

$$\begin{aligned}\mathbb{I}_1 &= \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor T\gamma\rfloor}\left(\left(y_{t,T}-x_{t,T}'\beta_1\right)^2-\mathrm{E}\left(y_{t,T}-x_{t,T}'\beta_1\right)^2\right)\\ &\quad -\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor T\gamma_0\rfloor}\left(\left(y_{t,T}-x_{t,T}'\beta_{10}\right)^2-\mathrm{E}\left(y_{t,T}-x_{t,T}'\beta_{10}\right)^2\right),\end{aligned}$$

and $\mathbb{I}_2$ is defined analogously for the summation after $\lfloor T\gamma\rfloor$ and $\lfloor T\gamma_0\rfloor$. Suppose that $\gamma>\gamma_0$. Then,

$$\mathbb{I}_1=A_1+B_1-2C_1,$$

where, with $b_1=(\beta_1-\beta_{10})$ and $g=(\gamma-\gamma_0)$,

$$\begin{aligned}A_1 &= \frac{1}{\sqrt{T}}\sum_{t=1+\lfloor T\gamma_0\rfloor}^{\lfloor T\gamma_0+Tg\rfloor}\left(e_{1t,T}^2-\mathrm{E}e_{1t,T}^2\right)\\ B_1 &= \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor T\gamma_0\rfloor}b_1'\left(x_{t,T}x_{t,T}'-\mathrm{E}x_{t,T}x_{t,T}'\right)b_1\\ C_1 &= \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor T\gamma_0\rfloor}\left(e_{1t,T}x_{t,T}-\mathrm{E}\left(e_{1t,T}x_{t,T}\right)\right)'b_1,\end{aligned}$$

where $e_{1t,T}=y_{t,T}-x_{t,T}'\beta_1$. As in the proof of Theorem 1, it is clear that

$$\mathrm{E}\left(\max_{|b_1|\leq\epsilon}|B_1|+\max_{|b_1|\leq\epsilon}|C_1|\right)=O(\epsilon).$$

Furthermore, another application of Lemma 2 or Lemma 3 to A_1 yields the desired result. That is, defining $S_g=T^{-1/2}\sum_{t=1+\lfloor T\gamma_0\rfloor}^{\lfloor T\gamma_0+Tg\rfloor}\left(e_{1t,T}^2-\mathrm{E}e_{1t,T}^2\right)$, we get for $1<q\leq 2$,

$$\mathrm{E}\left(\max_{g\leq\epsilon}|S_g|\right)\leq\mathrm{E}\left(\max_{g\leq\epsilon}|S_g|^q\right)^{1/q}\leq\left(\frac{K}{T}\sum_{t=1+\lfloor T\gamma_0\rfloor}^{\lfloor T\gamma_0+T\epsilon\rfloor}c_{t,T}^2\right)^{1/q}=O\left(\epsilon^{1/q}\right), \tag{A.8}$$

where K does not depend on T nor ϵ. The same argument applies for the other terms in A_1 as β_1 is bounded. Furthermore, as $|b_1|<\epsilon$, Lemma 3 again yields that

$$\mathrm{E}\left(\max_{|b_1|\leq\epsilon}|B_1|+\max_{|b_1|\leq\epsilon}|C_1|\right)=O(\epsilon).$$

Symmetric argument applies to $\mathbb{I}_2$. As a result, we only need to find r_T such that $r_T^2\phi_T(r_T^{-1}) \leq \sqrt{T}$ for every T. From (A.8), $\phi_T(\epsilon) = C\epsilon^{2/p}$ and hence r_T satisfies

$$r_T^2 r_T^{-2/p} = T^{1/2}.$$

Consequently, $r_T = T^{\frac{p}{4p-4}}$. Hence, the desired result follows. ■

Proof of Theorem 3. Due to Theorem 2 (with $p = 4$), $T^{1/3}(\hat{\theta}-\theta_0) = O_p(1)$ and we can apply the argmax continuous mapping theorem to a reparametrized criterion function. Specifically, let $b = r_T(\beta-\beta_0)$ and $g = r_T(\gamma - \gamma_0)$ where $r_T = T^{1/3}$, $h = (b', g)'$ and $|h| \leq K < \infty$. Also, $b = (b_1', b_2')'$ corresponds to $\beta = (\beta_1', \beta_2')'$, that is, $b_1 = r_T(\beta_1 - \beta_{10})$ and $b_2 = r_T(\beta_2 - \beta_{20})$. And consider the following map

$$h \mapsto \mathbb{M}_T(h) = r_T^2[\mathbb{S}_T(\theta_0 + \frac{h}{r_T}) - \mathbb{S}_T(\theta_0)] = \mathbb{J}_1(h) + \mathbb{J}_2(h),$$

where

$$\begin{aligned}
\mathbb{J}_1 &= \frac{r_T^2}{T}\left[\sum_{t=1}^{\lfloor T\gamma\rfloor}\left(y_{t,T} - x_t'\beta_1\right)^2 - \sum_{t=1}^{\lfloor T\gamma_0\rfloor}\left(y_{t,T} - x_t'\beta_{10}\right)^2\right], \\
\mathbb{J}_2 &= \frac{r_T^2}{T}\left[\sum_{t=1+\lfloor T\gamma\rfloor}^{T}\left(y_{t,T} - x_t'\beta_2\right)^2 - \sum_{t=1+\lfloor T\gamma_0\rfloor}^{T}\left(y_{t,T} - x_t'\beta_{20}\right)^2\right].
\end{aligned}$$

In the following, unless necessary, we mainly focus on the case of $g > 0$ since the case of $g < 0$ can be dealt with analogously.

For $g > 0$, and defining $e_{1t,T} = y_{t,T} - x_t'\beta_{10}$, write

$$\mathbb{J}_1 = A_1 + B_1 + C_1,$$

where

$$\begin{aligned}
A_1 &= \frac{r_T^2}{T}\sum_{t=1+\lfloor T\gamma_0\rfloor}^{\lfloor T\gamma_0 + Tr_T^{-1}g\rfloor}\left(e_{1t,T} - r_T^{-1}x_t'b_1\right)^2 \\
B_1 &= \frac{1}{T}\sum_{t=1}^{\lfloor T\gamma_0\rfloor} b_1'x_t x_t' b_1 \\
C_1 &= -2\frac{r_T}{T}\sum_{t=1}^{\lfloor T\gamma_0\rfloor}\left[\varepsilon_t + x_t'\left(\beta(t/T) - \beta_{10}\right)\right]x_t'b_1.
\end{aligned}$$

$\mathbb{J}_2$, A_2, B_2, and C_2 can be expressed analogously for the summation after $\left\lfloor T\gamma_0 + Tr_T^{-1}g\right\rfloor$. Applying the standard LLN and CLT,

$$B_1 \xrightarrow{p} \gamma_0 b_1' M b_1 \text{ and } C_1 - \mathrm{E}C_1 = o_p(1), \tag{A.9}$$

uniformly in b_1, where $M = \mathrm{E}(x_t x_t')$. And due to (7) and (9),

$$\sup_{b_1}\mathrm{E}C_1 = \sup_{b_1} b_1'\frac{r_T}{T}\sum_{t=1}^{\lfloor T\gamma_0\rfloor} M\left[(\beta(t/T) - \beta_{10})\right] = O(r_T/T).$$

Similarly, uniformly in b_2,

$$B_2 \xrightarrow{p} (1-\gamma_0)\, b_2' M b_2 \text{ and } C_2 = o_p(1). \tag{A.10}$$

Decompose A_1 further to

$$\begin{aligned} A_1 &= \frac{r_T^2}{T} \sum_{t=\lfloor T\gamma_0\rfloor+1}^{\lfloor T\gamma_0+Tr_T^{-1}g\rfloor} e_{1t,T}^2 - 2\frac{r_T}{T} \sum_{t=\lfloor T\gamma_0\rfloor+1}^{\lfloor T\gamma_0+Tr_T^{-1}g\rfloor} e_{1t,T} x_t' b_1 + \frac{1}{T} \sum_{t=\lfloor T\gamma_0\rfloor+1}^{\lfloor T\gamma_0+Tr_T^{-1}g\rfloor} b_1' x_t x_t' b_1 \\ &= D_1 + E_1 + o_p(1). \end{aligned} \tag{A.11}$$

Likewise, for A_2,

$$\begin{aligned} A_2 &= -\frac{r_T^2}{T} \sum_{t=\lfloor T\gamma_0\rfloor+1}^{\lfloor T\gamma_0+Tr_T^{-1}g\rfloor} e_{2t,T}^2 + 2\frac{r_T}{T} \sum_{t=\lfloor T\gamma_0\rfloor+1}^{\lfloor T\gamma_0+Tr_T^{-1}g\rfloor} e_{2t,T} x_t' b_2 - \frac{1}{T} \sum_{t=\lfloor T\gamma_0\rfloor+1}^{\lfloor T\gamma_0+Tr_T^{-1}g\rfloor} b_2' x_t x_t' b_2 \\ &= D_2 + E_2 + o_p(1) \end{aligned}$$

where $e_{2t,T} = y_{t,T} - x_t'\beta_{20}$. We start with E_1 and E_2. Proceeding similarly to the proof of Theorem 1,

$$E_1 - \mathrm{E}E_1 \xrightarrow{p} 0 \text{ and } E_2 - \mathrm{E}E_2 \xrightarrow{p} 0$$

and due to (7), the mean value theorem, (9), and the continuity of β at γ_0,

$$\begin{aligned} \mathrm{E}E_1 &= -\frac{2r_T}{T} \sum_{t=\lfloor T\gamma_0\rfloor+1}^{\lfloor T\gamma_0+Tr_T^{-1}g\rfloor} \mathrm{E}[(y_{t,T} - \beta_{10}' x_t) x_t' b_1] \\ &= -2r_T \left(\int_{\gamma_0}^{\gamma_0+r_T^{-1}g} \mathrm{E}[(\beta(u) - \beta_{10})' x_t x_t' b_1] du + O(T^{-1}) \right) \\ &= -2g\mathrm{E}[(\beta(\bar\gamma) - \beta_{10})' x_t x_t' b_1] + o(1) \\ &\to -2g\mathrm{E}[(\beta(\gamma_0) - \beta_{10})' x_t x_t' b_1] \text{ as } T \to \infty, \end{aligned} \tag{A.12}$$

for some $\bar\gamma \in [\gamma_0, \gamma_0 + r_T^{-1}g]$. Likewise,

$$\mathrm{E}E_2 \to 2g\mathrm{E}[(\beta(\gamma_0) - \beta_{20})' x_t x_t' b_2] \text{ as } T \to \infty. \tag{A.13}$$

For D_1 and D_2,

$$D_1 + D_2 = \underbrace{(D_1 - \mathrm{E}D_1) + (D_2 - \mathrm{E}D_2)}_{F.1} + \underbrace{\mathrm{E}(D_1 + D_2)}_{F.2}.$$

For $F.2$, proceed in the same way as for $\mathrm{E}E_1$ in $(A.12)$ so that

$$\begin{aligned} \mathrm{E}[D_1 + D_2] &= r_T^2 \sum_{t=\lfloor T\gamma_0\rfloor+1}^{\lfloor T\gamma_0+Tr_T^{-1}g\rfloor} \mathrm{E}e_{1t,T}^2 \frac{1}{T} - r_T^2 \sum_{t=\lfloor T\gamma_0\rfloor+1}^{\lfloor T\gamma_0+Tr_T^{-1}g\rfloor} \mathrm{E}e_{2t,T}^2 \frac{1}{T} \\ &= r_T^2 \int_{\gamma_0}^{\gamma_0+r_T^{-1}g} \{\mathrm{E}[\varepsilon_t + x_t'(\beta(u) - \beta_{10})]^2 - \mathrm{E}[\varepsilon_t + x_t'(\beta(u) - \beta_{20})]^2\} du + o(1) \\ &= r_T g \{\mathrm{E}[x_t'(\beta(\gamma_0) - \beta_{10})]^2 - \mathrm{E}[x_t'(\beta(\gamma_0) - \beta_{20})]^2\} \\ &\quad + g^2 \mathrm{E}[(\beta_{20} - \beta_{10})' x_t x_t' \frac{\partial}{\partial\gamma}\beta(\gamma_0)] + o(1) \end{aligned} \tag{A.14}$$

and the term $(A.14)$ is zero due to (8) . Regarding $F.1$, note that

$$F.1 = \frac{1}{r_T} \sum_{t=\lfloor T\gamma_0 \rfloor + 1}^{\lfloor T\gamma_0 + r_T^2 g \rfloor} \left[u_{t,T} - \mathrm{E}(u_{t,T}) \right],$$

where

$$u_{t,T} = 2\delta_0' x_t \left(\varepsilon_t + x_t' \left(\beta \left(\frac{t}{T} \right) - \frac{(\beta_{10} + \beta_{20})}{2} \right) \right). \tag{A.15}$$

The functional central limit Theorem (FCLT) for dependent heterogeneous sequences in Lemma 4 would yield the weak convergence of $F.1$, provided that for any $|g| \leq K$

$$\mathrm{E} \left(\frac{1}{r_T} \sum_{t=\lfloor T\gamma_0 \rfloor + 1}^{\lfloor T\gamma_0 + r_T^2 g \rfloor} \left[u_{t,T} - \mathrm{E}(u_{t,T}) \right] \right)^2 \to \omega^2 g.$$

To show this, note that $\mathrm{E} \left(\frac{1}{r_T} \sum_{t=\lfloor T\gamma_0 \rfloor + 1}^{\lfloor T\gamma_0 + r_T^2 g \rfloor} \left[u_{t,T} - \mathrm{E}(u_{t,T}) \right] \right)^2$ is bounded due to Assumption 3 (i) and (iii). And thus revoking the dominated convergence theorem, it is sufficient to show that for any $\lfloor T\gamma_0 \rfloor \leq s < t \leq \lfloor T\gamma_0 + r_n^2 g \rfloor$,

$$\mathrm{E}(u_{t,T}) = 2\delta_0' \mathrm{E} x_t x_t' \left(\beta \left(\frac{t}{T} \right) - \frac{(\beta_{10} + \beta_{20})}{2} \right) \to 2\delta_0' \mathrm{E} x_t x_t' \psi_0,$$

where $\psi_0 = \beta(\gamma_0) - (\beta_{10} + \beta_{20})/2$, and

$$\begin{aligned}
& \mathrm{E}\left(u_{t,T} u_{s,T} \right) \\
= \; & 4\mathrm{E}\delta_0' x_t \left(\varepsilon_t + x_t' \left(\beta \left(\frac{t}{T} \right) - \frac{(\beta_{10} + \beta_{20})}{2} \right) \right) \delta_0' x_s \left(\varepsilon_s + x_s' \left(\beta \left(\frac{s}{T} \right) - \frac{(\beta_{10} + \beta_{20})}{2} \right) \right) \\
= \; & 4\mathrm{E}\delta_0' x_t \delta_0' x_s x_t' \left(\beta \left(\frac{t}{T} \right) - \frac{(\beta_{10} + \beta_{20})}{2} \right) \left(\varepsilon_s + x_s' \left(\beta \left(\frac{s}{T} \right) - \frac{(\beta_{10} + \beta_{20})}{2} \right) \right) \\
\to \; & 4\mathrm{E}\delta_0' x_t \delta_0' x_s x_t' \psi_0 \left(\varepsilon_s + x_s' \psi_0 \right),
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left(u_{t,T}^2 \right) &= 4\mathrm{E} \left(\delta_0' x_t \right)^2 \left(\varepsilon_t^2 + \left(x_t' \left(\beta \left(\frac{t}{T} \right) - \frac{(\beta_{10} + \beta_{20})}{2} \right) \right)^2 \right) \\
&\to 4\mathrm{E} \left(\delta_0' x_t \right)^2 \left(\varepsilon_t^2 + \left(x_t' \psi_0 \right)^2 \right),
\end{aligned}$$

and finally

$$\begin{aligned}
& \lim_{T\to\infty} \frac{4}{r_T^2} \left(\sum_{t=\lfloor T\gamma_0 \rfloor + 1}^{\lfloor T\gamma_0 + r_T^2 g \rfloor} \mathrm{E} \left(\delta_0' x_t \right)^2 \varepsilon_t^2 + \sum_{t=\lfloor T\gamma_0 \rfloor + 1}^{\lfloor T\gamma_0 + r_T^2 g \rfloor} \mathrm{E} \left(\delta_0' x_t \right)^2 \left(x_t' \psi_0 \right)^2 \right) \\
& + \lim_{T\to\infty} \frac{8}{r_T^2} \sum_{s=\lfloor T\gamma_0 \rfloor + 1}^{\lfloor T\gamma_0 + r_T^2 g \rfloor} \sum_{t=s+1}^{\lfloor T\gamma_0 + r_T^2 g \rfloor} \mathrm{E}\delta_0' x_t \delta_0' x_s x_t' \psi_0 \left(\varepsilon_s + x_s' \psi_0 \right) - 4g \left(\delta_0' \mathrm{E} x_t x_t' \psi_0 \right)^2 \\
= \; & \lim_{T\to\infty} 4g \left[\gamma(0) + 2 \sum_{s=1}^{\lfloor r_T^2 g \rfloor} \gamma(s) - \frac{2}{r_T} \sum_{s=1}^{\lfloor r_T^2 g \rfloor} s\gamma(s) \right] = 4g\omega^2,
\end{aligned}$$

where γ is the autocovariance function of $\delta_0' x_s \left(\varepsilon_s + x_s'\psi_0\right)$, since $s\gamma(s)$ is absolutely summable due to Assumption 3 (iii).

The case with $g < 0$ is treated identically. Finally, note that for any $g_1 > 0$ and $g_2 < 0$,

$$\begin{aligned} & \frac{1}{r_T^2} \sum_{t=\lfloor T\gamma_0 \rfloor + 1}^{\lfloor T\gamma_0 + r_T^2 g_1 \rfloor} \sum_{s=\lfloor T\gamma_0 + r_T^2 g_2 \rfloor}^{\lfloor T\gamma_0 \rfloor} \operatorname{cov}\left(\delta_0' x_s \left(\varepsilon_s + x_s'\psi_0\right), \delta_0' x_t \left(\varepsilon_t + x_t'\psi_0\right)\right) \\ \leq & \frac{1}{r_T^2} \sum_{s=1}^{\infty} s \left|\operatorname{cov}\left(\delta_0' x_{t+s} \left(\varepsilon_{t+s} + x_{t+s}'\psi_0\right), \delta_0' x_t \left(\varepsilon_t + x_t'\psi_0\right)\right)\right| \\ \to & 0. \end{aligned}$$

Therefore,

$$F.1\,(g) \Rightarrow \omega \mathcal{B}_1(g) 1(g > 0) + \omega \mathcal{B}_2(-g) 1(g < 0),$$

where $\mathcal{B}_1$ and $\mathcal{B}_2$ are two independent standard Brownian motions.

Combining all the above results, the proof of Theorem 3 is complete due to the Argmax continuous mapping theorem. ■

Proof of Corollary 1. Let $\tilde{\theta} = \left(\hat{\beta}_1(\gamma_0)', \hat{\beta}_2(\gamma_0)', \gamma_0\right)$. Since $\beta(\gamma_0) = \beta_0 + O_p(T^{-1/2})$,

$$T(\mathbb{S}_T(\theta_0) - \mathbb{S}_T(\tilde{\theta})) = O_p(1), \tag{A.16}$$

which, in turn, implies that

$$\begin{aligned} T^{2/3}(\mathbb{S}_T(\tilde{\theta}) - \mathbb{S}_T(\hat{\theta})) &= T^{2/3}(\mathbb{S}_T(\tilde{\theta}) - \mathbb{S}_T(\theta_0) + \mathbb{S}_T(\theta_0) - \mathbb{S}_T(\hat{\theta})) \\ &= T^{2/3}(\mathbb{S}_T(\theta_0) - \mathbb{S}_T(\hat{\theta})) + o_p(1). \end{aligned}$$

From Theorem 3, we can get closed form expression for $\hat{b}$ for a given g. That is, the FOC yields

$$\frac{\partial}{\partial b}\left[\frac{1}{2}\begin{bmatrix} \hat{b}' & g \end{bmatrix}\begin{bmatrix} \mathcal{S}_{\beta\beta} & \mathcal{S}_{\beta\gamma} \\ \mathcal{S}_{\gamma\beta} & \mathcal{S}_{\gamma\gamma} \end{bmatrix}\begin{bmatrix} \hat{b} \\ g \end{bmatrix}\right] = 0.$$

And thus,

$$\begin{aligned} h'\mathcal{S}_{\theta\theta}h &= b'\mathcal{S}_{\beta\beta}b + b'\mathcal{S}_{\beta\gamma}g + g\mathcal{S}_{\gamma\beta}b + \mathcal{S}_{\gamma\gamma}g^2 \\ &= \left[\mathcal{S}_{\gamma\gamma} - \mathcal{S}_{\gamma\beta}\mathcal{S}_{\beta\beta}^{-1}\mathcal{S}_{\beta\gamma}\right]g^2. \end{aligned}$$

Theorem 3 can be restated as

$$T^{2/3}(\mathbb{S}_T(\theta_0) - \mathbb{S}_T(\hat{\theta})) \xrightarrow{d} \max_{g\in\mathbb{R}}\left[\omega\mathcal{B}(g) - \frac{1}{2}\phi g^2\right],$$

where $\phi = S_{\gamma\gamma} - S_{\gamma\beta}S_{\beta\beta}^{-1}S_{\beta\gamma}$. Since $\omega\mathcal{B}(g) = \mathcal{B}\left(\omega^2 g\right)$, setting

$$\xi^2\omega^2 g = z \text{ and } \xi\phi g^2 = z^2$$

with $\xi = \left(\phi\omega^{-4}\right)^{1/3}$ to apply the change-of-variables yields that

$$T^{2/3}\left[\mathbb{S}_T(\theta_0) - \mathbb{S}_T(\hat{\theta})\right] \xrightarrow{d} \max_{z\in\mathbb{R}} \xi_1^{-1}\left(\mathcal{B}(z) - \frac{1}{2}z^2\right).$$

Finally, due to the Δ-method,

$$\begin{aligned} \mathcal{LR}_T(\gamma_0) &= T^{2/3}(\ln\mathbb{S}_T(\gamma_0) - \ln\mathbb{S}_T(\hat{\gamma})) \\ &= T^{2/3}\frac{1}{\mathbb{S}_T(\gamma_0)}(\mathbb{S}_T(\gamma_0) - \mathbb{S}_T(\hat{\gamma})). \end{aligned}$$

This completes the proof. ■

B Lemmas

For the following Lemmas, we introduce the definition of mixingale arrays.

Definition 2 *The triangular array* $\{X_{t,T}, \mathcal{F}_{t,T}\}$ *is an* L_p*-mixingale for* $p \geq 1$ *if there exist nonnegative constants* $\{c_{t,T} : t = 1, ..., k_T, T = 1, 2, ...\}$ *and* $\{\psi_m : m = 0, 1, ...\}$ *such that* $\psi_m \to 0$ *as* $m \to \infty$ *and for all* $t = 1, ..., k_T$, $T \geq 1$, *and* $m \geq 0$, *we have*

$$\begin{aligned} (a)\ \left\|\mathrm{E}(X_{t,T}|\mathcal{F}_{t-m,T})\right\|_p &\leq c_{t,T}\psi_m \\ (b)\ \left\|X_{t,T} - \mathrm{E}(X_{t,T}|\mathcal{F}_{t+m,T})\right\|_p &\leq c_{t,T}\psi_{m+1} \end{aligned}$$

$\{c_{t,T}\}$ *and* $\{\psi_m\}$ *are called mixingale norms and coefficients respectively. If* $\psi_m = O(m^{-\kappa})$ *for* $\kappa > \kappa_0$, *the array is said to be an* L_p*-mixingale of size* $-\kappa_0$.

Lemma 1 *Suppose the triangular array* $\{X_{t,T}, \mathcal{F}_{t,T}\}$ *is a uniformly integrable* L^1*-mixingale. If* $\lim_{T\to\infty} \frac{1}{k_T} \sum_{t=1}^{k_T} c_{t,T} < \infty$, *then* $\mathrm{E}\left|\bar{X}_T\right| = \mathrm{E}\left\|\frac{1}{k_T} \sum_{t=1}^{k_T} X_{t,T}\right\| \to 0$ *as* $T \to \infty$ *and in consequence* $\bar{X}_T \xrightarrow{p} 0$.

Proof. See Andrews (1988). ■

Lemma 2 *Let* $\{X_{t,T}, \mathcal{F}_{t,T}\}$ *be an* L_p*-mixingale,* $1 < p < 2$, *of size* -1, *and let* $S_k = \sum_{t=1}^{k} X_{t,T}$. *Then,*

$$\mathrm{E}\left(\max_{1\leq j\leq n} |S_j|^p\right) \leq K \sum_{t=1}^{n} c_{t,T}^p$$

with $K = 4^p C_p \left(\frac{p}{p-1}\right)^p (\sum_{k=0}^{\infty} \psi_k)^p$ *where* C_p *is a positive constant and* ψ_k *is summable which is implied by the size of the mixingale.*

Proof. See Theorem 16.11. of Davidson (1994). ■

Lemma 3 *Let* $\{X_{t,T}, \mathcal{F}_{t,T}\}$ *be an* L_2*-mixingale of size* $-1/2$, *and* $S_k = \sum_{t=1}^{k} X_{t,T}$. *Then there exists a constant* K *depending only on* $\{\psi_m\}$ *such that*

$$\mathrm{E}\left(\max_{1\leq j\leq n} S_j^2\right) \leq K \sum_{t=1}^{n} c_{t,T}^2$$

Proof. See McLeish (1975). ■

Lemma 4 *Let* $\{X_{t,T}, \mathcal{F}_{t,T}\}$ *be a triangular array of real-valued random variables on the probability space* $(\Omega, \mathcal{F}, P)$ *and* $W_T(a) = \sum_{t=1}^{\lfloor Ta \rfloor} X_{t,T}$ *for* $a \in [0, 1]$. *Suppose the following conditions hold.*

(a) $\{X_{t,T}\}$ *is* L_p*-NED of size* $-1/2$ *with NED coefficients* $\{\nu_m\}$ *and NED norms* $\{d_{tT}\}$ *such that* $\|X_{t,T}\|_r < \infty$ *for some* $r > p \geq 2$ *and* $\mathrm{E}(X_{t,T}) = 0$ *for all* t, T.

(b) For $c_{t,T} = \max\{\|X_{t,T}\|_r, d_{tT}\}$,

$$\sup_{\substack{0<a<1,\\ 0<d<1-a}} \limsup_{T\to\infty} d^{-1} \sum_{t=\lfloor Ta \rfloor}^{\lfloor T(a+d) \rfloor} c_{t,T}^2 < \infty.$$

Then, the sequence $\{W_T\}$ *is tight in Stone's topology on* $\mathcal{D}[0,1]$ *where* $\mathcal{D}$ *is the Skorokhod space defined on* $[0,1]$*, i.e. the space of all right continuous function with left limits on* $[0,1]$*. In addition, any weak limit process is almost surely continuous. Moreover, in addition to (a) and (b), suppose the following condition holds.*

(c) For each $a \in [0,1]$*,*

$$\mathrm{E}\left(W_T^2(a)\right) \to a \text{ as } T \to \infty.$$

Then, $\{W_T\}$ *converges weakly to a standard Wiener process on* $\mathcal{D}$*.*

Proof. See Theorem 2.11 in Wooldridge and White (1988). Theorem 2.11 in Wooldridge and White (2008) is slightly modified here since Theorem 2.11 requires uniform integrability of $\left\{X_{t,T}^2/c_{t,T}^2\right\}$, which is however implied by (a) of this Theorem due to Proposition 2.9 in Wooldridge and White (1988). ■

C Data Appendix

C.1 USA (1960:1 to 2012:3)

Real Gross Domestic Product is based on billions of chained 2005 dollars, quarterly, seasonally adjusted at annual rates. The data source is the US Federal Reserve Economic Data Base (FRED).

Long-term interest rates are 10-Year Treasury Bond Constant Maturity Rates. The series are converted to quarterly through averages over business days. They are expressed as a percentage. The data source is FRED.

Short-term interest rates are 3-Month Treasury (Secondary) Bill Market Rates. The series are converted to quarterly through averages over business days at discount basis. They are expressed as a percentage. The data source is FRED.

C.2 Canada (1960:1 to 2012:3)

Gross Domestic Product is seasonally adjusted in constant 2007 prices. The data source is the OECD Economic Outlook database.

Long-term interest rates are long-term government bond yields, which are expressed as a percentage per annum. They are average yields of issues with original maturity of 10 years and over. The data source is the International Financial Statistics (IFS) database.

Short-term interest rates are 3-month treasury bill rates, which are expressed as a percentage per annum. They are weighted averages of the yields on successful bids for 3-month bills. The data source is the IFS database.

C.3 UK (1960:1 to 2012:3)

Gross Domestic Product is seasonally adjusted in constant 2009 prices. The data source is the OECD Economic Outlook database.

Long-term interest rates are long-term government bond yields, which are expressed as a percentage per annum. They are the gross redemption bond yields of government bonds with maturity of 20 years. The data source is the International Financial Statistics (IFS) database.

Short-term interest rates are 91-day treasury bill rates, which are expressed as a percentage per annum. They are the tender rates at which 91-day bills are allotted. The data source is the International Financial Statistics (IFS) database.

C.4 France (1969:4 to 2012:3)

Gross Domestic Product is seasonally adjusted in constant 2005 prices. The data source is the OECD Economic Outlook database.

Long-term interest rates are long-term government bond yields, which are expressed as a percentage per annum. They are the average yield of public sector bonds with original maturities of more than five years. The data source is the IFS database.

Short-term interest rates are 3-month treasury bill rates, which are expressed as a percentage per annum. The data source is the IFS database.

C.5 Germany (1970:2 to 2012:1)

The precentage changes of real Gross Domestic Product from the previous period are seasonally adjusted. The data source is the database of Main Economic Indicators, OECD.

Long-term interest rates are long-term government bond yields, which are expressed as a percentage per annum. They are average yields of bonds with remaining maturity of more than three years. The data source is the IFS database.

Short-term interest rates are money market rates, which are expressed as a percentage per annum. They are averages of ten daily average quotations for overnight credit. The data source is the IFS database.

C.6 Italy (1970:4 to 2012:2)

Gross Domestic Product is seasonally adjusted in constant 2005 prices. The data source is the OECD Economic Outlook database.

Long-term interest rates are long-term government bond yields, which are expressed as a percentage per annum. The data prior to 1991 are average yields to maturity on the treasury bonds with maturities of 15 to 20 years. The data between 1991 - 1998 are average yields to maturity on bonds with residual maturities between 9 to 10 years. The data from 1999 are average yields to maturity on the ten-year treasury bonds. The data source is the IFS database.

Short-term interest rates are money market rates, which are expressed as a percentage per annum. They are 3-month interbank rates. The data source is the IFS database.

C.7 Japan (1966:3 to 2012:3)

Gross Domestic Product is seasonally adjusted in constant 2005 prices. The data source is the OECD Economic Outlook database.

Long-term interest rates are long-term government bond yields, which are expressed as a percentage per annum. The data prior to 1999 are average yields to maturity of all ordinary government bonds, and after that period they are average yields of government bonds with 10-year maturity. The data source is the IFS database.

Short-term interest rates are money market rates, which are expressed as a percentage per annum. They are lending rates for collateral and overnight loans. The data source is the IFS database.

C.8 Australia (1969:2 to 2012:3)

Gross Domestic Product is seasonally adjusted in constant 2009-2010 prices. The data source is the OECD Economic Outlook database.

Long-term interest rates are 15-year treasury bond yields, which are expressed as a percentage per annum. The data source is the IFS database.

Short-term interest rates are money market rates, which are expressed as a percentage per annum. They are weighted average short-term rates of outstanding loans. The data source is the IFS database.

D Tables and Figures

Table 1. Asymptotic Critical Values for $\lambda^{\dagger}$

	.85	.90	.925	.95	.975	.99
$P(\lambda \leq x)$	1.582	1.775	1.906	2.085	2.369	2.728

$\dagger \lambda = \xi \mathcal{LR}_T(\gamma_0)$

Table 2. Percentage improvement of each averaging method[1)] over Post break (No averaging): Non-testing[3)]

	c = 5[2)]		c = 10		c = 20		Ave c	
AveLR	19.78	11.51	22.07	11.54	22.12	11.54	22.11	11.77
	(3.14)	(0.99)	(4.53)	(0.99)	(4.55)	(0.98)	(4.06)	(1.01)
Rwgt1	20.39	11.74	20.39	11.74	20.39	11.74	20.39	11.74
	(3.24)	(1.16)	(3.24)	(1.16)	(3.24)	(1.16)	(3.24)	(1.16)
Rwgt2	20.67	11.25	20.67	11.25	20.67	11.25	20.67	11.25
	(3.61)	(1.30)	(3.61)	(1.30)	(3.61)	(1.30)	(3.61)	(1.30)
AveW	21.43	11.19	21.43	11.19	21.43	11.19	21.43	11.19
	(3.86)	(0.98)	(3.86)	(0.98)	(3.86)	(0.98)	(3.86)	(0.98)
Optwgt	10.58	7.59	10.58	7.59	10.58	7.59	10.58	7.59
	(1.34)	(0.65)	(1.34)	(0.65)	(1.34)	(0.65)	(1.34)	(0.65)
ROC	20.37	10.21	20.37	10.21	20.37	10.21	20.37	10.21
	(3.96)	(0.91)	(3.96)	(0.91)	(3.96)	(0.91)	(3.96)	(0.91)
AveMal	8.96	6.25	8.96	6.25	8.96	6.25	8.96	6.25
	(0.57)	(0.41)	(0.57)	(0.41)	(0.57)	(0.41)	(0.57)	(0.41)

Notes: 1) AveLR: LR averaging; Rwgt1: Robust Weight1; Rwg2: Robust Weight2; AveW: Average Window; Optwgt: Optimal Weight; ROC: Reverse Ordered CUSUM AveMal: Average using a Mallows criterion.

2) Forecast errors are recursively computed beginning at $\lfloor \frac{T}{2} \rfloor + 1$ and $\lfloor \frac{2T}{3} \rfloor + 1$, which is reflected in two columns for each c and values in parentheses denote variance of percentage improvement of each averaging method over Post break.

3) For AveLR, we do not test the presence of a break.

Table 3. Percentage improvement of each averaging method[1)] over Post break (No averaging): Testing[3)]

	c = 5[2)]		c = 10		c = 20		Ave c	
AveLR	19.12	11.44	20.86	11.47	20.90	11.47	20.48	11.46
	(2.91)	(1.02)	(3.90)	(1.02)	(3.93)	(1.02)	(3.66)	(1.02)
Rwgt1	20.39	11.74	20.39	11.74	20.39	11.74	20.39	11.74
	(3.24)	(1.16)	(3.24)	(1.16)	(3.24)	(1.16)	(3.24)	(1.16)
Rwgt2	20.67	11.25	20.67	11.25	20.67	11.25	20.67	11.25
	(3.61)	(1.30)	(3.61)	(1.30)	(3.61)	(1.30)	(3.61)	(1.30)
AveW	21.43	11.19	21.43	11.19	21.43	11.19	21.43	11.19
	(3.86)	(0.98)	(3.86)	(0.98)	(3.86)	(0.98)	(3.86)	(0.98)
Optwgt	10.58	7.59	10.58	7.59	10.58	7.59	10.58	7.59
	(1.34)	(0.65)	(1.34)	(0.65)	(1.34)	(0.65)	(1.34)	(0.65)
ROC	20.37	10.21	20.37	10.21	20.37	10.21	20.37	10.21
	(3.96)	(0.91)	(3.96)	(0.91)	(3.96)	(0.91)	(3.96)	(0.91)
AveMal	8.96	6.25	8.96	6.25	8.96	6.25	8.96	6.25
	(0.57)	(0.41)	(0.57)	(0.41)	(0.57)	(0.41)	(0.57)	(0.41)

For notes 1) and 2), see Table 2.

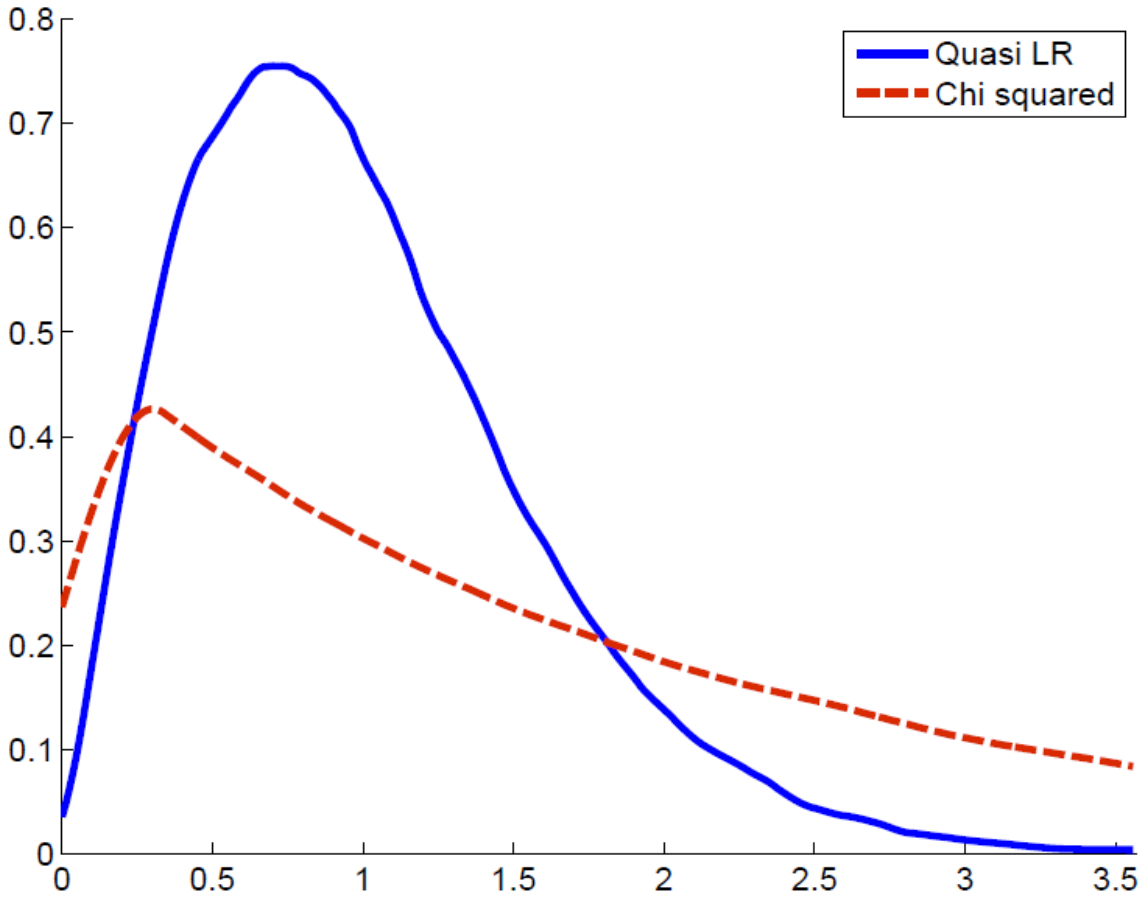
3) For AveLR, we test the presence of a break in every recursion.

Table 4. Comparison of AveLR with no break models[1)]

	c = 5[2)]		c = 10		c = 20		Ave c	
AveLR								
Test	19.12	11.44	20.86	11.47	20.90	11.47	20.48	11.46
	(2.91)	(1.02)	(3.90)	(1.02)	(3.93)	(1.02)	(3.66)	(1.02)
No Test	19.78	11.51	22.07	11.54	22.12	11.54	22.11	11.77
	(3.14)	(0.99)	(4.53)	(0.99)	(4.55)	(0.98)	(4.06)	(1.01)
AR(4)	16.28	6.56	16.28	6.56	16.28	6.56	16.28	6.56
	(6.41)	(2.50)	(6.41)	(2.50)	(6.41)	(2.50)	(6.41)	(2.50)
AR(1)	-4.90	-29.0	-4.90	-29.0	-4.90	-29.0	-4.90	-29.0
	(27.04)	(51.09)	(27.04)	(51.09)	(27.04)	(51.09)	(27.04)	(51.09)
UnitR	-90.93	-94.24	-90.93	-94.24	-90.93	-94.24	-90.93	-94.24
	(199.47)	(127.01)	(199.47)	(127.01)	(199.47)	(127.01)	(199.47)	(127.01)

Note 1) AveLR, Test: LR averaging with testing in every recursion; AveLR, No Test: LR averaging with no test of the presence of a break; UnitR: Unit Root process
For note 2), see Table 2.

Figure 1: Limit distribution of the rescaled QLR and the Chi-squared distribution



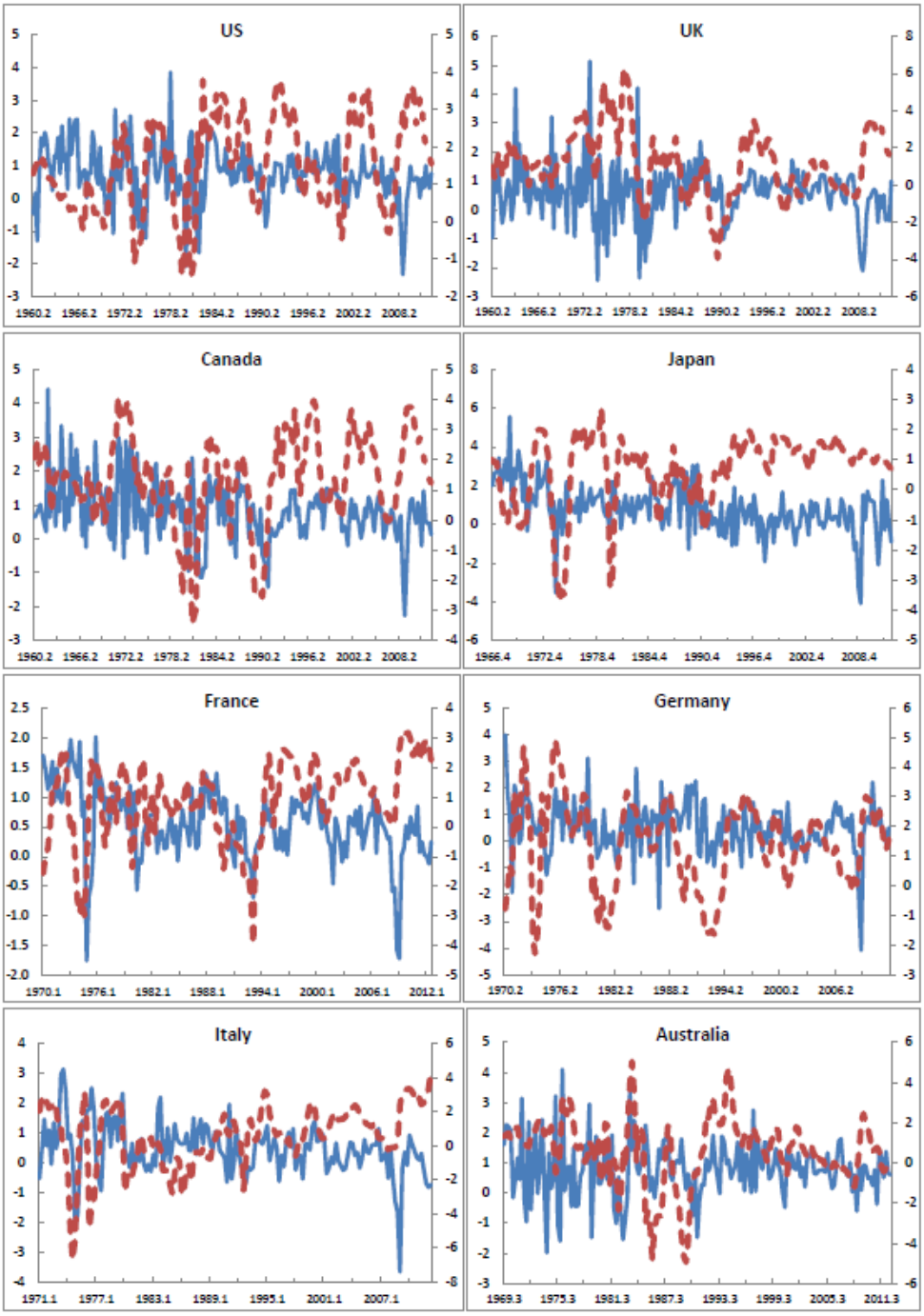


Figure 2: Movements of real GDP growth rate and yield spread

For each country, 1) The solid line and the dotted line denote the real GDP growth rate and the yield spread respectively ; 2) the left vertical axis is for the real GDP growth rate time series and the right vetical axis for the yield spread; 3) Both time series are in percentage.

Figure 3: Residual Sum of Squares

Note: For each country, the residual sum of squares are computed and plotted against a time point when this time point is used for a split point.

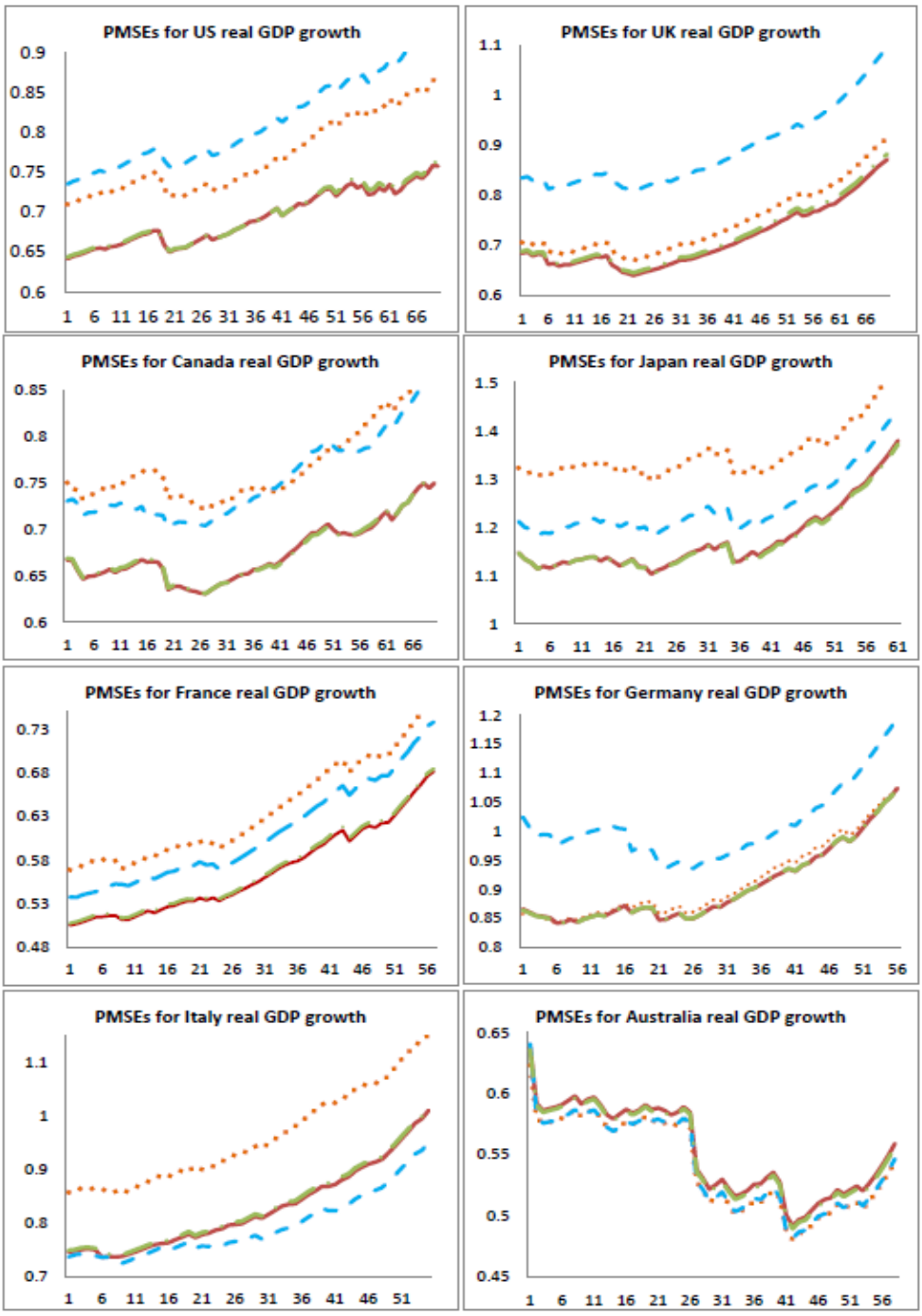


Figure 4: One Period Ahead PMSEs from different Forecasting Methods: Empirical Study II

Notes:1) These PMSEs are computed from one period ahead forecast errors as the estimation window increases by one. See Section 4.2.2 for more details; 2) The solid line is used for our method (AveLR), the short dashed line for Post break, a dotted line for a linear model without a break (FullW) and the dot-dash line for Averaging Window (AveW).