Supplementary Materials to "Multiscale Autoregression on Adaptively Detected Timescales"

Rafal Baranowski Yining Chen
Piotr Fryzlewicz

Department of Statistics, London School of Economics and Political Science

A Additional numerical experiments

A.1 Sensitivity analysis

Several tuning parameters are required in the algorithm of our approach. The notable ones are the maximum number of scales $q_{\max}$, and the autoregressive order p used in the initial step. Besides, the choice of number of intervals M would also be required, but it should be apparent from our algorithm that it only plays a minor role under a large p (which, in the setup of our current algorithm, would imply $T > 250000$).

Based on our experiments, we find that the proposed approach is not too sensitive to the choice of all the aforementioned tuning parameters. Detailed results are given below.

A.1.1 Maximum number of timescales - $q_{\max}$

Here we run the same experiments listed in the main manuscript, but set $q_{\max} = 5, 20$. The same evaluation metrics are used. Results are given in Table 4 and Table 5.

By comparing the results with those from Table 1 and Table 2 in the main manuscript (where by default $q_{\max} = 10$), it becomes evident that our approach does not appear to be sensitive to the choice of $q_{\max}$. In particular, for different choices of $q_{\max}$, every corresponding AMAR performs better than the competitors.

Model (M1)								
	$E\|\hat{q}-q\|$		$E(D_H)$		$E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$		$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$	
$q_{\max}$	5	20	5	20	5	20	5	20
$T=400$	0.164	0.17	0.539	0.589	0.016	0.0162	0.018	0.0134
	(0.013)	(0.013)	(0.043)	(0.046)	(0.00086)	(0.00082)	(0.0052)	(0.00092)
$T=800$	0.051	0.051	0.187	0.206	0.00351	0.00385	0.00446	0.00469
	(0.0072)	(0.0074)	(0.032)	(0.034)	(0.00026)	(0.00036)	(0.00049)	(0.00054)
$T=1500$	0.022	0.021	0.143	0.117	0.00116	0.00117	0.00138	0.00145
	(0.0046)	(0.0045)	(0.045)	(0.043)	(0.000088)	(0.000088)	(0.00024)	(0.00024)
$T=3000$	0.01	0.011	0.021	0.049	0.000546	0.000549	0.000671	0.000685
	(0.0031)	(0.0033)	(0.0083)	(0.029)	(0.000027)	(0.000027)	(0.00017)	(0.00017)

Model (M2)								
	$E\|\hat{q}-q\|$		$E(D_H)$		$E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$		$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$	
$q_{\max}$	5	20	5	20	5	20	5	20
$T=400$	0.251	0.289	1.07	1.27	0.0207	0.0206	0.0187	0.0243
	(0.017)	(0.017)	(0.064)	(0.071)	(0.0015)	(0.0014)	(0.002)	(0.0049)
$T=800$	0.134	0.149	0.463	0.538	0.00551	0.00574	0.00649	0.00911
	(0.011)	(0.012)	(0.044)	(0.049)	(0.0006)	(0.00059)	(0.0011)	(0.0013)
$T=1500$	0.125	0.136	1.18	1.26	0.00152	0.00142	0.00234	0.0025
	(0.011)	(0.011)	(0.13)	(0.13)	(0.00029)	(0.00026)	(0.00036)	(0.00039)
$T=3000$	0.064	0.069	0.673	0.663	0.000159	0.000181	0.000776	0.00118
	(0.008)	(0.008)	(0.1)	(0.1)	(0.00011)	(0.000074)	(0.0002)	(0.00034)

Model (M3)								
	$E\|\hat{q}-q\|$		$E(D_H)$		$E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$		$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$	
$q_{\max}$	5	20	5	20	5	20	5	20
$T=400$	0.511	0.706	1.46	1.35	0.0238	0.0209	0.0355	0.0289
	(0.023)	(0.035)	(0.054)	(0.046)	(0.00094)	(0.00075)	(0.0021)	(0.0016)
$T=800$	0.262	0.344	0.631	0.64	0.00756	0.00701	0.0103	0.00918
	(0.018)	(0.026)	(0.034)	(0.034)	(0.00039)	(0.00031)	(0.00088)	(0.00075)
$T=1500$	0.068	0.078	0.285	0.297	0.00197	0.00201	0.00341	0.00343
	(0.0089)	(0.011)	(0.04)	(0.042)	(0.0001)	(0.00011)	(0.00039)	(0.0004)
$T=3000$	0.052	0.054	0.192	0.196	0.000677	0.000671	0.00152	0.00151
	(0.0078)	(0.0082)	(0.04)	(0.04)	(0.000042)	(0.000041)	(0.00023)	(0.00023)

Table 4: Performance of AMAR using different $q_{\max}$ under (M1) – (M3), with estimated errors given in the brackets. Here $\hat{q}$ is the number of the fitted timescales, D_H is the Hausdorff distance between the fitted timescale locations $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}\}$ and the true ones $\{\tau_1, \ldots, \tau_q\}$, $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$ is the Euclidean distance between the fitted parameter vector and the true one, and MPSE is the mean squared prediction errors of different models.

Model (M4)								
	$E\lvert\hat{q}-q\rvert$		$E(D_H)$		$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$		$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$	
$q_{\max}$	5	20	5	20	5	20	5	20
$T=400$	0.065	0.106	0.154	0.252	0.0104	0.00932	0.015	0.0145
	(0.0079)	(0.013)	(0.022)	(0.031)	(0.00085)	(0.00061)	(0.0011)	(0.001)
$T=800$	0.041	0.061	0.129	0.131	0.00399	0.00489	0.00627	0.00722
	(0.0063)	(0.0098)	(0.024)	(0.022)	(0.0004)	(0.00042)	(0.00057)	(0.00063)
$T=1500$	0.041	0.035	0.274	0.202	0.0018	0.00193	0.00313	0.00352
	(0.0063)	(0.006)	(0.053)	(0.046)	(0.0001)	(0.00025)	(0.0004)	(0.00056)
$T=3000$	0.015	0.023	0.112	0.128	0.000753	0.000752	0.0017	0.00173
	(0.0038)	(0.0049)	(0.037)	(0.036)	(0.000023)	(0.000023)	(0.00024)	(0.00024)

Model (M5)								
	$E\lvert\hat{q}-q\rvert$		$E(D_H)$		$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$		$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$	
$q_{\max}$	5	20	5	20	5	20	5	20
$T=400$	0.211	0.222	1.63	1.66	0.0107	0.011	0.0129	0.0143
	(0.017)	(0.018)	(0.072)	(0.073)	(0.00043)	(0.00045)	(0.00092)	(0.0016)
$T=800$	0.137	0.137	0.86	0.859	0.0042	0.00421	0.00535	0.00533
	(0.013)	(0.013)	(0.055)	(0.055)	(0.00023)	(0.00022)	(0.00056)	(0.00056)
$T=1500$	0.101	0.104	0.729	0.736	0.00167	0.00171	0.0023	0.00223
	(0.012)	(0.013)	(0.078)	(0.078)	(0.00011)	(0.00012)	(0.00034)	(0.00034)
$T=3000$	0.052	0.052	0.327	0.336	0.000343	0.00034	0.000757	0.000761
	(0.0086)	(0.0086)	(0.054)	(0.056)	(0.000044)	(0.000043)	(0.00018)	(0.00018)

Model (M6)								
	$E\lvert\hat{q}-q\rvert$		$E(D_H)$		$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$		$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$	
$q_{\max}$	5	20	5	20	5	20	5	20
$T=400$	0.378	0.4	2.27	2.29	0.0133	0.013	0.0229	0.0221
	(0.022)	(0.023)	(0.054)	(0.053)	(0.00048)	(0.00045)	(0.0016)	(0.0013)
$T=800$	0.823	0.881	3.31	3.3	0.00909	0.009	0.0158	0.0152
	(0.031)	(0.035)	(0.072)	(0.071)	(0.00029)	(0.00027)	(0.00099)	(0.00097)
$T=1500$	0.428	0.462	3.09	3.1	0.00342	0.00341	0.00676	0.00667
	(0.025)	(0.028)	(0.1)	(0.1)	(0.00014)	(0.00013)	(0.00055)	(0.00055)
$T=3000$	0.533	0.644	3.57	3.52	0.00192	0.00178	0.00444	0.00394
	(0.029)	(0.038)	(0.12)	(0.11)	(0.000084)	(0.000064)	(0.00045)	(0.00038)

Table 5: Performance of AMAR using different $q_{\max}$ under (M4) – (M6), with estimated errors given in the brackets. Here $\hat{q}$ is the number of the fitted timescales, D_H is the Hausdorff distance between the fitted timescale locations $\{\hat{\tau}_1,\ldots,\hat{\tau}_{\hat{q}}\}$ and the true ones $\{\tau_1,\ldots,\tau_q\}$, $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$ is the Euclidean distance between the fitted parameter vector and the true one, and MPSE is the mean squared prediction errors of different models.

A.1.2 The initial order of AR - p

We run the same experiments listed in the main manuscript, but use a fixed $p = 25$. The same evaluation metrics are used. Results are given in Table 6 and Table 7. For the ease of comparison, here we also recall the performance results of the default AMAR that uses p selected via SIC, for which details can be founded in Section 3.1 of the main manuscript.

Here we carefully fixed p at 25, so that it is larger than the timescales among all cases. Here the largest timescale is equal to $\lfloor 3000^{0.4} \rfloor = 24$, from Model (M6) with $T = 3000$. It can be seen that for most cases, both approaches perform similarly. Indeed, AMAR with a fixed p might lead to some very moderate improvement over our current approach of selection via SIC in a few settings. Still, not surprisingly, using a fixed p could be quite problematic when the chosen p is close to or bigger than $\tau_{q_{\max}}$, as is evident in the setting of Model (M6) with $T = 3000$, where its performance is more than 100% worse in every evaluation metric.

A.2 (More conventional) higher-order AR

In this part, we compare AMAR and conventional AR models (selected both by AIC and BIC) over the data that are generated from more conventional high-order stationary AR models. In particular, we consider the following settings, with $\tau_q = 16$ and $q = 16, 12, 8$.

(M7) $q = 16$ and $\tau_i = i$ for $i = 1, \ldots, 16$, with the corresponding AR coefficients

$$\boldsymbol{\beta} = (0.2, -0.2, 0.2, -0.2, \ldots, 0.2, -0.2)^T.$$

(M8) $q = 12$ and $\{\tau_1, \ldots, \tau_{12}\} = \{1, \ldots, 16\} \backslash \{2, 6, 10, 14\}$, with the corresponding AR coefficients

$$\boldsymbol{\beta} = (0.2, 0, 0, -0.2, 0.2, 0, 0, -0.2, \ldots, 0.2, 0, 0, -0.2)^T.$$

(M9) $q = 8$ and $\tau_i = 2i$ for $i = 1, \ldots, 8$, with the corresponding AR coefficients

$$\boldsymbol{\beta} = (0.2, 0.2, -0.2, -0.2, \ldots, 0.2, 0.2, -0.2, -0.2)^T.$$

Here Model (M7) is a conventional high-order AR. Models (M8) and (M9) are also high-order, but their AR coefficients are more structured (though $\tau_q = 16$ and q are still at the same order). In particular, all three models are stationary.

For these models, we run the experiments using the same settings as listed in the main manuscript, but set $q_{\max} = 20$ (as here q can be as high as 16). For the evaluation metrics, we look at the accuracy of the estimated order of AR, denoted by $|\hat{\tau}_{\hat{q}} - \tau_q|$, the Euclidean distance between the

p	$E\lvert\hat{q}-q\rvert$ SIC	$E\lvert\hat{q}-q\rvert$ fixed	$E(D_H)$ SIC	$E(D_H)$ fixed	$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$ SIC	$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$ fixed	$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$ SIC	$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$ fixed
Model (M1)								
$T=400$	0.172	0.196	0.593	0.738	0.0159	0.0178	0.0133	0.0147
	(0.014)	(0.016)	(0.047)	(0.07)	(0.0008)	(0.00091)	(0.00093)	(0.001)
$T=800$	0.051	0.047	0.181	0.252	0.0035	0.00401	0.0046	0.00483
	(0.0072)	(0.0071)	(0.03)	(0.046)	(0.00026)	(0.00032)	(0.00048)	(0.00052)
$T=1500$	0.018	0.02	0.085	0.073	0.00116	0.00115	0.00138	0.0014
	(0.0042)	(0.0046)	(0.03)	(0.027)	(0.000088)	(0.000084)	(0.00024)	(0.00025)
$T=3000$	0.012	0.009	0.072	0.016	0.000546	0.000546	0.000662	0.000681
	(0.0034)	(0.003)	(0.035)	(0.0063)	(0.000027)	(0.000027)	(0.00017)	(0.00017)
Model (M2)								
$T=400$	0.303	0.235	1.33	1.67	0.02	0.0233	0.0281	0.0259
	(0.018)	(0.017)	(0.072)	(0.11)	(0.0013)	(0.0018)	(0.01)	(0.007)
$T=800$	0.194	0.154	0.764	1.13	0.00635	0.00638	0.00852	0.00815
	(0.014)	(0.012)	(0.06)	(0.1)	(0.00071)	(0.00065)	(0.0013)	(0.0011)
$T=1500$	0.108	0.122	0.921	0.821	0.00171	0.000986	0.00666	0.00386
	(0.01)	(0.011)	(0.11)	(0.092)	(0.00038)	(0.00019)	(0.0038)	(0.0019)
$T=3000$	0.07	0.056	0.646	0.446	0.0000979	0.0000896	0.000793	0.000687
	(0.0081)	(0.0073)	(0.099)	(0.072)	(0.000021)	(0.000021)	(0.0002)	(0.00017)
Model (M3)								
$T=400$	0.711	0.314	1.37	1.17	0.0211	0.0187	0.0296	0.0297
	(0.035)	(0.024)	(0.046)	(0.057)	(0.00076)	(0.00072)	(0.0016)	(0.0017)
$T=800$	0.344	0.146	0.643	0.481	0.00699	0.00571	0.00922	0.00825
	(0.026)	(0.015)	(0.034)	(0.036)	(0.00031)	(0.00029)	(0.00075)	(0.00069)
$T=1500$	0.083	0.087	0.31	0.254	0.00203	0.0022	0.0034	0.00359
	(0.011)	(0.011)	(0.043)	(0.03)	(0.00011)	(0.00018)	(0.0004)	(0.00044)
$T=3000$	0.054	0.055	0.219	0.126	0.000673	0.000685	0.0015	0.00147
	(0.0082)	(0.0091)	(0.045)	(0.023)	(0.000041)	(0.000041)	(0.00023)	(0.00023)

Table 6: Performance of AMAR under (M1) – (M3) with the initial AR order p either selected via SIC, or fixed at $p = 25$. The estimated errors given in the brackets. Here $\hat{q}$ is the number of the fitted timescales, D_H is the Hausdorff distance between the fitted timescale locations $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}\}$ and the true ones $\{\tau_1, \ldots, \tau_q\}$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ is the Euclidean distance between the fitted parameter vector and the true one, and MPSE is the mean squared prediction errors of different models.

Model (M4)								
	$E\lvert\hat{q}-q\rvert$		$E(D_H)$		$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$		$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$	
p	SIC	fixed	SIC	fixed	SIC	fixed	SIC	fixed
$T=400$	0.098	0.078	0.2	0.27	0.00892	0.0105	0.0145	0.0157
	(0.012)	(0.012)	(0.027)	(0.041)	(0.00065)	(0.00074)	(0.0011)	(0.0012)
$T=800$	0.044	0.039	0.092	0.172	0.00397	0.00446	0.00657	0.00653
	(0.0085)	(0.008)	(0.019)	(0.035)	(0.0003)	(0.0004)	(0.0006)	(0.00057)
$T=1500$	0.035	0.039	0.291	0.158	0.00179	0.00194	0.00333	0.00324
	(0.006)	(0.0063)	(0.059)	(0.03)	(0.00011)	(0.00011)	(0.0004)	(0.00039)
$T=3000$	0.023	0.024	0.129	0.077	0.000756	0.000753	0.0017	0.00162
	(0.0051)	(0.005)	(0.033)	(0.019)	(0.000023)	(0.000023)	(0.00024)	(0.00024)
Model (M5)								
	$E\lvert\hat{q}-q\rvert$		$E(D_H)$		$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$		$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$	
p	SIC	fixed	SIC	fixed	SIC	fixed	SIC	fixed
$T=400$	0.217	0.265	1.64	2.29	0.0109	0.0123	0.0164	0.0159
	(0.017)	(0.02)	(0.073)	(0.1)	(0.00045)	(0.00051)	(0.0028)	(0.0014)
$T=800$	0.133	0.144	0.858	1.05	0.00414	0.00447	0.00517	0.00588
	(0.013)	(0.014)	(0.056)	(0.073)	(0.00022)	(0.00026)	(0.00055)	(0.00065)
$T=1500$	0.099	0.092	0.704	0.574	0.00167	0.00168	0.00237	0.0023
	(0.012)	(0.011)	(0.076)	(0.058)	(0.00012)	(0.00011)	(0.00033)	(0.00034)
$T=3000$	0.052	0.049	0.331	0.271	0.000339	0.000346	0.000788	0.000746
	(0.0086)	(0.0082)	(0.054)	(0.042)	(0.000043)	(0.000043)	(0.00017)	(0.00017)
Model (M6)								
	$E\lvert\hat{q}-q\rvert$		$E(D_H)$		$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$		$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$	
p	SIC	fixed	SIC	fixed	SIC	fixed	SIC	fixed
$T=400$	0.407	0.43	2.3	3.9	0.0133	0.015	0.023	0.0279
	(0.024)	(0.024)	(0.054)	(0.12)	(0.00046)	(0.00058)	(0.0016)	(0.0016)
$T=800$	0.886	0.486	3.29	3.18	0.00902	0.00718	0.015	0.0129
	(0.035)	(0.026)	(0.071)	(0.083)	(0.00028)	(0.00023)	(0.00098)	(0.00086)
$T=1500$	0.455	0.639	3.08	3.04	0.00336	0.0038	0.00668	0.00712
	(0.028)	(0.035)	(0.1)	(0.085)	(0.00013)	(0.00014)	(0.00055)	(0.00056)
$T=3000$	0.642	2.04	3.52	6.8	0.00177	0.00392	0.00395	0.0071
	(0.037)	(0.063)	(0.11)	(0.12)	(0.000064)	(0.000096)	(0.00038)	(0.00055)

Table 7: Performance of AMAR under (M4) – (M6), with the initial AR order p either selected via SIC, or fixed at $p = 25$. Here $\hat{q}$ is the number of the fitted timescales, D_H is the Hausdorff distance between the fitted timescale locations $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}\}$ and the true ones $\{\tau_1, \ldots, \tau_q\}$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ is the Euclidean distance between the fitted parameter vector and the true one, and MPSE is the mean squared prediction errors of different models.

fitted parameter vector and the true one, denoted by $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$, and the mean squared prediction errors (MSPE) of different models. Results are given in Table 8.

We see that with in Model (M7), unsurprisingly AR with order selected via BIC performs the best among all the evaluation measures. However, the performance of AMAR is only slightly worse (and better than AR with order selected via AIC). In particular, it tends to estimates the number of scales (which is the same as the AR order) correctly when T is reasonably large, implying little efficiency loss for using AMAR even when there is no meaningful AMAR-type structure in the parameter vector of AR coefficients. On the other hand, as we move to Model (M8) and Model (M9) where the AR parameter vectors have more structures embedded (though here τ_q and q are still at the same order), AMAR tends to perform better than its competitors in terms of both the parameter estimation and prediction accuracy. The improvement is more visible in the setting of Model (M9), as it has less scales than Model (M8), so is intuitively more favourable to AMAR.

A.3 Non-stationary AR

Here we report the results from experiments with series simulated from non-stationary AR models with unit roots. The scenarios we consider are similar to (M1) – (M6) listed in the main manuscript, with their details outlined below.

(M1') Same as (M1) but with $\alpha_1 = 0.4$, $\alpha_2 = 0.6$ (i.e. $\boldsymbol{\beta} = (0.6, 0.2, 0.2)^T$).

(M2') Same as (M2) but with $\alpha_1 = 1.5$, $\alpha_2 = -0.5$ (i.e. $\boldsymbol{\beta} = (0.65, 0.65, -0.1, -0.1, -0.1)^T$).

(M3') Same as (M3) but with $\alpha_1 = 0.5$, $\alpha_2 = -1$, $\alpha_3 = 1.4$ (i.e. $\boldsymbol{\beta} = (0.5, -0.1, -0.1, -0.1, -0.1, 0.1, \ldots, 0.1)^T$).

(M4') Same as (M4) but with $\alpha_1 = 1$, $\alpha_2 = -4.8$, $\alpha_3 = 10.2$, $\alpha_4 = -6.4$ (i.e. $\boldsymbol{\beta} = (1, 0, \ldots, 0, 0.8, -0.8)^T$, so $\varepsilon_t = (1 - 0.8B^7)(1 - B)X_t$).

(M5') Same as (M5) but with $\alpha_1 = 1$ (i.e. $\boldsymbol{\beta} = (0.1, \ldots, 0.1)^T$).

(M6') Same as (M6) but with $\alpha_1 = \alpha_2 = 0.5$ (i.e. $\boldsymbol{\beta} = (0.5 + 0.5/\lfloor T^{0.4} \rfloor, 0.5/\lfloor T^{0.4} \rfloor \ldots, 0.5/\lfloor T^{0.4} \rfloor)^T$).

Here we use AMAR with default choice of its tuning parameters outlined in Section 3.1. The corresponding results are summarised in Table 9 and Table 10, where as before, we report the estimates for $|q - \hat{q}|$, with $\hat{q}$ being the number of the fitted timescales, the Hausdorff distance D_H between the fitted timescale locations $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}\}$ and the true ones $\{\tau_1, \ldots, \tau_q\}$, the Euclidean distance between the fitted parameter vector and the true one, denoted by $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$, and the ratio between the mean squared prediction error (MPSE) using the fitted model and that with the oracle over the next $T^* = 100$ unseen observations.

	$E\lvert\hat{\tau}_{\hat{q}}-\tau_q\rvert$			$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Model (M7)									
Method	AMAR	AIC	BIC	AMAR	AIC	BIC	AMAR	AIC	BIC
$T=400$	6.73	0.877	12.6	0.577	0.051	0.543	0.204	0.0534	0.177
	(0.14)	(0.042)	(0.17)	(0.0038)	(0.00084)	(0.0068)	(0.0031)	(0.0018)	(0.003)
$T=800$	2.05	1.68	0.23	0.139	0.027	0.0293	0.0826	0.0274	0.0265
	(0.1)	(0.09)	(0.055)	(0.0053)	(0.00053)	(0.0023)	(0.0026)	(0.0012)	(0.0016)
$T=1500$	0.916	2	0.014	0.012	0.0148	0.0105	0.0124	0.0141	0.0114
	(0.09)	(0.11)	(0.0037)	(0.00026)	(0.0004)	(0.00014)	(0.00074)	(0.00079)	(0.00071)
$T=3000$	0.662	2.14	0.015	0.00582	0.00733	0.00532	0.00597	0.00701	0.00555
	(0.077)	(0.13)	(0.0038)	(0.00015)	(0.00017)	(0.000074)	(0.00052)	(0.00057)	(0.00049)
Model (M8)									
	$E\lvert\hat{\tau}_{\hat{q}}-\tau_q\rvert$			$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	AIC	BIC	AMAR	AIC	BIC	AMAR	AIC	BIC
$T=400$	2.93	0.895	7.45	0.151	0.0503	0.175	0.132	0.051	0.14
	(0.085)	(0.04)	(0.21)	(0.0019)	(0.00075)	(0.0038)	(0.003)	(0.0017)	(0.0039)
$T=800$	1.62	1.69	0.307	0.08	0.0263	0.025	0.0767	0.0281	0.0273
	(0.034)	(0.091)	(0.041)	(0.001)	(0.00042)	(0.00077)	(0.002)	(0.0011)	(0.0014)
$T=1500$	0.669	2.02	0.019	0.00969	0.014	0.0109	0.00957	0.0122	0.0102
	(0.077)	(0.12)	(0.0052)	(0.00029)	(0.00025)	(0.00015)	(0.0007)	(0.00075)	(0.00068)
$T=3000$	0.597	1.95	0.014	0.00401	0.00687	0.00533	0.00452	0.0062	0.00516
	(0.077)	(0.14)	(0.004)	(0.00011)	(0.00013)	(0.000068)	(0.00041)	(0.0005)	(0.00045)
Model (M9)									
	$E\lvert\hat{\tau}_{\hat{q}}-\tau_q\rvert$			$E\lVert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rVert$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	AIC	BIC	AMAR	AIC	BIC	AMAR	AIC	BIC
$T=400$	1.37	0.849	0.161	0.104	0.0499	0.0485	0.117	0.0493	0.0487
	(0.015)	(0.039)	(0.022)	(0.0011)	(0.00082)	(0.0012)	(0.003)	(0.0017)	(0.0017)
$T=800$	1.04	1.77	0.021	0.0738	0.0269	0.0199	0.0695	0.0253	0.0212
	(0.0061)	(0.093)	(0.0048)	(0.00069)	(0.00051)	(0.00029)	(0.002)	(0.0011)	(0.001)
$T=1500$	0.214	1.93	0.017	0.00469	0.0147	0.0107	0.00691	0.0131	0.011
	(0.042)	(0.12)	(0.0041)	(0.00039)	(0.0003)	(0.00016)	(0.00067)	(0.0008)	(0.00073)
$T=3000$	0.174	1.75	0.018	0.00215	0.00703	0.00528	0.00384	0.00646	0.00551
	(0.041)	(0.12)	(0.0047)	(0.00022)	(0.00015)	(0.000079)	(0.00044)	(0.00051)	(0.00047)

Table 8: Performance of different methods under (M7) – (M9), with estimated errors given in the brackets. Here $|\hat{\tau}_{\hat{q}}-\tau_q|$ is the difference between the estimated and true order of AR, $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$ is the Euclidean distance between the fitted parameter vector and the true one, and MPSE is the mean squared prediction errors of different models.

Model (M1')										
	$E\|\hat{q}-q\|$		$E(D_H)$		$E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	Fused	AMAR	Fused	AMAR	Fused	AIC	AMAR	Fused	AIC
$T=400$	0.469	1.21	2.31	18.6	0.0449	0.381	0.0268	38.6	28.7	32.5
	(0.019)	(0.049)	(0.11)	(0.042)	(0.0022)	(0.0014)	(0.0008)	(23)	(1.2)	(3.1)
$T=800$	0.33	1.43	1.97	26.4	0.0347	0.398	0.0201	1.45	29	7.79
	(0.016)	(0.081)	(0.11)	(0.063)	(0.0022)	(0.001)	(0.00068)	(1.4)	(1.3)	(0.6)
$T=1500$	0.367	2.13	4.47	36.1	0.0302	0.409	0.017	0.0231	26.9	1.95
	(0.016)	(0.14)	(0.25)	(0.088)	(0.0022)	(0.00073)	(0.00059)	(0.0018)	(1.1)	(0.16)
$T=3000$	0.249	2.61	3.44	51.8	0.0192	0.42	0.0161	0.0148	29.8	0.536
	(0.014)	(0.18)	(0.23)	(0.094)	(0.0019)	(0.00026)	(0.00063)	(0.0016)	(1.3)	(0.043)

Model (M2')										
	$E\|\hat{q}-q\|$		$E(D_H)$		$E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	Fused	AMAR	Fused	AMAR	Fused	AIC	AMAR	Fused	AIC
$T=400$	0.399	3.85	2.09	13.5	0.0254	0.609	0.116	0.0374	4.41	0.226
	(0.019)	(0.1)	(0.075)	(0.12)	(0.00093)	(0.0092)	(0.0031)	(0.0024)	(0.14)	(0.014)
$T=800$	0.269	6.75	1.52	18.5	0.0134	0.69	0.0892	0.0175	6	0.0971
	(0.017)	(0.16)	(0.073)	(0.21)	(0.00061)	(0.0073)	(0.0025)	(0.0015)	(0.2)	(0.0055)
$T=1500$	0.187	10	2.11	24.5	0.00572	0.733	0.0681	0.00601	7.33	0.0506
	(0.013)	(0.22)	(0.16)	(0.33)	(0.00036)	(0.0059)	(0.0021)	(0.00074)	(0.25)	(0.0024)
$T=3000$	0.086	14.2	0.985	35.3	0.0018	0.774	0.0399	0.0023	8.7	0.0246
	(0.009)	(0.31)	(0.12)	(0.5)	(0.0002)	(0.0039)	(0.0014)	(0.00051)	(0.25)	(0.0013)

Model (M3')										
	$E\|\hat{q}-q\|$		$E(D_H)$		$E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	Fused	AMAR	Fused	AMAR	Fused	AIC	AMAR	Fused	AIC
$T=400$	0.747	2.47	1.37	17.4	0.0246	0.297	0.0606	0.0351	0.766	17.4
	(0.034)	(0.058)	(0.043)	(0.13)	(0.0011)	(0.003)	(0.00089)	(0.0021)	(0.022)	(2.1)
$T=800$	0.499	1.72	0.897	24.7	0.0201	0.318	0.0342	0.0365	0.767	3.76
	(0.029)	(0.089)	(0.039)	(0.11)	(0.0015)	(0.0027)	(0.00059)	(0.0036)	(0.018)	(0.34)
$T=1500$	0.177	2.4	0.744	34.1	0.0047	0.323	0.0216	0.0104	0.788	0.954
	(0.015)	(0.16)	(0.073)	(0.13)	(0.00076)	(0.0031)	(0.00043)	(0.0019)	(0.019)	(0.09)
$T=3000$	0.104	1.93	0.506	49.2	0.00336	0.339	0.0141	0.00609	0.854	0.288
	(0.011)	(0.16)	(0.069)	(0.15)	(0.00073)	(0.0027)	(0.00035)	(0.0015)	(0.02)	(0.028)

Table 9: Performance of different methods under (M1') – (M3'), with estimated errors given in the brackets. Here $\hat{q}$ is the number of the fitted timescales, D_H is the Hausdorff distance between the fitted timescale locations $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}\}$ and the true ones $\{\tau_1, \ldots, \tau_q\}$, $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$ is the Euclidean distance between the fitted parameter vector and the true one, and MPSE is the mean squared prediction errors of different models.

Model (M4')										
	$E\|\hat{q}-q\|$		$E(D_H)$		$E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	Fused	AMAR	Fused	AMAR	Fused	AIC	AMAR	Fused	AIC
$T=400$	0.527	2.64	1.45	18.1	0.161	2.22	0.888	0.254	216	28.1
	(0.021)	(0.024)	(0.07)	(0.074)	(0.013)	(0.0019)	(0.013)	(0.021)	(9.3)	(2)
$T=800$	0.27	2.55	0.702	25.4	0.21	2.24	0.84	0.295	227	7.28
	(0.017)	(0.026)	(0.052)	(0.11)	(0.014)	(0.00068)	(0.012)	(0.022)	(10)	(0.5)
$T=1500$	0.225	2.4	2.02	34.2	0.124	2.25	0.847	0.189	267	2.79
	(0.014)	(0.041)	(0.15)	(0.17)	(0.011)	(0.00035)	(0.011)	(0.018)	(13)	(0.13)
$T=3000$	0.175	2.97	1.69	48.2	0.0678	2.26	0.841	0.0921	293	1.4
	(0.012)	(0.092)	(0.14)	(0.22)	(0.0085)	(0.00026)	(0.0096)	(0.012)	(14)	(0.052)

Model (M5')										
	$E\|\hat{q}-q\|$		$E(D_H)$		$E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	Fused	AMAR	Fused	AMAR	Fused	AIC	AMAR	Fused	AIC
$T=400$	0.354	0.575	1.94	9.74	0.0134	0.0461	0.0464	5.81	0.439	10
	(0.022)	(0.037)	(0.077)	(0.039)	(0.00061)	(0.00035)	(0.00082)	(4.5)	(0.032)	(0.89)
$T=800$	0.322	1.33	1.44	17.5	0.00602	0.0591	0.0276	0.254	0.467	2.47
	(0.02)	(0.065)	(0.074)	(0.07)	(0.00052)	(0.00046)	(0.00062)	(0.19)	(0.027)	(0.19)
$T=1500$	0.154	2.3	1.2	27.6	0.00212	0.069	0.0203	2.32	0.569	0.692
	(0.013)	(0.12)	(0.11)	(0.062)	(0.0003)	(0.00044)	(0.0006)	(2.2)	(0.034)	(0.064)
$T=3000$	0.114	4.19	0.818	43.4	0.000744	0.0776	0.014	0.72	0.622	0.197
	(0.01)	(0.19)	(0.088)	(0.1)	(0.00024)	(0.00036)	(0.00048)	(0.54)	(0.03)	(0.022)

Model (M6')										
	$E\|\hat{q}-q\|$		$E(D_H)$		$E\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$			$\frac{\text{MSPE(fitted)}}{\text{MSPE(oracle)}}-1$		
Method	AMAR	Fused	AMAR	Fused	AMAR	Fused	AIC	AMAR	Fused	AIC
$T=400$	0.889	1.32	3.67	17.5	0.0177	0.234	0.0441	0.0251	0.762	10.4
	(0.034)	(0.064)	(0.076)	(0.11)	(0.0007)	(0.002)	(0.00068)	(0.0013)	(0.038)	(0.87)
$T=800$	1.54	1.52	7.63	24.2	0.0184	0.228	0.0329	0.0336	0.631	3.37
	(0.049)	(0.11)	(0.072)	(0.17)	(0.00073)	(0.0022)	(0.00041)	(0.0015)	(0.031)	(0.3)
$T=1500$	0.931	2.01	5.6	33.4	0.00501	0.23	0.0229	0.01	0.499	1.07
	(0.04)	(0.15)	(0.13)	(0.22)	(0.00025)	(0.0022)	(0.00025)	(0.00075)	(0.018)	(0.094)
$T=3000$	2.09	2.95	12.4	47.9	0.00515	0.236	0.0159	0.0122	0.427	0.328
	(0.066)	(0.18)	(0.13)	(0.28)	(0.00011)	(0.0019)	(0.00015)	(0.00076)	(0.012)	(0.036)

Table 10: Performance of different methods under (M4') – (M6'), with estimated errors given in the brackets. Here $\hat{q}$ is the number of the fitted timescales, D_H is the Hausdorff distance between the fitted timescale locations $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{q}}\}$ and the true ones $\{\tau_1, \ldots, \tau_q\}$, $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|$ is the Euclidean distance between the fitted parameter vector and the true one, and MPSE is the mean squared prediction errors of different models.

We see that even in the setting of non-stationary observations, AMAR still performs much better than its competitors in most settings, even though all methods seem to perform worse as compared to the stationary settings. Unsurprisingly, here the reported results are associated with larger estimation errors.

In addition, we note that the fused LASSO approach performs much worse than its competitors in terms of MSPE, especially in (M1') and (M4'). This is because the fused LASSO approach tends to over-estimate the number of scales, resulting in less accurate $\hat{\boldsymbol{\beta}}$, which could greatly affect the corresponding MSPE when the series is non-stationary.

B Additional real data example: well-log

We consider the well-log data from O Ruanaidh and Fitzgerald (1996). Prior to use, the data is cleaned by removing outliers, taken here to be the observations that differ from the median-fliter fit to the data (with span 25) by at least 7500. This retains 97.7% of the data points. The cleaned data, denoted as $\{X_t\}_{t=1}^{3956}$, is shown in the left plot of Figure 5.

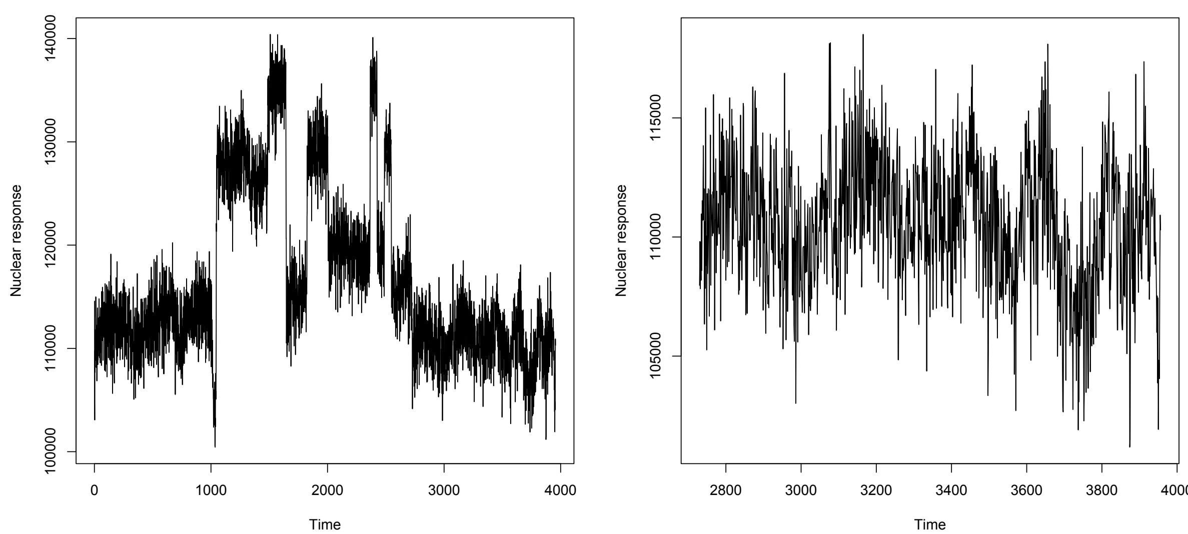

Figure 5: Left: the well-log data from O Ruanaidh and Fitzgerald (1996), cleaned as described in the text. Right: the end part of the data, from time location 2730.

As summarised in Fearnhead and Clifford (2003), the data represents measurements of the nuclear magnetic response of underground rocks. The underlying (unobserved) signal is assumed to be piecewise constant, with each constant segment representing a stratum of rock. The jumps occur when a new rock stratum is met. The problem of detecting these change-points in the underlying signal is of practical importance in oil drilling.

It is known (for instance, see Cho and Fryzlewicz (2021) and the references therein) that the problem of multiple change-point detection in a piecewise-constant signal observed in noise is much more challenging if the noise displays autocorrelation, as the natural fluctuations of the autocorrelated process can be mistaken for change-points, and vice versa. This appears to be the case in the well-log data: the right-hand plot of Figure 5 shows the end portion of the data, from the observation after the last visually obvious change-point (at location 2729) to the end. As discussed earlier, the visual appearance of the data fluctuations in this region of the dataset suggests that the AMAR model may be appropriate. Our aim is therefore to: (a) estimate the appropriate AMAR model on $\{X_{2730}, \ldots, X_{3956}\}$, (b) fit the estimated model from the previous step on the entire dataset (i.e. $\{X_1, \ldots, X_{3956}\}$) to remove the autocorrelations in the data, and (c) estimate change-point locations in the thus-decorrelated dataset using a method suitable for multiple change-point detection in uncorrelated (Gaussian) noise.

We start with a preliminary time series analysis of $\{X_{2730}, \ldots, X_{3956}\}$. The unconstrained AR fit to this subset of the data, with the AR order chosen via AIC yields order 17, and the estimated coefficients are shown in the left panel of Figure 6. The appearance of the vector of the estimated coefficients suggests that a piecewise-constant model (as dictated by AMAR) may be suitable here. The fitted AMAR model returns estimated scales $1, 9, 13, 16, 17$ (see Figure 6).

Prior to fitting the estimated AMAR model to the entire dataset, however, we shrink the estimated AMAR coefficients by a factor of $\rho \in (0, 1)$, i.e. we replace each estimated AMAR coefficient $\hat{\alpha}_r$ by $\rho\hat{\alpha}_r$. This is done because the original estimated AMAR coefficients sum up to practically 1 (0.9998), and therefore fitting such a "near-unit-root" AMAR model has a strong differencing effect, which as well as successfully removing the autocorrelations, could also potentially remove too much of the structure of the signal for successful detection of change-points in the levels.

We choose ρ as follows. Starting with $\rho = 0$, we increase ρ in steps of 0.01 until our selected procedure(s) for change-point detection under lack of serial correlation do not indicate any change-points after time $t = 2730$ (since we initially fitted an AMAR model on this portion of the data under the assumption of stationarity there). This is first achieved for $\rho = 0.78$, for both Wild Binary Segmentation (Fryzlewicz, 2014) and Narrowest-Over-Threshold (Baranowski et al., 2019), both with model selection via the strengthened Schwarz Information Criterion, and using the implementation from the R package **breakfast** (Anastasiou et al., 2021) with otherwise default parameters. These two procedures indicate, respectively, 12 and 10 change-points in the signal. The change-point locations estimated via Wild Binary Segmentation are shown in Figure 7. With the exception of the possible double detection at times $t = 1043, 1056$, the estimated change-point locations visually align with the signal very well.

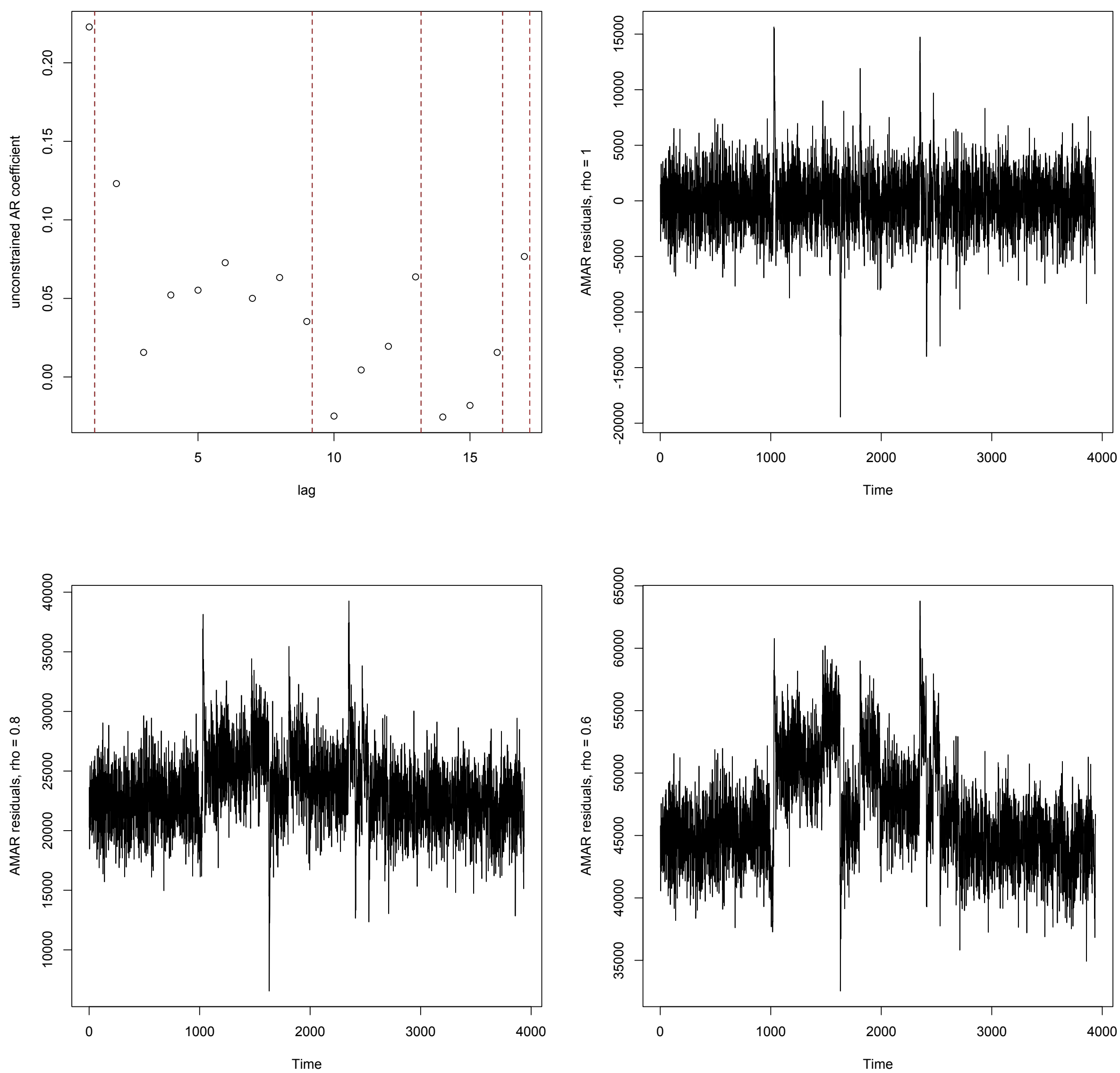

Figure 6: Top left: the unconstrained estimated AR coefficients for $X_{2730:3956}$ (circles); extents of estimated AMAR scales (dashed lines). Top right: unshrunk AMAR residuals. Bottom left: AMAR residuals with $\rho = 0.8$. Bottom right: AMAR residuals with $\rho = 0.6$.

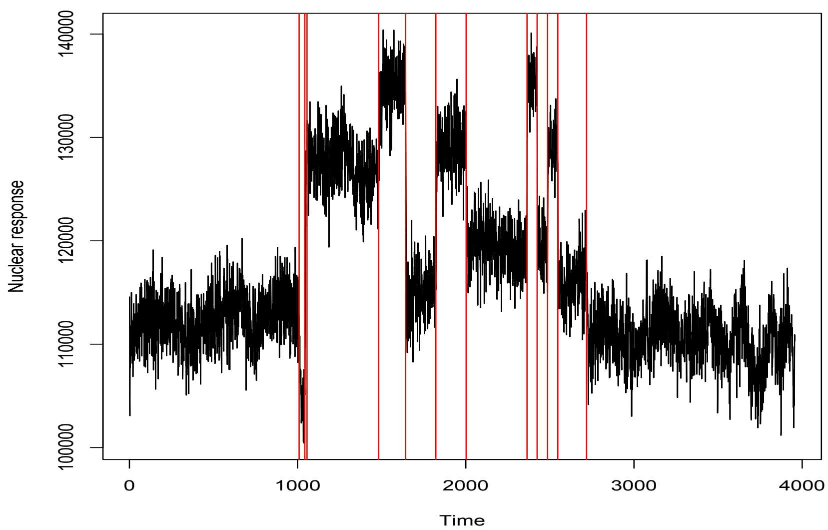

Figure 7: The well-log data with the change-point locations estimated via the shrunk AMAR fit with $\rho = 0.78$ and using WBS+sSIC on the residuals.

C Proofs

C.1 Proof of Proposition 2.1

For AR(p) processes, it has a stationary and causal solution if and only if all the roots of $b(z) = 0$ lie outside $\mathbb{T}$.

For any AMAR(q) with $\alpha_1, \ldots, \alpha_q$ (and the corresponding AR parameters $\beta_1, \ldots, \beta_p$), $\sum_{j=1}^q |\alpha_j| < 1$ under the AMAR framework is equivalent to

$$\sum_{j=1}^{p} |\beta_j| = \sum_{j=1}^{p} \Big(\sum_{k:\tau_k \geq j} \frac{|\alpha_k|}{\tau_k} \Big) = \frac{|\alpha_1|}{\tau_1}\tau_1 + \cdots + \frac{|\alpha_q|}{\tau_q}\tau_q < 1$$

in view of Equations (1) and (2) and (3). Now since $\sum_{j=1}^p |\beta_j| < 1$, $b(z) := 1 - \beta_1 z - \cdots - \beta_p z^p \geq 1 - |\beta_1|\|z\| - \cdots - |\beta_p|\|z\|^p \geq 1 - \sum_{j=1}^p |\beta_j| > 0$ for any $z \in \mathbb{T}$. As such, all the roots of $b(z) = 0$ lie outside $\mathbb{T}$, which implies the existence of a causal stationary solution.

Next, given $\alpha_1, \ldots, \alpha_q \geq 0$, we have that $\beta_1, \ldots, \beta_p \geq 0$. The existence of a causal stationary solution implies that all the roots of $b(z) = 0$ lie outside $\mathbb{T}$. Since $b(0) = 1$ and $b(\cdot)$ is continuous, one would necessarily require $b(1) > 0$. i.e. $\beta_1 + \cdots + \beta_p < 1$. This condition under the AMAR framework is equivalent to

$$\sum_{j=1}^{p}\Big(\sum_{k:\tau_k\geq j}\frac{\alpha_k}{\tau_k}\Big)=\frac{\alpha_1}{\tau_1}\tau_1+\cdots+\frac{\alpha_q}{\tau_q}\tau_q<1,$$

which is the same as $\sum_{j=1}^{q}|\alpha_j|<1$ under non-negativity. □

C.2 Proof of Theorem 2.1

We write the AR(p) model as

$$\mathbf{Y}_t=\mathbf{B}\mathbf{Y}_{t-1}+\varepsilon_t\mathbf{u},\quad t=1,\ldots,T, \tag{17}$$

where $\mathbf{Y}_t=(X_t,X_{t-1},...,X_{t-p+1})'$, the matrix of the coefficients

$$\mathbf{B}=\begin{pmatrix}\beta_1 & \beta_2 & \cdots & \beta_p\\ & \mathbf{I}_{p-1} & & \mathbf{0}\end{pmatrix} \tag{18}$$

and $\mathbf{u}=(1,0,\ldots,0)'\in\mathbb{R}^p$. We start with a few auxiliary results.

Lemma C.1 (Parseval's identity, Theorem 1.9 in Duoandikoetxea (2001)) *For any complex-valued sequence $\{f_k\}_{k\in\mathbb{Z}}$ such that $\sum_{k\in\mathbb{Z}}|f_k|^2<\infty$, the following identity holds*

$$\sum_{k\in\mathbb{Z}}|f_k|^2=\int_{\mathbb{T}}|f(z)|^2dm(z), \tag{19}$$

where $f(z)=\sum_{k\in\mathbb{Z}}f_kz^k$, $\mathbb{T}=\{z\in\mathbb{C}:\ |z|=1\}$, $dm(z)=\frac{d|z|}{2\pi}$.

Lemma C.2 (Cauchy's integral formula) *Let $\mathbf{M}\in\mathbb{R}^{p\times p}$ be a real- or complex- valued matrix. Then for any curve Γ enclosing all eigenvalues of $\mathbf{M}$ and any $j\in\mathbb{N}$ the following holds*

$$\mathbf{M}^j=\frac{1}{2\pi i}\int_{\Gamma}z^j(z\mathbf{I}_p-\mathbf{M})^{-1}dz=\frac{1}{2\pi i}\int_{\Gamma}z^{j-1}(\mathbf{I}_p-z^{-1}\mathbf{M})^{-1}dz. \tag{20}$$

Lemma C.3 *Let $\mathbf{B}$ given by (18) be the matrix of coefficients of a stationary AR(p) process and let $\mathbf{v}=(v_1,\ldots,v_p)'\in\mathbb{R}^p$. For all $z\in\mathbb{C}$ such that $\sum_{i=0}^{\infty}|\langle\mathbf{v},\mathbf{B}^i\mathbf{u}\rangle||z^i|<\infty$, we have*

$$b(z)\sum_{i=0}^{\infty}\left\langle\mathbf{v},\mathbf{B}^i\mathbf{u}\right\rangle z^i=b(z)\left\langle\mathbf{v},(\mathbf{I}_p-z\mathbf{B})^{-1}\mathbf{u}\right\rangle=v(z), \tag{21}$$

where $v(z) = v_1 + v_2 z + \ldots + v_p z^{p-1}$*, and where* $b(z)$ *is the AR polynomial.*

Proof. As $\sum_{i=0}^{\infty} |\langle \mathbf{v}, \mathbf{B}^i \mathbf{u}\rangle||z^i| < 0$, we can change the order of summation in the left-hand side of (21)

$$(1 - \beta_1 z - \ldots - \beta_p z^p) \sum_{i=0}^{\infty} \left\langle \mathbf{v}, \mathbf{B}^i \mathbf{u} \right\rangle z^i = \left\langle \mathbf{v}, \left(\sum_{i=0}^{\infty} (1 - \beta_1 z - \ldots - \beta_p z^p) z^i \mathbf{B}^i \right) \mathbf{u} \right\rangle.$$

Define $\beta_0 = -1$, $\beta_k = 0$ for $k > p$. By direct algebraic manipulation,

$$\sum_{i=0}^{\infty} (1 - \beta_1 z - \ldots - \beta_p z^p) z^i \mathbf{B}^i = - \sum_{i=0}^{\infty} \left(\sum_{k=0}^{i} \beta_k \mathbf{B}^{i-k} \right) z^i := - \sum_{i=0}^{\infty} \mathbf{D}_i z^i.$$

The characteristic polynomial of $\mathbf{B}$ is given by $\phi(z) = \sum_{k=0}^{p} \beta_k z^{p-k}$. From the Cayley–Hamilton theorem, $\mathbf{B}$ is a root of ϕ, and, consequently for $i \geq p$,

$$\mathbf{D}_i = \mathbf{B}^{i-p} \sum_{k=0}^{i} \beta_k \mathbf{B}^{p-k} = \mathbf{B}^{i-p} \sum_{k=0}^{p} \beta_k \mathbf{B}^{p-k} = 0.$$

It remains to demonstrate that $\langle \mathbf{v}, \mathbf{D}_i \mathbf{u} \rangle = -v_{i+1}$ for $i = 0, \ldots, p-1$, which we show by induction. For $i = 0$, $\langle \mathbf{v}, \mathbf{D}_i \mathbf{u} \rangle = \beta_0 \langle \mathbf{v}, \mathbf{u} \rangle = -v_1$. When $i \geq 1$, matrices $\mathbf{D}_i$ satisfy $\mathbf{D}_i = \mathbf{B}\mathbf{D}_{i-1} + \beta_i \mathbf{I}_p$, therefore

$$\begin{aligned}
\langle \mathbf{v}, \mathbf{D}_i \mathbf{u} \rangle &= \langle \mathbf{v}, \mathbf{B}\mathbf{D}_{i-1} \mathbf{u} \rangle + \beta_i \langle \mathbf{v}, \mathbf{u} \rangle = \langle \mathbf{B}'\mathbf{v}, \mathbf{D}_{i-1} \mathbf{u} \rangle + \beta_i \langle \mathbf{v}, \mathbf{u} \rangle \\
&= \langle v_1 (\beta_1, \ldots, \beta_p)' + (0, v_2, \ldots, v_p)', \mathbf{D}_{i-1} \mathbf{u} \rangle + \beta_i \langle \mathbf{v}, \mathbf{u} \rangle = -v_1 \beta_i - v_{i+1} + v_1 \beta_i \\
&= -v_{i+1},
\end{aligned}$$

which completes the proof. □

Lemma C.4 *Let* $Z_1, Z_2, \ldots$ *be a sequence of i.i.d.* $\mathcal{N}(0,1)$ *random variables. Then for any integers* $l \neq 0$ *and* $k > 0$*, the following exponential probability bound holds for any* $x > 0$*:*

$$\mathbb{P}\left(\left| \sum_{t=1}^{k} Z_t Z_{t+l} \right| > kx \right) \leq 2 \exp\left(-\frac{1}{8} \frac{kx^2}{4+x} \right). \tag{22}$$

Proof. We will show that $\mathbb{P}\left(\sum_{t=1}^{k} Z_t Z_{t+l} > kx \right) \leq \exp\left(-\frac{1}{8} \frac{kx^2}{4+x} \right)$, which would then imply (22) by symmetry. By Markov's inequality, for any $x > 0$ and $\lambda > 0$, it holds that

$$\mathbb{P}\left(\sum_{t=1}^{k} Z_t Z_{t+l} > kx\right) \leq \exp\left(-kx\lambda\right) \mathrm{E}\exp\left(\lambda \sum_{t=1}^{k} Z_t Z_{t+l}\right).$$

By the convexity of $y \mapsto \exp(\lambda y)$ for any $\lambda > 0$, Theorem 1 in Vershynin (2011) implies

$$\mathrm{E}\exp\left(\lambda \sum_{t=1}^{k} Z_t Z_{t+l}\right) \leq \mathrm{E}\exp\left(4\lambda \sum_{t=1}^{k} Z_t \tilde{Z}_t\right),$$

where $\tilde{Z}_1, \ldots, \tilde{Z}_k$ are independent copies of $Z_1, \ldots, Z_k$. Using the independence and by direct computation (see also Craig (1936)), we get

$$\mathrm{E}\exp\left(4\lambda \sum_{t=1}^{k} Z_t \tilde{Z}_t\right) = \left(\mathrm{E}\exp\left(4\lambda Z_1 \tilde{Z}_1\right)\right)^k = \left(\mathrm{E}\exp\left(8\lambda^2 \tilde{Z}_1^2\right)\right)^k = \left(1 - 16\lambda^2\right)^{-\frac{1}{2}k}$$

provided that $0 < \lambda < \frac{1}{4}$, therefore $\mathbb{P}\left(\sum_{t=1}^{k} Z_t Z_{t+l} > kx\right) \leq \exp\left(-kx\lambda - \frac{k}{2}\log\left(1 - 16\lambda^2\right)\right)$. Taking $\lambda = \frac{-2+\sqrt{4+x^2}}{4x}$ minimises the right-hand side of this inequality. With this value of λ and using $\log(x) \leq x - 1$, we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{t=1}^{k} Z_t Z_{t+l} > kx\right) &\leq \exp\left(\frac{k}{4}\left(2 - \sqrt{x^2+4} + 2\log\left(\frac{1}{4}\left(\sqrt{x^2+4} + 2\right)\right)\right)\right) \\
&\leq \exp\left(\frac{k}{4}\left(2 - \sqrt{x^2+4} + \frac{1}{2}\left(\sqrt{x^2+4} + 2\right) - 2\right)\right) \\
&= \exp\left(\frac{k}{8}\left(2 - \sqrt{x^2+4}\right)\right) = \exp\left(-\frac{1}{8}\frac{kx^2}{2 + \sqrt{x^2+4}}\right) \\
&\leq \exp\left(-\frac{1}{8}\frac{kx^2}{4+x}\right),
\end{aligned}$$

which completes the proof. □

Lemma C.5 (Lemma 1 in Laurent and Massart (2000)) *Let $Z_1, Z_2, \ldots$ be a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables. For any integer $k > 0$ and $x > 0$, the following exponential probability* *bounds hold*

$$\mathbb{P}\left(\sum_{t=1}^{k} Z_t^2 \geq k + 2\sqrt{kx} + 2x\right) \leq \exp(-x), \tag{23}$$

$$\mathbb{P}\left(\sum_{t=1}^{k} Z_t^2 \leq k - 2\sqrt{kx}\right) \leq \exp(-x). \tag{24}$$

Proof of Theorem 2.1. For $\mathbf{C}_T = \sum_{t=1}^{T-1} \mathbf{Y}_t \mathbf{Y}_t'$ and $\mathbf{A}_T = \sum_{t=1}^{T-1} \varepsilon_{t+1} \mathbf{Y}_t$, we have $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{C}_T^{-1} \mathbf{A}_T$. Here the distribution of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ is invariant to the value of σ. As such, in the following, we assume $\sigma = 1$ for notational convenience. Consequently,

$$\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| \leq \lambda_{max}(\mathbf{C}_T^{-1}) \left\|\mathbf{A}_T\right\| = \lambda_{min}^{-1}(\mathbf{C}_T) \left\|\mathbf{A}_T\right\|, \tag{25}$$

where $\lambda_{min}(\mathbf{M})$ and $\lambda_{max}(\mathbf{M})$ denote, respectively, the smallest and the largest eigenvalues of a symmetric matrix $\mathbf{M}$. To provide an upper bound on $\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|$ given in Theorem 2.1, we will bound $\lambda_{min}(\mathbf{C}_T)$ from below and $\|\mathbf{A}_T\|$ from above, working on a set whose probability is large.

In the calculations below, we will repeatedly use the following representation of $\mathbf{Y}_t$, which follows from applying (17) recursively:

$$\mathbf{Y}_t = \mathbf{B}^t \mathbf{Y}_0 + \sum_{j=1}^{t} \varepsilon_{t-j+1} \mathbf{B}^{j-1} \mathbf{u}, \quad t = 1, \ldots, T. \tag{26}$$

In addition, to improve the presentational aspect of the proof, here we shall take $\mathbf{Y}_0 = \mathbf{0}$. All the results would go through (with minor modifications to handle the extra terms) if one instead assumes that $\mathbf{Y}_0$ is a realization from a stationary solution.

In the arguments below, we will show result more specific than (9), i.e.

$$\|\mathbf{A}_T\| \leq \left(32 \underline{b}^{-2} \sqrt{1 + \|\boldsymbol{\beta}\|^2}\right) p \log(T) \sqrt{(1 + \log(T + p))T}, \tag{27}$$

$$\lambda_{min}(\mathbf{C}_T) \geq \bar{b}^{-2} \left(T - p(1 + 32 \log(T)\sqrt{T})\right), \tag{28}$$

on the event

$$\mathcal{E}_T = \mathcal{E}_T^{(1)} \cap \mathcal{E}_T^{(2)} \cap \mathcal{E}_T^{(3)}, \tag{29}$$

where

$$\mathcal{E}_T^{(1)} = \bigcap_{1\le i<j\le p} \left\{ \left| \sum_{t=1}^{T-\max(i,j)} \varepsilon_t \varepsilon_{t+|i-j|} \right| < 32\log(T)\sqrt{T-\max(i,j)} \right\},$$
$$\mathcal{E}_T^{(2)} = \bigcap_{j=1}^{T} \left\{ \left| \sum_{t=1}^{T-j} \varepsilon_t \varepsilon_{t+j} \right| < 32\log(T)\sqrt{T-j} \right\},$$
$$\mathcal{E}_T^{(3)} = \left\{ \sum_{t=1}^{T-p} \varepsilon_t^2 > T-p-2\sqrt{\log(T)(T-p)} \right\}.$$

Finally, we will demonstrate that $\mathcal{E}_T$ satisfies

$$\mathbb{P}\left(\mathcal{E}_T\right) \ge 1 - \frac{5}{T}. \tag{30}$$

Thus, (25), (27), (28) and (30) combined together imply the statement of Theorem 2.1. The remaining part of the proof is split into three parts, in which we show (27), (28) and (30) in turn.

Upper bound for $\|\mathbf{A}_T\|$. The Euclidean norm satisfies $\|\mathbf{A}_T\| = \sup_{\mathbf{v}:\in\mathbb{R}^p,\|\mathbf{v}\|=1} |\langle \mathbf{v}, \mathbf{A}_T\rangle|$, therefore we consider inner products $\langle \mathbf{v}, \mathbf{A}_T\rangle$ where $\mathbf{v}\in\mathbb{R}^p$ is any unit vector. By (26),

$$\begin{aligned}\langle \mathbf{v}, \mathbf{A}_T\rangle &= \sum_{t=1}^{T-1} \langle \mathbf{v}, \mathbf{Y}_t\rangle \varepsilon_{t+1} = \sum_{t=1}^{T-1}\sum_{j=1}^{t} \left\langle \mathbf{v}, \mathbf{B}^{j-1}\mathbf{u}\right\rangle \varepsilon_{t-j+1}\varepsilon_{t+1} \\ &= \sum_{j=1}^{T-1} \left\langle \mathbf{v}, \mathbf{B}^{j-1}\mathbf{u}\right\rangle a_j,\end{aligned}$$

where $a_j = \sum_{t=j}^{T-1} \varepsilon_{t-j+1}\varepsilon_{t+1} = \sum_{t=1}^{T-j} \varepsilon_t\varepsilon_{t+j}$.

Lemma C.2 and Lemma C.3 applied to the right-hand side of the above equation yield

$$\begin{aligned}\sum_{j=1}^{T-1}\left\langle \mathbf{v}, \mathbf{B}^{j-1}\mathbf{u}\right\rangle a_j &= \frac{1}{2\pi i}\int_{\mathbb{T}}\left(\sum_{j=1}^{T-1} z^{j-1}a_j\right)\left\langle \mathbf{v}, (z\mathbf{I}_p - \mathbf{B})^{-1}\mathbf{u}\right\rangle dz \\ &= \frac{1}{2\pi i}\int_{\mathbb{T}}\left(\sum_{j=1}^{T-1} z^{j-1}a_j\right)\left(\sum_{j=1}^{p} z^{p-j}v_j\right) q(z)dz \\ &= \frac{1}{2\pi i}\int_{\mathbb{T}}\left(\sum_{j=0}^{T+p-1} z^{j}c_j\right) q(z)dz,\end{aligned}$$

where $q(z) = (z^p b(z^{-1}))^{-1}$ and $c_j = \sum_{i=0}^{j} a_{i+1}v_{p-j+i}$. Integrating by parts, we get

$$\frac{1}{2\pi i}\int_{\mathbb{T}}\left(\sum_{j=0}^{T+p-1} z^{j}c_j\right) q(z)dz = -\frac{1}{2\pi i}\int_{\mathbb{T}}\left(\sum_{j=0}^{T+p-1} z^{j+1}\frac{c_j}{j+1}\right) q'(z)dz,$$

where $q'(\cdot)$ is the derivative of $q(\cdot)$. Combining the calculations above and using the fact that $\mathbb{T} = \{z \in \mathbb{C} : \ |z| = 1\}$, Cauchy's inequality and Lemma C.1, we obtain

$$\left|\sum_{j=1}^{T-1}\left\langle \mathbf{v}, \mathbf{B}^{j-1}\mathbf{u}\right\rangle a_j\right| \le \sqrt{\sum_{j=0}^{T+p-1}\left(\frac{c_j}{j+1}\right)^2}\sqrt{\int_{\mathbb{T}}|q'(z)|^2 dm(z)}, \tag{31}$$

where we recall that $dm(z) = \frac{d|z|}{2\pi}$. To further bound the first term on the right-hand side of (31), we recall that on the event $\mathcal{E}_T$ coefficients $|a_j| \le 32\log(T)\sqrt{T}$, hence

$$\begin{aligned}\sqrt{\sum_{j=0}^{T+p-1}\left(\frac{c_j}{j+1}\right)^2} &= \sqrt{\sum_{j=0}^{T+p-1}\frac{1}{(j+1)^2}\left(\sum_{i=0}^{j} a_{i+1}v_{p-j+i}\right)^2} \\ &\le \max_{j=0,\ldots,T+p-1}|a_j|\sqrt{\sum_{j=0}^{T+p-1}\frac{1}{(j+1)^2}\left(\sum_{i=0}^{j}|v_{p-j+i}|\right)^2} \\ &\le 32\log(T)\sqrt{T}\sqrt{\sum_{j=0}^{T+p-1}\frac{j+1}{(j+1)^2}} \\ &\le 32\log(T)\sqrt{(1+\log(T+p))T}.\end{aligned}$$

For the second term in (31), we calculate the derivative

$$q'(z) = -\frac{pz^{p-1} - \sum_{j=1}^{p}(p-j)\beta_j z^{p-j-1}}{(z^p b(z^{-p}))^2}$$

and use Lemma C.1 to bound

$$\begin{aligned}
\sqrt{\int_{\mathbb{T}} |q'(z)|^2 dm(z)} &= \sqrt{\int_{\mathbb{T}} \left| \frac{pz^{p-1} - \sum_{j=1}^{p}(p-j)\beta_j z^{p-j}}{(z^p b(z^{-p}))^2} \right|^2 dm(z)} \\
&\leq \frac{\sqrt{\int_{\mathbb{T}} \left| pz^{p-1} - \sum_{j=1}^{p}(p-j)\beta_j z^{p-j} \right|^2 dm(z)}}{\min_{|z|=1} |(z^p b(z^{-p}))|^2} \\
&= \underline{b}^{-2} \sqrt{\left(p^2 + \sum_{j=1}^{p} (p-j)^2 \beta_j^2 \right)} \leq \underline{b}^{-2} p \sqrt{1 + \|\boldsymbol{\beta}\|^2}.
\end{aligned}$$

Combining the bounds on the two terms, we obtain

$$\sum_{j=1}^{T-1} \left\langle \mathbf{v}, \mathbf{B}^{j-1}\mathbf{u} \right\rangle a_j \leq \left(32\underline{b}^{-2} \sqrt{1 + \|\boldsymbol{\beta}\|^2} \right) p \log(T) \sqrt{(1 + \log(T+p))T}.$$

Taking supremum over $\mathbf{v} \in \mathbb{R}^p$ such that $\|\mathbf{v}\| = 1$ proves (27).

Lower bound for $\lambda_{min}(\mathbf{C}_T)$. Let $\mathbf{v} = (v_1, \ldots, v_p)'$ be a unit vector in $\mathbb{R}^p$. We begin the proof by establishing the following inequality

$$\langle \mathbf{v}, \mathbf{C}_T \mathbf{v} \rangle \geq \overline{b}^{-2} \sum_{i,j=1}^{p} v_i v_j \sum_{t=1}^{T-1} \varepsilon_{t-j+1} \varepsilon_{t-i+1}, \tag{32}$$

where $\varepsilon_t = 0$ for $t \leq 0$ and $\overline{b} = \max_{z \in \mathbb{T}} |b(z)|$. By Lemma C.1 and (26), we rewrite the quadratic form on the left-hand side of (32) to

$$\langle \mathbf{v}, \mathbf{C}_T \mathbf{v} \rangle = \sum_{t=1}^{T-1} \langle \mathbf{v}, \mathbf{Y}_t \rangle^2 \tag{33}$$

$$= \int_{\mathbb{T}} \left| \sum_{t=1}^{T-1} \left\langle \mathbf{v}, \sum_{j=1}^{t} \varepsilon_j \mathbf{B}^{t-j} \mathbf{u} \right\rangle z^t \right|^2 dm(z) \tag{34}$$

$$= \int_{\mathbb{T}} \left| \sum_{t=1}^{T-1} \sum_{j=1}^{T-1} \varepsilon_j \omega_{t-j} z^t \right|^2 dm(z) \tag{35}$$

where $\omega_j = \langle \mathbf{v}, \mathbf{B}^j \mathbf{u} \rangle$ for $j \geq 0$, $\omega_j = 0$ for $j < 0$. Changing the order of summation and by a simple substitution we get

$$\sum_{t=1}^{T-1} \sum_{j=1}^{T-1} \varepsilon_j \omega_{t-j} z^t = \sum_{j=1}^{T-1} \varepsilon_j z^j \sum_{t=1}^{T-1} \omega_{t-j} z^{t-j} = \sum_{j=1}^{T-1} \varepsilon_j z^j \sum_{t=0}^{T-j-1} \omega_t z^t. \tag{36}$$

Using the definition of ω_j, the fact that all eigenvalues of $\mathbf{B}$ have modulus strictly lower than one and Lemma C.3, (36) simplifies to

$$\begin{aligned}
\sum_{j=1}^{T-1} \varepsilon_j z^j \sum_{t=0}^{T-j-1} \omega_t z^t &= \sum_{j=1}^{T-1} \varepsilon_j z^j \left\langle \mathbf{v}, (\mathbf{I}_p - (\mathbf{B}z)^{T-j})(\mathbf{I}_p - \mathbf{B}z)^{-1} \mathbf{u} \right\rangle \\
&= \sum_{j=1}^{T-1} \varepsilon_j \left(z^j \left\langle \mathbf{v}, (\mathbf{I}_p - \mathbf{B}z)^{-1} \mathbf{u} \right\rangle - z^T \left\langle \mathbf{B}^{T-j} \mathbf{v}, (\mathbf{I}_p - \mathbf{B}z)^{-1} \mathbf{u} \right\rangle \right) \\
&= b(z)^{-1} \sum_{j=1}^{T-1} \varepsilon_j \left(z^j v(z) - z^T w_j(z) \right),
\end{aligned}$$

where $v(z) = \sum_{k=1}^{p} v_k z_{k-1}$ and $w_j(z) = \sum_{k=1}^{p} (\mathbf{B}^{T-j} v)_k z^{k-1}$ for $j = 0, \ldots, T-1$. The equation above, (33) and (36) combined together imply the following inequality

$$\begin{aligned}
\langle \mathbf{v}, \mathbf{C}_T \mathbf{v} \rangle &= \int_{\mathbb{T}} \left| b(z)^{-1} \sum_{j=1}^{T-1} \varepsilon_j \left(z^j v(z) - z^T w_j(z) \right) \right|^2 dm(z) \\
&\geq \bar{b}^{-2} \int_{\mathbb{T}} \left| \sum_{j=1}^{T-1} \varepsilon_j \left(z^j v(z) - z^T w_j(z) \right) \right|^2 dm(z).
\end{aligned}$$

Observe that $\sum_{j=1}^{T-1} \varepsilon_j \left(z^j v(z) - z^T w_j(z)\right) = \sum_{j=1}^{T-1} \varepsilon_j \left(z^j v(z) - z^T w_j(z)\right) = \sum_{t=1}^{T+p-1} c_t z^t$ is a trigonometric polynomial, therefore by Lemma C.1 and simple algebra

$$\int_{\mathbb{T}} \left| \sum_{j=1}^{T-1} \varepsilon_j \left(z^j v(z) - z^T w_j(z)\right) \right| dm(z) = \sum_{t=1}^{T+p-1} |c_t|^2 \geq \sum_{t=1}^{T-1} |c_t|^2 = \sum_{t=1}^{T-1} \left(\sum_{j=1}^{p} v_j \varepsilon_{t-j+1} \right)^2 = \\ = \sum_{i,j=1}^{p} v_j v_i \sum_{t=1}^{T-1} \varepsilon_{t-j+1} \varepsilon_{t-i+1},$$

which proves (32).

We are now in a position to bound $\langle \mathbf{v}, \mathbf{C}_T \mathbf{v} \rangle$ from below. Rearranging terms in (32) yields

$$\begin{aligned} \langle \mathbf{v}, \mathbf{C}_T \mathbf{v} \rangle &\geq \bar{b}^{-2} \left(\sum_{i=1}^{p} v_i^2 \sum_{t=1}^{n-i} \varepsilon_t^2 + \sum_{1 \leq i < j \leq p} v_i v_j \sum_{t=1}^{T-\max(i,j)} \varepsilon_t \varepsilon_{t+|j-i|} \right) \\ &\geq \bar{b}^{-2} \left(\sum_{t=1}^{T-p} \varepsilon_t^2 \sum_{i=1}^{p} v_i^2 - \max_{1 \leq i < j \leq p} \left| \sum_{t=1}^{T-\max(i,j)} \varepsilon_t \varepsilon_{t+|j-i|} \right| \left(\left(\sum_{i=1}^{p} |v_i| \right)^2 - \sum_{i=1}^{p} v_i^2 \right) \right) \\ &\geq \bar{b}^{-2} \left(\sum_{t=1}^{T-p} \varepsilon_t^2 - (p-1) \max_{1 \leq i < j \leq p} \left| \sum_{t=1}^{T-\max(i,j)} \varepsilon_t \varepsilon_{t+|j-i|} \right| \right). \end{aligned}$$

Recalling the definition of $\mathcal{E}_T$, we conclude that on this event

$$\begin{aligned} \langle \mathbf{v}, \mathbf{C}_T \mathbf{v} \rangle &\geq \bar{b}^{-2} \left(T - p - 2\sqrt{\log(T)(T-p)} - (p-1) 32 \log(T) \sqrt{T} \right) \\ &\geq \bar{b}^{-2} \left(T - p(1 + 32 \log(T) \sqrt{T}) \right). \end{aligned}$$

Taking infimum over $\mathbf{v} \in \mathbb{R}^p$ such that $\|\mathbf{v}\| = 1$ in the inequality above proves (28).

Lower bound for $\mathbb{P}(\mathcal{E}_T)$. Recalling (29) and using a simple Bonferroni bound, we get

$$\begin{aligned}\mathbb{P}(\mathcal{E}_T^c) \leq p^2 \max_{1\leq i<j\leq p} \mathbb{P}\left(\left|\sum_{t=1}^{T-\max(i,j)} \varepsilon_t\varepsilon_{t+|i-j|}\right| \geq 32\log(T)\sqrt{T-\max(i,j)}\right) \\ + T\max_{1\leq j\leq T} \mathbb{P}\left(\left|\sum_{t=1}^{T-j}\varepsilon_t\varepsilon_{t+j}\right| < 32\log(T)\sqrt{T-j}\right) \\ + \mathbb{P}\left(\sum_{t=1}^{T-p}\varepsilon_t^2 > T-p-2\sqrt{\log(T)(T-p)}\right) \\ := p^2\max_{1\leq i<j\leq p} P_{i,j}^{(1)} + T\max_{1\leq j\leq T} P_j^{(2)} + P^{(3)}.\end{aligned}$$

Lemma C.4 implies that

$$\begin{aligned}P_{i,j}^{(1)} &\leq 2\exp\left(-\frac{1}{8}\frac{(32\log(T))^2}{4+(\sqrt{T-\max(i,j)})^{-1}32\log(T)}\right) \leq 2\exp\left(-2\log(T)\right) = \frac{2}{T^2}, \\ P_j^{(2)} &\leq 2\exp\left(-\frac{1}{8}\frac{(32\log(T))^2}{4+(\sqrt{T-j})^{-1}32\log(T)}\right) \leq 2\exp\left(-2\log(T)\right) = \frac{2}{T^2}.\end{aligned}$$

Moreover, by Lemma C.5, $P^{(3)} \leq \exp(-\log(T)) = \frac{1}{T}$, hence, given that $p^2 < T$, we have $\mathbb{P}(\mathcal{E}_T^c) \leq \frac{5}{T}$, which completes the proof. □

C.3 Proof of Theorem 2.2

In the proof below, we shall focus on the case where F_T^M consists of randomly drawn intervals (which is what Algorithm 2 does when p is large). For the case where all sub-intervals of $[1,p]$ are used, the same arguments would go through, because Algorithm 2 then produces a larger set F_T^M compared to the approach of random drawing.

We now split the proof into four steps.

Step 1. Consider the event $\left\{\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\| \leq \kappa_1(\underline{b}/\overline{b})^2\left\|\boldsymbol{\beta}\right\| \frac{p\log(T)\sqrt{\log(T+p)}}{\sqrt{T}-\kappa_2 p\log(T)}\right\}$ where κ_1, κ_2 are as in Theorem 2.1. Assumption (A3) implies that $\underline{b}/\overline{b}$ and $\|\boldsymbol{\beta}\|$ are bounded from above by constants. Furthermore, by Assumption (A2), $p \leq c_1 T^{\theta}$, which implies that

$$\kappa_1(\underline{b}/\overline{b})^2 \|\boldsymbol{\beta}\| \frac{p\log(T)\sqrt{\log(T+p)}}{\sqrt{T}-\kappa_2 p\log(T)} \leq c_3 T^{\theta-1/2}(\log(T))^{3/2} = c_3\underline{\lambda}_T =: \lambda_T \tag{37}$$

for some constant $c_3 > 0$ and a sufficiently large T. Define now

$$A_T = \left\{ \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\| \leq \lambda_T \right\} \tag{38}$$

By Theorem 2.1,

$$\mathbb{P}(A_T) \geq \mathbb{P}\left(\left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\| \leq \kappa_1(\underline{b}/\overline{b})^2 \|\boldsymbol{\beta}\| \frac{p\log(T)\sqrt{\log(T+p)}}{\sqrt{T}-\kappa_2 p\log(T)} \right) \geq 1 - \kappa_3 T^{-1}, \tag{39}$$

for some constant $\kappa_3 > 0$.

Step 2. For $j = 1, \ldots, q$, define the intervals

$$\mathcal{I}_j^L = (\tau_j - \delta_T/3, \tau_j - \delta_T/6) \tag{40}$$

$$\mathcal{I}_j^R = (\tau_j + \delta_T/6, \tau_j + \delta_T/3) \tag{41}$$

Recall that F_T^M is the set of M randomly drawn intervals with endpoints in $\{1, \ldots, p\}$. Denote by $[s_1, e_1], \ldots, [s_M, e_M]$ the elements of F_T^M and let

$$D_T^M = \left\{ \forall j = 1, \ldots, q, \exists k \in \{1, \ldots, M\}, \text{ s.t. } s_k \times e_k \in \mathcal{I}_j^L \times \mathcal{I}_j^R \right\}. \tag{42}$$

We have that

$$\begin{aligned} \mathbb{P}\left((D_T^M)^c\right) &\leq \sum_{j=1}^{q} \Pi_{m=1}^{M} \left(1 - \mathbb{P}\left(s_m \times e_m \in \mathcal{I}_j^L \times \mathcal{I}_j^R\right)\right) \\ &\leq q\left(1 - \frac{\delta_T^2}{6^2 p^2}\right)^M \leq \frac{p}{\delta_T}\left(1 - \frac{\delta_T^2}{36p^2}\right)^M. \end{aligned}$$

Therefore, $\mathbb{P}\left(A_T \cap D_T^M\right) \geq 1 - \kappa_3 T^{-1} - p\delta_T^{-1}(1 - \delta_T^2 p^{-2}/36)^M \to 1$. Note that the same conclusion still holds if F_T^M contains all the intervals with endpoints in $\{1, \ldots, p\}$. In the remainder of the proof, assume that A_T and D_T^M all hold.

Note that Assumption (A4) implies that there exists $\underline{c} > 0$ such that $\delta_T^{1/2}\underline{\alpha}_T > \underline{c}\lambda_T$ for all sufficiently large T. We are now in the position to specify the constants explicitly as

$$C_1 = 2\sqrt{C_3} + c_3, \quad C_2 = \frac{1}{\sqrt{6}} - \frac{1}{\underline{c}}, \quad C_3 = (4\sqrt{2}+6)c_3^2,$$

where c_3 is in Equation (37).

Step 3. We focus on a generic interval $[s,e]$ such that

$$\exists j \in \{1,\dots,q\},\ \exists k \in \{1,\dots,M\},\ \text{s.t. } [s_k, e_k] \subset [s,e] \text{ and } s_k \times e_k \in \mathcal{I}_j^L \times \mathcal{I}_j^R. \tag{43}$$

Fix such an interval $[s,e]$ and let $j \in \{1,\dots,q\}$ and $k \in \{1,\dots,M\}$ be such that (43) is satisfied. Let $b_k^* = \operatorname{argmax}_{s_k \le b \le e_k} \mathcal{C}_{s_k,e_k}^b\left(\hat{\boldsymbol{\beta}}\right)$. By construction, $[s_k, e_k]$ satisfies $\tau_j - s_k + 1 \ge \delta_T/6$ and $e_k - \tau_j > \delta_T/6$. Let

$$\mathcal{M}_{s,e} = \left\{m : [s_m, e_m] \in F_T^M, [s_m, e_m] \subset [s,e]\right\},$$
$$\mathcal{O}_{s,e} = \{m \in \mathcal{M}_{s,e} : \max_{s_m \le b < e_m} \mathcal{C}_{s_m,e_m}^b\left(\hat{\boldsymbol{\beta}}\right) > \zeta_T\}.$$

Our first aim is to show that $\mathcal{O}_{s,e}$ is non-empty. This follows from Lemma 2 in Baranowski et al. (2019), the Cauchy–Schwarz inequality, and the calculation below, as

$$\begin{aligned}
\mathcal{C}_{s_k,e_k}^{b_k^*}\left(\hat{\boldsymbol{\beta}}\right) &\ge \mathcal{C}_{s_k,e_k}^{\tau_j}\left(\hat{\boldsymbol{\beta}}\right) \\
&\ge \mathcal{C}_{s_k,e_k}^{b_k^*}(\boldsymbol{\beta}) - \lambda_T \ge \left(\frac{\delta_T}{6}\right)^{1/2} |\alpha_j \tau_j^{-1}| - \lambda_T \ge \left(\frac{\delta_T}{6}\right)^{1/2} \underline{\alpha}_T - \lambda_T \\
&= \left(\frac{1}{\sqrt{6}} - \frac{\lambda_T}{\delta_T^{1/2}\underline{\alpha}_T}\right) \delta_T^{1/2}\underline{\alpha}_T \ge \left(\frac{1}{\sqrt{6}} - \frac{1}{\underline{c}}\right) \delta_T^{1/2}\underline{\alpha}_T = C_2 \delta_T^{1/2}\underline{\alpha}_T > \zeta_T.
\end{aligned}$$

Let $m^* = \operatorname{argmin}_{m \in \mathcal{O}_{s,e}} (e_m - s_m + 1)$ and $b^* = \operatorname{argmax}_{s_{m^*} \le b < e_{m^*}} \mathcal{C}_{s_{m^*},e_{m^*}}^b\left(\hat{\boldsymbol{\beta}}\right)$. Observe that $[s_{m^*}, e_{m^*})$ must contain at least one change in $\hat{\boldsymbol{\beta}}$. Indeed, if this were not the case, we would have $\mathcal{C}_{s_{m^*},e_{m^*}}^b(\boldsymbol{\beta}) = 0$ and

$$\mathcal{C}_{s_{m^*},e_{m^*}}^{b^*}\left(\hat{\boldsymbol{\beta}}\right) = |\mathcal{C}_{s_{m^*},e_{m^*}}^{b^*}\left(\hat{\boldsymbol{\beta}}\right) - \mathcal{C}_{s_{m^*},e_{m^*}}^{b^*}(\boldsymbol{\beta})| \le \lambda_T < \frac{C_1}{c_3}\lambda_T = C_1 \underline{\lambda}_T \le \zeta_T,$$

which contradicted $\mathcal{C}_{s_{m^*},e_{m^*}}^{b^*}\left(\hat{\boldsymbol{\beta}}\right) > \zeta_T$. On the other hand, $[s_{m^*}, e_{m^*})$ cannot contain more than one change-points, because $e_{m^*} - s_{m^*} + 1 \le e_k - s_k + 1 \le \delta_T$.

Without loss of generality, assume $\tau_j \in [s_{m^*}, e_{m^*})$. Let $\eta_L = \tau_j - s_{m^*} + 1$, $\eta_R = e_{m^*} - \tau_j$ and $\eta_T = (C_1/c_3 - 1)^2 \alpha_j^2 \tau_j^{-2} \lambda_T^2$. We claim that $\min(\eta_L, \eta_R) > \eta_T$, because otherwise $\min(\eta_L, \eta_R) \leq \eta_T$ and Lemma 2 in Baranowski et al. (2019) would have implied

$$\begin{aligned}\mathcal{C}_{s_{m^*},e_{m^*}}^{b^*}\left(\hat{\boldsymbol{\beta}}\right) &\leq \mathcal{C}_{s_{m^*},e_{m^*}}^{b^*}(\boldsymbol{\beta}) + \lambda_T \leq \mathcal{C}_{s_{m^*},e_{m^*}}^{\tau_j}(\boldsymbol{\beta}) + \lambda_T \leq \eta_T^{1/2}|\alpha_j \tau_j^{-1}| + \lambda_T \\ &= (C_1/c_3 - 1 + 1)\lambda_T = C_1 \underline{\lambda}_T < \zeta_T,\end{aligned}$$

which contradicted $\mathcal{C}_{s_{m^*},e_{m^*}}^{b^*}\left(\hat{\boldsymbol{\beta}}\right) > \zeta_T$.

We are now in the position to prove $|b^* - \tau_j| \leq C_3 \underline{\lambda}_T \underline{\alpha}_T^{-2}$. Our aim is to find ϵ_T such that for any $b \in \{s_{m^*}, s_{m^*}+1, \ldots, e_{m^*}-1\}$ with $|b - \tau_j| > \epsilon_T$, we always have

$$\left\{\mathcal{C}_{s_{m^*},e_{m^*}}^{\tau_j}\left(\hat{\boldsymbol{\beta}}\right)\right\}^2 - \left\{\mathcal{C}_{s_{m^*},e_{m^*}}^{b}\left(\hat{\boldsymbol{\beta}}\right)\right\}^2 > 0. \tag{44}$$

This would then imply that $|b^* - \tau_j| \leq \epsilon_T$. By expansion and rearranging the terms, we see that (44) is equivalent to

$$\begin{aligned}\langle \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{\tau_j}\rangle^2 - \langle \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{b}\rangle^2 &> \langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{b}\rangle^2 - \langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{\tau_j}\rangle^2 \\ &+ 2\Big\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{b}\langle \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{b}\rangle - \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{\tau_j}\langle \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{\tau_j}\rangle\Big\rangle.\end{aligned} \tag{45}$$

Here $\boldsymbol{\psi}_{s,e}^b$ (with $1 \leq s < b < e \leq p$) is a p-dimensional vector, with its s-th to b-th component being $\sqrt{\frac{e-b}{(e-s+1)()b-s+1}}$, its $b+1$-th to e-th component being $\sqrt{\frac{b-s+1}{(e-s+1)(e-b)}}$, and the remaining elements being 0. In the following, we assume that $b \geq \tau_j$. The case that $b < \tau_j$ can be handled in a similar fashion. By Lemma 4 in Baranowski et al. (2019), we have

$$\begin{aligned}\langle \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{\tau_j}\rangle^2 - \langle \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{b}\rangle^2 &= (\mathcal{C}_{s^*,e^*}^{\tau_j}(\boldsymbol{\beta}))^2 - (\mathcal{C}_{s_{m^*},e_{m^*}}^{b}(\boldsymbol{\beta}))^2 \\ &= \frac{|b - \tau_j|\eta_L}{|b - \tau_j| + \eta_L}(\alpha_j \tau_j^{-1})^2 =: \kappa.\end{aligned}$$

In addition, since we assume event A_T,

$$\begin{aligned}&\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{b}\rangle^2 - \langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{\tau_j}\rangle^2 \leq \lambda_T^2, \\ &2\Big\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{b}\langle \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{b}\rangle - \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{\tau_j}\langle \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{\tau_j}\rangle\Big\rangle \\ &\qquad \leq 2\|\boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{b}\langle \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{b}\rangle - \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{\tau_j}\langle \boldsymbol{\beta}, \boldsymbol{\psi}_{s_{m^*},e_{m^*}}^{\tau_j}\rangle\|_2 \lambda_T = 2\kappa^{1/2}\lambda_T,\end{aligned}$$

where the final equality is also implied by Lemma 4 in Baranowski et al. (2019). Consequently, (45) can be deducted from the stronger inequality $\kappa - 2\lambda_T \kappa^{1/2} - \lambda_T^2 > 0$. This quadratic inequality is implied by $\kappa > (\sqrt{2}+1)^2\lambda_T^2$, and could be restricted further to

$$\frac{2|b-\tau_j|\eta_L}{|b-\tau_j|+\eta_L} \geq \min(|b-\tau_j|, \eta_L) > (4\sqrt{2}+6)(\alpha_j\tau_j^{-1})^{-2}\lambda_T^2 = C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2. \tag{46}$$

But since

$$\eta_L \geq \eta_T = (C_1/c_3-1)^2(\alpha_j\tau_j^{-1})^{-2}\lambda_T^2 = (2\sqrt{C_3}/c_3)^2(\alpha_j\tau_j^{-1})^{-2}\lambda_T^2 > C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2,$$

we see that (46) is implied by $|b-\tau_j| > C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2$. To sum up, $|b^*-\tau_j| > C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2$ would result in (44), a contradiction. So we have proved that $|b^*-\tau_j| \leq C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2$.

Step 4. With the arguments above valid on the event $A_T \cap B_T \cap D_T^M$, we can now proceed with the proof of the theorem. At the start of Algorithm 1, we have $s=1$ and $e=p$ and, provided that $q \geq 1$, condition (43) is satisfied. Therefore the algorithm detects a change-point b^* in that interval such that $|b^*-\tau_j| \leq C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2$. By construction, we also have that $|b^*-\tau_j| < 2/3\delta_T$. This in turn implies that for all $l=1,\ldots,q$ such that $\tau_l \in [s,e]$ and $l \neq j$ we have either $\mathcal{I}_l^L, \mathcal{I}_l^R \subset [s,b^*]$ or $\mathcal{I}_l^L, \mathcal{I}_l^R \subset [b^*+1,e]$. Therefore (43) is satisfied within each segment containing at least one change-point. Note that before all q change points are detected, each change point will not be detected twice. To see this, we suppose that τ_j has already been detected by b, then for all intervals $[s_k,e_k] \subset [\tau_j - C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2+1, \tau_j - C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2 + 2/3\delta_T + 1] \cup [\tau_j + C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2 - 2/3\delta_T, \tau_j + C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2]$, Lemma 2 in Baranowski et al. (2019), together with the definition of A_T, guarantee that

$$\begin{aligned}\max_{s_k \leq b < e} \mathcal{C}_{s_k,e_k}^b\left(\hat{\boldsymbol{\beta}}\right) &\leq \max_{s \leq b < e} \mathcal{C}_{s_k,e_k}^b\left(\boldsymbol{\beta}\right) + \lambda_T \\ &\leq \sqrt{C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2}\alpha_j\tau_j^{-1} + \sqrt{C_3(\alpha_{j+1}\tau_{j+1}^{-1})^{-2}\underline{\lambda}_T^2}\alpha_{j+1}\tau_{j+1}^{-1} + \lambda_T \\ &< (2\sqrt{C_3}/c_3+1)\lambda_T = C_1\underline{\lambda}_T < \zeta_T.\end{aligned}$$

Once all the change-points have been detected, we then only need to consider $[s_k,e_k]$ such that

$$[s_k,e_k] \subset [\tau_j - C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2+1, \tau_{j+1} + C_3(\alpha_{j+1}\tau_{j+1}^{-1})^{-2}\underline{\lambda}_T^2]$$

for $j=1,\ldots,q$. For such intervals, we have, by Lemmas 2 and 3 of Baranowski et al. (2019)

$$\begin{aligned}\max_{s_k \leq b < e_k} \mathcal{C}_{s_k,e_k}^b\left(\hat{\boldsymbol{\beta}}\right) &\leq \max_{s \leq b < e} \mathcal{C}_{s_k,e_k}^b\left(\boldsymbol{\beta}\right) + \lambda_T \\ &\leq \sqrt{C_3(\alpha_j\tau_j^{-1})^{-2}\underline{\lambda}_T^2}\alpha_j\tau_j^{-1} + \sqrt{C_3(\alpha_{j+1}\tau_{j+1}^{-1})^{-2}\underline{\lambda}_T^2}\alpha_{j+1}\tau_{j+1}^{-1} + \lambda_T \leq C_1\underline{\lambda}_T < \zeta_T.\end{aligned}$$

Hence no further scales is detected and the algorithm terminates. □

C.4 Proof of Theorem 2.3

The proof of Theorem 2.3 is similar to that of Theorem 2.2. In the following, we shall still divide the proof into four steps as before, but focus on the main differences.

Step 1. Let $\{\rho_h : h \in \mathbb{Z}\}$ the true auto-correlation function of $\{X_t\}$ and $\hat{\rho}_h$ be its sample version (without de-meaning). Let $\boldsymbol{\rho} = (\rho_0, \dots, \rho_p)'$. First, we note that for $\alpha > 2$, the distribution of the innovations has finite second moment. It then follows from Anderson and Walker (1964) that $\hat{\boldsymbol{\rho}} - \boldsymbol{\rho} = O_p(T^{-1/2})$. The least-square estimator for AR(p) can be written as

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \sum_{i=p}^{T-1} X_i^2 & \dots & \sum_{i=p}^{T-1} X_i X_{i-p+1} \\ & \ddots & \\ \sum_{i=1}^{T-p} X_i X_{i+p-1} & \dots & \sum_{i=1}^{T-p} X_i^2 \end{bmatrix}_{p\times p}^{-1} \begin{bmatrix} \sum_{i=p}^{T-1} X_i X_{i+1} \\ \vdots \\ \sum_{i=1}^{T-p} X_i X_{i+p} \end{bmatrix}.$$

This is asymptotically equivalent to

$$\begin{bmatrix} \hat{\rho}_0 & \dots & \hat{\rho}_{p-1} \\ & \ddots & \\ \hat{\rho}_{p-1} & \dots & \hat{\rho}_0 \end{bmatrix}^{-1} \begin{bmatrix} \hat{\rho}_p \\ \vdots \\ \hat{\rho}_1 \end{bmatrix},$$

which converges to $\boldsymbol{\beta}$ at $O_p(T^{-1/2})$ in view of the Yule–Walker equations. Now for $0 < \alpha \le 2$, despite infinite second moment in the innovations thus the time series, the auto-correlation function is still well-defined, in the sense of Davis and Resnick (1986). It follows from Hannan and Kanter (1977) that for any sufficiently small $\epsilon > 0$, $T^{1/\alpha-\epsilon}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \to 0$ in probability. See also Yohai and Maronna (1977) and Davis and Resnick (1986). In conclusion, we have that $T^{\max(1/2,1/\alpha)-\epsilon}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \to 0$ in probability.

Steps 2 and 3. The following arguments are simpler, due to the fact that p is fixed. Because we go through all the intervals $[s, e]$ over $\{1, \dots, p\}$, we could see that under the event that $T^{\max(1/2,1/\alpha)-\epsilon}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty < 1$ (N.B. here the norm does not matter, as p is fixed), for any $j = 1, \dots, q$, and taking $C_1 = \sqrt{p}$ and $C_2 = 1/2$,

$$\max_{\tau_j \le b < \tau_j+1} \mathcal{C}^b_{\tau_j,\tau_j+1}(\boldsymbol{\beta}) \ge \underline{\alpha}_T/\sqrt{2} - 2\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty > C_2 \underline{\alpha}_T.$$

On the other hand, for all the intervals $[s, e]$ that do not include any of the change-points $\{\tau_1, \dots, \tau_q\}$,

$$\max_{s \le b < e} \mathcal{C}^b_{s,e}(\boldsymbol{\beta}) \le \sqrt{p}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty < C_1 T^{-\max(1/2,1/\alpha)+\epsilon}.$$

Step 4. We shall now proceed with the proof under the event that $T^{\max(1/2,1/\alpha)-\epsilon}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty < 1$, which happens with probability one as $T \to \infty$. At the start of Algorithm 1, we have $s = 1$ and $e = p$. Since we pick the threshold $\zeta_T < C_2\underline{\alpha}_T$, and we consider only the narrowest intervals with the corresponding contrasts (i.e. CUSUM-type statistic) over the threshold, we would end up considering all $[\tau_j, \tau_j + 1]$ for $j \in \{1, \dots, q\}$. Notice that before all the q change-points are detected, we would not consider other longer intervals, because of the nature of Algorithm 1. In addition, we will not consider intervals without any change because their corresponding contrast values would be below the threshold, as proved in the previous step. Once all the changes are detected, we then only need to consider the intervals located in between consecutive change-points, which all have corresponding contrast values smaller than $C_1 T^{\epsilon-\max(1/2,1/\alpha)}$, thus the threshold ζ_T. Hence the algorithm would terminate with no further scales detected.

References

A. Anastasiou, Y. Chen, H. Cho, and P. Fryzlewicz. **breakfast**: Methods for fast multiple change-point detection and estimation, 2021. URL `https://CRAN.R-project.org/package=breakfast`. R package version 2.2.

T. W. Anderson and A. M. Walker. On the Asymptotic Distribution of the Autocorrelations of a Sample from a Linear Stochastic Process. *The Annals of Mathematical Statistics*, 35(3):1296 – 1303, 1964.

R. Baranowski, Y. Chen, and P. Fryzlewicz. Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society Series B*, 81:649–672, 2019.

H. Cho and P. Fryzlewicz. Multiple change point detection under serial dependence: Wild contrast maximisation and gappy Schwarz algorithm. *Preprint*, 2021.

C.C. Craig. On the frequency function of xy. *The Annals of Mathematical Statistics*, 7(1):1–15, 1936.

Richard Davis and Sidney Resnick. Limit Theory for the Sample Covariance and Correlation Functions of Moving Averages. *The Annals of Statistics*, 14(2):533 – 558, 1986.

J. Duoandikoetxea. *Fourier Analysis*, volume 29 of *Graduate Studies in Mathematics*. American Mathematical Society, 2001.

P. Fearnhead and P. Clifford. On-line inference for hidden Markov models via particle filters. *J. R. Statist. Soc. B*, 65:887–899, 2003.

P. Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, 42:2243–2281, 2014.

Edward James Hannan and Marek Kanter. Autoregressive processes with infinite variance. *Journal of Applied Probability*, 14(2):411–415, 1977.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28:1302–1338, 2000.

J. O Ruanaidh and W. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer, 1996.

R. Vershynin. A simple decoupling inequality in probability theory. Technical report, University of Michigan, 2011. URL https://www.math.uci.edu/~rvershyn/papers/decoupling-simple.pdf.

Victor J Yohai and Ricardo A Maronna. Asymptotic behavior of least-squares estimates for autoregressive processes with infinite variances. *The Annals of Statistics*, 5(3):554–560, 1977.