Title
Model Instability and Choice of Observation Window

Permalink
https://escholarship.org/uc/item/8zx626k6

Authors
Pesaran, Hashem
Timmermann, Allan

Publication Date
1999-09-01

99-19

UNIVERSITY OF CALIFORNIA, SAN DIEGO

DEPARTMENT OF ECONOMICS

MODEL INSTABILITY AND CHOICE OF OBSERVATION WINDOW

BY

M. HASHEM PESARAN

AND

ALLAN TIMMERMANN

DISCUSSION PAPER 99-19
SEPTEMBER 1999

Model Instability and Choice of Observation Window

M. Hashem Pesaran
University of Cambridge and USC

Allan Timmermann
University of California, San Diego

September 29, 1999

Abstract

Recent evidence suggests that many economic time series are subject to structural breaks. In the presence of breaks, including historical data prior to the most recent break to estimate a forecasting model will lead to prediction errors that are biased but also may have a smaller variance. This paper examines the trade-off between the bias and variance of forecast errors and proposes a new set of reversed Cusum procedures to determine the window size that minimizes mean squared forecast error. This window size varies over time and depends on the size of the break, the distance to the break and the squared correlation coefficient between predicted and realized values. The forecasting performances of several procedures for determination of window size are compared in a simulation experiment and in a recursive prediction exercise using data on US stock returns. We find evidence that out-of-sample forecasting performance can be improved by explicitly accounting for breaks and adopting the proposed method for optimally determining the window size.

JEL Classifications: C22, C53, G10.

Key Words: Parameter instability, forecasting, expanding and rolling windows, reversed Cusum or Cusum squared tests, multiple breaks, choice of observation window, predictability of US stock returns.

1. **Introduction**

Structural breaks have been an important preoccupation of economists for a long time. In the context of linear regression models Chow (1960) derived an F-test for a structural break when the point of the break is given, while Brown, Durbin and Evans (1975) derive Cusum and Cusum Squared tests that are also applicable when the time of the break is unknown. More recently, the literature has extended the earlier tests to dynamic models with unit roots and tests for consistent estimation of the size and timing of multiple break points have also been developed.[1] Following these developments, applied studies have reported evidence of breaks in several economic time series.[2]

If structural breaks characterize a particular time series, using the full historical data series to estimate a forecasting model will lead to forecast errors that are no longer unbiased, although they may also have a lower variance. Given this trade-off, it is far from clear how much of the data should be used to estimate a prediction model that minimizes out-of-sample mean squared forecast errors. This question, which we address in the current paper, is clearly an important issue. In the context of forecasting performance, Clements and Hendry (1999) go as far as stating that "deterministic shifts are a primary source of serious forecast failure" (p. 28) and that other problems related to model misspecification, failure to impose true restrictions, measurement errors etc are of relatively less significance.

Several informal procedures have been developed to handle non-stationarities in time series analysis. Most widespread is perhaps the approaches of using a rolling window of fixed size or to apply exponentially declining weights to past observations. Neither of these methods is likely to work well if the underlying time series undergoes sudden breaks in its conditional mean. Using a short fixed window size or a small degree of smoothing may work well immediately after a break but will discard valuable information as the distance to the break grows. Similarly, if a large fixed window size or a large degree of smoothing is used, it will take longer for the estimated model to recognize that a break has occurred, producing biased

[1] See, for example, Ploberger, Kramer and Kontrus (1989), Hansen (1992), Andrews (1993), Inclan and Tiao (1994), Andrews and Ploberger (1996), Chu, Stinchcombe and White (1996) and Bai and Perron (1998a).

[2] See, for example, Alogoskoufis and Smith (1991), Garcia and Perron (1996), Bai and Perron (1998a,b), Clements and Hendry (1998, 1999) and Timmermann (1998).

forecasts in the interim. Considerations such as these suggest that a time-varying window size is called for. Ideally, the window size should be large far away from the most recent break to allow for efficient estimation of the forecasting model. Closer to the most recent break the window size should be short to avoid using too much data prior to the occurrence of the break which will bias the estimates of the forecasting model.

In this paper we compare the forecasting performance of unconditional and conditional approaches to determination of window size. Unconditional methods such as a rolling or an expanding window at most let the window size vary as a deterministic function of time. In contrast, conditional approaches treat the window size as a parameter and attempt to estimate the point of the most recent break as well as its size. Based on these parameter estimates, the possible trade-off between bias and variance of the forecast error is next explored in selecting an appropriate window size. A particularly simple approach to the estimation of the window size is to apply the Cusum or Cusum squared procedures to observations reversed in time so that the last observation is placed first, the penultimate observation second and so on. We refer to this as the "reversed" Cusum or Cusum squared tests. As we shall argue below, some of the undesirable properties of the standard Cusum type tests in identifying break points now to some extent become strengths when the test is applied to observations reversed in time to select an observation window for forecasting purposes.

The plan of the paper is as follows. Section 2 derives formal results on the optimal choice of observation window in the presence of a structural break when the objective is to minimize the mean squared forecast error. Section 3 considers the case where the objective is to maximize the market timing value of the forecast. Section 4 presents further theoretical results demonstrating that the basic trade-off involved in determining the optimal window size is not confined to a "rare structural break" model but holds more generally for models with time-varying parameters. Section 5 discusses several approaches to determination of the observation window when breaks are present in the data generating process and develops the reversed Cusum procedures for real time forecasting. Section 6 conducts a Monte Carlo experiment on the performance of the alternative methods. Section 7 presents empirical evidence on structural breaks in a model for US stock returns and examines the empirical performance of the procedures in a recursive prediction experiment.

Section 8 concludes.

2. **Optimal Window Size under a Single Structural Break**

Consider the simple linear regression model subject to a single structural break

$$\begin{aligned} y_t &= \boldsymbol{\beta}_1'\mathbf{x}_t + u_t, \qquad u_t \sim IID(0, \sigma_1^2), \qquad t = 1, 2, ..., T_1 \\ & \boldsymbol{\beta}_2'\mathbf{x}_t + u_t, \qquad u_t \sim IID(0, \sigma_2^2), \qquad t = T_1 + 1, ..., T + 1, \end{aligned} \tag{1}$$

where y_t is some univariate stochastic process, $\mathbf{x}_t$ is a $p \times 1$ vector of known regressors, $\boldsymbol{\beta}_i$ $(i = 1, 2)$ are $p \times 1$ vectors of regression coefficients, and u_t is a serially uncorrelated error term that is independently distributed of $\mathbf{x}_s$ for all t and s, possibly with a shift in its variance from σ_1^2 to σ_2^2 at the time of the break point. Assuming that $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ or $\sigma_1^2 \neq \sigma_2^2$, it follows that there is a structural break in the data generating process at time T_1. Suppose that we know that $\boldsymbol{\beta}$ has changed at T_1 and our interest lies in forecasting y_{T+1} given the observations $\{y_t, t = 1, ..., T\}$ and $\{\mathbf{x}_t,\ t = 1, ..., T, T + 1\}$. Which fraction of the observations should we use to estimate a model that, when used to generate forecasts, will minimize the expected mean squared forecast error? Here we are not concerned with the classical problem of identifying the exact point of the break, but rather the fraction of the sample information that it is optimal to use in order to forecast out of sample on the assumption that a structural break has in fact occurred. The standard solution is to only use observations over the post-break period $(t = T_1 + 1, ..., T)$ to estimate the model. For the purpose of forecasting, we shall see that this will not necessarily be optimal.

Let Λ denote the fraction of sample observations to be used in estimation for the purpose of forecasting y_{T+1}, and denote by $m = T - [T\Lambda] + 1$, where $[T\Lambda]$ stands for the integer part of $T\Lambda$. Naturally we assume that the most recent fraction of observations is used for forecasting. Let $\mathbf{X}_{m,T}$ be the $(T - m + 1) \times p$ matrix of observations on the x-variables, while $\mathbf{Y}_{m,T}$ is the $(T - m + 1)$ vector of observations on the dependent variable whose value for period $T + 1$ we are interested in forecasting. Defining the quadratic form $\mathbf{Q}_{\tau,T_i} = \mathbf{X}_{\tau,T_i}'\mathbf{X}_{\tau,T_i}$ so that $\mathbf{Q}_{\tau,T_i} = 0$ if $\tau > T_i$, the OLS estimator of $\boldsymbol{\beta}$ based on using the fraction Λ of the observations is given by

$$\widehat{\boldsymbol{\beta}}_T(m) = \mathbf{Q}_{m,T}^{-1}\mathbf{X}'_{m,T}\mathbf{Y}_{m,T}. \tag{2}$$

The forecast error in the prediction of y_{T+1} will be a function of the data sample used to estimate β and is given by

$$e_{T+1}(m) = y_{T+1} - \widehat{y}_{T+1} = \left(\boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_T(m)\right)' \mathbf{x}_{T+1} + u_{T+1}. \tag{3}$$

Notice that we implicitly assume that it is known that there is no break in the regression model in period $T+1$. Otherwise the best forecast would need to consider the distribution from which new regression parameters are drawn after a break. Since typically there are relatively few breaks in most economic time series, we do not believe that information is readily available on the meta distribution determining the size and frequency of the breaks.

2.1. *Conditional MSFE results*

We first consider the case where the prediction can be conditioned on the sequence of x_t values. Taking expectations conditional on $\mathbf{X}_{T+1} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{T+1}\}$, we get the conditional bias in the forecast error:

$$bias(m|\mathbf{X}_{T+1}) \equiv E[e_{T+1}(m)|\mathbf{X}_{T+1}] = \left(\boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_T(m)\right)' \mathbf{x}_{T+1}. \tag{4}$$

Furthermore, it can be shown that

$$\begin{aligned} e_{T+1}(m|\mathbf{X}_{T+1}) &= \left(\boldsymbol{\beta}'_2 - \mathbf{Y}'_{m,T}\mathbf{X}_{m,T}\mathbf{Q}_{m,T}^{-1}\right)\mathbf{x}_{T+1} + u_{T+1} \\ &= (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)'\,\mathbf{Q}_{m,T_1}\mathbf{Q}_{m,T}^{-1}\mathbf{x}_{T+1} - \mathbf{u}'_{m,T}\mathbf{X}_{m,T}\mathbf{Q}_{m,T}^{-1}\mathbf{x}_{T+1} + u_{T+1}, \end{aligned} \tag{5}$$

where $\mathbf{u}_{m,T} = (u_m, u_{m+1}, .., u_T)'$. Squaring this expression and taking expectations, the conditional mean squared forecast error (MSFE) can be computed as follows:

$$\begin{aligned} MSFE(m|\mathbf{X}_{T+1}) &= E\left[e_{T+1}^2(m)|\mathbf{X}_{T+1}\right] \\ &= \sigma_2^2 + \sigma_2^2\boldsymbol{\mu}'\mathbf{Q}_{m,T_1}\mathbf{Q}_{m,T}^{-1}\mathbf{x}_{T+1}\mathbf{x}'_{T+1}\mathbf{Q}_{m,T}^{-1}\mathbf{Q}_{m,T_1}\boldsymbol{\mu} \\ &\quad + tr(\mathbf{X}_{m,T}\mathbf{Q}_{m,T}^{-1}\mathbf{x}_{T+1}\mathbf{x}'_{T+1}\mathbf{Q}_{m,T}^{-1}\mathbf{X}'_{m,T}\boldsymbol{\Sigma}_{m,T}), \end{aligned} \tag{6}$$

where $\boldsymbol{\mu} = (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)/\sigma_2$, $tr(.)$ is the trace operator and $\boldsymbol{\Sigma}_{m,T} = E[\mathbf{u}_{m,T}\mathbf{u}'_{m,T}]$ is a $(T-m+1) \times (T-m+1)$ diagonal matrix with σ_1^2 in the first $T_1 - m + 1$ diagonal places and σ_2^2 in the remaining $T - T_1$ places.

Intuition is gained by considering the case with a single regressor ($p = 1$):

$$e_{T+1}(m) = u_{T+1} + \theta_m(\beta_2 - \beta_1)x_{T+1} - v_T(m)x_{T+1}, \tag{7}$$

where

$$\theta_m = \frac{\sum\limits_{t=m}^{T_1} x_t^2}{\sum\limits_{t=m}^{T} x_t^2},\ v_T(m) = \frac{\sum\limits_{t=m}^{T} x_t u_t}{\sum\limits_{t=m}^{T} x_t^2},$$

and where $\theta_m \equiv \theta_m(T_1, T) = \frac{Q_{m,T_1}}{Q_{m,T}}$. Hence the conditional MSFE becomes

$$E\left[e_{T+1}^2(m)|X_{T+1}\right] = \sigma_2^2 + \sigma_2^2 x_{T+1}^2 \left\{\mu^2\theta_m^2 + \frac{\psi\theta_m + 1}{\Sigma_{t=m}^T x_t^2}\right\}, \tag{8}$$

where $\psi = (\sigma_1^2 - \sigma_2^2)/\sigma_2^2$ is the proportional decline in the variance after the break. Since $m = (T+1) - [T\Lambda]$, the MSFE is clearly a function of Λ and need not be monotonic in m. Suppose Λ is increased (m is decreased), so that a larger fraction of the recent observations is used for forecasting. The result is a higher θ_m, leading to an increased squared bias. However, $\left(\Sigma_{t=m}^T x_t^2\right)^{-1}$ decreases and the total effect on the MSFE depends on the balance between these two factors as well as on the extent of the squared structural break in the regression parameters (μ) and in the variances (ψ). Clearly there is a trade-off between using a biased estimate, which results from including observations from the first regime, and getting lower variability of the forecast error by including observations from the first regime.

The optimal window size can be determined from the value of m that minimizes the conditional MSFE:

$$m^* = \underset{m=1,..,T_1+1}{\arg\min} \left\{E\left[e_{T+1}^2(m)|X_{T+1}\right]\right\}. \tag{9}$$

The fraction of observations from the first regime used in forecasting depends in a complicated manner on the degree of the structural change, as measured by $\mu^2 = (\beta_2 - \beta_1)^2/\sigma_2^2$ and $\psi = (\sigma_1^2 - \sigma_2^2)/\sigma_2^2$, and the relative sample variation in x_t during the second as compared to the first regime.

A simple recursive decision rule can be developed to determine the window size that minimizes the conditional MSFE. The choice between whether to start with observation $t = m + 1$ rather than $t = m$, depends on whether

$$E\left[e_{T+1}^{2}(m)|\mathbf{X}_{T+1}\right] > E\left[e_{T+1}^{2}(m+1)|\mathbf{X}_{T+1}\right], \quad m = T_1, T_1-1, ..., 1. \tag{10}$$

This condition is satisfied if

$$\begin{aligned} & \sigma_2^2\boldsymbol{\mu}'\left(\mathbf{Q}_{m,T_1}\mathbf{Q}_{m,T}^{-1} - \mathbf{Q}_{m+1,T_1}\mathbf{Q}_{m+1,T}^{-1}\right)\mathbf{x}_{T+1}\mathbf{x}_{T+1}'\left(\mathbf{Q}_{m,T}^{-1}\mathbf{Q}_{m,T_1} + \mathbf{Q}_{m+1,T}^{-1}\mathbf{Q}_{m+1,T_1}\right)\boldsymbol{\mu} \\ > & \ tr\left(\mathbf{X}_{m,T}\mathbf{Q}_{m,T}^{-1}\mathbf{x}_{T+1}\mathbf{x}_{T+1}'\mathbf{Q}_{m,T}^{-1}\mathbf{X}_{m,T}'\boldsymbol{\Sigma}_{m,T}\right) \\ & -tr\left(\mathbf{X}_{m+1,T}\mathbf{Q}_{m+1,T}^{-1}\mathbf{x}_{T+1}\mathbf{x}_{T+1}'\mathbf{Q}_{m+1,T}^{-1}\mathbf{X}_{m+1,T}'\boldsymbol{\Sigma}_{m+1,T}\right). \end{aligned} \tag{11}$$

Again the expression simplifies somewhat in the case with a single regressor ($p = 1$):

$$\mu^2\left(\theta_m^2 - \theta_{m+1}^2\right) > \frac{\psi\theta_{m+1} + 1}{Q_{m+1,T}} - \frac{\psi\theta_m + 1}{Q_{m,T}}. \tag{12}$$

To enhance the intuition from this expression, we initially analyze whether it is optimal to include in the estimation of β only observations from the second regime. This will be the case if the following condition holds:

$$E\left[e_{T+1}^{2}(T_1)|\mathbf{X}_{T+1}\right] > E\left[e_{T+1}^{2}(T_1+1)|\mathbf{X}_{T+1}\right]. \tag{13}$$

Setting $m = T_1$ we see that $\theta_{T_1} = x_{T_1}^2 / \sum\limits_{t=T_1}^{T} x_t^2$, and $\mathbf{Q}_{m+1,T_1} = \theta_{T_1+1} = 0$. Since x_{T_1} is the last observation in the first regime, only observations after the break point will be used if

$$\mu^2 x_{T_1}^2 + \psi > \frac{Q_{T_1,T}}{Q_{T_1+1,T}}, \tag{14}$$

or, equivalently, if

$$\mu^2 x_{T_1}^2 + \psi > \frac{\sigma_2^2 / \sum\limits_{t=T_1+1}^{T} x_t^2}{\sigma_2^2 / \sum\limits_{t=T_1}^{T} x_t^2}. \tag{15}$$

This last form has a particularly intuitive interpretation. The first term on the left hand side is the squared bias in the estimate of y_T induced by including the extra information from the first regime, while the second term is the proportional decrease in volatility after the break. The term on the right hand side is the ratio of the efficiency of the parameter estimates, as measured by the two variances of β_2 based on the longer sample $[T_1, T]$ and the smaller sample $[T_1 + 1, T]$, respectively.

Some special cases are of immediate interest. It is not optimal to include pre-break observations when either μ^2 is very large or σ_1^2 is much larger than σ_2^2 so that ψ is large. Hence if the break in the mean parameters is high or the pre-break error variance is much higher than the post-break error variance, then only post-break observations should be used in the estimation. However, even if a sizeable break in the mean has occurred, it may still be optimal to include pre-break data provided that the variance of the regression equation is smaller before the break occurred. More generally, from (15) the following comparative statics results can be obtained:

Proposition 1 *Suppose the objective is to minimize mean squared forecast error. Then it is more likely that it is optimal to include observations prior to the break to estimate the parameters of the regression model if*

(i) the break in the mean parameters (μ) is small

(ii) the variance parameter increases at the point of the break ($\sigma_1^2 < \sigma_2^2$)

(iii) the post-break window size ($\nu_2 = T - T_1$) is small.

Provided it is optimal to use pre-break observations to estimate β, the next issue that naturally arises is how many pre-break observations to use. In the following proposition we establish conditions under which a recursive stopping rule can be used to determine the optimal window size.

Proposition 2 *Consider the univariate regression model with a single break point occurring in the regression coefficient ($\beta_1 \neq \beta_2$). Then there exists an optimal stopping rule for determining the window size that minimizes the MSFE conditional on X_{T+1}. This rule is to choose a window of observations $[m^*, ..., T]$ where m^* is the largest value of m for which the following condition holds*

$$MSFE(m-1) > MSFE(m)$$

for $m = T_1, T_1 - 1, ..., 1$ arranged in declining order.

This proposition, which is proved in the Appendix, suggests that even if we knew the point at which a structural break has taken place, it may still be worthwhile to utilize information before the break to forecast in the second regime.

A key property of the expression for the MSFE is the difference in the rates of dependence of the squared bias and the variance of the forecast error with respect to the window size $(T - m + 1)$. This suggests that, as one moves further away from the break point, the optimal window size need not expand uniformly, but may initially decline. Hence the optimal window, when plotted against the post-break window size $(T - T_1)$ may be U-shaped. It also means that a rolling window is likely to be suboptimal.

Proposition 3 *Suppose a break has occurred in the conditional mean* $(\beta_1 \neq \beta_2)$*. Then the optimal window size conditional on* X_{T+1} *never expands by more than a single observation as the sample size,* T*, increases. Furthermore, the optimal window size need not be a monotonically increasing function of* $(T - T_1)$*, but may initially decrease before it eventually increases.*

2.2. *Unconditional MSFE results*

The decision rule developed above conditions the optimal window size on the sequence of realizations of $\mathbf{x}_t$. However, it is also of interest to investigate which factors determine the optimal window size on average, i.e. across the possible realizations of x_t. Provided a process is postulated for $\{\mathbf{x}_t\}$ one can integrate out $\mathbf{X}_{T+1}$ in the expression for the optimal window size and the resulting MSFE. In general this can be done through Monte Carlo simulation. However, if the joint process generating $\{u_t, \mathbf{x}_t\}$ is sufficiently simple, analytical results can also be obtained. Considering once again the case with a single regressor, from (7) the first two unconditional moments of $e_{T+1}(m)$ are

$$\begin{aligned} E[e_{T+1}(m)] &= (\beta_1 - \beta_2)E[\theta_m x_{T+1}], \qquad (16) \\ E[e^2_{T+1}(m)] &= (\beta_1 - \beta_2)^2 E[\theta_m^2 x^2_{T+1}] + \sigma_2^2 + \sigma_2^2 E\left[\frac{x^2_{T+1}}{\sum_{t=m}^{T} x_t^2} + \frac{x^2_{T+1}\theta_m\psi}{\sum_{t=m}^{T} x_t^2}\right]. \end{aligned}$$

In the interesting special case where u_t and x_t are *i.i.d.* and normally distributed

$$\begin{pmatrix} u_t \\ x_t \end{pmatrix} \sim IIN \left[\begin{pmatrix} 0 \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \omega^2 \end{pmatrix} \right],$$

we have

$$E[x_{T+1}^2 \theta_m^2] = E[x_{T+1}^2] E[\theta_m^2].$$

Furthermore,

$$\theta_m \sim \frac{\chi^2_{\nu_1}(\lambda_1)}{\chi^2_{\nu_1}(\lambda_1) + \chi^2_{\nu_2}(\lambda_2)}, \tag{17}$$

where we have split the total window size $\nu = T - m + 1$ into a pre-break window size $\nu_1 = T_1 - m + 1$, and a post-break window size $\nu_2 = T - T_1$. $\chi^2_{\nu_1}(\lambda_1)$ is a non-central chi-squared distribution with non-centrality parameter $\lambda_1 = \nu_1 \mu_x^2$ and ν_1 degrees of freedom. Likewise, $\chi^2_{\nu_2}(\lambda_2)$ is a non-central chi-squared distribution (independent of $\chi^2_{\nu_1}(\lambda_1)$) with non-centrality parameter $\lambda_2 = \nu_2 \mu_x^2$ and ν_2 degrees of freedom. Hence θ_m follows a doubly non-central beta distribution with parameters $\nu_1/2$ and $\nu_2/2$ and non-centrality parameters λ_1 and λ_2. We show in the appendix that approximately we have

$$\begin{aligned} E[\theta_m] &= \frac{\nu_1}{\nu_1 + \nu_2} = \frac{\nu_1}{\nu} < 1, \qquad (18) \\ E[\theta_m^2] &= \left(\frac{\nu_1}{\nu}\right) \frac{(1 + k\nu_1)}{(1 + k\nu)}, \end{aligned}$$

where $k = (1+2\mu_x^2)^2/(2+8\mu_x^2)$. Assuming that $\psi = 0$, so that there is only a break in the conditional mean ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) we also show that the expected value of the last term in (16) is given by

$$\begin{aligned} E\left(\frac{x_{T+1}^2}{\sum_{t=m}^T x_t^2}\right) &= \left(\frac{1}{2}\right) \exp(-\frac{1}{2}\lambda)(1 + \delta^2) \times \qquad (19) \\ & \sum_{j=0}^{\infty} \frac{(\frac{1}{2}\lambda)^j}{j!} \frac{\Gamma(\frac{1}{2}(\nu - 2) + j)}{\Gamma(\frac{1}{2}\nu + j)}, \end{aligned}$$

where $\lambda = \nu \mu_x^2$ and $\delta = \mu_x/\omega$. This expression, which we denote by $\varkappa(\lambda, \delta, \nu)$, can easily be evaluated numerically. Hence the unconditional MSFE is approximately given by

$$
\begin{aligned}
E[e_{T+1}^2(m)] &= \sigma^2 + (\beta_1-\beta_2)^2(\omega^2+\mu_x^2)\left(\frac{\nu_1}{\nu}\right)\frac{(1+k\nu_1)}{(1+k\nu)} \\
&\quad + \left(\frac{\sigma^2}{2}\right)\exp(-\frac{1}{2}\lambda)(1+\delta^2)\times\sum_{j=0}^{\infty}\frac{(\frac{1}{2}\lambda)^j}{j!}\frac{\Gamma(\frac{1}{2}(\nu-2)+j)}{\Gamma(\frac{1}{2}\nu+j)}.
\end{aligned}
$$

Tractable exact analytical results can be obtained when $\mu_x = \psi = 0$. In this case it readily follows from (17) that the non-centrality parameters are zero so

$$\theta_m \sim Beta(\frac{\nu_1}{2}, \frac{\nu_2}{2}),$$

and the first two moments of θ_m are now given exactly by

$$
\begin{aligned}
E[\theta_m] &= \frac{\nu_1}{\nu}, \\
E[\theta_m^2] &= \frac{\nu_1(\nu_1+2)}{\nu(\nu+2)}.
\end{aligned}
$$

Also,

$$E\left(\frac{x_{T+1}^2}{\sum_{t=m}^{T} x_t^2}\right) = \frac{1}{\nu-2},$$

and, unconditionally,

$$E[\theta_m x_{T+1}] = E[x_{T+1}]E[\theta_m] = 0.$$

In total, we obtain the following expression for the unconditional MSFE:

$$E[e_{T+1}^2(m)] = \sigma^2 + \omega^2(\beta_1-\beta_2)^2\frac{\nu_1(\nu_1+2)}{\nu(\nu+2)} + \frac{\sigma^2}{\nu-2}. \tag{20}$$

One measure of the standardized decrease in the unconditional MSFE as pre-break information is used is suggested by comparing the case with $\nu_1 = 0$ and $m = T_1+1$ (i.e. use no pre-break information) to the case where $\nu_1 > 0$ so that $m < T_1+1:$

$$
\begin{aligned}
V &= E[e_{T+1}^2(T_1+1)] - E[e_{T+1}^2(m)] \\
&= (\sigma^2 + \frac{\sigma^2}{\nu_2-2}) - \left\{(\beta_1-\beta_2)^2\omega^2\frac{\nu_1(\nu_1+2)}{\nu(\nu+2)} + \sigma^2 + \frac{\sigma^2}{\nu-2}\right\},
\end{aligned}
$$

or

$$\begin{aligned}\frac{V}{\sigma^2} &= \frac{1}{\nu_2-2} - \frac{1}{\nu-2} - \mu^2\omega^2\frac{\nu_1(\nu_1+2)}{\nu(\nu+2)} \\ &= \nu_1\left\{\frac{1}{(\nu-2)(\nu_2-2)} - \frac{\mu^2\omega^2(\nu_1+2)}{\nu(\nu+2)}\right\}.\end{aligned}$$

Noting that $\mu^2\omega^2 = \frac{R^2}{1-R^2}\left(\frac{\beta_1-\beta_2}{\beta_2}\right)^2$, where R is the regressions's population correlation coefficient after the break, i.e. $R^2 = 1-\sigma^2/\left(\sigma^2+\beta_2^2\omega^2\right)$, we have

$$\frac{V}{\sigma^2} = \nu_1\left\{\frac{1}{(\nu-2)(\nu_2-2)} - \frac{(\nu_1+2)}{\nu(\nu+2)}\frac{R^2}{1-R^2}\left(\frac{\beta_1-\beta_2}{\beta_2}\right)^2\right\}. \tag{21}$$

We summarize these findings and state the obvious comparative statics results in the following proposition:

Proposition 4 *Consider the structural break data generating process (1) and suppose that $\mu_x = \psi = 0$. Then the unconditional MSFE is given by*

$$E[e_{T+1}^2(m)] = \sigma^2 + (\beta_1-\beta_2)^2\omega^2\frac{\nu_1(\nu_1+2)}{\nu(\nu+2)} + \frac{\sigma^2}{\nu-2}.$$

and the optimal pre-break window that minimizes the MSFE (ν_1^*) is larger, the
(i) smaller the regression R^2 (i.e. the lower the ω^2 and the higher the σ^2)
(ii) smaller the break $(\beta_1-\beta_2)^2$.

To demonstrate the magnitude of some of these effects in the iid case, we plot in Figure 1 the unconditional MSFE as a function of the break size $\mu = (\beta_1-\beta_2)/\sigma$ and the pre-break window size (ν_1). To construct this graph we have set $\sigma^2 = 9$ and $\omega^2 = 1$ (i.e. $R^2 = 0.1$) and $\nu_2 = 4$.[3] When the break is small, the MSFE declines uniformly as ν_1 increases and more information prior to the break is used to estimate β. However, for larger break sizes and larger biases, it is optimal to only use very few observations prior to the break. Conditional on a given value of μ and ν_2, Figure 1 can be used to determine the optimal value of the pre-break window (ν_1^*). Figure 2 shows how the optimal pre-break window (ν_1^*) depends

[3] We only plot the MSFE for positive values of the break since the MSFE is symmetric around zero as a function of the break size.

more generally on the break size (μ) and post-break window (ν_2). Independent of the post-break sample size, it is always optimal to use the full set of pre-break information (limited to at most 25 observations) when the break is small. However, as the break gets larger or the post-break window size increases, the optimal pre-break window size rapidly declines and the MSFE evaluated at ν_1^* increases.

3. Window Size and Market Timing Skills

In many applications in economics and finance, interest lies in market timing skills where the primary aim is to correctly predict turning points or sign of some variable such as asset returns. For example, market timing attempts by fund managers depend on their prediction of the direction of the market. Confirming the value to investment strategies of the proportion of correctly predicted signs, Leitch and Tanner (1991) find that the correlation between this statistic and the profits made from following investment advice dominates the correlation between profits and standard statistical measures of prediction such as MSFE. In this section we derive some results on the optimal window size when the objective is to maximize the (unconditional) probability of correctly predicting the sign of y_{T+1}. As one would expect, the decision rule is quite different from the stopping rule derived with the objective of minimizing MSFE.

Again let the forecast of y_{T+1} be given by $\widehat{y}_{T+1} = \widehat{\boldsymbol{\beta}}_T' \mathbf{x}_{T+1}$. The unconditional probability of correctly predicting the sign of y_{T+1} depends on the product of two random variables

$$\Pr(y_{T+1}\widehat{y}_{T+1} > 0). \tag{22}$$

To simplify the exposition, let $p = 1$, and recall that these variables have the following representation

$$\begin{aligned}
y_{T+1} &= \beta_2 x_{T+1} + u_{T+1}, \\
\widehat{y}_{T+1} &= \beta_2 x_{T+1} + (\beta_1 - \beta_2)\theta_m x_{T+1} + x_{T+1}\left(\frac{\sum_{t=m}^{T} x_t u_t}{\sum_{t=m}^{T} x_t^2}\right), \\
\theta_m &= \frac{\sum_{t=m}^{T_1} x_t^2}{\sum_{t=m}^{T} x_t^2}.
\end{aligned}$$

To derive an expression for the market timing test of Pesaran and Timmermann (1992), we proceed as follows. Granger and Pesaran (1999) show that this market timing test can be rewritten as

$$PT = \frac{\sqrt{n}KS}{\left(\frac{\widehat{p}_{\pi}(1-\widehat{p}_{\pi})}{\overline{z}(1-\overline{z})}\right)^{1/2}}, \tag{23}$$

where n is the number of observations in the forecast period, $KS = H - F$, the hit rate minus the false alarm rate which are defined as[4]

$$\begin{aligned} H &= \frac{\Pr(\widehat{y}_{T+1} > 0, y_{T+1} > 0)}{\Pr(y_{T+1} > 0)}, && (24) \\ F &= \frac{\Pr(\widehat{y}_{T+1} > 0, y_{T+1} < 0)}{\Pr(y_{T+1} < 0)}, && (25) \end{aligned}$$

and $\overline{z} = \Pr(y_{T+1} > 0)$, and $\widehat{p}_{\pi} = \Pr(\widehat{y}_{T+1} > 0)$ are the probabilities that the realization and predicted values are positive, respectively.

In this simple example, the sign test is not interesting if it is computed conditional on X_{T+1}. The reason is easy to see. Conditional on $\mathbf{X}_{T+1}$, and assuming that $\psi = 0$, we have

$$\begin{pmatrix} y_{T+1} & \\ \widehat{y}_{T+1} & \end{pmatrix}|\mathbf{X}_{T+1} \Bigg) \sim IIN(\mathbf{d}_{T+1}, \boldsymbol{\Omega}_{T+1}),$$

where

$$\begin{aligned} \mathbf{d}_{T+1} &= \begin{pmatrix} \beta_2 x_{T+1} \\ \beta_2 x_{T+1} + (\beta_1 - \beta_2)\theta_m x_{T+1} \end{pmatrix}, \\ \boldsymbol{\Omega}_{T+1} &= \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{x_{T+1}^2 \sigma^2}{\sum_{t=m}^{T} x_t^2} \end{pmatrix}. \end{aligned}$$

Conditional on $\mathbf{X}_{T+1}$, y_{T+1} and $\widehat{y}_{T+1}$ are independent, therefore using (24) and (25) the sign test will take a value of zero as $H = F = \Pr(\widehat{y}_{T+1} > 0)$ and $KS = 0$. Hence we concentrate on the unconditional results.

[4] The KS statistic is known by the Kuipers score in the meteorology literature. See Granger and Pesaran (1999) for the references to the relevant literature.

In general it is complicated to derive an analytical expression for the probability of correctly predicting the sign. However, in the simple case where u_t and x_t are serially uncorrelated and normally distributed we can derive an expression that demonstrates how the probability of correctly predicting the sign of y_{T+1} depends on the window size. To state the result we first introduce some notations. Let

$$\begin{aligned} \mu_1 &= \beta_2 \mu_x, \\ \mu_2 &= \frac{\mu_x(\beta_1 \nu_1 + \beta_2 \nu_2)}{\nu}. \end{aligned} \tag{26}$$

Also let

$$\mathbf{\Sigma} = \begin{pmatrix} \beta_2^2 \omega^2 + \sigma^2 & g \\ g & h^2 \end{pmatrix},$$

be a 2×2 covariance matrix, where g and h are constants defined by

$$\begin{aligned} h^2 \equiv\ & V(\widehat{y}_{T+1}) = \sigma^2 \varkappa(\lambda, \delta, \nu) + \beta_2^2 \omega^2 + \\ & (\beta_1 - \beta_2)^2 \omega^2 (\frac{\nu_1}{\nu}) \varsigma + 2\beta_2(\beta_1 - \beta_2)\omega^2(\frac{\nu_1}{\nu}), \end{aligned} \tag{27}$$

and

$$\begin{aligned} g &\equiv Cov(y_{T+1}, \widehat{y}_{T+1}) = \beta_2^2 \omega^2 + \beta_2(\beta_1 - \beta_2)\omega^2(\nu_1/\nu), \\ &= \beta_2^2 \omega^2 \left(1 + \left(\frac{\nu_1}{\nu}\right)\left(\frac{\beta_1 - \beta_2}{\beta_2}\right)\right), \end{aligned} \tag{28}$$

where $\varkappa(\lambda, \delta, \nu)$ is defined by (19), and

$$\varsigma = \left(\frac{1 + k\nu_1}{1 + k\nu}\right) + \frac{\mu_x^2}{\omega^2}\left(\frac{\nu_2}{\nu(1 + k\nu)}\right). \tag{29}$$

We have the following result:

Proposition 5 *Suppose that u_t and x_t are serially uncorrelated and normally distributed*

$$\begin{pmatrix} u_t \\ x_t \end{pmatrix} \sim IIN\left\{\begin{pmatrix} 0 \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \omega^2 \end{pmatrix}\right\}.$$

Then the Kuipers Score (KS) associated with the realizations and forecasts $(y_{T+1}, \widehat{y}_{T+1})$ is given by

$$KS = \frac{A_H}{\Phi(\tilde{\mu}_1)} - \frac{A_F}{1 - \Phi(\tilde{\mu}_1)}, \tag{30}$$

where $\tilde{\mu}_1 = \mu_1/\sqrt{\sigma^2 + \beta_2^2\omega^2}$,

$$\Phi(\tilde{\mu}_1) = \int_{-\infty}^{\tilde{\mu}_1} (2\pi)^{-1/2} \exp(\frac{-1}{2}a^2)da,$$

$$A_H = \Pr(y_{T+1} > 0, \widehat{y}_{T+1} > 0) = \int_{a_2=-\mu_2}^{\infty} \int_{a_1=-\mu_1}^{\infty} f(a_1, a_2)da_1da_2,$$

$$A_F = \Pr(y_{T+1} < 0, \widehat{y}_{T+1} > 0) = \int_{a_2=-\mu_2}^{\infty} \int_{a_1=-\infty}^{-\mu_1} f(a_1, a_2)da_1da_2,$$

$$f(a_1, a_2) = (2\pi)^{-1} |\mathbf{\Sigma}|^{-1/2} \exp(-\frac{1}{2}\mathbf{a}'\mathbf{\Sigma}^{-1}\mathbf{a}),$$

$\mathbf{a} = (a_1, a_2)'$*, and* μ_1 *and* μ_2 *are defined by (26). In the case where* $\mu_x = 0$*, we have* $KS = 2(A_H - A_F)$*.*

The more familiar measure of association between forecasts $(y_{T+1})$and realizations $(\widehat{y}_{T+1})$, namely their correlation coefficient, ρ, is given by:

$$\rho = \frac{\gamma^2}{\sqrt{1+\gamma^2}} \times \frac{1 + \phi(\nu_1/\nu)}{\left[\varkappa(\lambda, \delta, \nu) + \gamma^2 + \gamma^2\phi^2\varsigma + 2\gamma^2\phi(\nu_1/\nu)\right]^{1/2}}, \tag{31}$$

where $\phi = (\beta_1 - \beta_2)/\beta_2$, is the rate of change in β, and $\gamma^2 = \omega^2\beta_2^2/\sigma^2$ is the signal-to-noise ratio. The expressions for $\varkappa(\lambda, \delta, \nu)$ and ς are defined by (19) and (29), respectively. Notice also that $\gamma^2 = R^2/(1 - R^2)$, where as before R is the regression population correlation coefficient after the break.

Using the above results, and setting $\sigma^2 = 3$, $\omega^2 = 1$, $\mu_x = 2$, Figure 3 plots the optimal window size determined by the unconditional MSFE and market timing criteria. If the objective is to minimize MSFE, it is optimal to use the full sample of pre-break observations only when ν_2 and μ are small. In contrast, to maximize the KS statistic, no pre-break observations are used except for when the break is large. This example demonstrates that the window sizes determined to be optimal under the MSFE and sign criteria can be very different. To underline this point, the right-most window of Figure 3 plots the difference between the optimal window sizes chosen under the MSFE and KS criteria. For this particular parameter configuration the window size under the market timing criterion is almost always smaller than that selected under the MSFE criterion. The intuition for this finding is as follows. Under the MSFE criterion, smoothness of the forecast matters in determining the forecast and this is obtained by adopting a large window size.

In contrast, all that matters for the sign criterion is whether the prediction has the right sign and so a shorter window that quickly picks up a change may be chosen. These findings should be treated as preliminary and could be sensitive to the underlying parameter values.

Finally, to demonstrate how the market timing value of a prediction depends on the post-break information and the break size, Figure 4 shows the resulting values of the KS statistic evaluated at the optimal choice of ν_1. Not surprisingly, the KS statistic is larger, the longer the post-break window (ν_2) and the smaller the break (μ).

4. **Further Theoretical Results**

The above results, and particularly the trade-off between the squared bias and variance of the forecasting error, may appear to be specific to the structural break model considered so far. However, as we show in this section, this is far from the case. To demonstrate this point, we first relate our earlier results to the standard weighted least squares approach that is often adopted to regressions with heteroskedastic errors and then consider an alternative time-varying coefficients model in which the regression parameters change every period.

4.1. *Optimal weights under a single structural break*

Suppose again that data is being generated by the linear regression model (1) with a single structural break, but now consider the weighted least-squares estimator that uses constant weights w_1 and w_2 on the pre-break and post-break observations:

$$\begin{aligned}\widehat{\beta}_T &= \frac{\sum_{t=1}^{T}\lambda_t^2 x_t y_t}{\sum_{t=1}^{T}\lambda_t^2 x_t^2}, \text{ where} \\ \lambda_t &= w_1,\, t = 1, 2, ..., T_1 \\ \lambda_t &= w_2,\, t = T_1+1, 2, ..., T.\end{aligned} \tag{32}$$

Using this estimator, the prediction error $(y_{T+1} - \widehat{y}_{T+1})$ becomes

$$e_{T+1} = u_{T+1} - \frac{x_{T+1}(\beta_1 - \beta_2)\sum_{t=1}^{T_1}\lambda_t^2 x_t^2}{\sum_{t=1}^{T}\lambda_t^2 x_t^2} - x_{T+1}\frac{\sum_{t=1}^{T}\lambda_t^2 x_t u_t}{\sum_{t=1}^{T}\lambda_t^2 x_t^2}. \tag{33}$$

From (32), and defining $\xi = w_1/w_2$ and $\varphi^2(T_1,T_2) \equiv \varphi^2 = \sum_{t=1}^{T_1} x_t^2 / \sum_{t=T_1+1}^{T} x_t^2$, we have

$$e_{T+1} = u_{T+1} - x_{T+1}(\beta_1 - \beta_2)\left(\frac{\xi^2\varphi^2}{\xi^2\varphi^2+1}\right) - x_{T+1}\left(\frac{\xi^2 \sum_{t=1}^{T_1} x_t u_t + \sum_{t=T_1+1}^{T} x_t u_t}{\xi^2 \sum_{t=1}^{T_1} x_t^2 + \sum_{t=T_1+1}^{T} x_t^2}\right).$$

The first two conditional moments of e_{T+1} thus become

$$\begin{aligned} E[e_{T+1}|X_{T+1}] &= -x_{T+1}(\beta_1 - \beta_2)\left(\frac{\xi^2\varphi^2}{\xi^2\varphi^2+1}\right), \\ V(e_{T+1}|X_{T+1}) &= \sigma_2^2 + x_{T+1}^2 \left(\frac{\xi^4\sigma_1^2 \sum_{t=1}^{T_1} x_t^2 + \sigma_2^2 \sum_{t=T_1+1}^{T} x_t^2}{\left(\xi^2 \sum_{t=1}^{T_1} x_t^2 + \sum_{t=T_1+1}^{T} x_t^2\right)^2}\right). \end{aligned}$$

Letting $\vartheta = \sigma_1/\sigma_2$,[5] we can obtain the conditional MSFE:

$$\begin{aligned} MSFE(\xi|X_{t+1}) &= V(e_{T+1}|X_{T+1}) + (E[e_{T+1}|X_{T+1}])^2 \qquad (34) \\ &= \sigma_2^2 + x_{T+1}^2(\beta_1-\beta_2)^2\left(\frac{\xi^2\varphi^2}{\xi^2\varphi^2+1}\right)^2 + \frac{\sigma_2^2 x_{T+1}^2}{\sum_{t=T_1+1}^{T} x_t^2}\left(\frac{\xi^4\vartheta^2\varphi^2+1}{\left(\xi^2\varphi^2+1\right)^2}\right) \\ &= \sigma_2^2 + V(\widehat{\beta}_2)x_{T+1}^2\left\{d^2\left(\frac{\xi^2\varphi^2}{1+\xi^2\varphi^2}\right)^2 + \frac{1+\vartheta^2\xi^4\varphi^2}{(1+\xi^2\varphi^2)^2}\right\}, \end{aligned}$$

where $V(\widehat{\beta}_2) = \sigma_2^2/\sum_{t=T_1+1}^{T} x_t^2$, and $d^2 = (\beta_1-\beta_2)^2/V(\widehat{\beta}_2)$. The part of $MSFE(\xi|X_{T+1})$ that depends on ξ is given by

$$f(\xi^2) = d^2\left(\frac{\xi^2\varphi^2}{1+\xi^2\varphi^2}\right)^2 + \frac{1+\vartheta^2\xi^4\varphi^2}{(1+\xi^2\varphi^2)^2}. \qquad (35)$$

The familiar weighted least squares result is obtained when $\beta_1 = \beta_2$, namely when there is no break in the mean and $d = 0$. For this case we have

$$f(\xi^2) = \frac{1+\vartheta^2\xi^4\varphi^2}{(1+\xi^2\varphi^2)^2},$$

which is minimized with respect to ξ^2 for $\xi^* = 1/\vartheta$, namely for

$$w_1^*/w_2^* = \sigma_2/\sigma_1.$$

[5] Notice that in terms of the notations of the previous sections, $\vartheta = \sqrt{1+\psi}$.

When $d \neq 0$, there is a trade-off between squared bias and volatility of the forecast error. Differentiating $f(\xi^2)$ with respect to ξ^2, we obtain the following first-order condition:

$$\frac{\partial f(\xi^2)}{\partial \xi^2} = \frac{2\varphi^2\left((d^2\varphi^2 + \vartheta^2)\xi^2 - 1\right)}{(1 + \xi^2\varphi^2)^3} = 0, \tag{36}$$

which means that the optimal value of ξ is

$$\xi^* = \sqrt{\frac{1}{d^2\varphi^2 + \vartheta^2}}. \tag{37}$$

The larger the break in the mean parameters, the higher the pre-break variance relative to the post-break variance, and the larger the pre-break window size relative to the post-break window size, the lower the weight on the pre-break observations will be. Notice that the solution for ξ^* does not depend on the distance to the break point, so this procedure is not capable of exploiting the time-varying nature of the bias-forecast error variance tradeoff.

4.2. *Time-varying Coefficient Model*

An alternative approach to dealing with structural change is to consider regression models where the coefficients are time-varying:

$$y_t = \beta_t x_t + u_t,\ u_t \sim iid(0, \sigma_u^2), \tag{38}$$

for $t = 1, .., , T$. This literature, generally assumes that the regression coefficient, β_t, follows a mean reverting process:

$$\beta_t - \beta_{t-1} = \theta(\beta_{t-1} - \overline{\beta}) + \eta_t,\ \eta_t \sim iid(0, \sigma_\eta^2) \tag{39}$$

where $-1 < \theta < 0$ and $E[u_t\eta_s] = 0$, for all t, s. Following Cooley and Prescott (1976), it is also common to set $\theta = 0$, so shocks to the coefficients are permanent:[6]

$$\beta_t = \beta_{t-1} + \eta_t.$$

[6] This model and various extensions of it has been studied under the heading 'structural time series' by Harvey (1989).

In this section we briefly examine the implication of this time-varying model for the trade-off between bias and squared forecast error involved in choosing m. Notice that

$$\beta_t = \beta_m + S_{t,m},$$

where $S_{t,m} = \eta_{m+1} + \eta_{m+2} + ... + \eta_{m+t}$ is a partial sum and $S_{m,m} = 0$.

Once again we assume that the investigator is interested in forecasting y_{T+1} by some weighted least squares procedure, using the last ν observations ($\nu \leq T$), where ν is the size of the observation window and applying the smoothing weights $\{\lambda_t\}$ to past observations. The estimator thus becomes

$$\widehat{\beta}_\nu = \frac{\sum_{t=m}^T \lambda_t^2 x_t^2 (\beta_m + S_{t,m}) + \sum_{t=m}^T \lambda_t^2 x_t u_t}{\sum_{t=m}^T \lambda_t^2 x_t^2}.$$

Defining the regression weights

$$\omega_t = \frac{\lambda_t^2 x_t^2}{\sum_{i=m}^T \lambda_i^2 x_i^2},$$

we can write the forecast error as follows

$$e_{T+1} = u_{T+1} - x_{T+1}\left(S_{T+1,m} - \sum_{t=m}^T \omega_t S_{t,m}\right) - x_{T+1}\left(\frac{\sum_{t=m}^T \lambda_t^2 x_t u_t}{\sum_{t=m}^T \lambda_t^2 x_t^2}\right). \tag{40}$$

It is easily seen that $E[e_{T+1}|X_{T+1}] = 0$ so that $E[e_{T+1}] = 0$. Hence the conditional $MSFE$ given X_{T+1} is

$$MSFE(\omega_m, .., \omega_T|X_{T+1}) = \sigma_u^2 + x_{T+1}^2 V(S_{T+1,m} - \sum_{t=m}^T \omega_t S_{t,m}) + \frac{\sigma_u^2 x_{T+1}^2 \sum_{t=m}^T \lambda_t^4 x_t^2}{\left(\sum_{t=m}^T \lambda_t^2 x_t^2\right)^2}. \tag{41}$$

Notice that

$$\begin{aligned}
& S_{T+1,m} - \sum_{t=m}^T \omega_t S_{t,m} \\
= \; & (1 - \omega_{m+1} - \omega_{m+2} - ... - \omega_T)\eta_{m+1} \\
& +(1 - \omega_{m+2} - \omega_{m+3} - ... - \omega_T)\eta_{m+2} \\
& +(1 - \omega_{m+3} - \omega_{m+4} ... - \omega_T)\eta_{m+3} \\
& +(1 - \omega_T)\eta_{m+T} + \eta_{m+T+1}.
\end{aligned}$$

Hence

$$\begin{aligned} V(S_{T+1,m}-\sum_{t=m}^{T}\omega_t S_{t,m}) &= \sigma_\eta^2\{1+(1-\omega_T)^2+(1-\omega_{T-1}-\omega_T)^2+ \\ &\quad ....+(1-\omega_{m+1}-\omega_{m+2}-..-\omega_T)^2\} \\ &= \sigma_\eta^2\left\{1+\sum_{t=1}^{T-m}\left[1-\sum_{j=T-t+1}^{T}\omega_j\right]^2\right\}. \end{aligned}$$

¿From this we get the standardized conditional MSFE:

$$\chi(m|\mathbf{X}_{T+1})=\frac{MSFE-\sigma_u^2}{\sigma_u^2 x_{T+1}^2}=\left(\frac{\sigma_\eta^2}{\sigma_u^2}\right)\left\{1+\sum_{t=1}^{T-m}\left[1-\sum_{j=T-t+1}^{T}\omega_j\right]^2\right\}+\frac{\sum_{t=m}^{T}\lambda_t^4 x_t^2}{\left(\sum_{t=m}^{T}\lambda_t^2 x_t^2\right)^2}. \tag{42}$$

Clearly $\chi(m|\mathbf{X}_{T+1})$ provides a trade-off across the two terms in (42). Allowing m to go to one, i.e. increasing ν to T, causes the second term in (42) to decrease, but at the same time can cause the first term to increase. The optimal value of m depends on the variance ratio σ_η^2/σ_u^2 and the $\{x_t\}$ process. Under parameter stability, $\sigma_\eta^2=0$, and the optimal value for m (say m^*) is $m^*=1$. However, when $\sigma_\eta^2>0$ and $\{\lambda_t x_t\}$ is a stationary process the first term in (42) is of order $\nu=T-m+1$, while the second term is of order ν^{-1}, and depending on the value of the ratio σ_η^2/σ_u^2, and the weights ω_j, m^* could be much larger than 1.

5. **Determining the Observation Window**

Several methods have been proposed for selecting a sample period, or an observation window, for estimation and forecasting. If parameter breaks are thought either to be very rare or of a very small magnitude, the usual method is to use an expanding window, by augmenting an already selected sample period with new observations. The aim here is to obtain a more efficient estimate of the same fixed coefficients by using more information as they become available. However, if the parameters of the regression model are not believed to be constant over time, frequently a rolling window of observations with a fixed size is used to generate forecasts. Weighted regressions are also often employed by professional economists where exponentially declining weights are applied to the full set of observations, assigning smaller weights to observations further away from the point of prediction.

None of these methods attempts to explicitly detect and condition on the parameters of one or several break points. The second set of methods we explore estimates the time and size of a break. Bai and Perron (1998a) propose a method that consistently estimates the number of breaks (as well as their size) in the context of a linear regression model. While this method can handle multiple breaks, the standard Cusum tests are only designed to detect a single break. For this reason we propose a new reversed Cusum test for detection of the most recent break point. Having obtained an estimate of the time of the most recent break, the methods derived in the previous section are then used to determine the window size and estimate a forecasting model.

In all cases we consider simple OLS estimators of the form

$$\widehat{\boldsymbol{\beta}}_T(\tau) = \mathbf{Q}_{\tau,T}^{-1}\mathbf{X}_{\tau,T}'\mathbf{Y}_{\tau,T}, \tag{43}$$

where $\mathbf{Q}_{\tau,T}$, $\mathbf{X}_{\tau,T}$ and $\mathbf{Y}_{\tau,T}$ are as defined in section 2, and $T-\tau+1$ is the chosen window size for estimation. The forecast of y_{T+1} conditional on information at time T is then computed as

$$\widehat{y}_{T+1} = \mathbf{x}_{T+1}'\widehat{\boldsymbol{\beta}}_T(\tau). \tag{44}$$

The objective of the exercise is to provide a plausible procedure for choosing τ.

5.1. *Expanding Window*

In the absence of breaks in the data generating process, $\boldsymbol{\beta}$ can be consistently estimated by OLS. If interest lies in computing recursive forecasts of y_t, it is, subject to standard assumptions, efficient to use an expanding window of the data and ignoring parts of the data will lead to efficiency losses. The regression model used to forecast y_{T+1} is hence based on the following data

$$\begin{aligned} \mathbf{Y}_{1,T} &= (y_1, y_2, ..., y_T)' \\ \mathbf{X}_{1,T} &= (x_1, x_2, ..., x_T)'. \end{aligned} \tag{45}$$

If breaks in the data generating process are a possibility but no information is available about the time of the last break point, breaks have to be treated as a possible source of misspecification. Two approaches are popular, namely using a fixed window size of the data ('rolling window') or exponentially smoothing the data. We briefly describe these approaches.

5.2. *Rolling Window*

Let c be the window size. Then the rolling window regressions used to forecast y_{T+1} is based on the following data

$$\begin{aligned}\mathbf{Y}_{T-c+1,T} &= (y_{T-c+1}, y_{T-c+2}, ..., y_T)' \\ \mathbf{X}_{T-c+1,T} &= (x_{T-c+1}, x_{T-c+2}, ..., x_T)'\end{aligned} \tag{46}$$

There are several problems with this approach if the regression vector follows a step function. Immediately after a break the window will tend to be too long, while further away from the break the window will be too short. The problem is of course that no further information is used to determine possible time variation in the optimal window size.

5.3. *Exponential Weighting*

Another ad-hoc approach to account for nonstationarities often used by practitioners assigns exponentially declining weights to past observations. This is an special case of the weighted regression method discussed in section 4.2 where λ_t is chosen to be λ^{T-t}, for $t = 1, 2, ..., T$, and $\lambda \in [0, 1]$ is the smoothing parameter, c.f. Harvey (1989, Section 2.2). If $\lambda = 1$, an expanding window is obtained. Values of λ further away from one will put less weight on the earlier observations. The more frequent breaks are believed to occur in a given time-series, the lower the value of λ should be chosen.

$$\begin{aligned}\mathbf{Y}^{\gamma}_{1,T} &= (\lambda^{T-1}y_1, \lambda^{T-2}y_2, ..., \lambda^0 y_T)' \\ \mathbf{X}^{\gamma}_{1,T} &= (\lambda^{T-1}x_1, \lambda^{T-2}x_2, ..., \lambda^0 x_T)'.\end{aligned} \tag{47}$$

Again the problem with this discounted least squares approach is that it ignores possible information about the time of the break. Immediately after a break, too much weight is likely to be put on observations prior to the break, while the opposite will be true further away from the most recent break point.

5.4. *Reversed Cusum and Cusum Squared*

The above window size selection procedures are based on *ad hoc* rules and need not be efficient. What we need is an optimal stopping rule. Such a rule is, for example, embodied in the Cusum or Cusum squared procedures proposed by Brown et al. (1975) as a recursive structural stability test. The test is usually applied to observations running forward from start to finish of a given time interval.[7] However, the application of such a forward Cusum or Cusum squared test to our problem will not be appropriate for two reasons. First, even if the test is successful at identifying the time of the first break it will not be effective when there are multiple breaks. Also because it takes some time before the test can confidently identify a break its forward application will result in smaller than optimal window sizes for forecasting purposes even if it is known that the model is at most subject to a single break over the entire sample under investigation. Both of these shortcomings can to some extent be alleviated by simply reversing the observations in time before proceeding with the application of the Cusum testing procedures. We refer to this use of the test as the "reversed" Cusum or Cusum squared tests. In fact the known undesirable property of the Cusum type tests in identifying break points with long delays now becomes an advantage when the test is applied to observations reversed in time. The end break point identified by the reversed Cusum or Cusum squared tests will tend to be biased upward, viewed from the point of the forecast, but as the above discussion shows this is likely to be often desirable if one is interested in forecasting. Finally, although it may appear that important information is lost by not attempting to use the full available information set at a given point in time to identify all possible breaks, methods that attempt to identify multiple break points often lead to very imprecise estimates of break points and can be prone to problems with over-fitting.

We use the following notation to denote the observation matrices with the orders of the observations reversed in time, starting from the mth observation (so that the observation window is given by $T-m+1$):

$$\tilde{\mathbf{Y}}_{T,m} \quad = \quad (y_T, y_{T-1}, ..., y_{m+1}, y_m)' \qquad (48)$$

[7] Typically, a few observations are dropped from the start date to ensure that the regressions can be computed and very uncertain results can be avoided.

$$\tilde{\mathbf{X}}_{T,m} \quad = \quad (\mathbf{x}_T, \mathbf{x}_{T-1}, ..., \mathbf{x}_{m+1}, \mathbf{x}_m)'$$

and define the (backward) recursive least squares estimates as

$$\widehat{\boldsymbol{\beta}}_r = (\tilde{\mathbf{X}}'_{T,r}\tilde{\mathbf{X}}_{T,r})^{-1}\tilde{\mathbf{X}}'_{T,r}\tilde{\mathbf{Y}}_{T,r},\ r = \tilde{T}, \tilde{T}-1, ..., 2, 1.$$

The choice of the shortest observation window selected, namely $T - \tilde{T} + 1$, is arbitrary but one would expect it be set around 2 to 3 times the dimension of $\boldsymbol{\beta}$.

The standardized recursive residuals from the regression that is reversed in time are

$$v_r = (y_r - \widehat{\boldsymbol{\beta}}'_{r-1}\mathbf{x}_r)/d_r,\ r = \tilde{T}, \tilde{T}-1, ..., 2, 1,$$

where

$$d_r = (1 + \mathbf{x}'_r(\tilde{\mathbf{X}}'_{T,r}\tilde{\mathbf{X}}_{T,r})^{-1}\mathbf{x}_r),\ r = \tilde{T}, \tilde{T}-1, ..., 2, 1,$$

and the reversed Cusum and Cusum Squared tests can be based on

$$\begin{aligned} W_r &= \frac{1}{\widehat{\sigma}}\sum_{j=p+1}^{r} v_j, \\ WW_{r,T} &= \sum_{j=p+1}^{r} v_j^2 / \sum_{j=p+1}^{T} v_j \end{aligned}$$

Critical values from Brown et al. (1975) can be used to decide if a break has occurred.

6. **Monte Carlo Simulations**

This section demonstrates the use of our recursive stopping rule for selection of the optimal window size in estimating a forecasting model. Our simulation assumes a break occurs in a regression model after 100 periods and we track the one-step ahead forecasting performance during 50 periods after the break. To demostrate the factors determining the optimal window size, we consider a simple model with a single regressor (x_t) which we assume follows a persistent first order autoregressive process. The break takes the form of a shift in the linear regression coefficient that relates x_t to y_t. All shocks in the system are identically and independently, normally distributed:

$$
\begin{aligned}
y_t &= \beta_t x_t + u_t, \\
x_t &= 0.95 x_{t-1} + e_t, \\
e_t &\sim IIN(0,1), u_t \sim IIN(0,1), \\
Cov(u_t,\ e_t) &= 0, \\
\beta_t &= 1 \text{ for } t \leq 100, \\
&= 1-\delta,\ \text{ for } 101 \leq t \leq 200.
\end{aligned} \tag{49}
$$

Simulations are carried out under three different values of the change in the regression parameter $\{\delta = 0.1, 0.5, 2.0\}$. Since $\sigma_2 = 1$, $\mu = (\beta_2 - \beta_1)/\sigma_2 = \delta$. Although the break sizes may seem large, they are quite small when compared to the unconditional variability in x_t. One thousand simulations were conducted and we report the forecasting performance averaged across simulations.[8]

6.1. *Known Size and Time of the Break*

To demonstrate the trade-off between bias and efficiency in the choice of window size, we first simulate the case where δ, σ^2, T_1 are assumed to be known. Figure 5 shows the average value of the optimal pre-break window size ($\overline{\nu}_1^*$) computed over 1000 replications as determined by our stopping rule[9] while Figure 6 plots the average full window size ($\bar{\nu}$). Finally, Figure 7 plots the resulting MSFE for the stopping rule and for the expanding and rolling window procedures. The rolling window has a fixed size of 50 observations.

First consider the case where the break is small ($\mu = 0.1$). In this case the bias arising from including observations prior to T_1 to estimate β is very small. Consequently it is optimal to use most of the pre-break data to estimate β and the pre-break window size only declines very slowly as a function of ν_2 (the post-break window size). Although the reversed window stopping rule produces lower MSFE values than when expanding and rolling windows are used, the difference between MSFEs is quite small and declines as the distance to the break point increases.

[8] Values of the MSFE are computed based on the expressions derived in section 2.

[9] The average value of the optimal pre-break window size is determined as follows. Let $\nu_{1,i}^*$ be the optimal window size for the i'th Monte Carlo simulation, and let n be the number of simulations. Then $\overline{\nu}_1^* = (1/n)\sum_{i=1}^n \nu_{1,i}^*$. Other statistics are calculated in a similar way.

Turning to the case with a medium-sized break ($\mu = 0.5$), the second graph in Figure 5 shows that the optimal window includes up to 26 pre-break observations immediately after the break. At this point, the trade-off between efficiency gains and squared bias is high. However, as ν_2 increases, the trade-off worsens and the number of observations from the first regime used in estimating β rapidly declines. Now that the bias is larger, the stopping rule produces a substantially lower MSFE than the rolling or expanding window procedures.

Finally, in the case where the break is very large ($\mu = 2$), both the expanding window and the rolling window methods do very poorly and generate values of the MSFE that are orders of magnitude higher than those produced by the stopping rule. Interestingly, Figure 5 shows that even for this case it is still optimal in the first three or so periods after the break to include a small number of pre-break observations to estimate β.

These figures do not show the MSFE generated by the procedure that only uses observations after the break. This procedure generates a far higher MSFE than the other three methods whenever the break is small ($\mu = 0.1$ or $\mu = 0.5$). Only when the break is very large ($\mu = 2$) does this procedure generate MSFEs that are comparable to the reversed window method. However, even in this case the method that only uses data after the break generates far higher MSFEs than the reversed method during the first couple of observations after the break.

All diagrams in Figure 6 demonstrate the features identified in Proposition 2. When the break is small (the first graph), the optimal window size declines very gradually as a function of the distance to the break point and pre-break observations are only dropped very gradually. When the break is relatively large (the second graph), the optimal window size is 30 right after the break, but decreases rapidly to a minimum of less than 10 observations five periods after the break. From this point onwards, the window size increases in proportion with the distance to the time of the break. Finally, when the break is very large, a non-increasing relation between optimal window size and distance to the break is only observed immediately after the break.

6.2. *Unknown Size of the Break*

In many situations the time of a possible break will be known while its size is unknown. For example, knowledge of institutional shifts (such as the change in the

Federal Reserve Bank's operating procedures from 1979 to 1982) can help identify the date, though not the size of a break. In these situations T_1 will be known, while μ is unknown and has to be estimated.

To investigate the performance of our stopping rule in the presence of uncertainty about μ, we repeated our simulation experiments. The parameters of the data generating process are the same as above, but now the choice of m is based on the least-squares estimate of μ, $\widehat{\mu}$. First consider the window size (ν) shown in Figure 8. Comparing Figure 6 (μ known) and Figure 8 (μ unknown) it is clear that, when μ is small, the average optimal window size based on $\widehat{\mu}$ tends to be much smaller than the optimal window that assumes μ to be known. This effect disappears for larger values of the break: when $\mu = 0.5$ or $\mu = 2$, the average optimal window sizes are very similar independently of whether μ is known.

Figure 9 plots the MSFE generated by the optimal window sizes determined conditional on $\widehat{\mu}$. Interestingly, when μ and ν_2 are small and after introducing estimation uncertainty in determining μ, the average MSFE is now larger under the optimal stopping rule than under the expanding or rolling window methods. This results from the simulations in which μ is overestimated which leads to too short a window that does not correctly exploit the bias-variance trade-off. However, once the break gets larger, as in the second and third windows, the MSFE determined under the optimal stopping rule is much lower than under the two alternatives.

6.3. *Unknown Size and Time of the Break*

Finally, in some cases both the time of the break and its size are unknown to the researcher. This raises the issue of whether our stopping rule is of any use when all parameters characterizing the breaks are unknown. To investigate the effects of introducing estimation uncertainty into our decision rule, we adopt the stopping rule from section 2, now using estimated parameters and an estimate of the time of the break point. Our procedure is as follows:

1. Use the reversed Cusum Squared procedure to determine an estimate of T_1, say $\widehat{T}_1$.
2. Based on this estimate, let $\widehat{\theta}_m = \sum_{t=m}^{\widehat{T}_1} x_t^2 / \sum_{t=m}^{T} x_t^2$ and $\widehat{\mu} = (\widehat{\beta}_1 - \widehat{\beta}_2)/\widehat{\sigma}$, where

$$\widehat{\beta}_1 \;=\; (\mathbf{X}'_{1,\widehat{T}_1}\mathbf{X}_{1,\widehat{T}_1})^{-1}\mathbf{X}'_{1,\widehat{T}_1}\mathbf{Y}_{1,\widehat{T}_1},$$

$$\begin{aligned}\widehat{\beta}_2 &= (\mathbf{X}'_{\widehat{T}_1+1,T}\mathbf{X}_{\widehat{T}_1+1,T})^{-1}\mathbf{X}'_{\widehat{T}_1+1,T}\mathbf{Y}_{\widehat{T}_1,T} \\ \widehat{\sigma}^2 &= (\mathbf{e}'_{1,\widehat{T}_1}\mathbf{e}_{1,\widehat{T}_1} + \mathbf{e}'_{\widehat{T}_1+1,T}\mathbf{e}_{\widehat{T}_1+1,T})/(T-2),\end{aligned} \tag{50}$$

and $\mathbf{e}_{1,\widehat{T}_1}$ is the $\widehat{T}_1$ vector comprising the residuals from the first regression, while $\mathbf{e}_{\widehat{T}_1+1,T}$ is a $T-\widehat{T}_1$ vector comprising the residuals from the second regression. The estimated MSFE is now

$$\widehat{MSFE}(m) = \widehat{\sigma}^2\left(1+\widehat{\theta}_m^2\widehat{\mu}^2x_{T+1}^2 + x_{T+1}^2/\sum_{t=m}^{T}x_t^2\right).$$

3. Choose m as the largest value, m^*, for which

$$\widehat{MSFE}(m-1) > \widehat{MSFE}(m)$$

Of course, the performance of this procedure depends on how precisely the first step determines the time and size of the break.

We consider the same parameter values as in the first simulation experiment. Figure 10 provides a plot of the estimated break points ($\widehat{T}_1$) averaged across 1000 simulations. Recall that the true value of T_1 is 100. When $\widehat{T}_1 = 1$, no break is identified by the Cusum procedure. The reversed Cusum test has a particularly low power against small breaks. As the break gets larger, the last two graphs in Figure 10 show that the Cusum squared method gets a little better at identifying that a break has occurred. When $\mu = 0.5$, the break is estimated to have occurred around observation 30 while when $\mu = 2$, the estimated break point is between observation 70 and 100, depending on the distance to the break. As a result of the delay in identifying a break, the average optimal window size ($\bar{\nu}^*$) plotted in Figure 11 tends to be higher than the optimal size under no estimation uncertainty. For example, the reversed Cusum squared method effectively uses most of the pre-break data and hence closely resembles an expanding window when $\mu = 0.1$.

The tendency to use too long a window of data means that the reversed Cusum Squared procedure performs as well as or marginally better (according to the MSFE criterion) than the expanding window method when the break size is small or moderate, c.f. Figure 12. In these cases, the rolling window method performs the best. But in the case of large breaks, the reversed Cusum Squared rule nevertheless outperforms both the expanding and rolling window methods.

These findings suggest that our reversed Cusum squared procedure performs relatively well even when there is substantial uncertainty about the time of the break. However, comparing the MSFE plots in Figures 7, 9 and 12, our findings also demonstrate the importance of correctly identifying the time of the break point in order to successfully exploit the potential gains from using a time-varying window size.

7. Structural Breaks in US Stock Returns

In this section we consider a recursive prediction experiment for monthly stock returns in the US. US stock prices have been the subject of numerous studies and, with very few exceptions, these assume that the same data generating process stays in effect over the sample.[10] In this section we use full-sample information to investigate the extent to which structural breaks have been a problem over our sample period. The next section studies the performance of the recursive forecasting procedures discussed above.

Our data set consists of monthly returns on the equal-weighted NYSE price indices reported by the Center for Research in Securities Prices (CRSP). The equal-weighted index is dominated by small firms which have been found to be more susceptible to changes in the underlying economic state and hence are particularly suited for this type of study. As the dependent variable we use excess returns (EXR_t), defined as the difference between monthly stock returns and the one-month T-bill rate calculated at a monthly rate.

As forecasting variables we include a constant, the dividend-price ratio ($Yield_{t-1}$), the one-month T-bill rate (I_{t-1}) and the default premium (Def_{t-1}) defined as the yield spread of Baa-rated bonds over Aaa rated bonds. All of these regressors are standard forecasting variables from the empirical finance literature. The dividend yield is computed from the CRSP data and is defined as dividends over the previous twelve months divided by the stock price index at the end of the month. The T-bill rate is obtained from the Fama-Bliss files on the CRSP tapes. Yields on the Baa and Aaa rated bonds as well as data on the money stock are obtained from Citibase. We follow studies in the literature and model the excess return on stocks

[10] See, e.g., Campbell (1987), Breen, Glosten and Jagannathan (1989) and Whitelaw (1994). Studies that allow the regression model to change over time include Pesaran and Timmermann (1995) and Perez-Quiros and Timmermann (1999).

defined as the difference between the monthly stock returns and the risk-free rate (I_{t-1}), both continuously compounded.

7.1. *Full Sample Estimation Results*

An alternative to the reversed Cusum and Cusum Squared tests that attempt to identify only the most recent break point is to adopt the Bai-Perron (1998a,1998b) procedure for estimation of multiple break points. Suppose the full data set up to time T is used to estimate up to q regression models.

$$\begin{array}{ll} y_\tau = \widehat{\boldsymbol{\beta}}_1^{\prime}\mathbf{x}_\tau + \widehat{u}_\tau & \tau = 1, 2, ..., T_1, \\ y_\tau = \widehat{\boldsymbol{\beta}}_2^{\prime}\mathbf{x}_\tau + \widehat{u}_\tau & \tau = T_1 + 1, ..., T_2, \\ \vdots & \vdots \\ y_\tau = \widehat{\boldsymbol{\beta}}_{q+1}^{\prime}\mathbf{x}_\tau + \widehat{u}_\tau & \tau = T_q + 1, ..., T. \end{array} \tag{51}$$

Bai and Perron (1998a,1998b) develop tests for the consistent estimation of the number and location of break points $(T_1, ..., T_q)$ and the parameters $(\beta_1^{\prime}, ..., \beta_{q+1}^{\prime})$.[11] Using data from the full sample, the AIC and sequential break procedure proposed by Bai and Perron (1998a) select three breaks, while the BIC and the information criterion proposed by Liu, Wu and Zidek (1997), denoted by LWZ, identify a single break point. Assuming a single break point (estimated to have occurred at 1962:10), the following estimates were obtained for the sample period 1954:1 to 1962:10

$$\begin{array}{lllll} \widehat{EXR}_t = & 0.016 & +0.468 Yield_{t-1} & \text{-}22.17 I_{t-1} & +26.29 Def_{t-1}, \\ & (0.048) & (0.972) & (8.21) & (36.70) \end{array} \tag{52}$$

while for 1962:11 to 1997:12 the model is

$$\begin{array}{lllll} \widehat{EXR}_t = & \text{-}0.020 & +1.635 Yield_{t-1} & \text{-}10.56 I_{t-1} & +26.63 Def_{t-1}. \\ & (0.011) & (0.471) & (1.63) & (8.72) \end{array} \tag{53}$$

The 90 percent confidence interval for the time of occurrence of this break point goes from 1962:04 to 1963:04.

[11] Bai and Perron consider two separate break point specifications. If lagged dependent variables are included as regressors, then u_t must be a martingale difference sequence and hence cannot be autocorrelated. However, if no lagged dependent variables are included as regressors u_t can be serially correlated and heteroskedastic.

The parameter estimates corresponding to the model with three break points are as follows. From 1954:1 to 1962:10 (with a 90 % confidence interval for the second date given by [1962:07;1963:01]) the following model is estimated

$$\begin{array}{llllll} \widehat{EXR}_t = & 0.016 & +0.468 Yield_{t-1} & \text{-}22.17 I_{t-1} & +26.29 Def_{t-1}, \\ & (0.046) & (0.947) & (7.99) & (35.72) \end{array} \tag{54}$$

while from 1962:11 to 1969:01 ([1967:12; 1970:02]) the parameters change to the following:

$$\begin{array}{llllll} \widehat{EXR}_t = & \text{-}0.014 & +0.586 Yield_{t-1} & \text{-}6.72 I_{t-1} & +76.29 Def_{t-1}. \\ & (0.096) & (2.95) & (9.75) & (47.23) \end{array} \tag{55}$$

¿From 1969:02 to 1990:12 ([1990:01;1991:11]) the parameters are

$$\begin{array}{llllll} \widehat{EXR}_t = & \text{-}0.085 & +3.066 Yield_{t-1} & \text{-}11.44 I_{t-1} & +35.25 Def_{t-1}. \\ & (0.016) & (0.555) & (1.80) & (9.48) \end{array} \tag{56}$$

Finally, from 1991:01 to 1997:12 we obtain

$$\begin{array}{llllll} \widehat{EXR}_t = & \text{-}0.022 & \text{-}1.196 Yield_{t-1} & \text{-}1.20 I_{t-1} & +100.39 Def_{t-1}. \\ & (0.042) & (1.498) & (6.54) & (48.29) \end{array} \tag{57}$$

These results point to the following conclusions. First, and most importantly, there is little doubt that there has been at least one break in the model for stock returns during the sample period under consideration. Even criteria as conservative as BIC and LWZ identify at least one break and less conservative procedures select three breaks. Second, at least in this application there seems to be some uncertainty about the number of break points, although conditional on having chosen the number of breaks, their time of occurrence is reasonably precisely estimated. Third, the size of the parameter variation between the break points is very large. For example, the coefficient on the interest rate is around twice as high in the sample prior to 1962:10 as in any subsequent sample. Also, the coefficient of the dividend yield which have been positive and highly significant before 1991, has become negative and statistically significant since 1991.

7.2. *Recursive Forecasting Results*

The recursive forecasting experiment begins in 1970:1 and extends to 1997:12. Initially there are 192 monthly observations. As explained in section 2, the key distinction between the approaches to determining window size lies in how much data they use to estimate the prediction model. Hence it is natural to begin our empirical analysis by showing time-series plots of the window length determined recursively through time.

Figure 13 converts the time series for the recursive estimates of the break point ($\widehat{T}_1$) identified by the reversed Cusum, Cusum Squared and recursive Bai-Perron procedures into calendar time. The breaks identified by the reversed Cusum method change very erratically and it is only up to around 1982 that it manages to identify any breaks. This finding is not surprising since it is well known that the Cusum test loses power very rapidly as the distance to a break point increases. The main breaks identified by the Cusum Squared method occur in 1968/69, 1974 and 1988. Accounting for the delays in the break point detection, the last two breaks thus correspond to the oil price shocks and the October 1987 stock market crash. The recursively identified break point series selected by the Bai Perron method is somewhat less stable than that produced by the reversed Cusum Squared method. While 1969 is identified as the break point for most of the sample, 1974 also gets selected during a five year spell from 1977 to 1982 and 1963 is chosen during three separate blocks of time from 1970 to 1972, from 1975 to 1977 and again from 1995 to 1997.[12]

Figure 14 plots the sequence of optimal window sizes determined by the three procedures conditional on the parameter estimates. For reference we also show the expanding window which gets selected if no break is identified. A sharp break in the regression model should show up as a drop in the window size, followed by a smoothly increasing window size until a subsequent break. First consider the reversed Cusum Squared criterion. Initially a small window of 20 or so observations

[12] Plots of recursive parameter estimates also indicated that there are breaks around 1974, 1979 and after 1994. These plots demonstrate the trade-off between bias and efficiency implied by the estimation methods. Methods that attempt to identify break dates and condition on this information pick up breaks quickly. However, they also produce very volatile regression estimates and forecasts as a result of only using a small window size immediately after a break is perceived to have occurred.

is adopted by this method rather than the full window of 192 observations which would be chosen in the absence of a break. The window then expands up to around 1975 at which point it again drops to 20 or so observations. From this point the window increases more or less in line with the data set up to 1994 when another sharp drop is registered. A somewhat smaller decline in the window size also occurs around 1997. Fairly long windows of the data are thus selected for most of the sample. On average the window size chosen by the Bai-Perron method is slightly longer than the corresponding window for the Cusum Squared method.

Our interest lies of course in the precision of the recursive forecasts produced by the alternative methods and plots of which are provided in Figure 15. Sample correlations between the forecasts vary between 0.69 and 0.85. Interestingly, the methods that do not put full weight on the earliest data points generate higher predictions than the expanding window method towards the end of the sample. This finding sheds light on some problems recently debated by economists. At the end of the sample the dividend yield was at a historical low and economists were speculating whether the relationship between the yield and stock returns had broken down. Indeed, the expanding window estimates suggest that the dividend yield coefficient has been in a systematic decline since 1994. This is picked up dramatically by the Cusum Squared, exponential smoothing and fixed window size methods all of which give a negative regression coefficient on the dividend yield towards the end of the sample and is also confirmed by the full-sample Bai-Perron estimates that allow for three breaks.

Basic performance results for the recursive forecasts generated by the six procedures are presented in the following table.[13] We measure forecasting performance in two ways. First we adopt the mean squared forecast error (MSFE) criterion. Secondly, since the proportion of correctly predicted signs of the 'market direction' is important for market timing purposes, we also report this statistic and the market timing test proposed by Pesaran and Timmermann (1992).

[13] In the exponential smoothing we set $\lambda = 0.95$. The fixed window size (c) equals sixty and thus uses data over the most recent five years. These are not arbitrarily chosen values and reflect common industry practice. Furthermore, the relative performances of the methods were found to be quite robust to varying these parameters.

Method	MSFE	correct signs (%)	PT test
Reversed Cusum	.324	60.11	3.53
Reversed Cusum (+stopping rule)	.323	59.27	3.16
Reversed Cusum Squared	.340	64.29	5.23
Reversed Cusum Squared (+stopping rule)	.333	62.20	4.52
Exponential smoothing ($\lambda = 0.95$)	.395	59.52	3.22
Fixed window ($c = 60$)	.339	61.31	3.77
Expanding window	.307	60.11	3.55
Recursive Bai-Perron	.352	57.44	2.91
Recursive Bai-Perron (+stopping rule)	.348	58.33	3.21

First consider the predictions generated by the expanding window method. We would expect these to deliver the best performance if there are no breaks. Certainly, this procedure generates the lowest values of the MSFE reflecting the stability of the parameter estimates computed using an expanding window of data. However, the proportion of correctly predicted signs of excess returns and the value of the market timing test obtained from this procedure are relatively low.

The exponential smoothing, reversed Cusum squared and recursive Bai-Perron methods generate the highest mean squared forecast errors and the lowest proportion of correct signs. Low MSFE values and high values of the sign test are generated by the fixed window size and the reversed Cusum Squared approaches.[14]

8. Conclusion

This paper has derived new results on determination of the optimal window of the most recent data that should be used in out-of-sample forecasting based on linear regression models. A number of new and, perhaps, surprising, results are

[14] Notice that for this data the reversed Cusum rules with optimally determined window size tend to perform better on the MSFE criterion but worse on the sign criterion than when the window simply goes back to $\widehat{T}_1$, the most recent break point. This is because the stopping rule conditions on $\widehat{T}_1$ and hence ignores the upward bias in estimating the distance to the most recent break. The Bai-Perron method for determining break points estimates T_1 consistently and for this method the stopping rule does improve both the MSFE and sign criterion.

reported. First of all, there exists a recursive stopping rule for determination of the window size that minimizes the mean squared prediction error. Second, this window size does not grow monotonically as more data arrives immediately after a break. Compared to both fixed window size and expanding window size approaches, important gains in forecast accuracy can be obtained by attempting to identify the most recent break and applying our stopping rule to select the optimal window size.

Appendix

Proof of Proposition 2

For different values of m, the relative value of $MSFE(m)$ depends on $\theta_m^2\mu^2 + (\theta_m\psi + 1)/Q_{m,T}$. Hence the change in MSFE resulting from using observations back to period m instead of period $m+1$ is

$$\Delta(m) = \left(\theta_m^2 - \theta_{m+1}^2\right)\mu^2 + (\theta_m\psi + 1)/Q_{m,T} - (\theta_{m+1}\psi + 1)/Q_{m+1,T}. \tag{58}$$

To evaluate this expression, first notice that

$$\theta_m^2 - \theta_{m+1}^2 = \frac{Q_{m,T_1}^2}{Q_{m,T}^2} - \frac{Q_{m+1,T_1}^2}{Q_{m+1,T}^2}.$$

where $Q_{m,T} = Q_{m+1,T} + x_m^2$, so that $Q_{m,T}^2 = Q_{m+1,T}^2 + x_m^4 + 2x_m^2 Q_{m+1,T}$. Hence

$$\begin{aligned}
\theta_m^2 - \theta_{m+1}^2 &= \frac{(Q_{m+1,T_1}^2 + x_m^4 + 2x_m^2 Q_{m+1,T_1})Q_{m+1,T}^2 - (Q_{m+1,T}^2 + x_m^4 + 2x_m^2 Q_{m+1,T})Q_{m+1,T_1}^2}{Q_{m,T}^2 Q_{m+1,T}^2} \\
&= \frac{x_m^4(Q_{m+1,T}^2 - Q_{m+1,T_1}^2) + 2x_m^2 Q_{m+1,T_1}Q_{m+1,T}(Q_{m+1,T} - Q_{m+1,T_1})}{Q_{m,T}^2 Q_{m+1,T}^2} \\
&= \frac{2x_m^2\theta_{m+1}(1-\theta_{m+1})}{Q_{m,T}} + \frac{x_m^4(1-\theta_{m+1})^2}{Q_{m,T}^2}. \qquad (59)
\end{aligned}$$

Furthermore, the second term in (58) becomes

$$\begin{aligned}
\frac{\theta_m\psi + 1}{Q_{m,T}} - \frac{\theta_{m+1}\psi + 1}{Q_{m+1,T}} &= \psi\left(\frac{Q_{m,T_1}}{Q_{m,T}^2} - \frac{Q_{m+1,T_1}}{Q_{m+1,T}^2}\right) + \left(\frac{1}{Q_{m,T}} - \frac{1}{Q_{m+1,T}}\right) \\
&= \psi\left(\frac{Q_{m,T_1}Q_{m+1,T}^2 - Q_{m+1,T_1}Q_{m,T}^2}{Q_{m,T}^2 Q_{m+1,T}^2}\right) + \frac{Q_{m+1,T} - Q_{m,T}}{Q_{m,T}Q_{m+1,T}} \\
&= \frac{\psi\left(\left(Q_{m+1,T_1} + x_m^2\right)Q_{m+1,T}^2 - Q_{m+1,T_1}(Q_{m+1,T}^2 + x_m^4 + 2x_m^2 Q_{m+1,T})\right)}{Q_{m,T}^2 Q_{m+1,T}^2} \\
&\quad - \frac{x_m^2}{Q_{m,T}Q_{m+1,T}} \\
&= \frac{\psi x_m^2(1 - 2\theta_{m+1})}{Q_{m,T}^2} - \frac{\psi x_m^4\theta_{m+1}}{Q_{m,T}^2 Q_{m+1,T}} - \frac{x_m^2}{Q_{m,T}Q_{m+1,T}}. \qquad (60)
\end{aligned}$$

Combining (59) and (60) and setting $\psi = 0$, we have

$$\begin{aligned}
\Delta(m) &= \frac{2x_m^2\theta_{m+1}(1-\theta_{m+1})}{Q_{m,T}} + \frac{x_m^4(1-\theta_{m+1})^2}{Q_{m,T}^2} - \frac{x_m^2}{Q_{m,T}Q_{m+1,T}} \\
&= \frac{x_m^2}{Q_{m,T}Q_{m+1,T}}\left\{\frac{\mu^2(Q_{m+1,T}-Q_{m+1,T_1})}{Q_{m,T}Q_{m+1,T}}\left(2Q_{m,T}Q_{m+1,T_1} + x_m^2(Q_{m+1,T}-Q_{m+1,T_1})\right) - 1\right\} \\
&= \frac{x_m^2}{Q_{m,T}Q_{m+1,T}}\left\{\frac{\mu^2 Q_{T_1+1,T}}{Q_{m,T}Q_{m+1,T}}\left(2Q_{m,T}Q_{m+1,T_1} + x_m^2 Q_{T_1+1,T}\right) - 1\right\}. \qquad (61)
\end{aligned}$$

To prove that an optimal stopping rule exists, it is sufficient to show that, for a given value of T, $\Delta(m) > 0$ implies that $\Delta(m-1) > 0$. Now $\Delta(m) > 0$ if

$$\frac{\mu^2 Q_{T_1+1,T}}{Q_{m,T}Q_{m+1,T}}\left(2Q_{m,T}Q_{m+1,T_1} + x_m^2 Q_{T_1+1,T}\right) > 1. \qquad (62)$$

Hence the result follows if the left hand side of (62), evaluated one period earlier, increases. This is easily demonstrated to hold:

$$\begin{aligned}
&\frac{\mu^2 Q_{T_1+1,T}}{Q_{m,T}Q_{m+1,T}}\left(2Q_{m,T}Q_{m+1,T_1} + x_m^2 Q_{T_1+1,T}\right) - \frac{\mu^2 Q_{T_1+1,T}}{Q_{m-1,T}Q_{m,T}}\left(2Q_{m-1,T}Q_{m,T_1} + x_{m-1}^2 Q_{T_1+1,T}\right) \\
= &\frac{\mu^2 Q_{T_1+1,T}}{Q_{m-1,T}Q_{m,T}Q_{m+1,T}}\{2Q_{m,T}Q_{m-1,T}Q_{m+1,T_1} + x_m^2 Q_{m-1,T}Q_{T_1+1,T} \\
&-2Q_{m-1,T}Q_{m,T_1}Q_{m+1,T} - x_{m-1}^2 Q_{m+1,T}Q_{T_1+1,T}\}
\end{aligned}$$

Using that $Q_{m,T} = Q_{m-1,T} + x_m^2$, the term inside the curly bracket simplifies to

$$\begin{aligned}
&2Q_{m-1,T}\left((Q_{m+1,T} + x_m^2)Q_{m+1,T_1} - (Q_{m+1,T_1} + x_m^2)Q_{m+1,T}\right) \\
&+Q_{T_1+1,T}(x_m^2 Q_{m-1,T} - x_{m-1}^2 Q_{m+1,T}) \\
= \; &-Q_{T_1+1,T}(x_m^2 Q_{m-1,T} + x_{m-1}^2 Q_{m+1,T}) < 0.
\end{aligned}$$

Hence we have shown that if $\Delta(m) > 0$ (so that MSFE increases by starting the window at $t = m$, rather than at $t = m + 1$, then the MSFE will also increase by starting the window at observation $t = m - 1$ rather than at observation $t = m$. By induction, the optimal window size will be the largest value $t = m$, for which $\Delta(m) > 0$.

Proof of Proposition 3

Suppose the sample is extended from T to $T+1$. Recall that the stopping rule determining m^* is the first value of m (arranged in decreasing order) for which $\Delta(m-1) > 0$. Hence we need to show that if $\Delta(m^*(T)) > 0$, then $\Delta(m^*(T+1)) > 0$. By assumption (and using (62)),

$$\frac{\mu^2 Q_{T_1+1,T}}{Q_{m,T}Q_{m+1,T}} \left(2Q_{m,T}Q_{m+1,T_1} + x_m^2 Q_{T_1+1,T}\right) > 1,$$

so we need to show that this implies that

$$\frac{\mu^2 Q_{T_1+1,T+1}}{Q_{m,T+1}Q_{m+1,T+1}} \left(2Q_{m,T+1}Q_{m+1,T_1} + x_m^2 Q_{T_1+1,T+1}\right) > 1. \tag{63}$$

This is easily demonstrated. A sufficient condition for the difference between the left hand sides of (62) and (63) to be positive is that

$$\begin{aligned} \frac{Q_{T_1+1,T+1}Q_{m+1,T_1}}{Q_{m+1,T+1}} &> \frac{Q_{T_1+1,T}Q_{m+1,T_1}}{Q_{m+1,T}}, \text{ and} \\ \frac{Q^2_{T_1+1,T+1}}{Q_{m,T+1}Q_{m+1,T+1}} &> \frac{Q^2_{T_1+1,T}}{Q_{m,T}Q_{m+1,T}} \end{aligned}$$

The first condition holds if

$$\begin{aligned} (Q_{T_1+1,T} + x^2_{T+1})Q_{m+1,T} &> Q_{T_1+1,T}(Q_{m+1,T} + x^2_{T+1}), \text{ ie. if} \\ x^2_{T+1}(Q_{m+1,T} - Q_{T_1+1,T}) &> 0, \end{aligned}$$

which holds as $m < T_1 + 1$. The second condition holds if

$$Q_{m,T}Q_{m+1,T}(Q^2_{T_1+1,T}+x^4_{T+1}+2x^2_{T+1}Q_{T_1+1,T}) > (Q_{m,T}+x^2_{T+1})(Q_{m+1,T}+x^2_{T+1})Q^2_{T_1+1,T}, \text{ or}$$

$$Q_{m,T}Q_{m+1,T}(x^4_{T+1} + 2x^2_{T+1}Q_{T_1+1,T}) > Q^2_{T_1+1,T}(x^4_{T+1} + Q_{m,T}x^2_{T+1} + Q_{m+1,T}x^2_{T+1}).$$

This in turn is satisfied provided that

$$x^4_{T+1}(Q_{m,T}Q_{m+1,T} - Q^2_{T_1+1,T}) + Q_{m,T}Q_{T_1+1,T}x^2_{T+1}(Q_{m+1,T} - Q_{T_1+1,T})$$

$$+Q_{m+1,T}Q_{T_1+1,T}x^2_{T+1}(Q_{m,T} - Q_{T_1+1,T}) > 0$$

which is easily seen to hold as each term is positive.

To prove that $\Delta(m,T) > 0$, does not rule out that $\Delta(m+2,T+1) > 0$, notice that $\Delta(m,T) > 0$ holds provided that

$$\frac{1}{Q_{m,T}Q_{m+1,T}}\left(2Q_{m,T}Q_{m+1,T_1} + x_m^2 Q_{T_1+1,T}\right) > \frac{1}{\mu^2 Q_{T_1+1,T}}, \tag{64}$$

while the second condition holds if

$$\frac{1}{Q_{m+2,T+1}Q_{m+3,T+1}}\left(2Q_{m+2,T+1}Q_{m+3,T_1} + x_{m+2}^2 Q_{T_1+1,T+1}\right) > \frac{1}{\mu^2 Q_{T_1+1,T+1}}. \tag{65}$$

It is easily seen that (64) does not rule out (65).

Proof of proposition 5

To derive the moments of θ_m we first note that

$$\begin{aligned}\theta_m &= \frac{\sum_{t=m}^{T_1} x_t^2}{\sum_{t=m}^{T_1} x_t^2 + \sum_{t=T_1+1}^{T} x_t^2}, \\ &\overset{d}{=} \frac{\chi_{\nu_1}^2(\lambda_1)}{\chi_{\nu_1}^2(\lambda_1) + \chi_{\nu_2}^2(\lambda_2)},\end{aligned}$$

where $\chi_{\nu_i}^2(\lambda_i)$ is distributed as a non-central chi–squared with ν_i degrees of freedom and the non-centrality parameter $\lambda_i = \nu_i \mu_x^2$. Recall that $\nu_1 = T_1 - m + 1$ and $\nu_2 = T - T_1$. Hence θ_m has a doubly non-central beta distribution with parameters $\nu_1/2$ and $\nu_2/2$ and non-centrality parameters λ_1 and λ_2.[15] Approximating each of the non-central χ^2 variables in θ_m and using Patnaik's approximation (Patnaik (1949)), we have

$$\theta_m \overset{appr}{\sim} \left(\frac{\frac{\nu_1}{2} + 2\lambda_1}{\frac{\nu_1}{2} + \lambda_1}\right)\left(\frac{\frac{\nu_2}{2} + \lambda_2}{\frac{\nu_2}{2} + 2\lambda_2}\right) Beta(f_1, f_2),$$

where

$$f_i = \frac{\left(\frac{\nu_i}{2} + \lambda_i\right)^2}{\frac{\nu_i}{2} + 2\lambda_1} = \frac{\nu_i(1+2\mu_x^2)^2}{2+8\mu_x^2} \equiv \nu_i k,$$

and $k = (1+2\mu_x^2)^2/(2+8\mu_x^2)$. Noting that since $\lambda_i = \nu_i \mu_x^2$ then

$$\left(\frac{\frac{\nu_1}{2} + 2\lambda_1}{\frac{\nu_1}{2} + \lambda_1}\right)\left(\frac{\frac{\nu_2}{2} + \lambda_2}{\frac{\nu_2}{2} + 2\lambda_2}\right) = 1,$$

[15] See, for example, Johnson and Kotz (1970) pages 197-198.

then $\theta_m \overset{appr}{\sim} Beta(f_1, f_2)$ and $E[\theta_m]$ and $E[\theta_m^2]$ can be directly calculated from the moments of the (central) beta distribution:

$$\begin{aligned} E[\theta_m] &\approx \frac{f_1}{f_1+f_2} = \frac{\nu_1}{\nu_1+\nu_2} = \frac{\nu_1}{\nu}, \\ E[\theta_m^2] &\approx \frac{f_1(f_1+1)}{(f_1+f_2)(f_1+f_2+1)}, \\ &= \frac{k\nu_1(1+k\nu_1)}{k(\nu_1+\nu_2)\left[k(\nu_1+\nu_2)+1\right]}, \\ &= \left(\frac{\nu_1}{\nu}\right)\frac{(1+k\nu_1)}{(1+k\nu)}. \end{aligned}$$

To derive the MSFE, notice that, under the assumptions stated in Proposition 3, the conditional distribution of $\widehat{y}_{T+1}$ given the sequence of x's $\{x_1, ..., x_{T+1}\}$ is

$$E[\widehat{y}_{T+1}|x_1, x_2, ..., x_{T+1}] = \beta_2 x_{T+1} + (\beta_1 - \beta_2)\theta_m x_{T+1}, \tag{66}$$

or, unconditionally,

$$E[\widehat{y}_{T+1}] = \beta_2 \mu_x + (\beta_1 - \beta_2)E[\theta_m x_{T+1}].$$

Since θ_m only depends on $x_1, .., x_T$, it is independent of x_{T+1} and we have by the law of iterated expectations

$$\begin{aligned} E[\widehat{y}_{T+1}] &= \beta_2\mu_x + (\beta_1-\beta_2)E[E[\theta_m x_{T+1}|x_1, ...x_T]], \\ &= \beta_2\mu_x + (\beta_1-\beta_2)E[\theta_m E[x_{T+1}|x_1, ...x_T]], \\ &= \beta_2\mu_x + (\beta_1-\beta_2)\frac{\nu_1}{\nu}\mu_x, \\ &= \mu_x\left(\frac{\beta_1\nu_1+\beta_2\nu_2}{\nu}\right). \end{aligned}$$

Hence

$$\begin{pmatrix} y_{T+1} \\ \widehat{y}_{T+1} \end{pmatrix} \sim IIN\left\{\begin{pmatrix} \beta_2\mu_x \\ \beta_2\mu_x + (\beta_1-\beta_2)\left(\frac{\mu_x\nu_1}{\nu}\right) \end{pmatrix}, \mathbf{\Sigma}\right\},$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \beta_2^2\omega^2+\sigma^2 & g \\ g & h^2 \end{pmatrix},$$

and $h^2 = V(\widehat{y}_{T+1})$ and $g = Cov(y_{T+1}, \widehat{y}_{T+1})$. First consider the unconditional variance of $\widehat{y}_{T+1}$:

$$h^2 = V(\widehat{y}_{T+1}) = E[V(\widehat{y}_{T+1}|x_1, .., x_T, x_{T+1})] + V\left[E\left(\widehat{y}_{T+1}|x_1, .., x_T, x_{T+1}\right)\right].$$

Using the assumption that $\psi = 0$, the conditional variance of $\widehat{y}_{T+1}$ is given by

$$V(\widehat{y}_{T+1}|x_1, .., x_T, x_{T+1}) = \frac{\sigma^2 x_{T+1}^2}{\sum_{t=m}^T x_t^2}.$$

Therefore, using (66) we have

$$h^2 = \sigma^2 E\left(\frac{x_{T+1}^2}{\sum_{t=m}^T x_t^2}\right) + V\left[\beta_2 x_{T+1} + (\beta_1 - \beta_2)\theta_m x_{T+1}\right].$$

The second term in this expression is given by

$$\begin{aligned} V(E[\widehat{y}_{T+1}|x_1, .., x_{T+1}]) &= \beta_2^2\omega^2 + (\beta_1 - \beta_2)^2 E[(\theta_m x_{T+1} - \mu_x \frac{\nu_1}{\nu})^2] \\ &\quad +2\beta_2(\beta_1 - \beta_2)E[(x_{T+1} - \mu_x)(\theta_m x_{T+1} - \mu_x \frac{\nu_1}{\nu})] \\ &= \beta_2^2\omega^2 + (\beta_1 - \beta_2)^2(\frac{\nu_1}{\nu})\left(\omega^2\left(\frac{1 + k\nu_1}{1 + k\nu}\right) + \mu_x^2(\frac{\nu_2}{\nu(1 + k\nu)})\right) \\ &\quad +2\beta_2(\beta_1 - \beta_2)(\frac{\nu_1}{\nu})\omega^2. \end{aligned}$$

To evaluate $E\left(x_{T+1}^2 / \sum_{t=m}^T x_t^2\right)$, we first note that

$$\frac{x_{T+1}^2}{\sum_{t=m}^T x_t^2} = \frac{(x_{T+1}^2/\omega)^2}{\nu\sum_{t=m}^T (x_t/\omega)^2/\nu} = \left(\frac{1}{\nu}\right)\left(\frac{(x_{T+1}^2/\omega)}{\chi_\nu^2(\lambda)/\nu}\right)^2$$

and hence

$$\frac{\nu x_{T+1}^2}{\sum_{t=m}^T x_t^2} = \left[t_\nu(\delta, \lambda)\right]^2,$$

where $t_\nu(\delta, \lambda) = (x_{T+1}/\omega)/(\chi_\nu(\lambda)/\sqrt{\nu})$ is distributed as a double non-central t-distribution with ν degrees of freedom and the non-centrality parameters $\delta = \mu_x/\omega$ and $\lambda = \nu\mu_x^2$. Using results in Johnson and Kotz (1970, p. 214, equation (25)) we now have

$$E\left(\frac{x_{T+1}^2}{\sum_{t=m}^{T} x_t^2}\right) = \varkappa_1(\lambda,\delta,\nu) = \left(\frac{1}{2}\right)\exp(-\frac{1}{2}\lambda)(1+\delta^2)\times \sum_{j=0}^{\infty}\frac{(\frac{1}{2}\lambda)^j}{j!}\frac{\Gamma(\frac{1}{2}(\nu-2)+j)}{\Gamma(\frac{1}{2}\nu+j)}. \tag{67}$$

Using this expression the total variance of $\widehat{y}_{T+1}$ can be written as:

$$\begin{aligned} h^2 &= V(\widehat{y}_{T+1}) = \sigma^2\varkappa_1(\lambda,\delta,\nu) + \beta_2^2\omega^2 + \\ &(\beta_1-\beta_2)^2\omega^2(\frac{\nu_1}{\nu})\left\{\left(\frac{1+k\nu_1}{1+k\nu}\right)+\frac{\mu_x^2}{\omega^2}\left(\frac{\nu_2}{\nu(1+k\nu)}\right)\right\} \\ &+2\beta_2(\beta_1-\beta_2)\omega^2(\frac{\nu_1}{\nu}). \end{aligned} \tag{68}$$

Also

$$\begin{aligned} g &\equiv Cov(y_{T+1},\widehat{y}_{T+1}) = E[(\beta_2(x_{T+1}-\mu_x)+u_{T+1})(\beta_2(x_{T+1}-\mu_x) \\ &+(\beta_1-\beta_2)(\theta_m x_{T+1}-\mu_x\frac{\nu_1}{\nu})+x_{T+1}\frac{\sum_{t=m}^{T}x_t u_t}{\sum_{t=m}^{T}x_t^2})], \\ &= \beta_2^2\omega^2+\beta_2(\beta_1-\beta_2)\omega^2(\nu_1/\nu). \end{aligned} \tag{69}$$

Using this result and noting that

$$\begin{aligned} \Pr(y_{T+1} &> 0) = \Phi\left(\frac{\beta_2\mu_x}{\sqrt{\beta_2^2\omega^2+\sigma^2}}\right), \\ \Pr(\widehat{y}_{T+1} &> 0) = \Phi\left(\frac{\mu_x(\beta_1\nu_1+\beta_2\nu_2)}{\nu h}\right), \end{aligned}$$

it is easy to compute the joint probability that $y_{T+1}>0$ and $\widehat{y}_{T+1}>0:$

$$A_H = \Pr(y_{T+1}>0,\widehat{y}_{T+1}>0) = \int_{a_2=-\mu_2}^{\infty}\int_{a_1=-\mu_1}^{\infty} f(a_1,a_2)da_1da_2,$$

where

$$\mu_1=\mu_x\beta_2,\ \ \mu_2=\mu_x(\beta_1\nu_1+\beta_2\nu_2)/\nu,$$

$$f(a_1,a_2) = (2\pi)^{-1}\left|\mathbf{\Sigma}\right|^{-1/2}\exp(-\frac{1}{2}\mathbf{a}'\mathbf{\Sigma}^{-1}\mathbf{a}).$$

and $\mathbf{a} = (a_1,a_2)'$. Similarly,

$$A_F = \Pr(y_{T+1} < 0, \widehat{y}_{T+1} > 0) = \int_{a_2=-\mu_2}^{\infty}\int_{a_1=-\infty}^{-\mu_1} f(a_1,a_2)da_1da_2.$$

Using the above results in (24) and (25) we have:

$$KS = \frac{A_H}{\Phi(\mu_1/\sqrt{\beta_2^2\omega^2+\sigma^2})} - \frac{A_F}{1-\Phi(\mu_1/\sqrt{\beta_2^2\omega^2+\sigma^2})}. \tag{70}$$

In the case where $\mu_x = 0$, $\Phi(\mu_1/\sqrt{\beta_2^2\omega^2+\sigma^2}) = 1/2$, and $KS = 2(A_H - A_F)$.

Under joint normality of y_{T+1} and $\widehat{y}_{T+1}$, a sufficient statistic that characterizes their distribution is the correlation coefficient, ρ given by

$$\rho = \frac{g}{h\sqrt{\sigma^2+\beta_2^2\omega^2}} = \frac{g}{\sigma h\sqrt{1+\gamma^2}},$$

where $\gamma^2 = \omega^2\beta_2^2/\sigma^2$ is the signal-to-noise ratio. When $\rho = 0$, we have $f(a_1,a_2) = f(a_1)f(a_2)$ so $KS = \Pr(\widehat{y}_{T+1} > 0) - \Pr(\widehat{y}_{T+1} > 0) = 0$. In general, however, using (68) and (69) we have

$$\rho = \frac{\gamma^2}{\sqrt{1+\gamma^2}} \times \frac{1+\phi(\nu_1/\nu)}{\left[\varkappa(\lambda,\delta,\nu)+\gamma^2+\gamma^2\phi^2\varsigma+2\gamma^2\phi(\nu_1/\nu)\right]^{1/2}}, \tag{71}$$

where $\phi = (\beta_1-\beta_2)/\beta_2$ is the rate of change in β, $\varkappa(\lambda,\delta,\nu)$ is defined by (67) and

$$\varsigma = \left(\frac{\nu_1}{\nu}\right)\left\{\left(\frac{1+k\nu_1}{1+k\nu}\right)+\frac{\mu_x^2}{\omega^2}\left(\frac{\nu_2}{\nu(1+k\nu)}\right)\right\}. \tag{72}$$

In the case where $\mu_x = 0$, then $k = 1/2$ and the expression for ρ simplifies and we have (exactly):

$$\rho = \left(\frac{\gamma^2}{\sqrt{1+\gamma^2}}\right)\left(\frac{1+\phi(\nu_1/\nu)}{\left\{\frac{1}{\nu-2}+\gamma^2+\gamma^2\phi^2\left(\frac{(2+\nu_1)\nu_1}{(2+\nu)\nu}\right)+2\gamma^2\phi(\nu_1/\nu)\right\}^{1/2}}\right). \tag{73}$$

Bibliography

Alogoskoufis, G.S. and R. Smith (1991) “The Phillips Curve, the Persistence of Inflation, and the Lucas Critique: Evidence from Exchange Rate Regimes”. American Economic Review 81, 1254-1275.

Andrews, D.W.K. (1993) “Tests for Parameter Instability and Structural Change with Unknown Change Point”. Econometrica 61, 821-856.

Andrews, D.W.K. and W. Ploberger (1996) “Optimal Changepoint Tests for Normal Linear Regression”. Journal of Econometrics 70, 9-38.

Bai, J. and P. Perron (1998a) “Estimating and Testing Linear Models with Multiple Structural Changes”. Econometrica 66, 47-78.

Bai, J. and P. Perron (1998b) “Computation and Analysis of Multiple Structural Change Models”. Manuscript, MIT and Boston University.

Breen, W., L.R. Glosten, and R. Jagannathan (1989) “The Economic Significance of Predictable Variations in Stock Index Returns”. Journal of Finance 44, 1177-1189.

Brown, R.L., J. Durbin, and J.M. Evans (1975) “Techniques for Testing the Constancy of Regression Relationships over Time”. Journal of the Royal Statistical Society, Series B, 37, 149-192.

Campbell, J.Y. (1987) “Stock Returns and the Term Structure”. Journal of Financial Economics 18, 373-399.

Chow, G. (1960) “Tests of Equality Between Sets of Coefficients in Two Linear Regressions”. Econometrica 28, 591-605.

Chu, C-S J., M. Stinchcombe, and H. White (1996) “Monitoring Structural Change”. Econometrica 64, 1045-1065.

Clements, M.P. and D.F. Hendry (1998) “Forecasting Economic Time Series”. Cambridge University Press.

Clements, M.P. and D.F. Hendry (1999) “Forecasting in a Non-stationary Economy”. Mimeo, Nuffield College.

Cooley, T.F. and E.C. Prescott (1976) ”Efficient Estimation in the Presence of Stochastic Parameter Variation”. Econometrica 44, 167-184.

Garcia, R.. and P. Perron, 1996 “An Analysis of the Real Interest Rate under Regime Shifts”. Review of Economics and Statistics 78, 111-125.

Granger, C.W.J., and M.H. Pesaran (1999) “Economic and Statistical Measures of Forecast Accuracy”. Manuscript, University of Cambridge.

Hansen, B.E. (1992) “Tests for Parameter Instability in Regressions with I(1) Processes”. Journal of Business and Economic Statistics 10, 321-335.

Harvey, A.C. (1989) “Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge.

Inclan, C. and G.C. Tiao (1994) ”Use of Cumulative Sums of Squares for Retrospective Detection of Changes of Variance”. Journal of the American Statistical Association 89, 913-923.

Johnson, N.L. and S. Kotz (1970) “Continuous Univariate Distributions - 2”. John Wiley.

Keim, D.B., and R.F. Stambaugh (1986) “Predicting Returns in the Stock and Bond Markets”. Journal of Financial Economics, 17, 357-390.

Leitch, G. and Tanner, J.E. (1991) “Economic Forecast Evaluation: Profits Versus the Conventional Error Measures”, American Economic Review, 81, 580-90.

Liu, J., S. Wu, and J.V. Zidek (1997) “On Segmented Multivariate Regressions”. Statistica Sinica 7, 497-525.

Patnaik, P.B. (1949) “The Non-central χ^2- and F-Distributions and Their Applications,”, Biometrika 36, 202-232.

Perez-Quiros, G. and A. Timmermann (1999) “Firm Size and Cyclical Variations in Stock Returns”. Forthcoming in Journal of Finance.

Pesaran, M.H. and A. Timmermann (1992) “A Simple Non-parametric Test of Predictive Performance”. Journal of Business and Economic Statistics, 10, 461-465.

Pesaran, M.H., and A. Timmermann (1995) “Predictability of Stock Returns: Robustness and Economic Significance”. Journal of Finance 50, 1201-1228.

Ploberger, W., W. Kramer, and K. Kontrus (1989) ”A New Test for Structural Stability in the Linear Regression Model”. Journal of Econometrics 40, 307-318.

Timmermann, A. (1998) “Structural Breaks, Incomplete Information and Stock Prices”. Manuscript, Financial Markets Group, LSE.

Whitelaw, R.F. (1994) “Time Variations and Covariations in the Expectation and Volatility of Stock Market Returns”. Journal of Finance 49, 515-541.

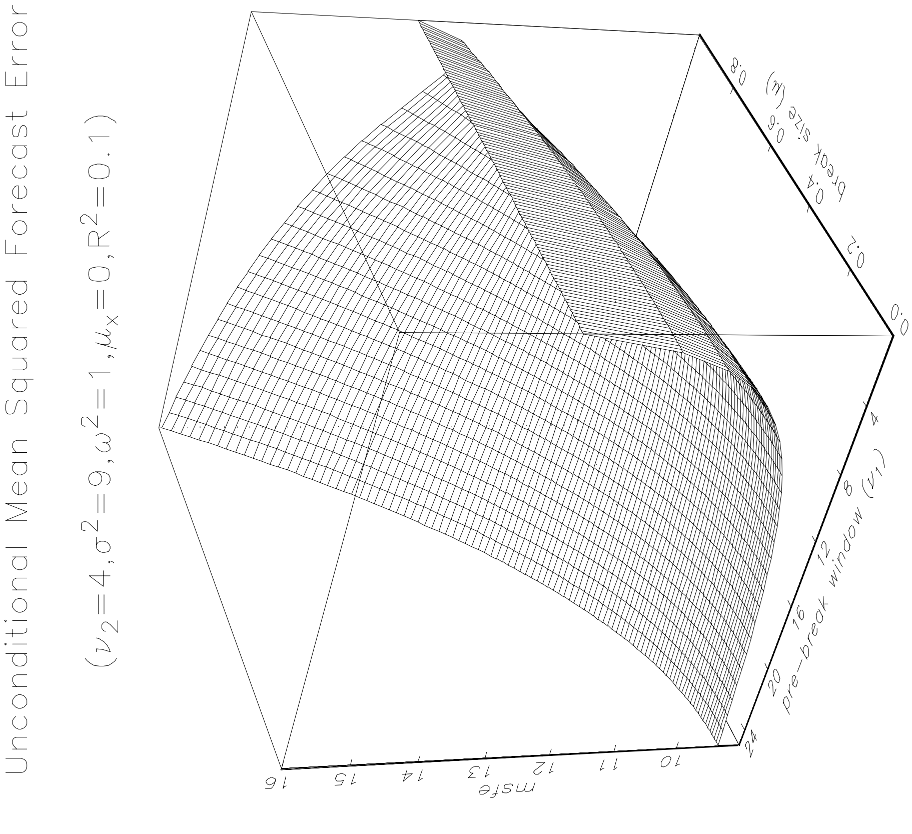

Unconditional Mean Squared Forecast Error
$(\nu_2=4,\sigma^2=9,\omega^2=1,\mu_x=0,R^2=0.1)$
msfe
16
15
14
13
12
11
10
pre-break window (ν_1)
24
20
16
12
8
4
break size (μ)
0.0
0.2
0.4
0.6
0.8

Figure 1:

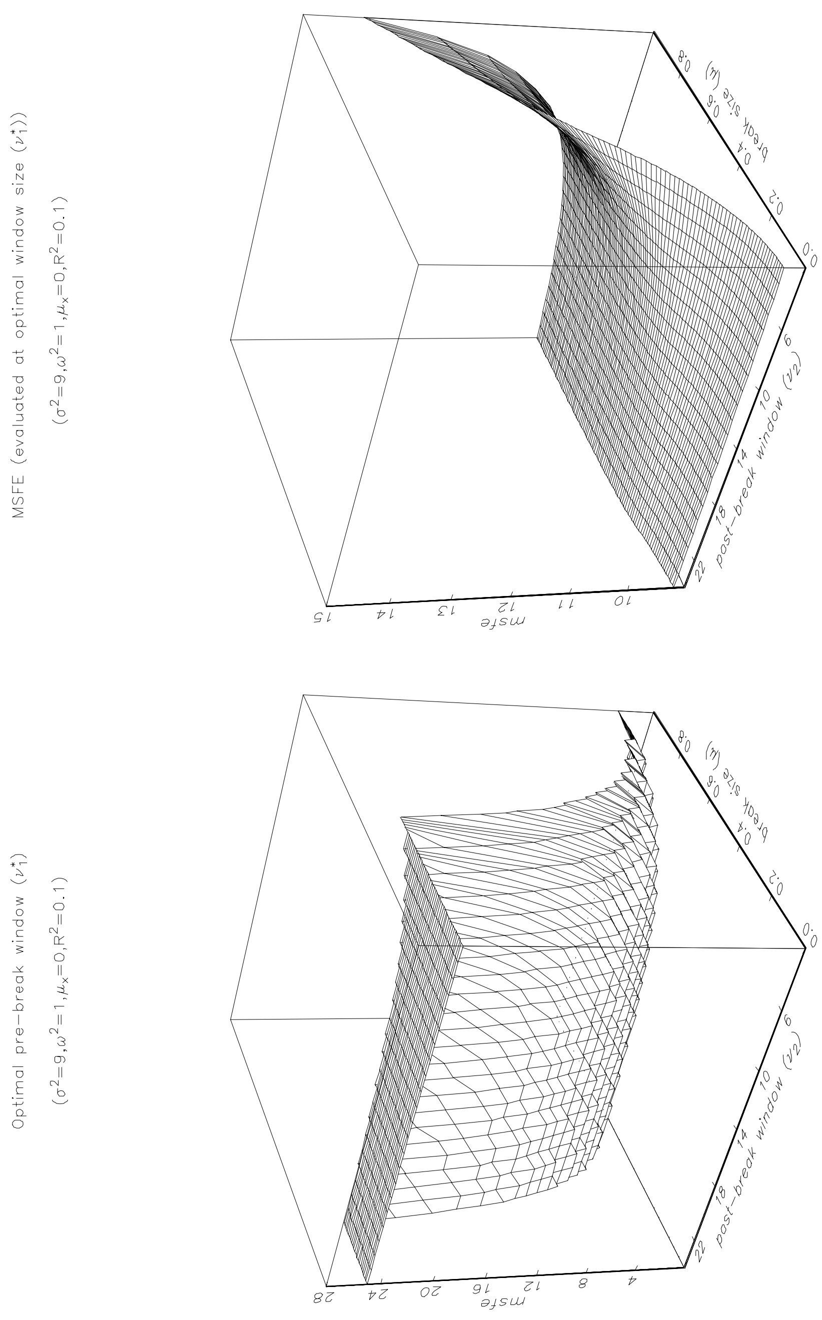

MSFE (evaluated at optimal window size (ν_1^*))
(σ^2=9,ω^2=1,μ_x=0,R^2=0.1)
msfe
10
11
12
13
14
15
post-break window (ν_2)
6
10
14
18
22
break size (μ)
0.0
0.2
0.4
0.6
0.8
Optimal pre-break window (ν_1^*)
(σ^2=9,ω^2=1,μ_x=0,R^2=0.1)
msfe
4
8
12
16
20
24
28
post-break window (ν_2)
6
10
14
18
22
break size (μ)
0.0
0.2
0.4
0.6
0.8

Figure 2:

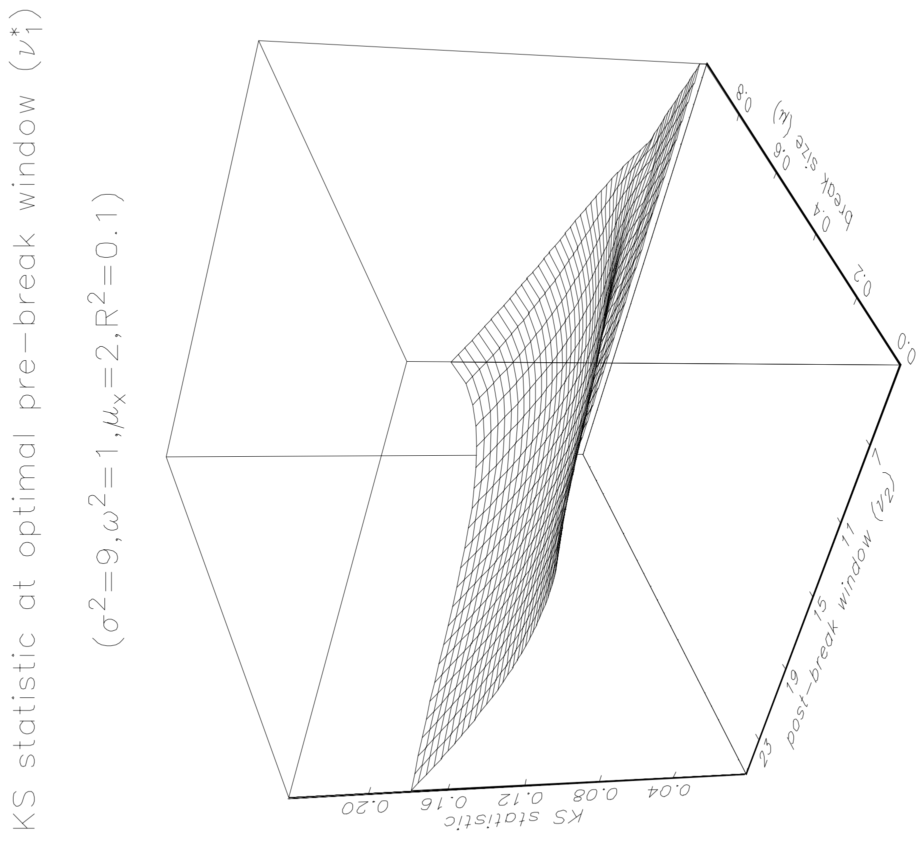

KS statistic at optimal pre-break window (ν_1^*)
(σ^2=9,ω^2=1,μ_x=2,R^2=0.1)
break size (μ)
0.0
0.2
0.4
0.6
0.8
post-break window (ν_2)
7
11
15
19
23
KS statistic
0.04
0.08
0.12
0.16
0.20

Figure 3:

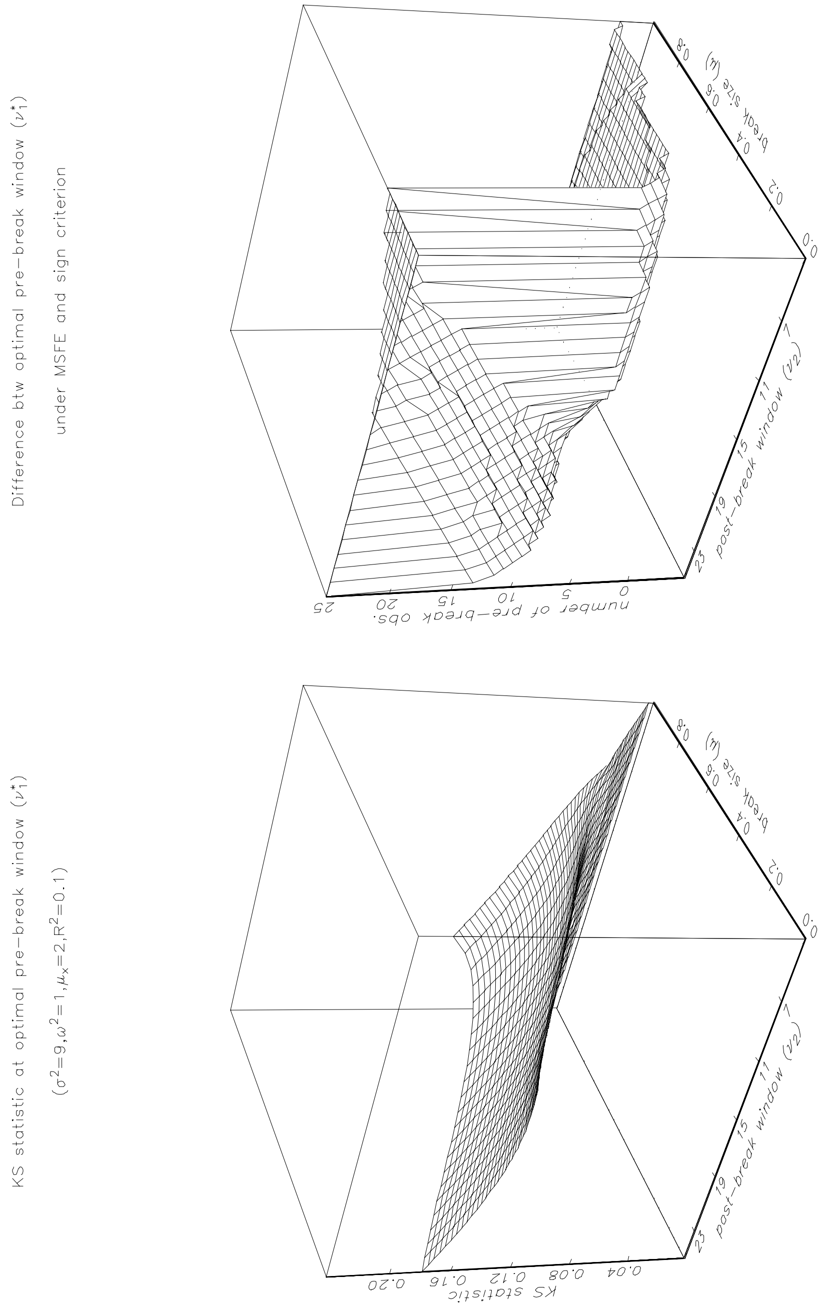
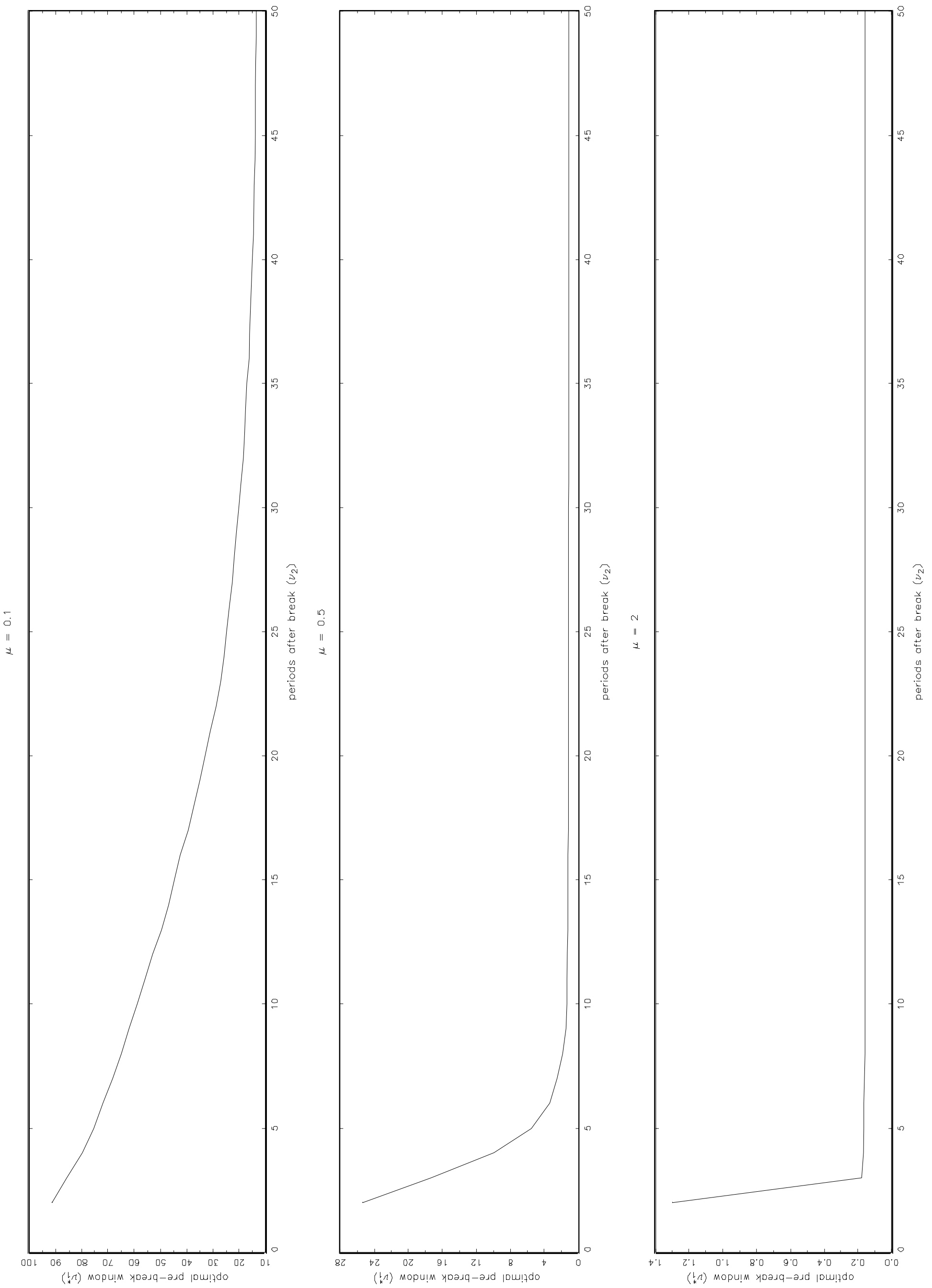

Difference btw optimal pre-break window (ν_1^*)
under MSFE and sign criterion
number of pre-break obs.
25
20
15
10
5
0
post-break window (ν_2)
23
19
15
11
7
break size (μ)
0.0
0.2
0.4
0.6
0.8
KS statistic at optimal pre-break window (ν_1^*)
(σ^2=9,ω^2=1,μ_x=2,R^2=0.1)
KS statistic
0.20
0.16
0.12
0.08
0.04
post-break window (ν_2)
23
19
15
11
7
break size (μ)
0.0
0.2
0.4
0.6
0.8

Figure 4:



μ = 0.1
μ = 0.5
μ = 2
periods after break (ν_2)
optimal pre−break window (ν_1^*)
0
5
10
15
20
25
30
35
40
45
50
10
20
30
40
50
60
70
80
90
100
0
4
8
12
16
20
24
28
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

Figure 5:

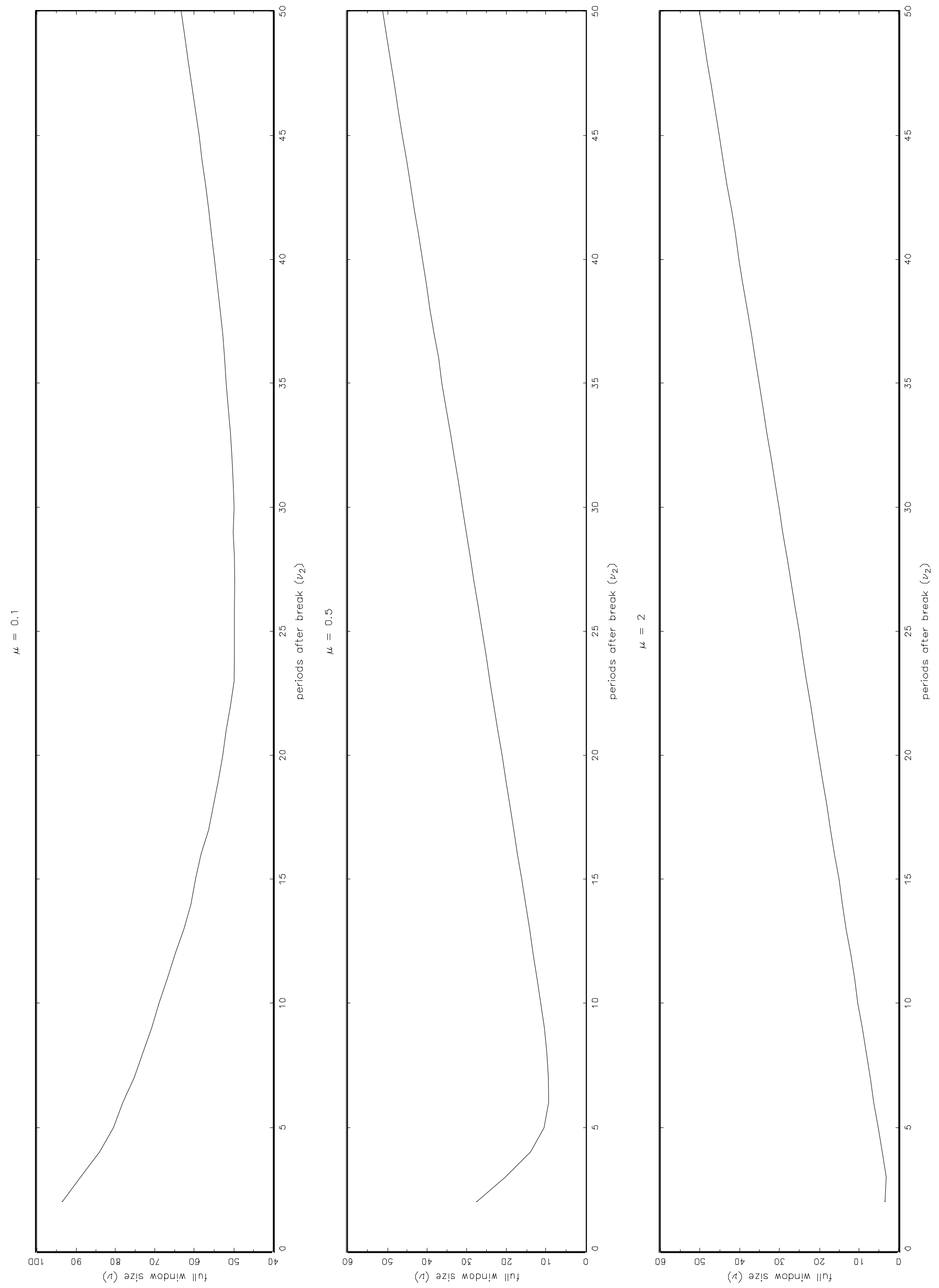

μ = 0.1
μ = 0.5
μ = 2
periods after break (ν_2)
periods after break (ν_2)
periods after break (ν_2)
full window size (ν)
full window size (ν)
full window size (ν)
0
5
10
15
20
25
30
35
40
45
50
40
50
60
70
80
90
100
0
10
20
30
40
50
60

Figure 6:

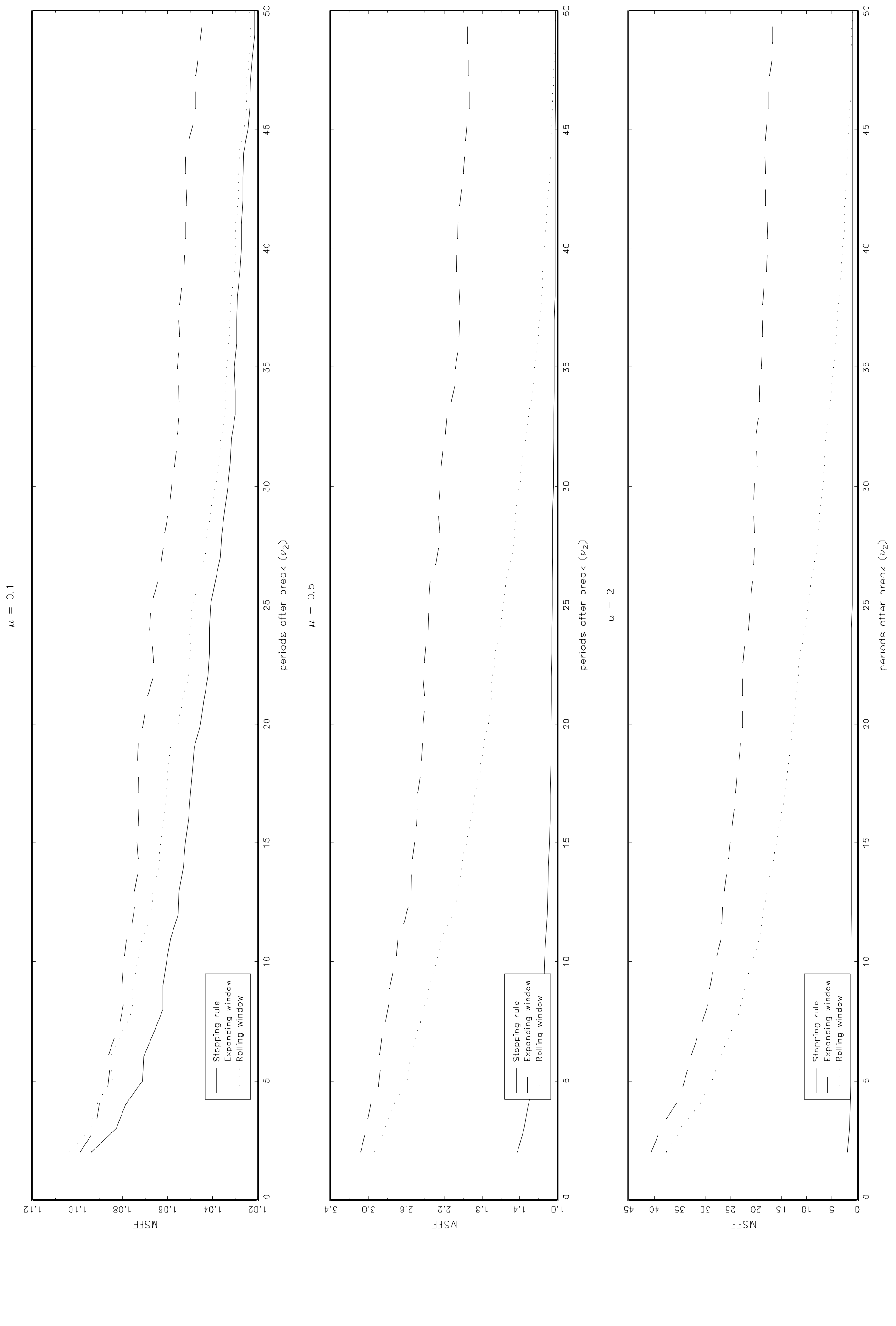
μ = 0.1
μ = 0.5
μ = 2
MSFE
periods after break (ν2)
Stopping rule
Expanding window
Rolling window
1.02
1.04
1.06
1.08
1.10
1.12
1.0
1.4
1.8
2.2
2.6
3.0
3.4
0
5
10
15
20
25
30
35
40
45
50

Figure 7:

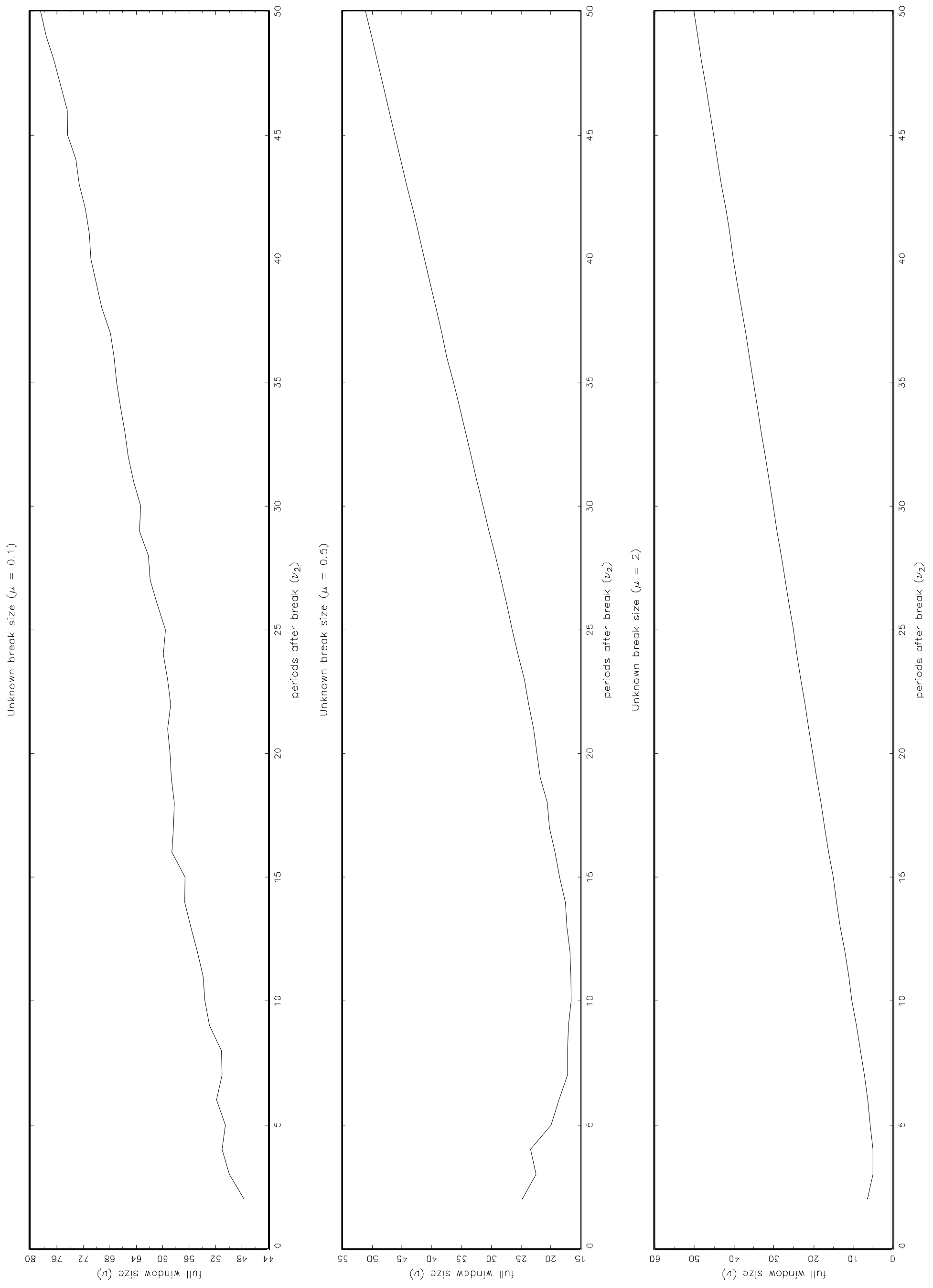

Unknown break size (μ = 0.1)
full window size (ν)
80
76
72
68
64
60
56
52
48
44
periods after break (ν_2)
0
5
10
15
20
25
30
35
40
45
50
Unknown break size (μ = 0.5)
full window size (ν)
55
50
45
40
35
30
25
20
15
periods after break (ν_2)
0
5
10
15
20
25
30
35
40
45
50
Unknown break size (μ = 2)
full window size (ν)
60
50
40
30
20
10
0
periods after break (ν_2)
0
5
10
15
20
25
30
35
40
45
50

Figure 8:

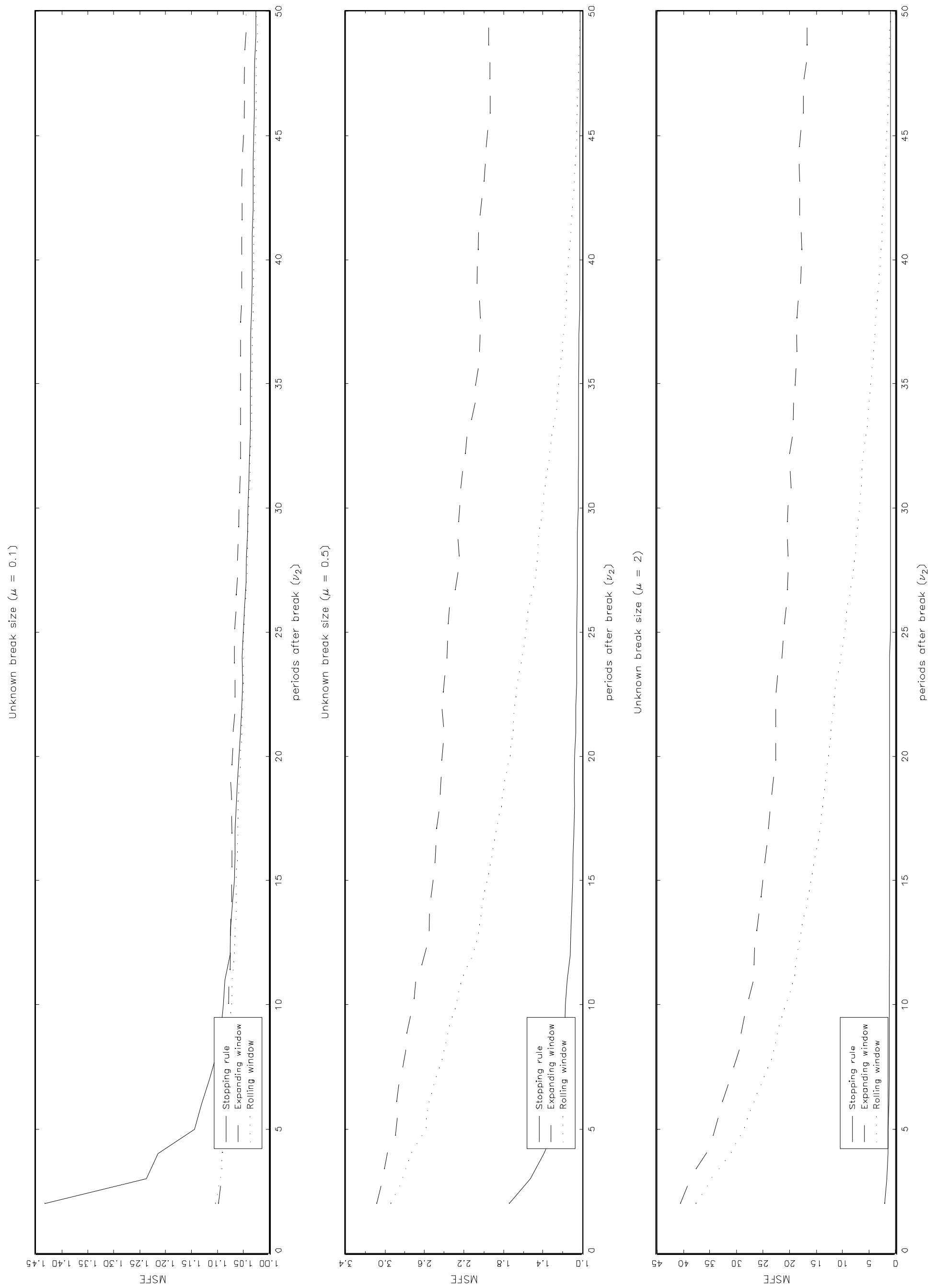

Unknown break size (μ = 0.1)
MSFE
1.00
1.05
1.10
1.15
1.20
1.25
1.30
1.35
1.40
1.45
periods after break (ν2)
0
5
10
15
20
25
30
35
40
45
50
Stopping rule
Expanding window
Rolling window
Unknown break size (μ = 0.5)
MSFE
1.0
1.4
1.8
2.2
2.6
3.0
3.4
periods after break (ν2)
0
5
10
15
20
25
30
35
40
45
50
Stopping rule
Expanding window
Rolling window
Unknown break size (μ = 2)
MSFE
0
5
10
15
20
25
30
35
40
45
periods after break (ν2)
0
5
10
15
20
25
30
35
40
45
50
Stopping rule
Expanding window
Rolling window

Figure 9:

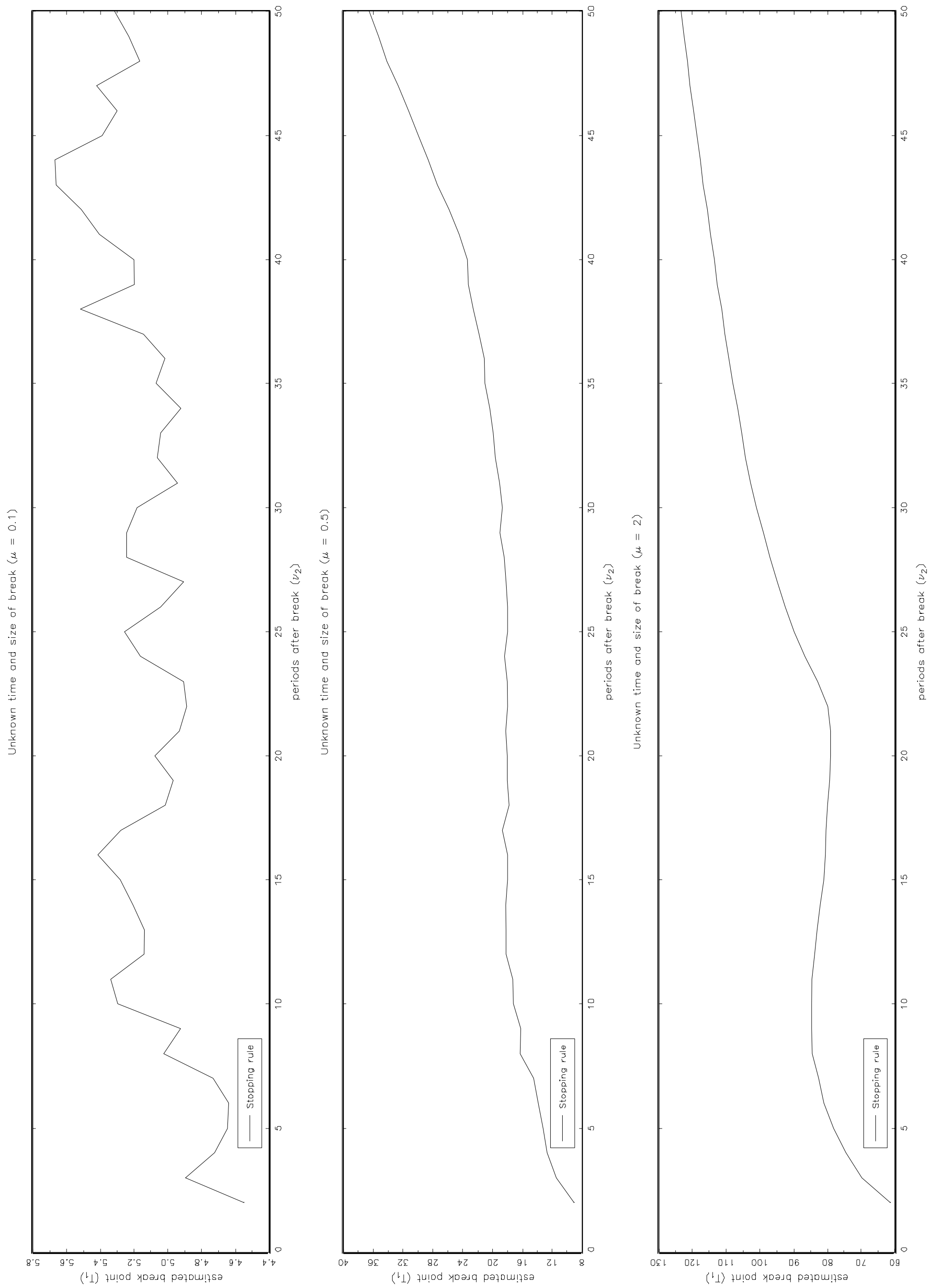

Unknown time and size of break (μ = 0.1)
Unknown time and size of break (μ = 0.5)
Unknown time and size of break (μ = 2)
periods after break (ν_2)
estimated break point (T_1)
Stopping rule
0
5
10
15
20
25
30
35
40
45
50
4.4
4.6
4.8
5.0
5.2
5.4
5.6
5.8
8
12
16
20
24
28
32
36
40
60
70
80
90
100
110
120
130

Figure 10:

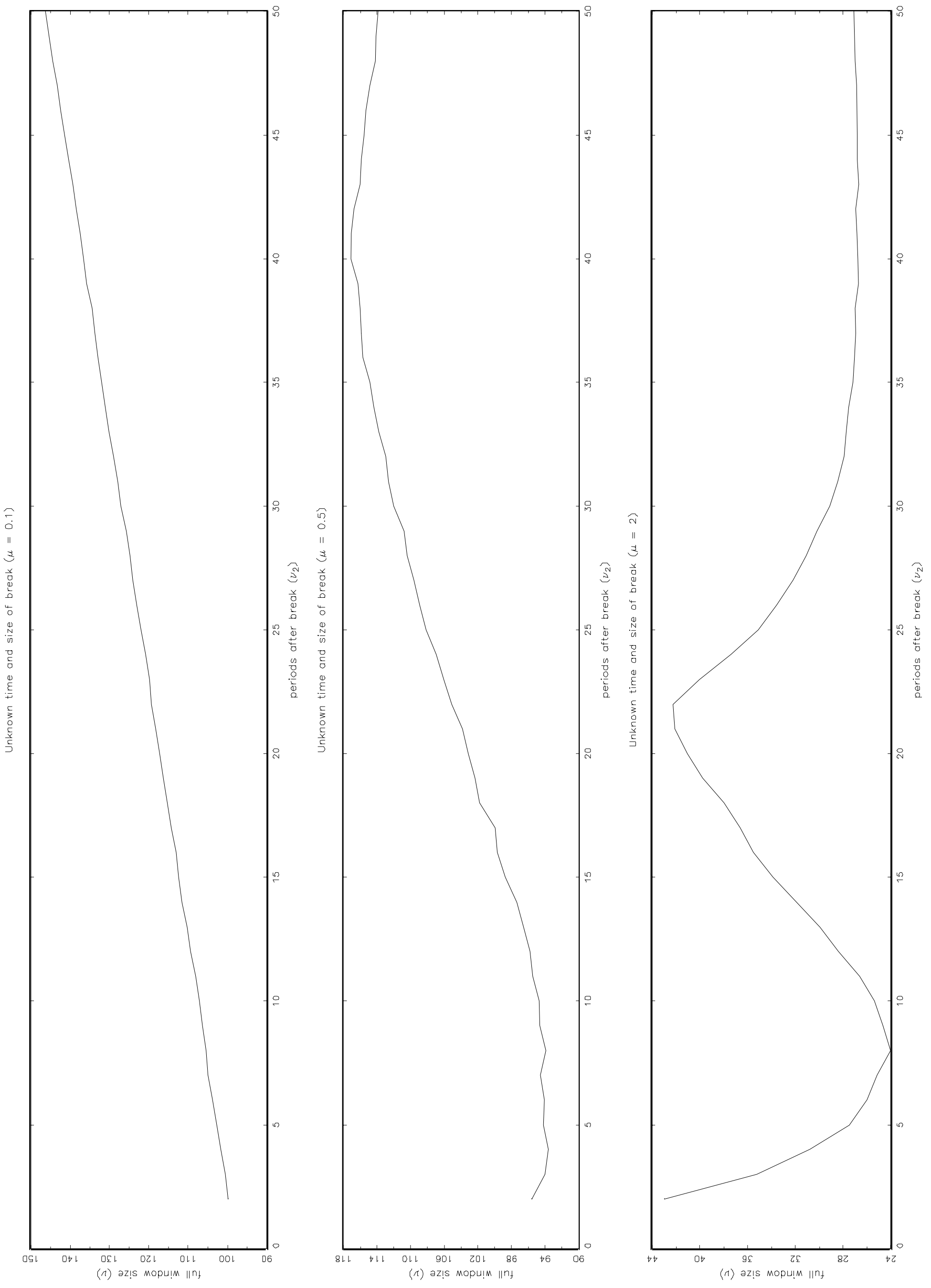

Unknown time and size of break (μ = 0.1)
periods after break (ν_2)
full window size (ν)
0
5
10
15
20
25
30
35
40
45
50
90
100
110
120
130
140
150
Unknown time and size of break (μ = 0.5)
periods after break (ν_2)
full window size (ν)
90
94
98
102
106
110
114
118
Unknown time and size of break (μ = 2)
periods after break (ν_2)
full window size (ν)
24
28
32
36
40
44

Figure 11:

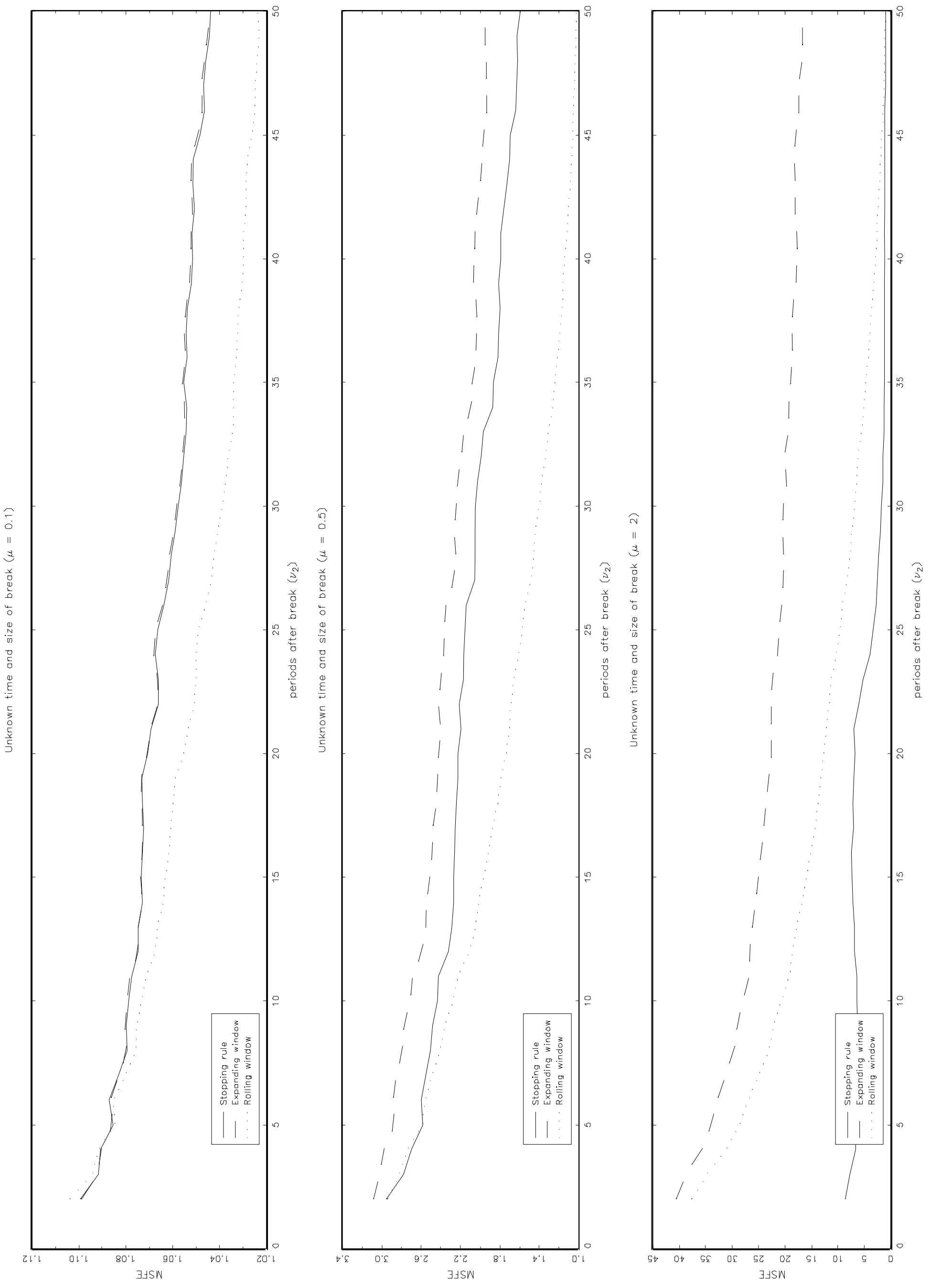

Unknown time and size of break (μ = 0.1)
Unknown time and size of break (μ = 0.5)
Unknown time and size of break (μ = 2)
MSFE
periods after break (ν2)
Stopping rule
Expanding window
Rolling window
1.02
1.04
1.06
1.08
1.10
1.12
1.0
1.4
1.8
2.2
2.6
3.0
3.4
0
5
10
15
20
25
30
35
40
45
50

Figure 12:

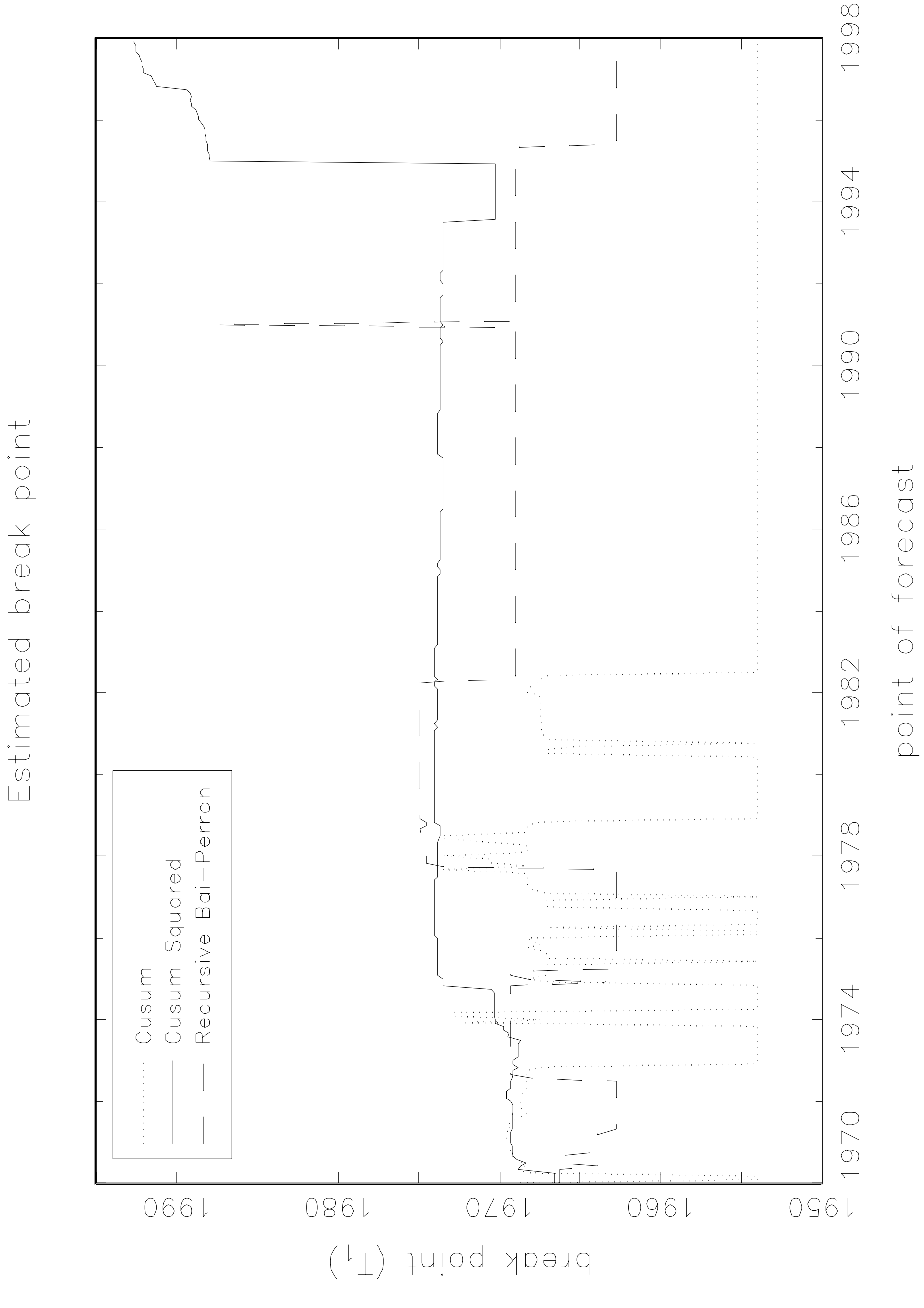

Estimated break point
Cusum
Cusum Squared
Recursive Bai-Perron
point of forecast
1970
1974
1978
1982
1986
1990
1994
1998
break point (T_1)
1950
1960
1970
1980
1990

Figure 13:

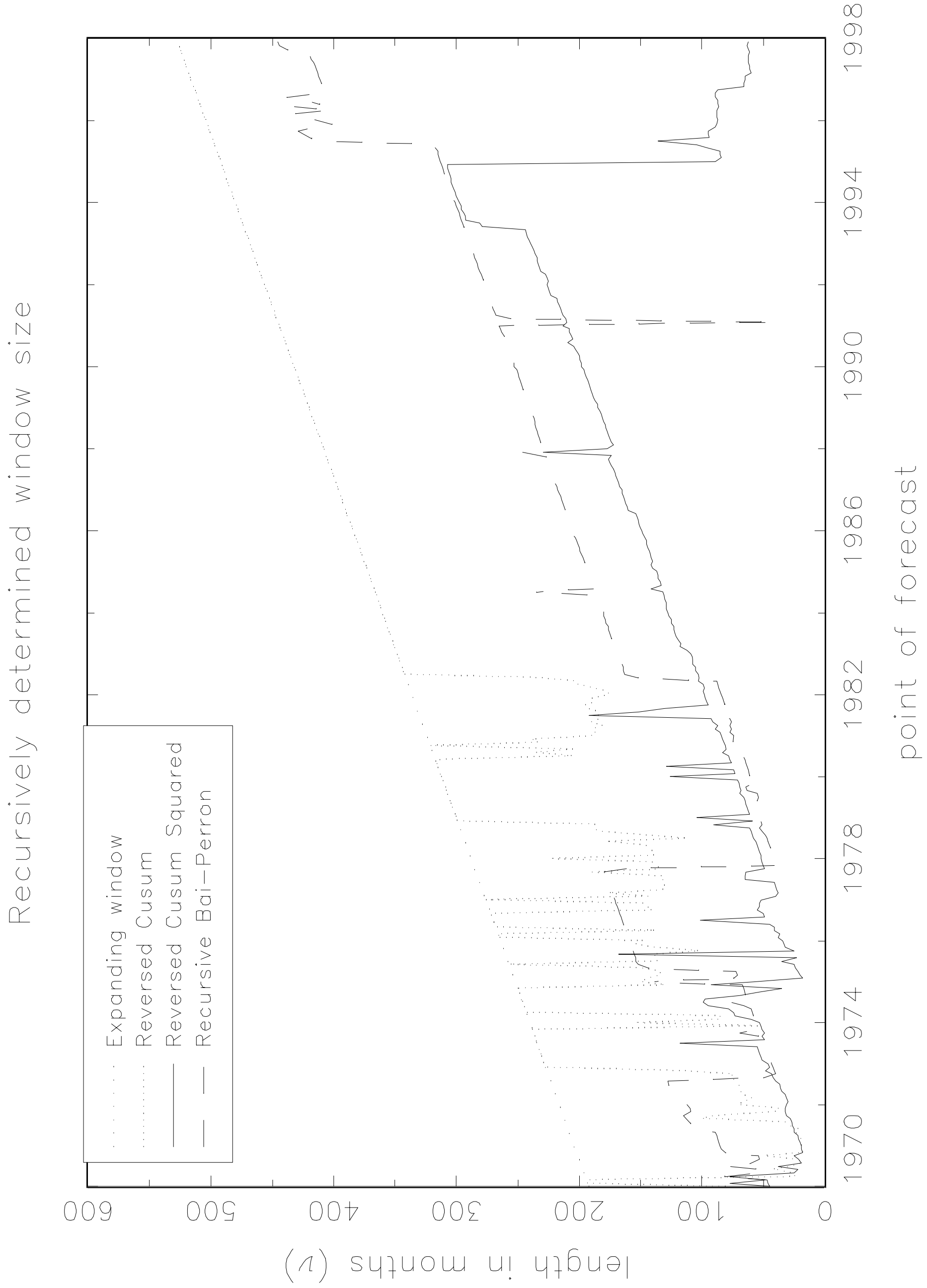

Recursively determined window size
Expanding window
Reversed Cusum
Reversed Cusum Squared
Recursive Bai-Perron
length in months (v)
0
100
200
300
400
500
600
point of forecast
1970
1974
1978
1982
1986
1990
1994
1998

Figure 14:

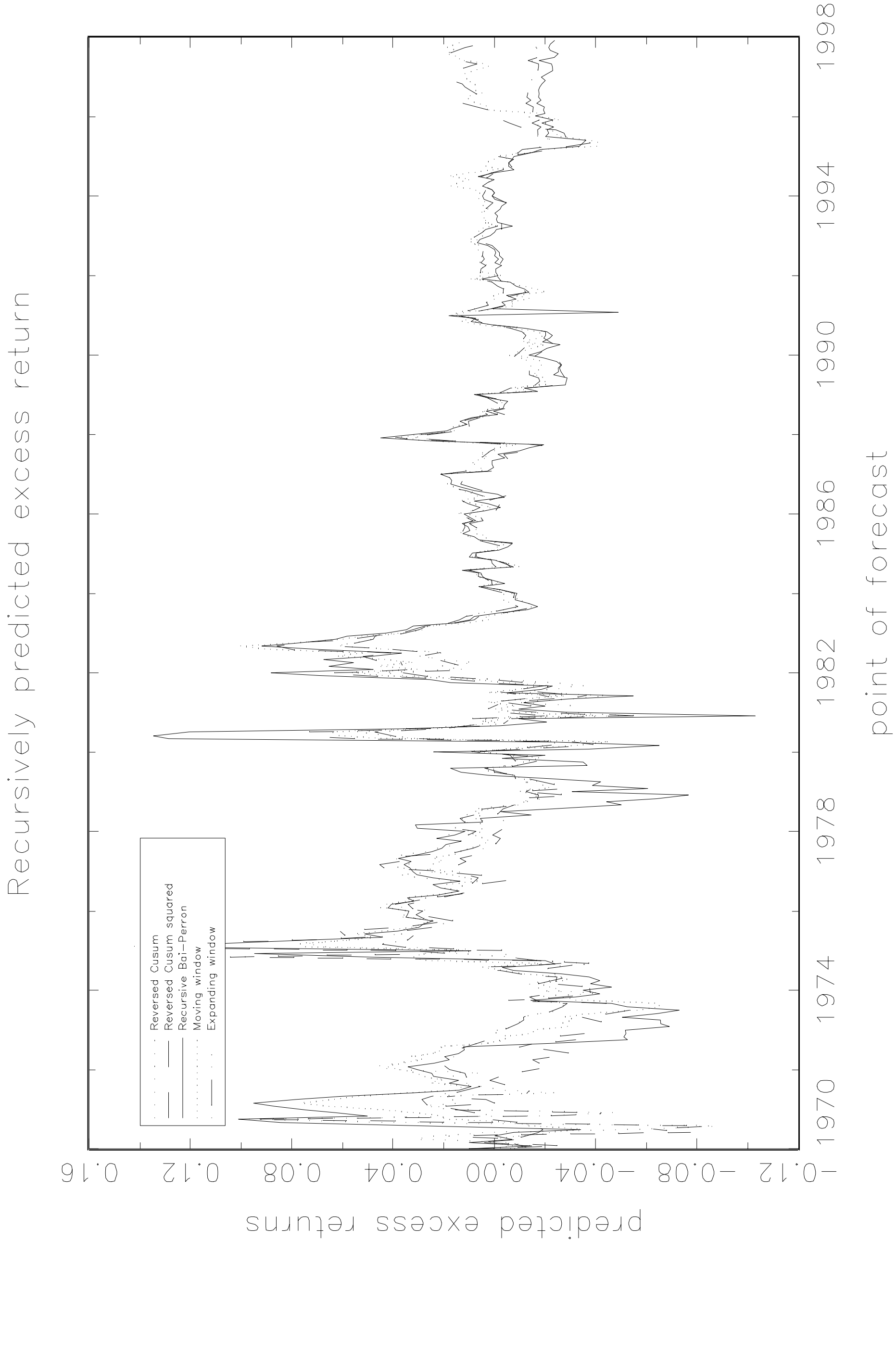

Recursively predicted excess return
Reversed Cusum
Reversed Cusum squared
Recursive Bai–Perron
Moving window
Expanding window
predicted excess returns
0.16
0.12
0.08
0.04
0.00
−0.04
−0.08
−0.12
point of forecast
1970
1974
1978
1982
1986
1990
1994
1998

Figure 15: