Econometric Theory, **25**, 2009, 1498–1514. Printed in the United States of America.
doi:10.1017/S0266466609990235

AVERAGING ESTIMATORS FOR REGRESSIONS WITH A POSSIBLE STRUCTURAL BREAK

BRUCE E. HANSEN
University of Wisconsin

This paper investigates selection and averaging of linear regressions with a possible structural break. Our main contribution is the construction of a Mallows criterion for the structural break model. We show that the correct penalty term is nonstandard and depends on unknown parameters, but it can be approximated by an average of limiting cases to yield a feasible penalty with good performance. Following Hansen (2007, *Econometrica* 75, 1175–1189) we recommend averaging the structural break estimates with the no-break estimates where the weight is selected to minimize the Mallows criterion. This estimator is simple to compute, as the weights are a simple function of the ratio of the penalty to the Andrews SupF test statistic.

To assess performance we focus on asymptotic mean-squared error (AMSE) in a local asymptotic framework. We show that the AMSE of the estimators depends exclusively on the parameter variation function. Numerical comparisons show that the unrestricted least-squares and pretest estimators have very large AMSE for certain regions of the parameter space, whereas our averaging estimator has AMSE close to the infeasible optimum.

1. INTRODUCTION

Structural change is an important issue in time series econometrics. Applied economists routinely test their models for the presence of structural change, typically using the Andrews (1993) and Andrews and Ploberger (1994) tests. Sometimes, when the evidence supports it, a structure break model is estimated. The breakdate may be estimated formally (as recommended by Bai, 1997) or may be selected informally, but the practical effects are rather similar. This means that applied econometricians may be de facto using a pretest estimator: using a restricted estimator (linear regression) when the structural change test is insignificant and using the unrestricted estimator (the structural change estimator) when the test is significant.

This practice is unfortunate because it is well known that pretest estimators generally have poor sampling properties. The squared error of pretest estimators is parameter-dependent and can be quite high relative to unrestricted estimation.

This research was supported by the National Science Foundation. I thank two referees for helpful comments. Address correspondence to Bruce E. Hansen. Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706, USA; e-mail: bhansen@ssc.wisc.edu.

  0266-4666/09 $15.00

This paper investigates the asymptotic performance of estimates of structural change regressions when a structural break is uncertain. We consider unrestricted least-squares estimation (as in Bai, 1997), estimation imposing the restriction of no-break, pretest estimation, selection estimators, and averaging estimators. Although the model to be estimated imposes a single structural break, we allow the truth to include general parameter variation (which includes a single structural break as a special case). We also assume that the magnitude of the parameter variation is inversely proportional to the square root of the sample size, so that the asymptotic distribution of the parameter estimates is continuous in the unknown functions. We assess the performance of the various estimators by asymptotic mean-squared error (AMSE) loss and focus on the estimate of the difference in regression slopes (the magnitude of structural change) as the leading parameter of interest.

In this framework, we find that both the unrestricted least-squares estimator and the pretest estimator can have quite large AMSE. The unrestricted estimator has large AMSE when the parameter variation magnitude is small, whereas the pretest estimator has large AMSE when the parameter magnitude is of moderate size.

Following Hansen (2007), we propose an averaging estimator with the weight selected to minimize a Mallows information criterion (Mallows, 1973). Averaging puts a weight w on the unrestricted estimator (Bai's least-squares estimator) and a weight $1 - w$ on the restricted estimator (the no-break regression). The Mallows criterion is constructed to be an approximately unbiased estimate of the in-sample fit. We show that in the structural break model the Mallows penalty takes a nonstandard form. Although the correct penalty depends on unknown parameters and is therefore infeasible, we show that approximating the penalty by an average of its limiting values (which are known) yields a feasible penalty with good performance. We recommend selecting the weight that minimizes the Mallows criterion, which in this case is a simple function of the sum of squared errors, the Andrews SupF test statistic, and the penalty term. The averaging estimator obtained using this weight is the Mallows model averaging (MMA) estimator.

We investigate the performance of the MMA estimator numerically, as the AMSE depends on the form of parameter variation. We show that when the true parameter variation is either a single structural break or a smooth structural break, the AMSE of the MMA estimator is very close to that of the infeasible optimal weighted average estimator. Based on the criterion of maximum regret, the MMA estimator has dramatically better performance than any of the other estimators considered.

The remainder of the paper is organized as follows. Section 2 introduces the structural change model and estimation methods. Section 3 presents the asymptotic distribution theory under the assumption of general parameter variation. Section 4 presents the Mallows criterion for the structural change model. Section 5 proposes selection and averaging estimators using the Mallows criterion. Section 6 presents formulas for the AMSE of the estimators. Section 7 numerically calculates these formulas and compares the estimators. Proofs of the theorems are presented in the Appendix.

2. MODEL AND ESTIMATION

The model to be estimated is a linear time-series regression with a possible structural break. The observations are $(y_t, \mathbf{x}_t)$ for $t = 1, \ldots, n$, where y_t is scalar and $\mathbf{x}_t$ is an m vector that may contain lagged values of y_t. The model for estimation is

$$y_t = \mathbf{x}_t'\boldsymbol{\beta}_1 1(t < k) + \mathbf{x}_t'\boldsymbol{\beta}_2 1(t \geq k) + e_t, \tag{1}$$

$$\mathrm{E}(e_t \mid \mathbf{x}_t) = 0,$$

$$\mathrm{E}\left(e_t^2 \mid \mathbf{x}_t\right) = \sigma^2,$$

which has parameters $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, k, \sigma^2)$. The breakdate k is constrained to satisfy the restriction $k_1 \leq k \leq k_2$. A parameter of interest is the difference in regression slopes $\boldsymbol{\theta} = \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$.

If there is no break in the slope coefficients then $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, and the model simplifies to

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + e_t, \tag{2}$$

where $\boldsymbol{\beta} = \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ and the breakdate k drops out.

When a structural break is uncertain it is common in applications to employ a two-step procedure where the first step is to test for the presence of a structural break and the second step is to estimate the model selected by the test. Let us describe this procedure in detail.

Estimation of (1) is easiest by concentration. First, fix k. Then equation (1) is estimated by least squares, which we write as

$$y_t = \mathbf{x}_t'\hat{\boldsymbol{\beta}}_1(k)1(t < k) + \mathbf{x}_t'\hat{\boldsymbol{\beta}}_2(k)1(t \geq k) + \hat{e}_t(k) \tag{3}$$

and set $\hat{\boldsymbol{\theta}}(k) = \hat{\boldsymbol{\beta}}_2(k) - \hat{\boldsymbol{\beta}}_1(k)$.

Let $\hat{\mathbf{e}}(k)$ denote the $n \times 1$ residuals from (3). The concentrated sum of squared errors given k is $\hat{\mathbf{e}}(k)'\hat{\mathbf{e}}(k)$. The least-squares estimate of k is found by numerically minimizing this criterion:

$$\hat{k} = \underset{k_1 \leq k \leq k_2}{\operatorname{argmin}}\, \hat{\mathbf{e}}(k)'\hat{\mathbf{e}}(k).$$

The remaining estimates are obtained using $\hat{k}$:

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1(\hat{k}),$$

$$\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_2(\hat{k}),$$

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}_2 - \hat{\boldsymbol{\beta}}_1.$$

We write the fitted model as

$$y_t = \mathbf{x}_t'\hat{\boldsymbol{\beta}}_1 1(t < \hat{k}) + \mathbf{x}_t'\hat{\boldsymbol{\beta}}_2 1(t \geq \hat{k}) + \hat{e}_t. \tag{4}$$

Let $\hat{\mathbf{e}} = \hat{\mathbf{e}}(\hat{k})$ denote the vector of fitted residuals.

The no-break model (2) is also estimated by least squares, which we write as

$$y_t = \mathbf{x}_t'\tilde{\boldsymbol{\beta}} + \tilde{e}_t. \tag{5}$$

As an estimator of the parameters in (1), we set $\tilde{\boldsymbol{\beta}}_1 = \tilde{\boldsymbol{\beta}}_2 = \tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\theta}} = \mathbf{0}$. Let $\tilde{\mathbf{e}}$ denote the $n \times 1$ vector of residuals from (5).

The standard test of model (2) against model (1) is the SupF test of Andrews (1993). The test statistic is the standard F-test

$$F_n = \frac{\left(\tilde{\mathbf{e}}'\tilde{\mathbf{e}} - \hat{\mathbf{e}}'\hat{\mathbf{e}}\right)}{s^2}, \tag{6}$$

where

$$s^2 = \frac{1}{n - 2m}\hat{\mathbf{e}}'\hat{\mathbf{e}} \tag{7}$$

is the bias-corrected estimator of the error variance from the full model (4).

Let $\pi = k/n$ denote the breakdate fraction, $\pi_1 = k_1/n$, and $\pi_2 = k_2/n$. Under the hypothesis $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$,

$$F_n \to_d \text{SupF} = \sup_{\pi_1 \leq \pi \leq \pi_2} J_0(\pi), \tag{8}$$

where

$$J_0(\pi) = \frac{\mathbf{W}^*(\pi)'\mathbf{W}^*(\pi)}{\pi(1-\pi)} \tag{9}$$

and $\mathbf{W}^*(\pi)$ is an m-dimensional standard Brownian bridge. An $\alpha\%$ asymptotic test rejects (2) in favor of (1) if $F_n > c_\alpha$ where c_α is the $(1-\alpha)\%$ upper quantile of the distribution of SupF. Critical values are tabulated in Andrews (2003) and depend on m and

$$\lambda = \pi_2(1 - \pi_1)/(1 - \pi_2)\pi_1. \tag{10}$$

For example, if $m = 1$, $\pi_1 = 0.15$, and $\pi_2 = 0.85$, then $c_{.05} = 8.68$.

As we mentioned previously, a conventional estimator of the model is to use the unrestricted estimator (4) when F_n is significant and otherwise to use the restricted estimator (5). We will call this the pretest estimator, and it can be written as

$$\begin{aligned}\hat{\boldsymbol{\theta}}^p &= \hat{\boldsymbol{\theta}}1(F_n \geq c_\alpha) + \tilde{\boldsymbol{\theta}}1(F_n < c_\alpha)\\ &= \hat{\boldsymbol{\theta}}1(F_n \geq c_\alpha).\end{aligned}$$

The motivation for the pretest estimator is fairly straightforward. A reasonable presumption is that unless there is evidence to the contrary we should use a standard linear regression model. We should use a breakdate estimator only if there is evidence of a structural break. As the most compelling evidence is a statistical test, this leads to two-step (or pretest) estimation.

3. ASYMPTOTIC DISTRIBUTION UNDER GENERAL PARAMETER VARIATION

In this section we review the asymptotic distribution theory for the parameter estimates allowing for general parameter variation (thus allowing for the possibility that the single structural break assumption in (1) is misspecified). We also assume that the parameter variation is of small magnitude so that the asymptotic distributions are asymptotically continuous. We believe that this is the appropriate framework in which to study model selection.

To be specific, we assume that the data satisfy the regression

$$
\begin{aligned}
y_t &= \mathbf{x}_t'\boldsymbol{\beta}_t + e_t, \\
\mathrm{E}(e_t \mid \mathbf{x}_t) &= 0, \\
\mathrm{E}\left(e_t^2 \mid \mathbf{x}_t\right) &= \sigma^2, \\
\boldsymbol{\beta}_t &= \boldsymbol{\beta} + n^{-1/2}\boldsymbol{\eta}(t/n)\delta\sigma,
\end{aligned}
\tag{11}
$$

where $\boldsymbol{\eta}(\cdot)$ is a bounded $\mathbb{R}^m$-valued Riemann integrable function on [0, 1] and δ is a scalar indexing the magnitude of parameter variation. This general specification includes single and multiple structural change as special cases.

Andrews (1993) derived the asymptotic distribution of the parameter estimates and test statistics under these assumptions. Define $\mathbf{M} = \mathrm{E}\left(\mathbf{x}_t\mathbf{x}_t'\right)$ and the functions

$$
\begin{aligned}
\bar{\boldsymbol{\eta}}_1(\pi) &= \frac{1}{\pi}\int_0^\pi \boldsymbol{\eta}(s)\,ds, \\
\bar{\boldsymbol{\eta}}_2(\pi) &= \frac{1}{1-\pi}\int_\pi^1 \boldsymbol{\eta}(s)\,ds, \\
\mathbf{g}(\pi) &= \mathbf{M}^{1/2}\left(\bar{\boldsymbol{\eta}}_2(\pi) - \bar{\boldsymbol{\eta}}_1(\pi)\right).
\end{aligned}
\tag{12}
$$

THEOREM 1 (Andrews, 1993). *Under model (11), as $n \to \infty$*

$$
\begin{aligned}
\frac{\sqrt{n}}{\sigma}\left(\hat{\boldsymbol{\beta}}_1(n\pi) - \boldsymbol{\beta}\right) &\xrightarrow{d} \mathbf{M}^{-1/2}\pi^{-1}\mathbf{W}(\pi) + \bar{\boldsymbol{\eta}}_1(\pi)\delta, \\
\frac{\sqrt{n}}{\sigma}\left(\hat{\boldsymbol{\beta}}_2(n\pi) - \boldsymbol{\beta}\right) &\xrightarrow{d} \mathbf{M}^{-1/2}(1-\pi)^{-1}(\mathbf{W}(1) - \mathbf{W}(\pi)) + \bar{\boldsymbol{\eta}}_2(\pi)\delta, \\
\frac{\sqrt{n}}{\sigma}\hat{\boldsymbol{\theta}}(n\pi) &\xrightarrow{d} \mathbf{M}^{-1/2}\mathbf{S}_\delta(\pi),
\end{aligned}
$$

$$\frac{\hat{k}}{n} \xrightarrow{d} \xi_\delta = \underset{\pi_1 \le \pi \le \pi_2}{\arg\max}\, J_\delta(\pi), \tag{13}$$

$$F_n \xrightarrow{d} \mathrm{SupF}_\delta = \sup_{\pi_1 \le \pi \le \pi_2} J_\delta(\pi), \tag{14}$$

where

$$J_\delta(\pi) = \pi(1-\pi)\mathbf{S}_\delta(\pi)'\mathbf{S}_\delta(\pi), \tag{15}$$

$$\mathbf{S}_\delta(\pi) = -\frac{\mathbf{W}^*(\pi)}{\pi(1-\pi)} + \mathbf{g}(\pi)\delta, \tag{16}$$

$\mathbf{W}^*(\pi) = \mathbf{W}(\pi) - \pi\mathbf{W}(1)$, *and* $\mathbf{W}(\pi)$ *is an m-dimensional standard Brownian motion.*

Under the general parameter variation assumption (11) there is not necessarily a true "structural break," but we can define the pseudo-true breakdate. Let

$$\pi_0 = \underset{\pi_1 \le \pi \le \pi_2}{\arg\max}\, \pi(1-\pi)\mathbf{g}(\pi)'\mathbf{g}(\pi), \tag{17}$$

which we assume is unique. We call π_0 the pseudo-true breakdate fraction and $[n\pi_0]$ the pseudo-true breakdate. These correspond to the true breakdate when (11) equals the structural change model (1). Given π_0 we can define the pseudo-true value for $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_0 = n^{-1/2}\mathbf{M}^{-1/2}\mathbf{g}(\pi_0)\delta\sigma.$$

This is the best fitting value of $\boldsymbol{\theta}$ for the structural change model (1) when the true parameter variation takes the form (11).

The limiting distribution of the breakdate estimator ξ_δ defined in (13) is random even in large samples. However, as the degree of parameter variation becomes stronger, the distribution of ξ_δ collapses to a point mass.

THEOREM 2. *Under model (11), if* π_0 *defined in (17) is unique then*

$$\operatorname*{plim}_{\delta\to\infty} \xi_\delta \equiv \xi_\infty = \pi_0. \tag{18}$$

4. MALLOWS CRITERION

In this section we develop Mallows criteria appropriate for the regressions (3) and (4). The Mallows criterion is a penalized sum of squared residuals designed to be approximately unbiased for the in-sample fit. The general approach is as follows. Write the model (11) in vector notation as $\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}$ where $\boldsymbol{\mu}$ is the regression function. Let $\hat{\boldsymbol{\mu}} = \mathbf{P}\mathbf{y}$ be an estimator of $\boldsymbol{\mu}$ with residual vector $\hat{\mathbf{e}} = \mathbf{y} - \hat{\boldsymbol{\mu}}$.

A measure of in-sample fit is $\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)'\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)$, and the goal is to estimate this quantity by the sum of squared errors $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ plus a penalty. By expanding the square,

$$\begin{aligned}\hat{\mathbf{e}}'\hat{\mathbf{e}} &= \left(\mathbf{e}+\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)'\left(\mathbf{e}+\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)\\ &= \left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)'\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)+\mathbf{e}'\mathbf{e}+2\mathbf{e}'\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)\\ &= \left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)'\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)+\mathbf{e}'\mathbf{e}+2\mathbf{e}'\left(\mathbf{I}-\mathbf{P}\right)\boldsymbol{\mu}-2\mathbf{e}'\mathbf{P}\mathbf{e}.\end{aligned}$$

Thus the sum of squared residuals $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ equals the in-sample fit $\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)'\left(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}\right)$ plus three terms. The first term $\mathbf{e}'\mathbf{e}$ is independent of the estimation method and therefore does not matter. The second term $2\mathbf{e}'\left(\mathbf{I}-\mathbf{P}\right)\boldsymbol{\mu}$ has an approximate mean of zero and is therefore also ignored. The final term $2\mathbf{e}'\mathbf{P}\mathbf{e}$ has a nonzero mean that is the traditional focus of attention. The Mallows criterion takes the general form

$$C=\hat{\mathbf{e}}'\hat{\mathbf{e}}+2s^2p,$$

where the penalty p is an estimate of the expectation of the asymptotic distribution of $\sigma^{-2}\mathbf{e}'\mathbf{P}\mathbf{e}$ and s^2 defined in (7) is an estimator of σ^2.

In the case of the restricted least-squares estimator (5) for model (2), $\mathbf{P}=\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ where $\mathbf{X}$ is the regressor matrix. Under the conditions of Theorem 1, $\sigma^{-2}\mathbf{e}'\mathbf{P}\mathbf{e}\rightarrow_d \chi^2_m$ and $\mathrm{E}\left(\chi^2_m\right)=m$. It follows that the Mallows criterion for this model takes the classic form

$$\tilde{C}=\tilde{\mathbf{e}}'\tilde{\mathbf{e}}+2s^2m. \tag{19}$$

If the breakdate k were known then a similar argument could be applied to the estimates (3), and we would obtain the Mallows criterion

$$\hat{C}(k)=\hat{\mathbf{e}}(k)'\hat{\mathbf{e}}(k)+4s^2m.$$

However the case with unknown breakdate is nonstandard. For fixed k we can write $\hat{\boldsymbol{\mu}}(k)=\mathbf{P}(k)\mathbf{y}$ where $\mathbf{P}(k)$ is the projection matrix onto the space of regressors $\mathbf{x}_t 1(t<k)$ and $\mathbf{x}_t 1(t\geq k)$. With estimated $\hat{k}$ we have $\hat{\boldsymbol{\mu}}=\hat{\boldsymbol{\mu}}(\hat{k})=\mathbf{P}(\hat{k})\mathbf{y}$, and thus the desired penalty is twice the expectation of the asymptotic distribution of $\sigma^{-2}\mathbf{e}'\mathbf{P}(\hat{k})\mathbf{e}$.

We calculate the asymptotic distribution of $\sigma^{-2}\mathbf{e}'\mathbf{P}(\hat{k})\mathbf{e}$ when the parameter variation takes the form (11). This is our main result.

THEOREM 3. *Under model (11), as $n\rightarrow\infty$,*

$$\sigma^{-2}\mathbf{e}'\mathbf{P}(\hat{k})\mathbf{e}\overset{d}{\longrightarrow}\chi^2_m+J_0(\xi_\delta), \tag{20}$$

where $J_0(\cdot)$ and ξ_δ are defined in (9) and (13).

The theorem shows that correct Mallows penalty is $2\sigma^2$ times $\mathrm{E}\left(\chi_m^2 + J_0(\xi_\delta)\right) = m + p_\delta$ where $p_\delta = \mathrm{E}\left(J_0(\xi_\delta)\right)$ depends on unknowns. However, in the limiting cases $\delta = 0$ and $\delta \to \infty$ it simplifies. When $\delta = 0$, $J_0(\xi_0) = \mathrm{SupF}$ as defined in (8), and so

$$p_0 = \mathrm{E}\left(J_0(\xi_0)\right) = \mathrm{E}(\mathrm{SupF}),$$

which is a function only of m and λ. As $\delta \to \infty$, Theorem 3 implies that $J_0(\xi_\delta) \to_d J_0(\xi_\infty) = J_0(\pi_0) \sim \chi_m^2$, and so

$$p_\infty = \mathrm{E}\left(J_0(\xi_\infty)\right) = m.$$

As a practical solution, we recommend approximating p_δ by an average of these limiting cases:

$$\bar{p} = \frac{1}{2}\left(p_0 + p_\infty\right) = \frac{1}{2}(\mathrm{E}(\mathrm{SupF}) + m). \tag{21}$$

Using this approximation we obtain a practical Mallows criterion for the structural change model.

PROPOSITION 1. *A Mallows criterion for the structural change model is*

$$\hat{C} = \hat{\mathbf{e}}'\hat{\mathbf{e}} + 2s^2\left(m + p_\delta\right),$$

where $p_\delta = \mathrm{E}\left(J_0(\xi_\delta)\right)$*, which is infeasible. An approximate Mallows criterion is*

$$\hat{C}^* = \hat{\mathbf{e}}'\hat{\mathbf{e}} + 2s^2\left(m + \bar{p}\right), \tag{22}$$

where $\bar{p}$ *defined in (21) depends only on* m *and* λ.

The penalty coefficients $\bar{p}$ are displayed in Table 1 as a function of m and the trimming parameter π_1.

5. SELECTION AND AVERAGING

Model selection based on the Mallows criterion (22) picks the structural change estimates (4) if $\hat{C}^* < \tilde{C}$ (equivalently if $F_n \geq 2\bar{p}$) and picks the restricted estimates (5) if $\tilde{C} < \hat{C}^*$ (equivalently if $F_n < 2\bar{p}$). This is similar to the pretest estimator but replaces the critical value c_α from Andrews's table with the value $2\bar{p}$. The Mallow selection estimator for $\boldsymbol{\theta}$ is thus

$$\begin{aligned}\hat{\boldsymbol{\theta}}^m &= \hat{\boldsymbol{\theta}}1\left(F_n \geq 2\bar{p}\right) + \tilde{\boldsymbol{\theta}}1\left(F_n < 2\bar{p}\right)\\ &= \hat{\boldsymbol{\theta}}1\left(F_n \geq 2\bar{p}\right).\end{aligned}$$

In a recent paper, Hansen (2007) has argued for averaging based on Mallows weights rather than selection. Averaging assigns a weight of w to model (1) and

TABLE 1. Penalty coefficients $\bar{p}$

	$\pi_1 = 1 - \pi_2$					
m	0.01	0.05	0.10	0.15	0.20	0.25
1	3.28	2.90	2.67	2.49	2.33	2.19
2	5.01	4.56	4.27	4.05	3.85	3.66
3	6.56	6.05	5.73	5.47	5.24	5.01
4	8.00	7.45	7.09	6.80	6.55	6.29
5	9.38	8.78	8.40	8.09	7.82	7.55
6	10.7	10.1	9.68	9.36	9.06	8.77
7	12.0	11.4	10.9	10.6	10.3	9.96
8	13.3	12.6	12.2	11.8	11.5	11.2
9	14.6	13.9	13.4	13.0	12.7	12.4
10	15.8	15.1	14.6	14.2	13.9	13.5
11	17.1	16.3	15.8	15.4	15.0	14.7
12	18.3	17.5	17.0	16.6	16.2	15.8
13	19.6	18.8	18.2	17.8	17.4	17.0
14	20.8	19.9	19.4	18.9	18.5	18.1
15	21.9	21.1	20.5	20.0	19.6	19.2
16	23.2	22.3	21.7	21.2	20.8	20.4
17	24.3	23.5	22.8	22.4	21.9	21.5
18	25.5	24.6	24.0	23.5	23.1	22.6
19	26.7	25.8	25.1	24.6	24.2	23.7
20	27.9	26.9	26.3	25.8	25.3	24.8

Note: $\lambda = (1-\pi_1)^2/\pi_1^2$. The coefficients were calculated by simulation, approximating the distribution of SupF in (8) taking the average of 200,000 random Gaussian samples of size 10,000.

a weight of $1-w$ to model (2). An averaging estimator of θ given the weight w is

$$\hat{\boldsymbol{\theta}}^{w} = w\hat{\boldsymbol{\theta}} + (1-w)\tilde{\boldsymbol{\theta}} = w\hat{\boldsymbol{\theta}}. \tag{23}$$

The Mallows criterion for the weighted average is

$$C(w) = \left(\hat{\mathbf{e}}w + \tilde{\mathbf{e}}(1-w)\right)'\left(\hat{\mathbf{e}}w + \tilde{\mathbf{e}}(1-w)\right) + 2s^2(m + \bar{p}w).$$

The Mallows weight is the value in $[0,1]$ that minimizes $C(w)$. The solution is

$$\hat{w} = \begin{cases} 0 & \text{if } F_n < \bar{p} \\ 1 - \dfrac{\bar{p}}{F_n} & \text{if } F_n \geq \bar{p}. \end{cases} \tag{24}$$

Viewed as a function of the test statistic F_n, the weight $\hat{w}$ is a smoothed version of the Mallows selection criterion.

Hansen's MMA estimates of the model parameters are the weighted averages using the weight $\hat{w}$:

$$\hat{\boldsymbol{\theta}}^{MMA} = \hat{\boldsymbol{\theta}}\left(1 - \frac{\bar{p}}{F_n}\right)1(F_n \geq \bar{p}).$$

6. ASYMPTOTIC MEAN-SQUARED ERROR

We investigate the performance of the different estimates of θ by focusing on the AMSE. For an estimator $\boldsymbol{\theta}^*$ of $\boldsymbol{\theta}_0$, we define the AMSE as

$$AMSE(\boldsymbol{\theta}^*) = \lim_{n\to\infty} \frac{n}{\sigma^2} \mathrm{Etr}\left(\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\right)\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\right)' \mathbf{M}\right), \tag{25}$$

where $\mathbf{M} = \mathrm{E}(\mathbf{x}_t\mathbf{x}_t')$ is used as a weighting matrix to reduce the dependence on nuisance parameters. Set $\mathbf{g}_\delta(\pi) = \mathbf{g}(\pi)\delta$ and $\mathbf{g}_\delta = \mathbf{g}_\delta(\pi_0)$.

THEOREM 4. *In model (11), the AMSE of the restricted, unrestricted, pretest, Mallows selection, and Mallows averaging estimators $\tilde{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}^p$, $\hat{\boldsymbol{\theta}}^m$, and $\hat{\boldsymbol{\theta}}^{MMA}$ are*

$$AMSE(\tilde{\boldsymbol{\theta}}) = \mathbf{g}_\delta'\mathbf{g}_\delta,$$

$$AMSE(\hat{\boldsymbol{\theta}}) = \mathrm{E}\left((\mathbf{S}_\delta(\xi_\delta) - \mathbf{g}_\delta)'(\mathbf{S}_\delta(\xi_\delta) - \mathbf{g}_\delta)\right),$$

$$AMSE\left(\hat{\boldsymbol{\theta}}^p\right) = \mathrm{E}\left(\left(\mathbf{S}_\delta^p - \mathbf{g}_\delta\right)'\left(\mathbf{S}_\delta^p - \mathbf{g}_\delta\right)\right),$$

$$AMSE(\hat{\boldsymbol{\theta}}^m) = \mathrm{E}\left(\left(\mathbf{S}_\delta^m - \mathbf{g}_\delta\right)'\left(\mathbf{S}_\delta^m - \mathbf{g}_\delta\right)\right),$$

$$AMSE(\hat{\boldsymbol{\theta}}^{MMA}) = \mathrm{E}\left(\left(\mathbf{S}_\delta^{MMA} - \mathbf{g}_\delta\right)'\left(\mathbf{S}_\delta^{MMA} - \mathbf{g}_\delta\right)\right),$$

where ξ_δ and $\mathbf{S}_\delta(\cdot)$ are defined in (13) and (16),

$$\mathbf{S}_\delta^p = \mathbf{S}_\delta(\xi_\delta) \cdot 1\left(\mathrm{SupF}_\delta > c_\alpha\right),$$

$$\mathbf{S}_\delta^m = \mathbf{S}_\delta(\xi_\delta) \cdot 1\left(\mathrm{SupF}_\delta > 2\bar{p}\right),$$

$$\mathbf{S}_\delta^{MMA} = \mathbf{S}_\delta(\xi_\delta) \cdot 1\left(\mathrm{SupF}_\delta > \bar{p}\right)\left(1 - \frac{\bar{p}}{\mathrm{SupF}_\delta}\right),$$

and SupF_δ *is defined in (14).*

One useful feature of these representations is that they are free of dependence on unknowns other than the function $\mathbf{g}_\delta(\cdot)$. (Recall that the pseudo-true breakdate fraction π_0 and the distribution of ξ_δ are determined by $\mathbf{g}_\delta$.)

Another useful feature is that we can compare the AMSE with that of the infeasible optimal weighted average.

THEOREM 5. *The AMSE of the optimal weighted average estimator $\hat{\boldsymbol{\theta}}^w$ (eqn. (23)) of $\boldsymbol{\theta}$ is*

$$AMSE(\hat{\boldsymbol{\theta}}^w) = \mathbf{g}_\delta'\mathbf{g}_\delta - \mathbf{g}_\delta'\mathrm{E}\mathbf{S}_\delta(\xi_\delta)\left(\mathrm{E}\left(\mathbf{S}_\delta(\xi_\delta)'\mathbf{S}_\delta(\xi_\delta)\right)\right)^{-1}\mathrm{E}\mathbf{S}_\delta(\xi_\delta)'\mathbf{g}_\delta. \tag{26}$$

The AMSE given in (26) is infeasible but gives a lower bound on the AMSE for estimators that take the weighted average form (which includes pretest and selection estimators).

7. NUMERICAL COMPARISON

We now illustrate the asymptotic efficiency gains achievable by Mallows averaging through numerical calculation of the AMSE. We compare the unrestricted, pretest, Mallows selection, Mallows averaging, and infeasible optimal averaging estimators of $\boldsymbol{\theta}$ using the formula[1] from Theorems 4 and 5 of the previous section. Normalizing $\mathbf{M} = \mathbf{I}_m$ the AMSE is completely determined by the parameter variation function $\mathbf{g}(\cdot)$ defined in (12).

We first consider the case of single structural change. We set

$$\boldsymbol{\eta}(\pi) = \boldsymbol{\psi} 1(\pi \geq \pi_0)$$

so that the (scaled) parameter vector jumps by the magnitude $\boldsymbol{\psi}$ at the break fraction π_0.

The asymptotic AMSEs of our estimators are fully determined by m (the number of regressors), π_0, and $\psi = \|\boldsymbol{\psi}\|$. We plot the AMSE in Figures 1, 2, and 3 for $m = 1$, 5, and 10, respectively. Each figure has four plots, for $\pi_0 = 0.2$, 0.3, 0.4, and 0.5. (The results are symmetric for $\pi_0 > 0.5$.) In each plot, ψ is varied on the x-axis and AMSE (eqn. (25)) shown on the y-axis. The AMSE of the unrestricted

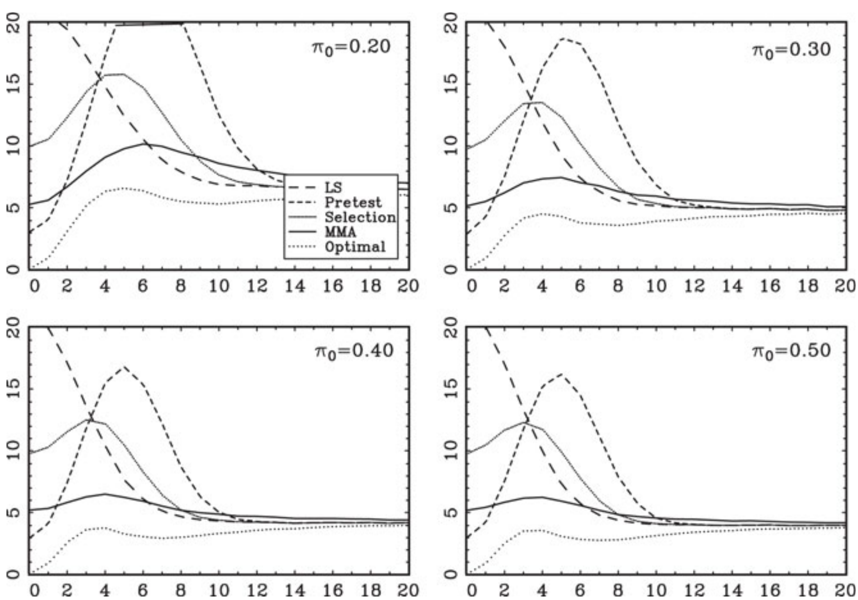

FIGURE 1. Pure structural break, $m = 1$.

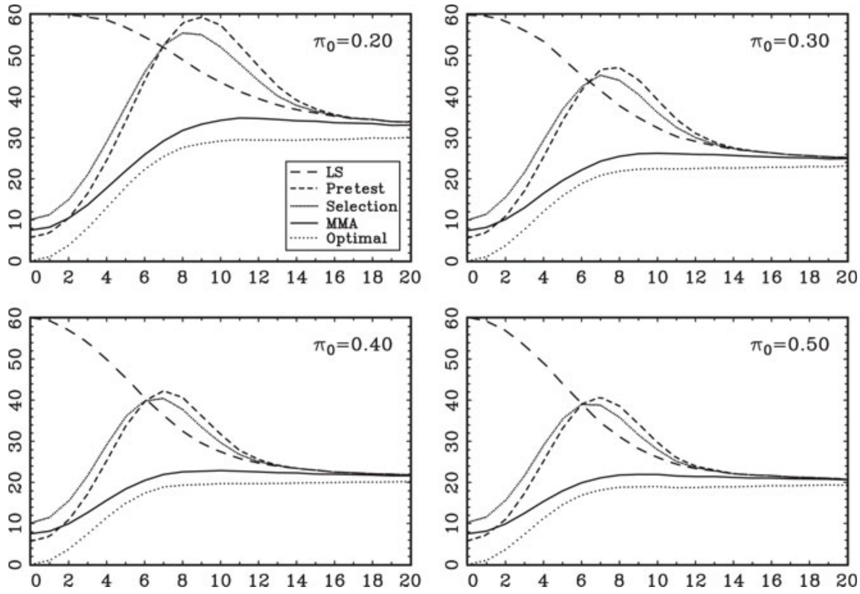

FIGURE 2. Pure structural break, $m = 5$.

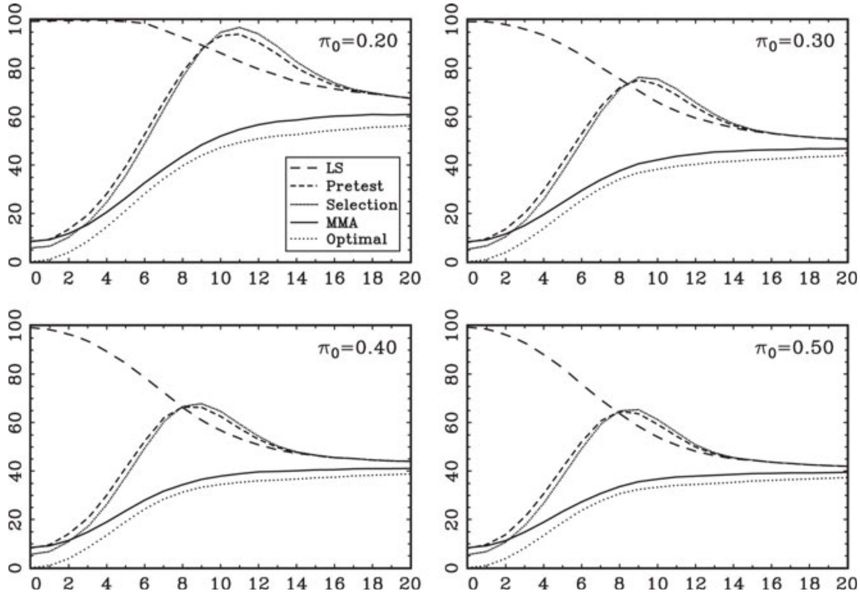

FIGURE 3. Pure structural break, $m = 10$.

TABLE 2. Maximum regret

m	Unrestricted LS	Pretest	Mallows selection	Mallows averaging
A. Pure structural change				
1	21.1	17.6	9.9	5.3
5	60.2	30.8	27.9	7.5
10	99.4	46.2	47.6	8.4
B. Smooth structural change				
1	21.1	16.7	10.5	5.2
5	60.1	31.2	29.1	7.5
10	99.5	47.9	48.7	8.4

least-squares estimator is shown with the long dashes, the pretest estimator with the short dashes, the Mallows selection estimator with the closely spaced dots, the MMA estimator with the solid line, and the infeasible optimal weighted average estimator with the lowest dotted line.

Looking across the plots, we can see that the least-squares and pretest estimators have very large values of AMSE for certain values of the parameters. The least-squares estimator has high AMSE for small ψ, whereas the pretest estimator has high AMSE for moderate values of ψ. For $m = 1$ the Mallows selection estimator is less severely affected by the parameters than the pretest estimator, but

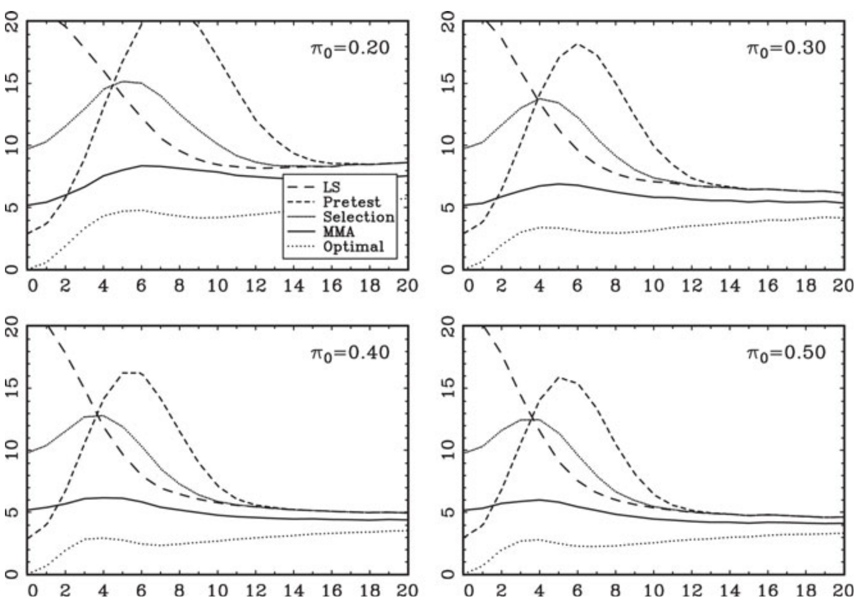

FIGURE 4. Smooth structural break, $m = 1$.

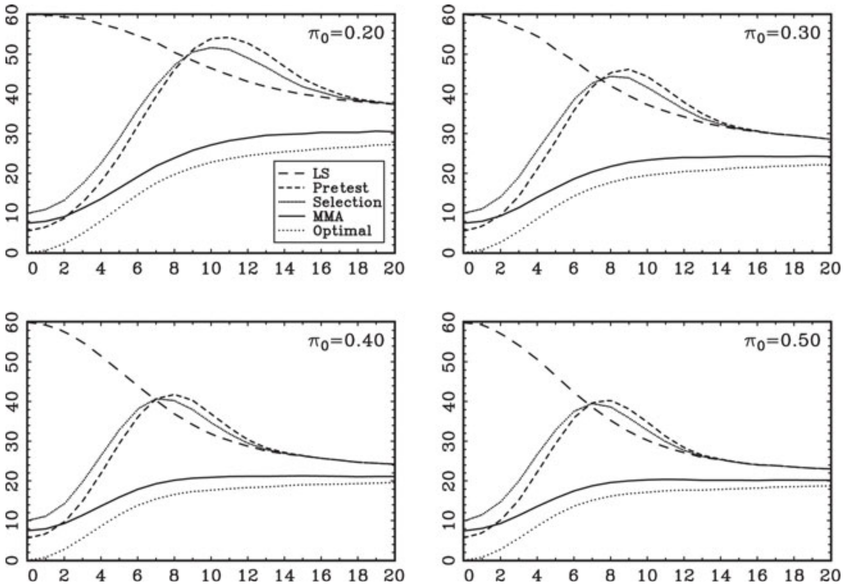

FIGURE 5. Smooth structural break, $m = 5$.

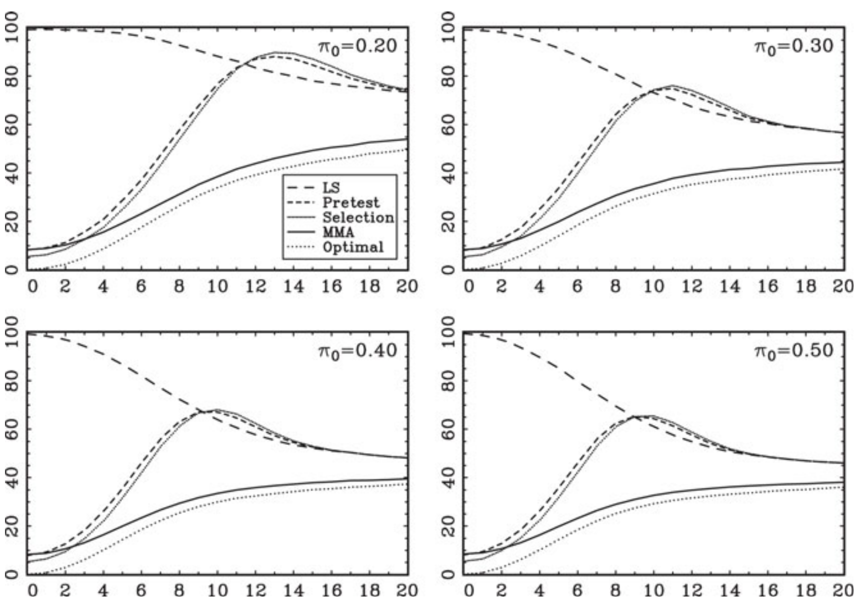

FIGURE 6. Smooth structural break, $m = 10$.

they are nearly equivalent for $m = 5$ and $m = 10$. The MMA estimator, however, has moderate AMSE for all values of ψ and π_0, and its AMSE closely tracks the optimal AMSE. This is especially the case for large m, where the AMSE of MMA is proportionately very close to the infeasible optimal AMSE. Furthermore, among the four feasible estimators, MMA either achieves the smallest AMSE or has AMSE close to the minimum for all parameter values. No other estimator behaves similarly. Again the comparison is most striking for large m, where MMA has the lowest AMSE for nearly all parameter values, the only exception occurring at the smallest values of ψ.

As a summary comparison of the estimators, we report the maximum regret of the estimators in Table 2A as a function of m. The regret is the difference between the AMSE of the estimator and the AMSE of the infeasible optimum. The maximum regret is the largest value of the regret across the parameter values (in this case, ψ and π_0). Table 2A shows that the estimators other than MMA can have extremely large regret.

Next, we allow for misspecification by allowing the true process to be smooth structural change. We set

$$\boldsymbol{\eta}(\pi) = \boldsymbol{\psi} G_\tau(\pi - \pi_0),$$

where $G_\tau(u) = G(u\tau)$, $G(u)$ is the logistic distribution, and $\tau = 20$. (Similar results are found using different smoothing parameters τ.). In this specification, the parameter vector smoothly changes by the magnitude $\boldsymbol{\psi}$ over a substantial period of observations. In this model (as discussed in Section 3) we are estimating the pseudo-true value of the parameter $\boldsymbol{\theta}$. The AMSE of the estimators is fully determined by π_0 and $\psi = \|\boldsymbol{\psi}\|$. The AMSE plots are shown in Figures 4, 5, and 6 for $m = 1$, 5, and 10, and the maximum regret is reported in Table 2B. The results are very similar to the case of pure structural change. Again we find that the MMA estimator is the preferred estimator based on asymptotic mean-square loss.

8. CONCLUSION

Common empirical practice is to test time series regressions for the presence of a structural break, and then if a break is detected account for this by allowing for structural change in estimation. This practice corresponds to a pretest estimator, and this estimator has poor sampling properties. An estimator with much better risk is the weighted average of the no-break and break estimates, where the weight is selected by minimizing a modified Mallows criterion. The latter is a simple function of the sum of squared errors, the Andrews SupF test statistic, and a penalty term.

A referee has suggested that it would be useful to extend the method to allow for two (or more) structural breaks, to allow weighted averages of the no-break, one-break, and two-break model estimates. I believe that this would be a difficult extension. One challenge is that the one-break and two-break models are nonnested (unless one of the breaks is held constant across models), a condition for which the MMA has not yet been developed. Furthermore, development of an appropriate penalty term for a two-break model would be a nontrivial generalization of the work presented in this paper. This would certainly be a challenging yet rewarding topic for future research.

NOTE

1. The AMSE was approximated by simulation, averaging across 50,000 random samples of size 1,000.

REFERENCES

Andrews, D.W.K. (1993) Tests for parameter instability and structural change with unknown change point. *Econometrica* 61, 821–856.

Andrews, D.W.K. (2003) Tests for parameter instability and structural change with unknown change point: A corrigendum. *Econometrica* 71, 395–397.

Andrews, D.W.K. & Ploberger, W. (1994) Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62, 1383–1414.

Bai, J. (1997) Estimation of a change point in multiple regression models. *Review of Economics and Statistics* 79, 674–679.

Hansen, B.E. (2007) Least squares model averaging. *Econometrica* 75, 1175–1189.

Mallows, C.L. (1973) Some comments on C_p. *Technometrics*, 15, 661–675.

APPENDIX: Proofs of Theorems

Proof of Theorem 2. Let $\bar{g}(\pi) = \pi(1-\pi)\mathbf{g}(\pi)'\mathbf{g}(\pi)$ and note that for $J_\delta(\pi)$ defined in (15), $\delta^{-1} J_\delta(\pi) = \bar{g}(\pi) + R(\pi)$ where

$$R(\pi) = -\frac{2}{\delta}\mathbf{g}(\pi)'\mathbf{W}^*(\pi) + \frac{1}{\delta^2}\frac{\mathbf{W}^*(\pi)'\mathbf{W}^*(\pi)}{\pi(1-\pi)}.$$

Because $\mathbf{g}(\pi)$ is bounded on $[\pi_1, \pi_2]$ and $\sup_{\pi_1 \le \pi \le \pi_2} \left|\mathbf{W}^*(\pi)\right| = O_p(1)$, it follows that as $\delta \to \infty$,

$$\sup_{\pi_1 \le \pi \le \pi_2} |R(\pi)| \to_p 0,$$

and so

$$\sup_{\pi_1 \le \pi \le \pi_2} \left|\delta^{-1} J_\delta(\pi) - \bar{g}(\pi)\right| \to_p 0.$$

Because $\bar{g}(\pi)$ is continuous and is uniquely maximized at π_0, it follows that as $\delta \to \infty$,

$$\xi_\delta = \underset{\pi_1 \le \pi \le \pi_2}{\operatorname{argmax}}\, \delta^{-1} J_\delta(\pi) \to_p \pi_0. \qquad \blacksquare$$

Proof of Theorem 3. Let $\mathbf{X}(k)$ be the matrix of stacked regressors $\mathbf{x}_t 1(t < k)$. Define $\mathbf{P} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$,

$$\begin{aligned}\mathbf{X}^*(k) &= \mathbf{X}(k) - \mathbf{P}\mathbf{X}(k)\\ &= \mathbf{X}(k) - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}(k)\\ &= \mathbf{X}(k) - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}(k)'\mathbf{X}(k),\end{aligned}$$

and $\mathbf{P}^*(k) = \mathbf{X}^*(k)\left(\mathbf{X}^*(\mathbf{k})'\mathbf{X}^*(k)\right)^{-1}\mathbf{X}^*(k)'$. By standard projection arguments, $\mathbf{P}(k) = \mathbf{P} + \mathbf{P}^*(k)$, so that

$$\mathbf{e}'\mathbf{P}(k)\mathbf{e} = \mathbf{e}'\mathbf{P}\mathbf{e} + \mathbf{e}'\mathbf{P}^*(k)\mathbf{e}.$$

Under the conditions of the theorem, $\mathbf{e}'\mathbf{P}\mathbf{e} \rightarrow_d \sigma^2\chi_m^2$, and it is not hard to see that $\mathbf{e}'\mathbf{P}^*(n\pi)\mathbf{e} \xrightarrow{d} \sigma^2 J_0(\pi)$. Combined with Theorem 1, (20) follows. ■

Proof of Theorem 4. We deduce from Theorem 1 that

$$\mathbf{M}^{1/2}\frac{\sqrt{n}}{\sigma}\hat{\boldsymbol{\theta}} \xrightarrow{d} \mathbf{S}_\delta(\xi_\delta). \qquad \textbf{(A.1)}$$

Recalling that $F_n \xrightarrow{d} \mathrm{SupF}_\delta$, it follows that

$$\mathbf{M}^{1/2}\frac{\sqrt{n}}{\sigma}\hat{\boldsymbol{\theta}}^p \xrightarrow{d} \mathbf{S}_\delta(\xi_\delta)1\left(\mathrm{SupF}_\delta > c_\alpha\right),$$

$$\mathbf{M}^{1/2}\frac{\sqrt{n}}{\sigma}\hat{\boldsymbol{\theta}}^m \xrightarrow{d} \mathbf{S}_\delta(\xi_\delta)1\left(\mathrm{SupF}_\delta > 2\bar{p}\right),$$

$$\mathbf{M}^{1/2}\frac{\sqrt{n}}{\sigma}\hat{\boldsymbol{\theta}}^{MMA} \xrightarrow{d} \mathbf{S}_\delta(\xi_\delta)1\left(\mathrm{SupF}_\delta > \bar{p}\right)\left(1 - \frac{\bar{p}}{\mathrm{SupF}_\delta}\right).$$

From the definition of $\boldsymbol{\theta}_0$ we have

$$\mathbf{M}^{1/2}\frac{\sqrt{n}}{\sigma}\boldsymbol{\theta}_0 = \mathbf{g}_\delta.$$

The AMSE expressions follow conventionally. ■

Proof of Theorem 5. For fixed w, (A.1) shows that for $\hat{\boldsymbol{\theta}}^w = w\hat{\boldsymbol{\theta}}$,

$$\mathbf{M}^{1/2}\frac{\sqrt{n}}{\sigma}\hat{\boldsymbol{\theta}}^w \xrightarrow{d} w\mathbf{S}_\delta(\xi_\delta).$$

Thus by the calculations of Theorem 3,

$$\begin{aligned}AMSE(\hat{\boldsymbol{\theta}}^w) &= \mathrm{E}\left((w\mathbf{S}_\delta(\xi_\delta) - \mathbf{g}_\delta)'(w\mathbf{S}_\delta(\xi_\delta) - \mathbf{g}_\delta)\right)\\ &= w^2\mathrm{E}\left(\mathbf{S}_\delta(\xi_\delta)'\mathbf{S}_\delta(\xi_\delta)\right) - 2w\mathrm{E}\mathbf{S}_\delta(\xi_\delta)'\mathbf{g}_\delta + \mathrm{E}\mathbf{g}_\delta'\mathbf{g}_\delta.\end{aligned}$$

This is minimized by setting

$$w = \mathrm{E}\left(\mathbf{S}_\delta(\xi_\delta)'\mathbf{S}_\delta(\xi_\delta)\right)^{-1}\mathrm{E}\mathbf{S}_\delta(\xi_\delta)'\mathbf{g}_\delta$$

and has minimized value (26). ■