Ann Inst Stat Math (2019) 71:275–306
https://doi.org/10.1007/s10463-017-0642-9

Frequentist model averaging for threshold models

Yan Gao[1,2] · Xinyu Zhang[2] · Shouyang Wang[2] · Terence Tai-leung Chong[3] · Guohua Zou[4]

Received: 1 July 2015 / Revised: 28 November 2017 / Published online: 12 January 2018

Abstract This paper develops a frequentist model averaging approach for threshold model specifications. The resulting estimator is proved to be asymptotically optimal in the sense of achieving the lowest possible squared errors. In particular, when combining estimators from threshold autoregressive models, this approach is also proved to be asymptotically optimal. Simulation results show that for the situation where the existing model averaging approach is not applicable, our proposed model averaging approach has a good performance; for the other situations, our proposed model averaging approach performs marginally better than other commonly used model selection and model averaging methods. An empirical application of our approach on the US unemployment data is given.

Keywords Asymptotic optimality · Generalized cross-validation · Model averaging, Threshold model

1 Introduction

Threshold models have developed rapidly over the past three decades since the pioneering studies of Tong and Lim (1980) and Tong (1983, 1990). Chan (1993) studied

✉ Xinyu Zhang
xinyu@amss.ac.cn

[1] Department of Statistics, College of Science, Minzu University of China, Beijing 100081, China

[2] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[3] Department of Economics, The Chinese University of Hong Kong, Shatin, Hong Kong

[4] School of Mathematical Sciences, Capital Normal University, Beijing 100048, China

the consistency and limiting distribution of the estimated parameters of threshold autoregressive (TAR) models. Hansen (2000) developed the asymptotic distribution for the threshold estimator with a shrinking threshold effect. Delgado and Hidalgo (2000) proposed estimators for the location and size of structural breaks in a nonparametric regression model. An important question in the study of threshold models is the selection of a candidate model. Kapetanios (2001) compared the small sample performance of different information criteria in threshold models. Model averaging (MA), as an alternative to the model selection (MS), considers model uncertainty by weighting estimators across different models, instead of relying entirely upon a single model. The MA estimator is generally more stable than the MS estimator, as a small change in data can lead to a significant change in the selection of the optimal model (Yang 2001; Shen and Huang 2006).

There are two strands of literature on model averaging: Bayesian model averaging (BMA) and frequentist model averaging (FMA). Cuaresma and Doppelhofer (2007) applied the BMA to take an average over possible threshold effects and associated threshold observations. From the frequentist perspective, there are two research fields on model averaging. One is on the limiting distribution theory of FMA estimator; see, for example, Hjort and Claeskens (2003) and Xu et al. (2013). The other is on how to choose weights in model averaging. Hansen (2009) applied Mallows model averaging (MMA) in weight choice of averaging threshold models. He performed averaging on models with and without a threshold effect, but did not consider models with different threshold parameters and explanatory variables.

In the current paper, we explore how the FMA approach can be used to obtain an average of threshold models. Two cases are considered. In Case I, we first estimate the threshold parameters of different candidate models and then perform averaging on these threshold models with different explanatory variables. In particular, we consider the averaging of TAR models. In Case II, models with a break at different observed threshold points are considered as different candidate models. We do not estimate the threshold values in this case. In MMA, the variance of random error σ^2 is estimated by the model with the largest number of variables (referred to as the largest model), which leads to the following two problems:

(i) For Case II, the largest model is not unique.
(ii) Even if there exists a unique largest model, using it to estimate σ^2 places too much confidence on a single model.

To address these two problems, this paper develops a new MA approach based on the approximate generalized cross-validation (GCV) method of Craven and Wahba (1979), for which the existence of a unique largest model is unnecessary and the estimation of σ^2 depends on the weights of MA. The resulting averaging estimator is proved to be asymptotically optimal in achieving the lowest possible squared error. In Case I, since the estimator of the threshold parameter is random, the associated coefficient estimator is not a linear combination of the dependent variable. As a result, the proof of asymptotic optimality is more challenging than the existing proofs for other MA methods, such as MMA and optimal frequentist model averaging (Liang et al. 2011).

We investigate the performance of the proposed averaging estimators numerically. The simulation results show that in most cases the new MA estimators have lower MSEs than the MS estimators and other MA estimators. We also apply our method to analyse the unemployment data for the USA and show that our model averaging estimator has better forecasting performance than its competitors.

The remainder of this paper is organized as follows. Section 2 introduces the threshold model and the estimation method. Section 3 provides the criterion for selecting weights and develops the asymptotic optimality theory of the averaging estimator. Section 4 compares our MA estimators with some commonly used MS and MA estimators. Section 5 presents an empirical application of our method. Section 6 concludes the paper. The technical proofs are are given in "Appendix".

2 The model

We consider a threshold regression model with a possible threshold effect,

$$y_i = \mu_i + e_i = x_i'\beta_1 \mathrm{I}(z_i \le \gamma) + x_i'\beta_2 \mathrm{I}(z_i > \gamma) + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where y_i is the dependent variable, $x_i = (x_{i1}, x_{i2}, \ldots)$ are the explanatory variables which can be countably infinite, β_1 and β_2 are two vectors of coefficients, $\mathrm{I}(\cdot)$ is an indicator function, z_i is the threshold variable and can be part of x_i, γ is the threshold parameter, and e_i's are errors with $E(e_i|x_i) = 0$ and $E(e_i^2|x_i) = \sigma^2$. Let $Y = (y_1, \ldots, y_n)'$, $e = (e_1, \ldots, e_n)'$ and $\mu = (\mu_1, \ldots, \mu_n)'$. In application, μ is generally approximated by

$$\mu \approx X(\gamma)\beta,$$

where $X(\gamma)$ is an $n \times 2\eta$ matrix with the ith row $((x_{i1}, \ldots, x_{i\eta})\mathrm{I}(z_i \le \gamma), (x_{i1}, \ldots, x_{i\eta})\mathrm{I}(z_i > \gamma))$ and β is the corresponding coefficient vector. Since the threshold models can be regarded as piecewise linear models, the estimation and averaging methods for linear models can be employed. In a similar way to Hansen (2000), we estimate the parameters by conditional least squares. Let

$$S(\beta, \gamma) = (Y - X(\gamma)\beta)'(Y - X(\gamma)\beta), \tag{2}$$

which is the sum of squared errors (SSE). By minimizing (2), we obtain all the estimators. We assume that γ belongs to a bounded set $\Gamma = [\underline{\gamma}, \bar{\gamma}]$. First, given γ, $\hat{\beta}(\gamma)$ can be obtained by minimizing $S(\beta, \gamma)$. We then replace β by $\hat{\beta}(\gamma)$, and the SSE becomes $S(\hat{\beta}(\gamma), \gamma)$, which is written as $S(\gamma)$. The estimate of γ is defined as:

$$\hat{\gamma} = \arg\min_{\gamma \in \Gamma_n} S(\gamma),$$

where $\Gamma_n = \{z_1, \ldots, z_n\} \cap \Gamma$. Let $z_{(i)}$ be the ith smallest element in $\{z_1, \ldots, z_n\}$. To ensure that the model is estimable, Γ is assumed to satisfy $\underline{\gamma} \ge z_{(\eta+1)}$ and $\bar{\gamma} \le z_{(n-\eta-1)}$. We also assume that Γ_n is non-empty.

3 Model averaging and weight choice

In this section, we propose a new criterion for selecting the optimal weights. Two cases are considered. For Case I, we consider the uncertainty caused only by different explanatory variables, and in Case II, we perform averaging on both different threshold parameters and different explanatory variables. All limiting processes discussed in this section are with respect to $n \to \infty$.

3.1 Averaging for models with estimated γ

In this subsection, we aim to average threshold models with different explanatory variables. We consider model averaging for threshold models that do not contain lagged dependent variables and model averaging for TAR models. Moreover, we show the asymptotic optimality of the proposed MA estimators in both cases under certain regularity conditions.

3.1.1 Averaging for threshold models without lagged dependent variables

Assume that the errors $(e_1, \ldots, e_n)$ are i.i.d. We consider a sequence of approximating models among which the mth model includes k_m explanatory variables that form the vector $x_{(m)i}$. Specifically, the mth model is:

$$Y = X_{(m)}(\gamma)\beta_{(m)} + e_{(m)}, \tag{3}$$

where $X_{(m)}(\gamma)$ is a matrix stacking the vectors $(x'_{(m)i}\mathrm{I}(z_i \leq \gamma), x'_{(m)i}\mathrm{I}(z_i > \gamma))$ and of full column rank, $\beta_{(m)}$ is the coefficient vector of $X_{(m)}(\gamma)$, $e_{(m)} = \mu^C_{(m)}(\gamma) + e$, and the term $\mu^C_{(m)}(\gamma) = \mu - X_{(m)}(\gamma)\beta_{(m)}$ of which is the approximation error of model (3).

Following the estimation method in Sect. 2, we can obtain the estimated threshold parameter $\hat{\gamma}_{(m)}$ and coefficient

$$\hat{\beta}_{(m)} = (X'_{(m)}(\hat{\gamma}_{(m)})X_{(m)}(\hat{\gamma}_{(m)}))^{-1}X'_{(m)}(\hat{\gamma}_{(m)})Y \tag{4}$$

under the mth model. Let $\hat{X}_{(m)} = X_{(m)}(\hat{\gamma}_{(m)})$ and $\hat{P}_{(m)} = \hat{X}_{(m)}(\hat{X}'_{(m)}\hat{X}_{(m)})^{-1}\hat{X}'_{(m)}$, so that the estimator of μ under the mth candidate model is given by $\hat{\mu}_{(m)} = \hat{P}_{(m)}Y$. Denote $w = (w_1, \ldots, w_M)'$, a weight vector in the unit simplex in R^M

$$\mathcal{H}_n = \left\{ w \in [0, 1]^M : \sum_{m=1}^{M} w_m = 1 \right\},$$

where M is the number of candidate models. Note that $\mathcal{H}_n$ is a continuous set and is different from the weight set in Hansen (2007), which is discrete. In addition, Cheng et al. (2015) used a continues weight set, which is more general than the discrete set of Hansen (2007) but is still a subset of $\mathcal{H}_n$. The MA estimator of μ can be expressed as

$$\hat{\mu}(w) = \sum_{m=1}^{M} w_m \hat{\mu}_{(m)} = \sum_{m=1}^{M} w_m \hat{P}_{(m)} Y \equiv \hat{P}(w)Y,$$

where $\hat{P}(w) = \sum_{m=1}^{M} w_m \hat{P}_{(m)}$ is symmetric but not necessarily idempotent. The squared error is $L_n(w) = \|\hat{\mu}(w) - \mu\|^2$, and the corresponding risk is $R_n(w) = E(L_n(w)|X, Z)$, where $X = (x_1, \ldots, x_n)'$ and $Z = (z_1, \ldots, z_n)'$.

When σ^2 is known, one may obtain weights by minimizing the following Mallows' criterion proposed by Hansen (2007):

$$C_n(w) = \|Y - \hat{\mu}(w)\|^2 + 2\sigma^2 tr \hat{P}(w).$$

Since σ^2 is usually unknown in practice, Hansen (2007) suggested estimating it by the largest candidate model, i.e.

$$\hat{\sigma}^2 = (n - k_{M^*})^{-1} \left\| Y - \hat{\mu}_{M^*} \right\|^2,$$

where $M^* = \arg\max_{m \in \{1,\ldots,M\}} k_m$. It is shown that as $n \to \infty$, if $k_{M^*} \to \infty$ and $k_{M^*}/n \to 0$, then $\hat{\sigma}^2$ is consistent and the asymptotic optimality result still holds for unknown σ^2.

In time series case, Hansen (2008) applied this criterion to averaging autoregressive models. However, the largest model may not be unique in practice. In fact, even if the largest model is unique, using the single model to estimate σ^2 may deviate, in some sense, from the objective of model averaging. Motivated by these concerns, we develop a new least squares MA estimator for threshold models. The criterion for selecting weights is as follows:

$$\mathcal{L}_n(w) = \|Y - \hat{\mu}(w)\|^2 \left(1 + 2\frac{tr \hat{P}(w)}{n}\right). \tag{5}$$

If we set one component of the weight vector w to be 1 and the others to be 0, then (5) reduces to a criterion for model selection. Therefore, one may approximate the GCV criterion by the MS version of (5) and use it to relate GCV to Mallows' C_p (Li 1987). For any fixed w in (5), $\left\| Y - \hat{\mu}(w) \right\|^2 / n$ is the mean of residual squared sums of the MA estimator $\hat{\mu}(w)$. If we take it as an estimator of σ^2, then $\mathcal{L}_n(w)$ can be regarded as another estimator of $C_n(w)$. As mentioned previously, Hansen (2007, 2008) estimated σ^2 based on the largest model. We use a averaging estimator of σ^2 instead. Thus, our criterion can be viewed as an adjusted Mallows criterion, which can be used in more general cases because MMA would be infeasible when the largest model is not unique, as is the case in Sect. 4.2. If the covariance matrix of the error term e is not diagonal, to estimate the inverse of the covariance matrix, we may use the estimators proposed by Cheng et al. (2014, 2015).

We rewrite $\mathcal{L}_n(w)$ as $\mathcal{L}_n(w) = w'\hat{e}'\hat{e}w(1 + 2w'K/n)$ for simplicity, where $K = (k_1, ..., k_M)'$, $\hat{e} = (\hat{e}_{(1)}, \ldots, \hat{e}_{(M)})$ and $\hat{e}_{(m)} = Y - \hat{\mu}_{(m)}$. When constraining w to

$\mathcal{H}_n$, we can obtain weights through minimizing $\mathcal{L}_n(w)$, i.e. $\hat{w} = \arg\min_{w\in\mathcal{H}_n} \mathcal{L}_n(w)$. The estimator $\hat{\mu}(\hat{w})$ is referred to as the adjusted Mallows model averaging (AMMA) estimator of μ hereafter. Note that although $\mathcal{L}_n(w)$ is a cubic function of w, the numerical algorithms for minimizing such a criterion are actually readily available. For example, one can use ‘solnp’ in the R package ‘Rsolnp’. Therefore, our AMMA approach can be easily performed in practice.

Note that for each candidate model, the estimator of μ depends on a random item $\hat{\gamma}_m$, thus causing problems for conducting the asymptotic optimality. So the theory in this subsection is not just an extension of that of Hansen (2007). To solve this problem, we try to find a properly defined limit for $\hat{\gamma}_{(m)}$ under each candidate model. We assume that there exists a constant $\gamma^*_{(m)}$ such that $\hat{\gamma}_{(m)} \xrightarrow{p} \gamma^*_{(m)}$, where $\gamma^*_{(m)}$ is not necessarily equal to the true value γ_0. If $z_i = i/n$ and k_m is bounded, the convergency was proved by Koo and Seo (2015). However, if k_m is related to n, further work is required.

Let $X^*_{(m)} = X_{(m)}(\gamma^*_{(m)})$, $P^*_{(m)} = X^*_{(m)}\left(X^{*\prime}_{(m)}X^*_{(m)}\right)^{-1}X^{*\prime}_{(m)}$, $P^*(w) = \sum_{m=1}^M w_m P^*_{(m)}$, $A^*(w) = I_n - P^*(w)$ and $L^*_n(w) = \|P^*(w)Y - \mu\|^2$. Then we have $R^*_n(w) \equiv E(L^*_n(w)|X, Z) = \|A^*(w)\mu\|^2 + \sigma^2 tr P^{*2}(w)$. Define $\xi^*_n = \inf_{w\in\mathcal{H}_n} R^*_n(w)$ and $\lambda_{\max}(A)$ as the maximum singular value of matrix A. The following theorem states the asymptotic optimality of the AMMA estimator.

Theorem 1 *For some finite integer $G \geq 1$, if*

$$E(e_i^{4G}|x_i) < \infty, \tag{6}$$

$$M\xi_n^{*-2G} \sum_{m=1}^{M} \left(R^*_n(w^0_m)\right)^G \xrightarrow{p} 0, \tag{7}$$

$$n\xi_n^{*-1} \max_{1\leq m\leq M} \lambda_{\max}(P^*_{(m)} - \hat{P}_{(m)}) \xrightarrow{p} 0, \tag{8}$$

$$k^2_{M^*}/n \leq a_1 < \infty, \tag{9}$$

and

$$\|\mu\|^2 = O_p(n), \tag{10}$$

then

$$\frac{L_n(\hat{w})}{\inf_{w\in\mathcal{H}_n} L_n(w)} \xrightarrow{p} 1, \tag{11}$$

where a_1 is a constant, and w^0_m is an $M \times 1$ vector in which the mth element is one and the others are zeros.

Proof See “Appendix.” □

Condition (6) is a moment condition and requires the regression error distribution to have sufficiently thin tails. For example, it excludes the Cauchy distribution and holds for Gaussian distribution. Condition (7) is a commonly used condition in the model averaging literature such as Wan et al. (2010) and Liu and Okui (2013). To explain this condition, we consider a situation with $\xi_n^* = n^a$, $\sup_{w\in\mathcal{H}_n} R_n^*(w) = n^b$ and $0 < a \le b < 1$; then, Condition (7) is implied by $M^2 n^{G(b-2a)} \to 0$, which holds when $b < 2a$ and M doest not increase with n too fast. Cheng et al. (2015) pointed out that Condition (7) will preclude some good models with smaller $L_n(w)$ in linear cases. Similarly, it still may happen in the threshold models. However, they select weights on a narrower set compared with our continuous set $\mathcal{H}_n$. Thus, we need to add Condition (7) to ensure the asymptotic optimality of AMMA, which means M cannot increase with n as fast as it in Cheng et al. (2015). Condition (8) puts some restrictions on the order of ξ_n^* and the convergence rate of the elements of matrix $\hat{P}_{(m)} - P_{(m)}^*$. Note that because $\hat{\gamma}_{(m)} \xrightarrow{p} \gamma_{(m)}^*$, the elements of matrix $\hat{P}_{(m)} - \hat{P}_{(m)}^*$ converge to zeros. The proof of (58) in "Appendix" shows that Condition (8) can be satisfied when k_{M^*} is bounded. Condition (9) requires that the numbers of covariates in candidate models do not increase faster than $n^{1/2}$. Condition (10) is on the sum of $\mu_1^2, \dots, \mu_n^2$ and need only that $\mu_1^2, \dots, \mu_n^2$ do not expand with n.

3.1.2 Averaging for TAR models

The TAR model is a special case among threshold models and is widely used in empirical analysis. However, when averaging TAR models, the asymptotic theory developed above is no longer valid due to serial dependence and the existence of lagged dependent variables. This subsection develops the asymptotic optimality for averaging TAR models.[1] In the same way as in Sect. 3.1.1, we have

$$\begin{aligned} y_i &= \mu_i + e_i \\ &= \left(\beta_{10} + \sum_{j=1}^{p_1} \beta_{1j} y_{i-j}\right) \mathrm{I}(z_i \le \gamma) + \left(\beta_{20} + \sum_{j=1}^{p_2} \beta_{2j} y_{i-j}\right) \mathrm{I}(z_i > \gamma) \\ &\quad + e_i, \qquad i = 1, \dots, n, \end{aligned}$$

where p_k is the lag order for regime k $(k = 1, 2)$, e_i's are white noise with mean zero and variance σ^2 and β_{kj}'s are autoregressive coefficients with $\sum_{j=1}^{p_k} |\beta_{kj}| < 1$ $(k = 1, 2)$. For simplicity, we set $p_1 = p_2 = p$, where p can be infinite. In this case, $x_i = (1, y_{i-1}, \dots, y_{i-p})'$ and each regime is an AR(k_m) process in the mth model. We assume that for each m, k_m is fixed, so M is bounded.

We focus on μ and apply the AMMA method to select the weights. Let $Q_n^*(w) = \|A^*(w)\mu\|^2 + \sigma^2 tr(P^{*2}(w))$ and $\zeta_n^* = \inf_{w\in\mathcal{H}_n} Q_n^*(w)$. To study the asymptotic optimality of the MA estimator, we make the following assumptions:

(a.1) $\{x_i, z_i, e_i\}$ is strictly stationary and ergodic, and $E(e_i|\sigma(x_i, x_{i-1}, \dots)) = 0$, where $\sigma(x_i, x_{i-1}, \dots)$ is the σ-algebra generated by $x_i, x_{i-1}, \dots$. (a.2) $E|y_i|^4 < \infty$ and $E|y_i e_i|^4 < \infty$.

[1] Although Hansen (2008, 2009) studied averaging estimators in time series models, they did not develop the asymptotic optimality.

(a.3) Let $f_2(z|\hat{\gamma}_{(m)})$ be the conditional density of z_i given $\hat{\gamma}_{(m)}$. Uniformly for $z \in \Gamma$ and $\hat{\gamma}_{(m)} \in \Gamma$, the conditional density $f_2(z|\hat{\gamma}_{(m)})$ is bounded by a finite constant $\bar{f}_2$, and the conditional expectation $E(|x_{ij}x_{ik}||z_i = \gamma, \hat{\gamma}_{(m)})$ with z_i and $\hat{\gamma}_{(m)}$ given is bounded.

(a.4) $E|\hat{\gamma}_{(m)} - \gamma^*_{(m)}| = O(n^{-\rho})$ for some constant $0 < \rho \leq 1, \quad m = 1, \ldots, M$.

Assumptions (a.1) and (a.2) are common assumptions for stationary processes. In real data analysis, if the series is non-stationary, we can use some data conversion methods, such as the differential operator and seasonal adjustment to get a stationary series. Assumption (a.3) requires the conditional density and expectation are bounded. Assumption (a.4) is based on the result of Koo and Seo (2015), who showed that the convergence rate of $\hat{\gamma}$ can be as fast as $T^{-1/3}$ for the structural break model. Under these assumptions, we have the following theorem.

Theorem 2 *If Assumptions (a.1)–(a.4) and Condition (10) are satisfied and*

$$n^{1-\rho/2}\zeta_n^{*-1} \xrightarrow{p} 0, \tag{12}$$

then (11) is valid.

Proof See “Appendix.” □

3.2 Averaging for models without estimating γ

In this subsection, we average models with different threshold parameters and different explanatory variables simultaneously using the models set up in Sect. 3.1.1. Let $|\Gamma_n|$ be the size of Γ_n. Since there are $|\Gamma_n|$ possible threshold points, there will be $|\Gamma_n|$ models with the same explanatory variables. Let $\gamma_{(s)}$ be the sth item of Γ_n. Assume that the m_sth candidate model contains k_m explanatory variables, with $\gamma_{(s)}$ being the threshold parameter. Then the threshold parameter in every candidate model can be regarded as a fixed constant. Therefore, the coefficient estimated by the m_sth model is:

$$\widetilde{\beta}_{(m_s)} = (X'_{(m)}(\gamma_{(s)})X_{(m)}(\gamma_{(s)}))^{-1}X'_{(m)}(\gamma_{(s)})Y,$$

and the estimator of μ is given by

$$\widetilde{\mu}_{(m_s)} = X_{(m)}(\gamma_{(s)})(X'_{(m)}(\gamma_{(s)})X_{(m)}(\gamma_{(s)}))^{-1}X'_{(m)}(\gamma_{(s)})Y \equiv P_{(m)}(\gamma_{(s)})Y.$$

Let $w = (w_{1_1}, \ldots, w_{M_{|\Gamma_n|}})'$ and $\widetilde{\mathcal{H}}_n = \left\{w \in [0, 1]^{M|\Gamma_n|} : \sum_{m=1}^{M}\sum_{s=1}^{|\Gamma_n|} w_{m_s} = 1\right\}$, which is also a continuous weight set, so that the averaging estimator of μ is:

$$\widetilde{\mu}(w) = \sum_{m=1}^{M}\sum_{s=1}^{|\Gamma_n|} w_{m_s}\widetilde{\mu}_{(m_s)} = \sum_{m=1}^{M}\sum_{s=1}^{|\Gamma_n|} w_{m_s}P_{(m)}(\gamma_{(s)})Y \equiv P(w)Y.$$

The squared error is $\widetilde{L}_n(w) = \|\widetilde{\mu}(w) - \mu\|^2$, and the corresponding risk is $\widetilde{R}_n(w) = E(\widetilde{L}_n(w)|X, Z)$. Let $\widetilde{\xi}_n = \inf_{w\in\widetilde{\mathcal{H}}_n} \widetilde{R}_n(w)$. In this subsection, the largest model is not unique, so the Mallows' criterion does not apply. In light of this concern, we make use of the AMMA idea; that is, we select weights by the following criterion:

$$\widetilde{\mathcal{L}}_n(w) = \|Y - \widetilde{\mu}(w)\|^2 \left(1 + 2\frac{tr\,P(w)}{n}\right).$$

Let $\widetilde{w} = \arg\min_{w\in\widetilde{\mathcal{H}}_n} \widetilde{\mathcal{L}}_n(w)$ and the corresponding AMMA estimator be $\widetilde{\mu}(\widetilde{w})$. The following theorem guarantees the asymptotic optimality of the AMMA estimator.

Theorem 3 *For some finite integer $G \geq 1$, if Conditions (6), (9) and*

$$M|\Gamma_n|\widetilde{\xi}_n^{-2G} \sum_{m=1}^{M}\sum_{s=1}^{|\Gamma_n|} \left(\widetilde{R}_n(w^0_{m_s})\right)^G \xrightarrow{p} 0, \tag{13}$$

hold, then

$$\frac{\widetilde{L}_n(\widetilde{w})}{\inf_{w\in\widetilde{\mathcal{H}}_n} \widetilde{L}_n(w)} \xrightarrow{p} 1. \tag{14}$$

In the current case, since the threshold parameter is known in every candidate model, the proof of Theorem 3 is more straightforward than that of Theorem 1. We only provide a simple explanation in "Appendix". The detailed proof is available on request from the authors. Note that Condition (13) is similar to Condition (7).

4 Simulations

In this section, we conduct three simulation studies to compare the performance of the MA estimator and the MS estimator. The first simulation performs averaging for models with different explanatory variables and i.i.d. errors, the second simulation performs averaging for models with different explanatory variables and threshold parameters, and the third simulation performs averaging for TAR models with different orders.

4.1 Simulation I: averaging for models with estimated $\boldsymbol{\gamma}$

The data generating process is:

$$y_i = \mu_i + e_i = \sum_{j=1}^{\infty} x_{ij}\beta_{1j}\mathrm{I}(x_{i3} \leq \gamma) + \sum_{j=1}^{\infty} x_{ij}\beta_{2j}\mathrm{I}(x_{i3} > \gamma) + e_i,\ i = 1, \ldots, n,$$

where $\gamma = 0$, $x_{i1} = 1$, all other x_{ij}'s and e_i's come from $N(0, 1)$ and are independent of one another, and the coefficients $\beta_{11} = c$, the remaining $\beta_{1j} = cj^{-\zeta}$ with $\zeta = 0.25, 0.5, 0.75$ controlling the decay rate of the coefficients, and $\beta_2 = a\beta_1$ with $a = 1.5$ and $c > 0$. The difference between coefficients is denoted by a. The parameter c is set to make the population $R^2 = \mathrm{var}(y_i - e_i)/\mathrm{var}(y_i)$ vary on a grid from 0.1 to

0.9. To let the threshold variable x_{i3} appear in each candidate model, we set the mth candidate model to include the first $m+2$ explanatory variables ($m = 1, \ldots, M$), and $M = 3n^{1/3}$. When estimating γ, we restrict it to the set containing the 20, 25, ..., 80% quantiles of $\{x_{i3}\}$ for decreasing computation time, as suggested by Hansen (2000). The sample size is set at 60, 100, 250 and 400. To evaluate the performance of the estimators, we simulate 500 replications and compute mean squared risk by

$$\frac{1}{500}\sum_{r=1}^{500}\sum_{i=1}^{n}(\hat{\mu}_i^{(r)} - \mu_i)^2, \tag{15}$$

where $\hat{\mu}_i^{(r)}$ is the estimates of μ in the rth replication. For each parameterization, we normalize the risks by dividing the risk by the infeasible optimal risk (the risk of the best single model).

We compare our averaging estimator with the AIC and BIC model selection estimators. The AIC score for the mth model is given by $\mathrm{AIC}_m = n\log\hat{\sigma}_m^2 + 2k_m$, where $\hat{\sigma}_m^2 = \|Y - \hat{\mu}_{(m)}\|^2/n$, and the BIC score for the mth model is $\mathrm{BIC}_m = n\log\hat{\sigma}_m^2 + k_m\log n$. We also compare our averaging estimator with the existing model averaging methods: MMA, smoothed AIC (S-AIC), and smoothed BIC (S-BIC), proposed in Buckland et al. (1997) and ARM (adaptive regression by mixing), an adaptive method developed by Yang (2001). The S-AIC method assigns weight $w_{\mathrm{AIC},m} = \exp(-\mathrm{AIC}_m/2)/\sum_{m=1}^{M}\exp(-\mathrm{AIC}_m/2)$ to the mth model and the S-BIC method assigns weight $w_{\mathrm{BIC},m} = \exp(-\mathrm{BIC}_m/2)/\sum_{s=1}^{M}\exp(-\mathrm{BIC}_m/2)$ to the mth model. The ARM method divides samples into a training part and a testing part. The parameters are estimated by the training samples, while the weights are obtained by the testing samples. For more details, one can refer to Yang (2001).

The simulation results are displayed in Figs. 1, 2, 3. In each panel, the relative risk is displayed on the y axis and the population R^2 is displayed on the x axis. Since the MA methods are always better than the MS methods, we only show the MA results to distinguish different lines clearly. In addition, we cut off part of the figures to make it easier to compare AMMA and MMA in some cases. Although some risks do not appear in the figures, they are all bounded actually. The factors that affect the relative performances of the competitors include n (sample size), ζ (the decay rate of the coefficient) and R^2 (population). First, in the majority of cases of $\{n, \zeta, R^2\}$, the AMMA outperforms S-AIC and S-BIC. Second, the AMMA performs better than the MMA and ARM when R^2 is large; while when R^2 is small, the AMMA performs worse than the MMA and ARM. Third, when n or ζ decreases, the region of R^2 where the AMMA outperforms the MMA and ARM becomes wider. Fourth, when n increases, the AMMA and MMA perform more closely. In addition, we also conduct simulations for $a = 0.2$ and $a = 3$. The corresponding results are qualitatively similar to those obtained for $a = 1.5$.

4.2 Simulation II: averaging for models without estimating γ

The setup of this simulation is the same as that in Sect. 4.1 However, in this subsection, we do not estimate the threshold parameter. We average or select among

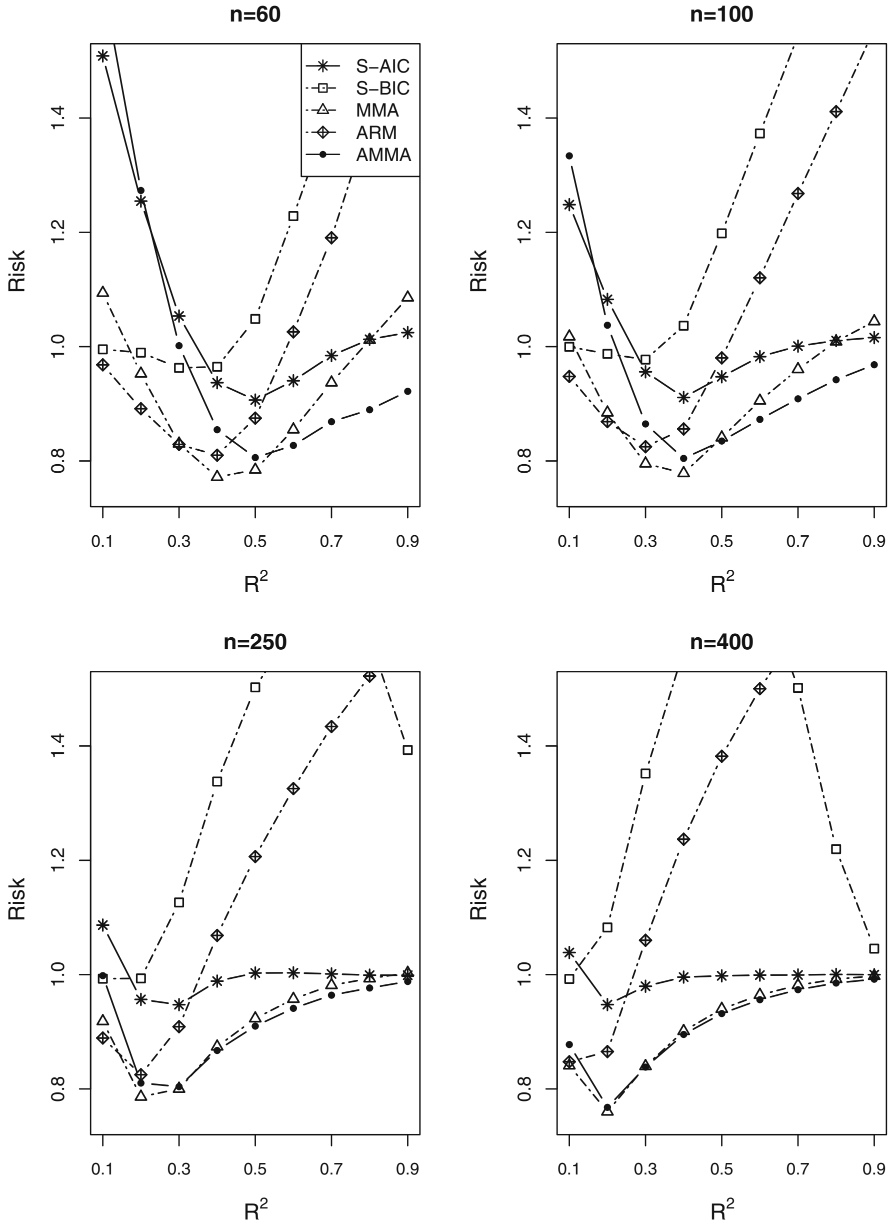

Fig. 1 Results of Simulation I. Risks of averaging models with estimated γ ($\zeta = 0.25$)

models with different explanatory variables at all possible threshold points, and do not compare the AMMA method with the MMA method as MMA is infeasible in this example.

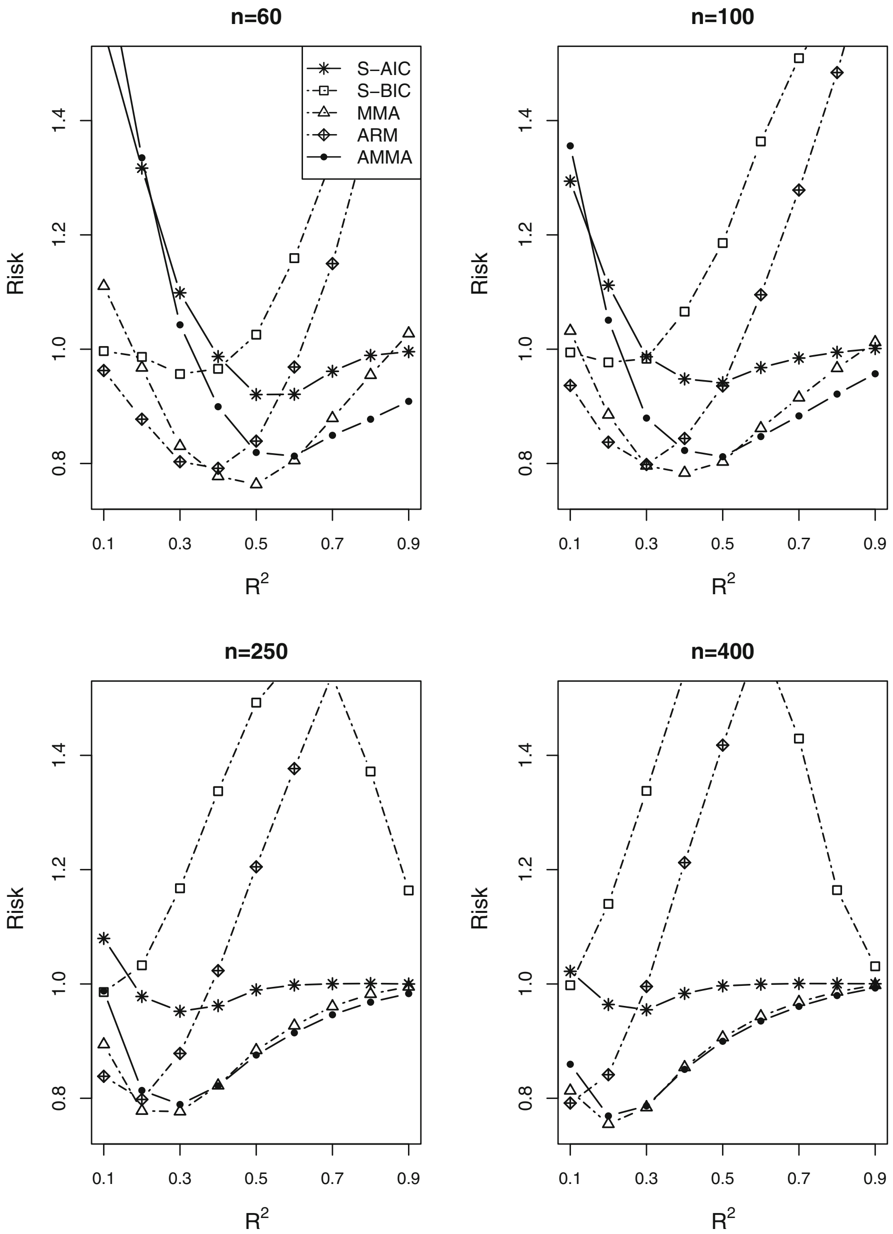

Fig. 2 Results of Simulation I. Risks of averaging models with estimated γ ($\zeta = 0.5$)

The simulation results are displayed in Figs. 4, 5, 6. Again, we can find the AMMA outperforms S-AIC, S-BIC and ARM. The detailed comparison findings are very similar to those in Simulation I.

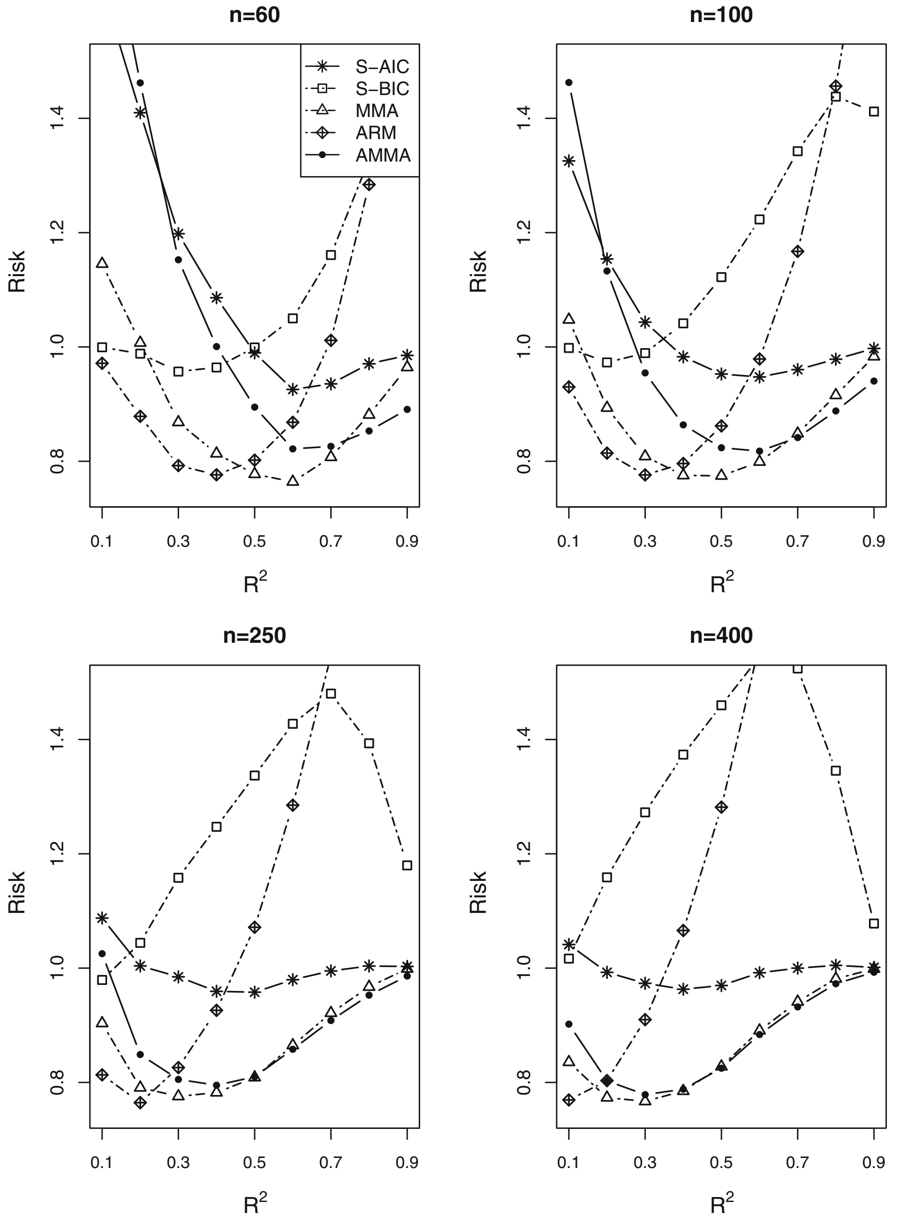

Fig. 3 Results of Simulation I. Risks of averaging models with estimated γ ($\zeta = 0.75$)

4.3 Simulation III: averaging for TAR models

We now investigate the performance of the averaging estimator for TAR models. The data generating process is as follows:

$$y_i = (\beta_{10} + \sum_{j=1}^{p} \beta_{1j} y_{i-j}) \mathrm{I}(y_{i-d} \le \gamma) + \left(\beta_{20} + \sum_{j=1}^{p} \beta_{2j} y_{i-j}\right) \mathrm{I}(y_{i-d} > \gamma) + e_i,\ i = 1, \dots, n,$$

where y_{i-d} is the threshold variable and d is the lag order. We set e_i to be i.i.d. $N(0, \sigma^2)$, $d = 3$, $\gamma = 0$, $p = 6$, $\beta_{10} = 0.5$ and $\beta_{20} = -0.5$. The coefficients are generated by the rule $\beta_{kj} = \dfrac{5(1+j)^{\alpha_k}(-\phi)^j}{6\sum_{i=1}^{p}(1+i)^{\alpha_k}\phi^i}$, where ϕ and α_k are constants and $k = 1, 2$, $j = 1, \dots, p$, which is similar to the setting in Hansen (2008). As $\sum_{j=1}^{p} |\beta_{kj}| < 1$, $\{y_n\}$ is stationary. Note that $\beta_{ki}/\beta_{kj} = \left(\frac{1+i}{1+j}\right)^{\alpha_k}(-\phi)^{i-j}$ $(i > j)$, so the item $(-\phi)^{i-j}$ determines the convergence rate of the coefficients. We let $\alpha_1 = 0.1$, $\alpha_2 = 0.3$, $n \in \{60, 100, 250, 400\}$, $\sigma^2 = 0.5, 1, 2$ and ϕ vary on a grid from 0.6 to 0.9.

Candidate models differ in their lag orders. Identical orders are used in the two regimes and the threshold parameter is estimated, so we have $M = p = 6$ candidate models. Unlike the previous simulations, we also need to estimate d here. Denote by $\hat{d}_m$ the estimator of d under the mth candidate model. According to the mth candidate model, the one-step-ahead out-of-sample forecast of y_{n+1} given $y_n, y_{n-1}, \dots$ is:

$$\hat{y}_{n+1}(m) = \left(\hat{\beta}_{(m)_{10}} + \sum_{j=1}^{m} \hat{\beta}_{(m)_{1j}} y_{n+1-j}\right) \mathrm{I}(y_{n+1-\hat{d}_m} \le \hat{\gamma}_{(m)}) + \left(\hat{\beta}_{(m)_{20}} + \sum_{j=1}^{m} \hat{\beta}_{(m)_{2j}} y_{n+1-j}\right) \mathrm{I}(y_{n+1-\hat{d}_m} > \hat{\gamma}_{(m)}),$$

where $\hat{\beta}_{(m)_{rj}}$ is the estimator of $\beta_{(m)_{rj}}$ for $r = 1, 2$ and $j = 0, \dots, p$. The combined forecast is given by $\hat{y}_{n+1}(w) = \sum_{m=1}^{M} w_m \hat{y}_{n+1}(m)$. To compare the performance of model selection and averaging methods, we use 500 replications. For each replication, we generate a series of size $n + 1$ and use the first n samples to get the averaged coefficients. Then we calculate the one-step-ahead out-of-sample prediction and get the mean squared forecast error (MSFE) given by

$$\frac{1}{500} \sum_{r=1}^{500} (y_{n+1}^{(r)} - \hat{y}_{n+1}^{(r)})^2, \tag{16}$$

where r denotes the rth replication.

Figures 7, 8, 9 show the simulation results. As the ARM method cannot be used for time series prediction, we choose another adaptive method, named AFTER (Aggregated Forecast Through Exponential Reweighting, Yang 2004) instead. We can see

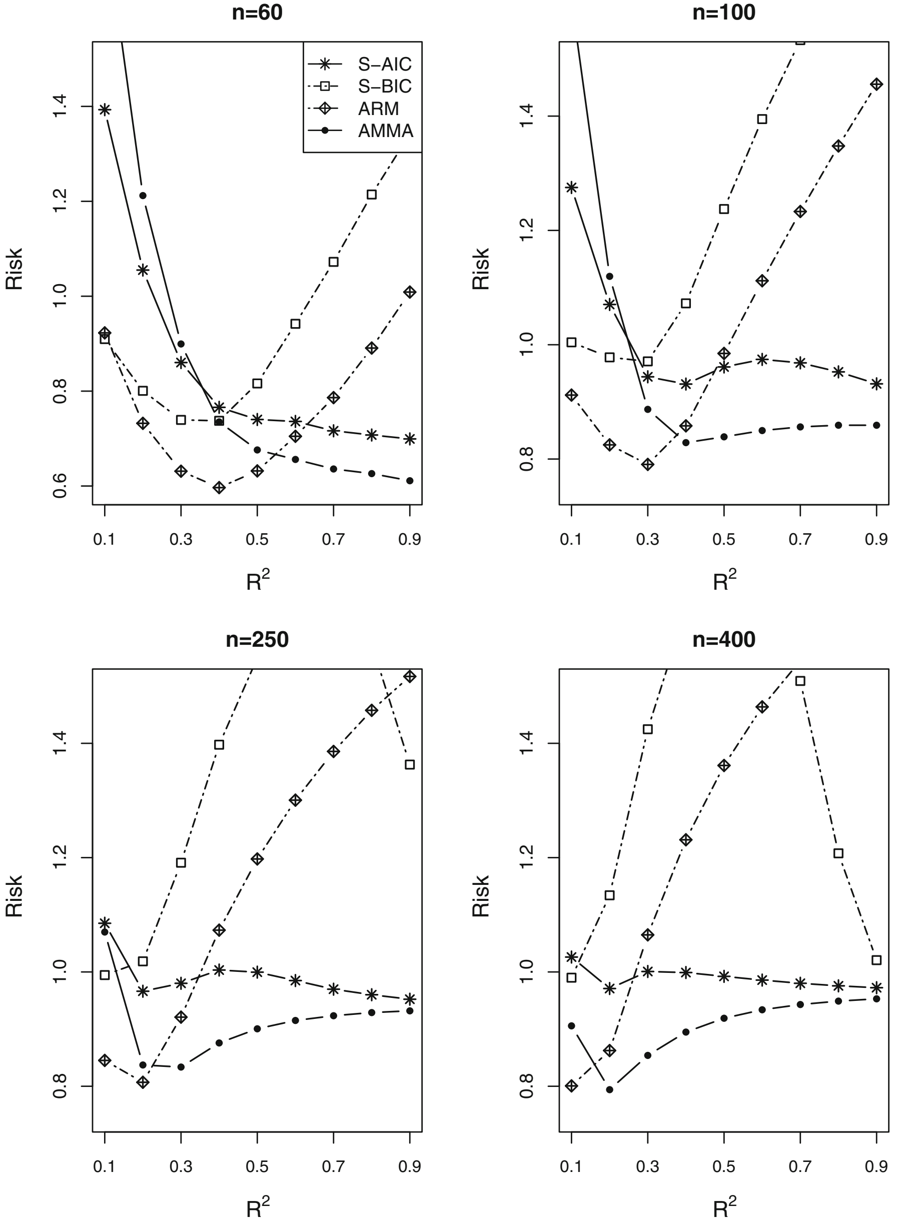

Fig. 4 Results of Simulation II. Risks of averaging models without estimating γ ($\zeta = 0.25$)

that the MMA and AMMA always perform better than the other methods. The factors that affect the relative performances of the competitors include n (sample size), σ^2 (noise level) and ϕ (the convergence rate of the coefficients). First, in the majority of cases of $\{n, \sigma^2, \phi\}$, the AMMA and MMA outperform S-AIC, S-BIC and

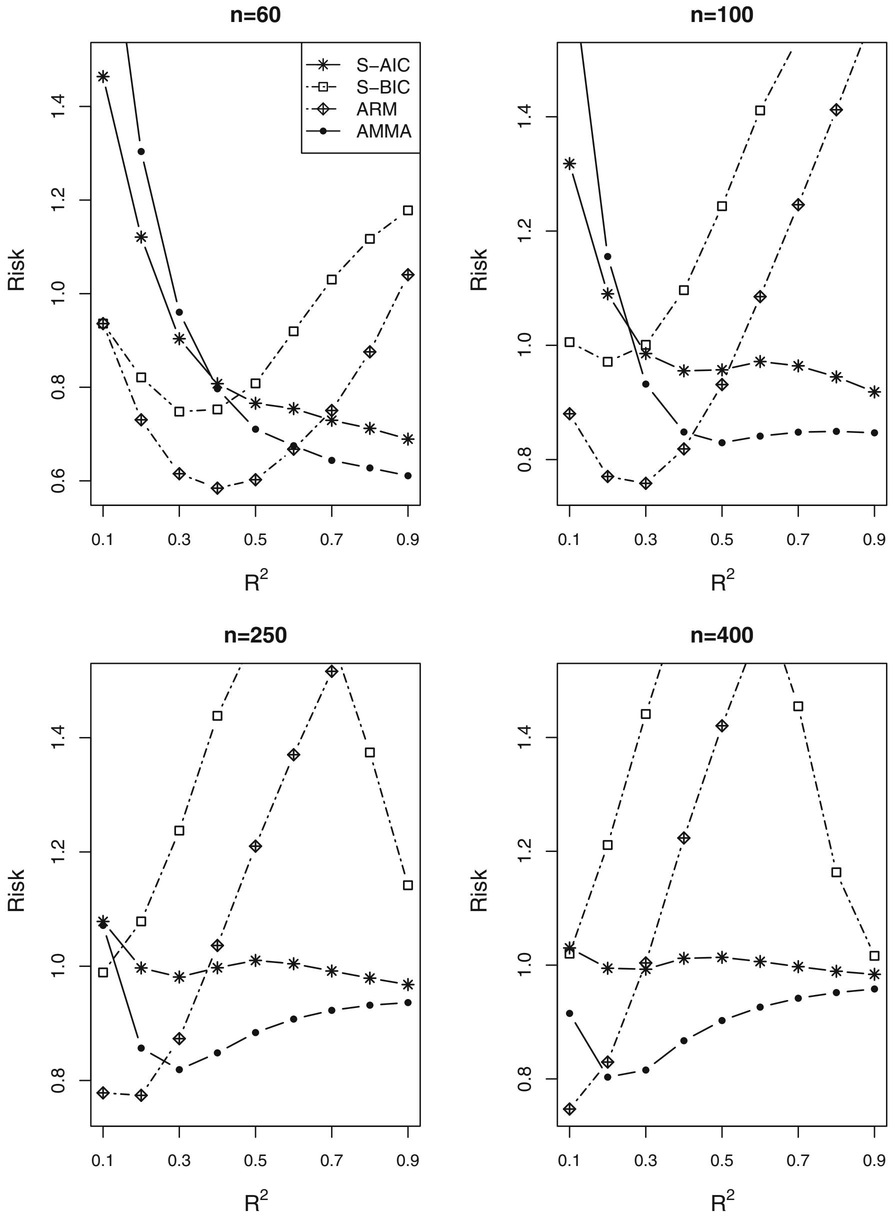

Fig. 5 Results of Simulation II. Risks of averaging models without estimating γ ($\zeta = 0.5$)

AFTER. Second, when $n = 60, 100$, the MMA performs better than AMMA in most of values of ϕ, while when $n = 250, 400$, the AMMA performs better than the MMA in most of values of ϕ. Third, for different σ^2, the comparison results are very similar.

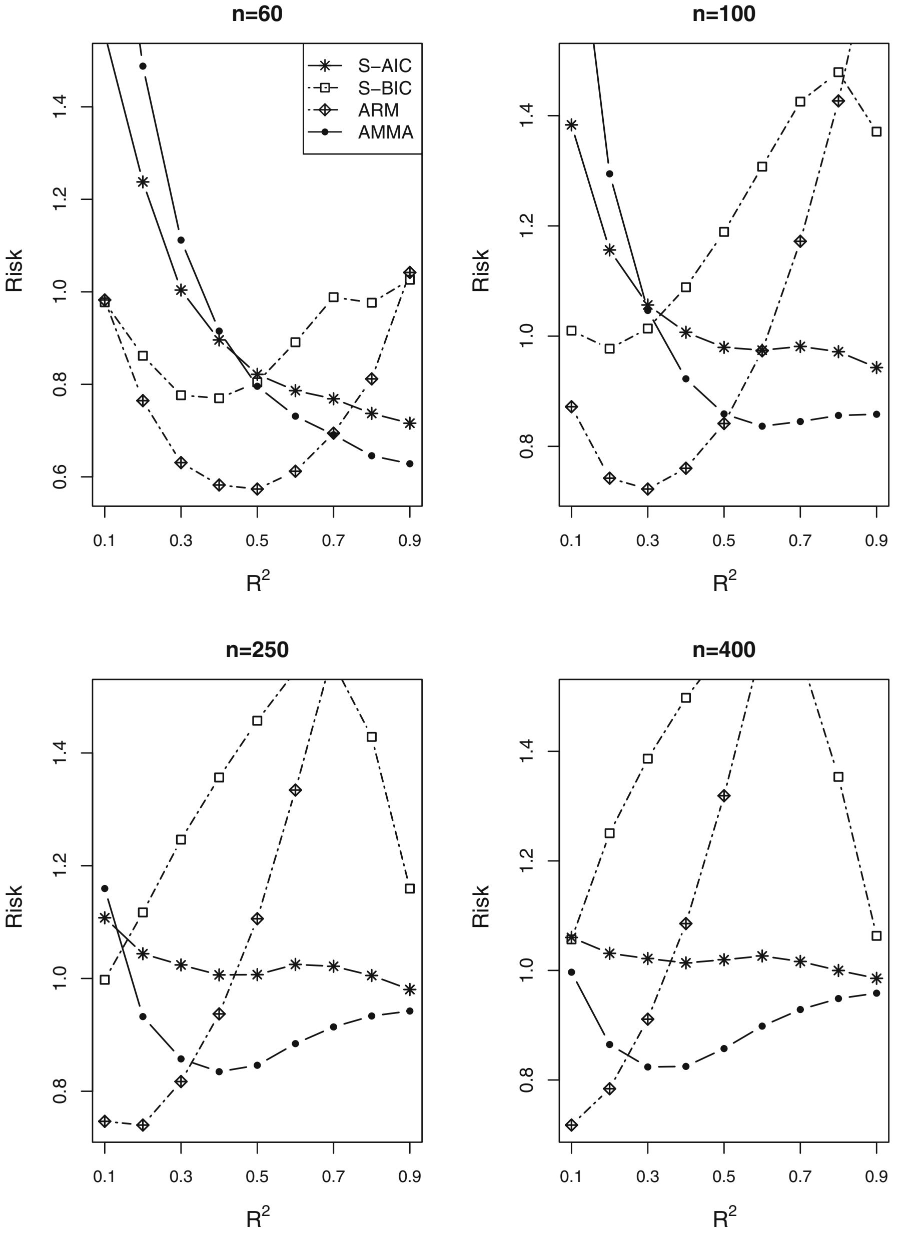

Fig. 6 Results of Simulation II. Risks of averaging models without estimating γ ($\zeta = 0.75$)

5 Empirical application

In this section, we apply the averaging approach to a monthly data set for US unemployment from January 1970 to Dec 2012. The sample size is 516 in total. The unit

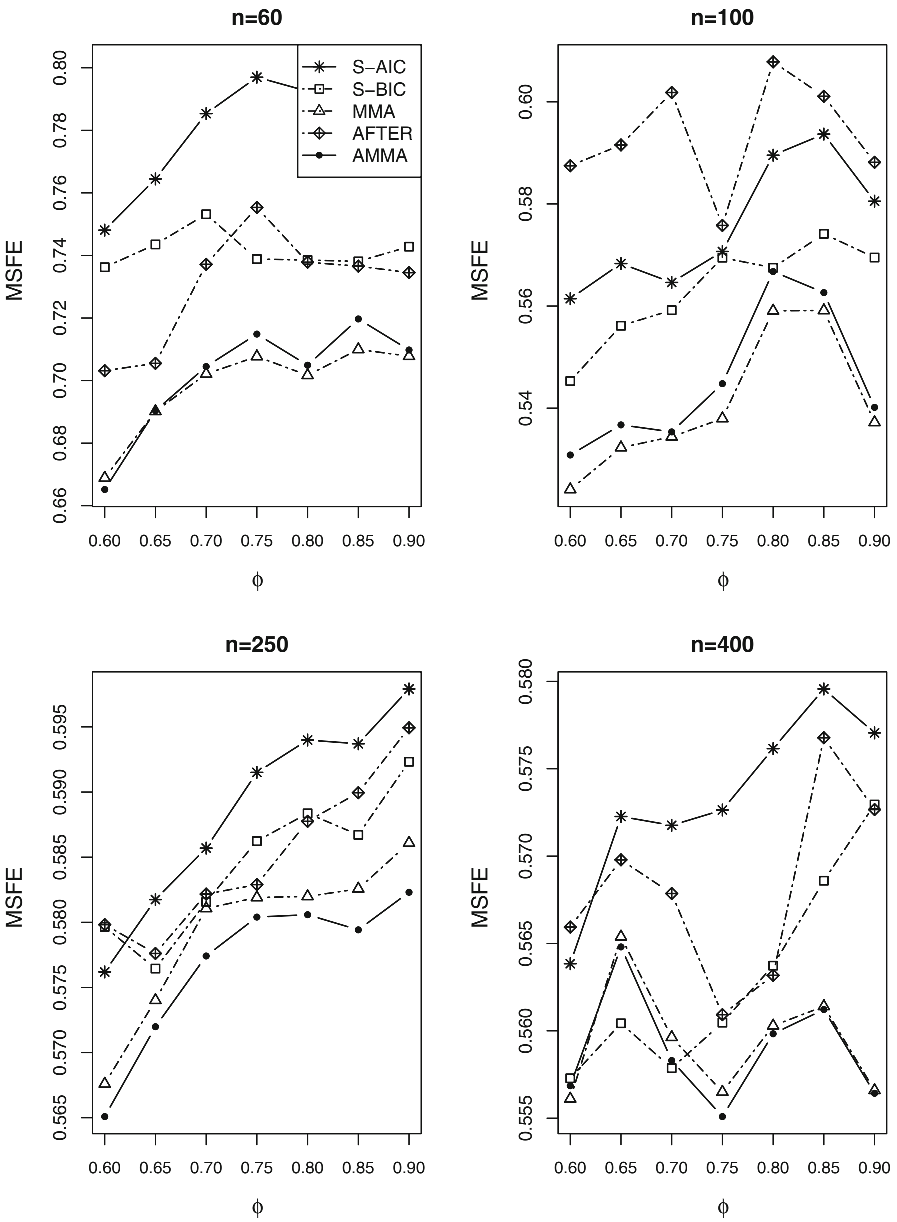

Fig. 7 Results of Simulation III. MSFEs for averaging TAR models with $\sigma^2 = 0.5$

root test for threshold model (Caner and Hansen 2001) suggests that the process is a stationary nonlinear threshold autoregression. The model selection and averaging methods are the same as those in Simulation III, with the largest order set to be 12. The candidate set for d is $\{1, 2, \ldots, 12\}$. We use $\{y_1, \ldots, y_n\}$ to fit the model and predict

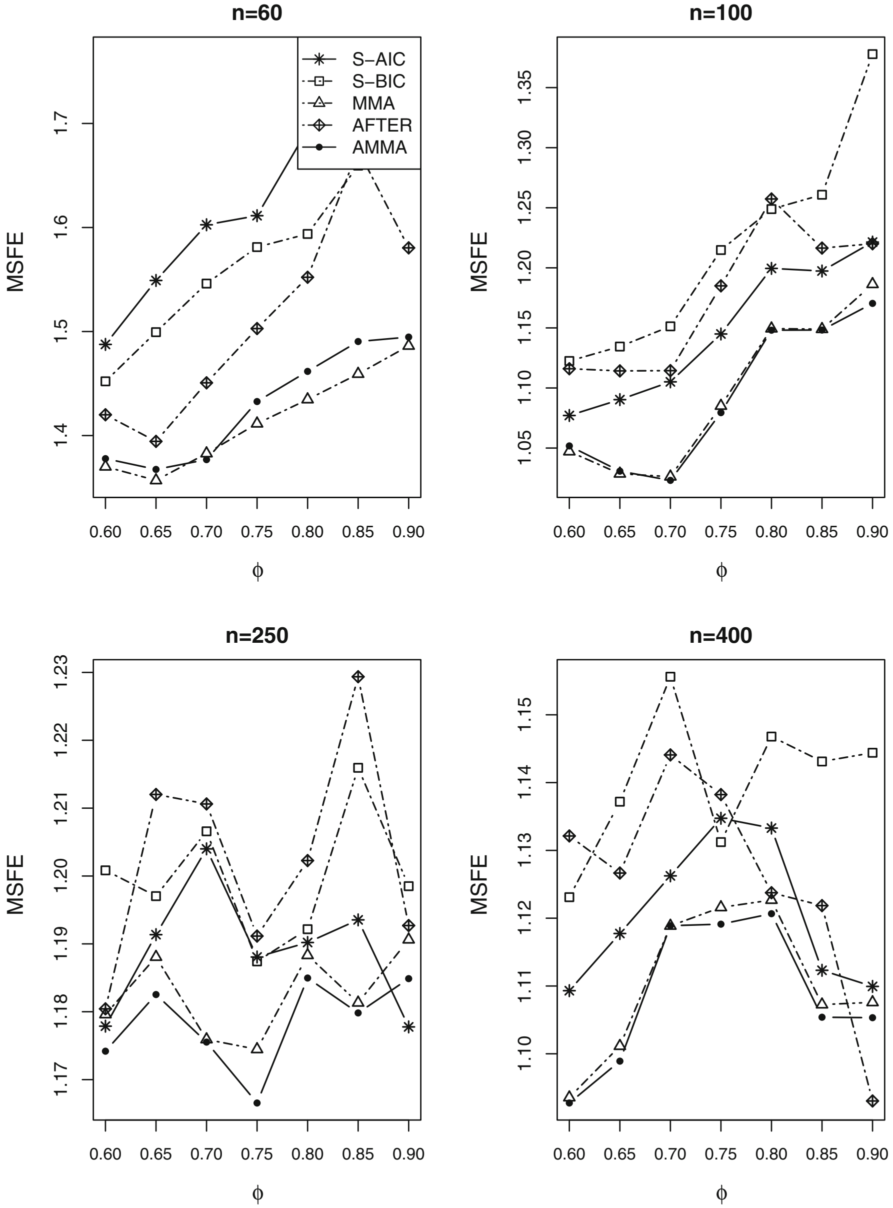

Fig. 8 Results of Simulation III. MSFEs for averaging TAR models with $\sigma^2 = 1$

y_{n+1}. Then, we use $\{y_2, \ldots, y_{n+1}\}$ to fit the model and predict y_{n+2}. By pushing on this procedure step by step, we can get $516 - n$ predictions at last. n is set at 60, 150, 250 and 400. We compare the AMMA method with the AIC, BIC, S-AIC, S-BIC, AFTER and MMA methods using the MSFE. We also report the standard deviation (SD) of the squared forecast error. The results are shown in Table 1.

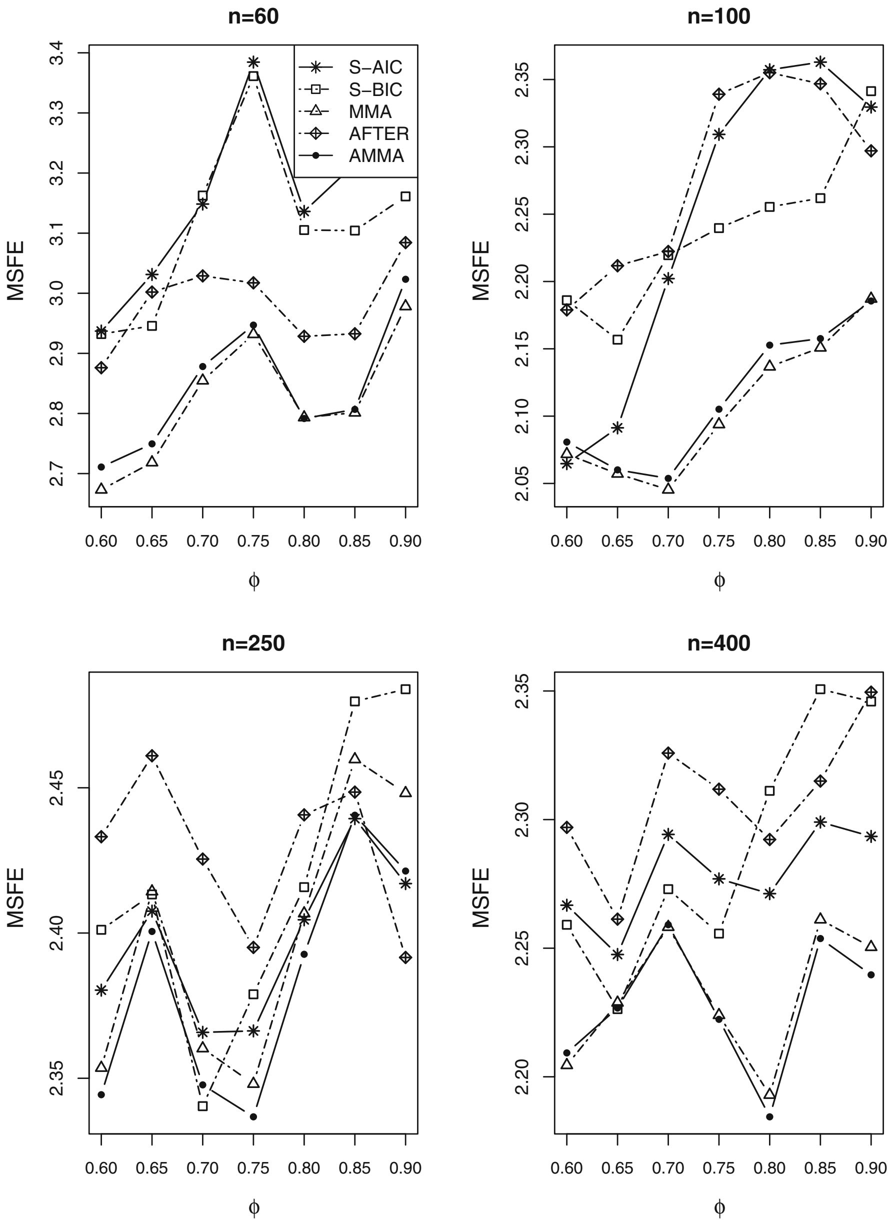

Fig. 9 Results of Simulation III. MSFEs for averaging TAR models with $\sigma^2 = 2$

The performance of the AMMA estimation is always better than that of the AIC, BIC, S-AIC and S-BIC methods, since its means are the lowest. When $n = 250$ and $n = 400$, the AMMA estimator has lower means than the MMA estimator, while the MMA performs better when $n = 60$ and $n = 150$.

Table 1 Squared forecast errors of different methods ($\times 10^{-2}$)

Method	$n = 60$		$n = 150$		$n = 250$		$n = 400$	
	MSFE	SD	MSFE	SD	MSFE	SD	MSFE	SD
AIC	9.6844	32.27	2.8071	4.999	2.1979	3.276	2.6816	3.610
BIC	5.4289	15.92	2.9072	5.284	2.5954	4.913	2.8980	3.894
S-AIC	7.8667	26.22	2.7287	4.872	2.1697	3.316	2.6597	3.540
S-BIC	5.5677	15.34	2.8495	5.209	2.5803	4.857	2.7529	3.850
AFTER	5.9782	17.15	2.7696	4.900	2.3260	3.708	2.7379	3.796
MMA	4.7168	8.714	2.5690	4.612	2.1750	3.401	2.5248	3.406
AMMA	5.3363	8.683	2.6127	4.647	2.1662	3.354	2.5193	3.396

6 Conclusion remarks

Threshold models have wide empirical applications. In this paper, two cases of averaging are considered: Case I studies models with different explanatory variables and a given estimated threshold parameter and Case II studies models with different explanatory variables at all possible threshold parameters. A new least squares MA estimator—the AMMA estimator—based on an approximation of GCV is developed. Compared with the MMA, our AMMA method has wider application because it does not require a unique largest model. When the threshold is estimated, the coefficient estimator in each candidate model is not a linear combination of the dependent variable Y, and the proof of asymptotic optimality is challenging. Both the simulations and the empirical analysis show the superiority of the AMMA estimator over some commonly used MS and MA estimators.

For future research along this line, one could extend our method to allow for multiple thresholds. For the case of TAR model averaging, one could allow the largest lag order of the TAR model to be unbounded asymptotically. As this paper mainly focuses on the asymptotic optimality of the AMMA estimator, the derivation of the consistency and asymptotic distribution of the AMMA estimator would also be an interesting future research topic. Hansen and Racine (2012) developed a jackknife model averaging (JMA) estimator under heteroscedastic error settings, and Zhang et al. (2013) studied the JMA in models with dependent data. Therefore, the development of a model averaging method for threshold models with heteroscedastic errors also warrants future research. Lastly, although we have developed theoretical properties for our model averaging method, they only hold in large sample sense. Understanding the asymptotic results when the sample size is limited and developing finite sample properties are also very necessary in the future research.

Acknowledgements We thank the editor and the two anonymous referees for their constructive comments. Zhang's work was partially supported by the National Natural Science Foundation of China (Grant Nos. 71522004, 11471324, 71463012 and 71631008) and a grant from the Ministry of Education of China (Grant No. 17YJC910011). Zou's work was partially supported by the National Natural Science Foundation of China (Grant Nos. 11529101 and 11331011) and a Grant from the Ministry of Science and Technology of China (Grant No. 2016YFB0502301).

Appendix

Lemma 1 *Let $\mathcal{W}$ be a weight vector set which can be related to the sample size n. Define*

$$w^* = \underset{w \in \mathcal{W}}{argmin} \left(L_n(w) + a_n(w) \right). \tag{17}$$

If

$$\sup_{w \in \mathcal{W}} \frac{|a_n(w)|}{R_n(w)} \xrightarrow{p} 0, \tag{18}$$

$$\sup_{w \in \mathcal{W}} \left| \frac{L_n(w)}{R_n(w)} - 1 \right| \xrightarrow{p} 0, \tag{19}$$

and there exists a constant κ_3 such that

$$\inf_{w \in \mathcal{W}} R_n(w) \geq \kappa_3 > 0, \tag{20}$$

then

$$\frac{L_n(w^*)}{\inf_{w \in \mathcal{W}} L_n(w)} \xrightarrow{p} 1. \tag{21}$$

Proof From the definition of the infimum, there exist a non-negative series ϑ_n and a vector $w(n) \in \mathcal{W}$ such that $\vartheta_n \to 0$ and

$$\inf_{w \in \mathcal{W}} L_n(w) = L_n(w(n)) - \vartheta_n. \tag{22}$$

In addition, it follows from (19) that

$$\begin{aligned} \inf_{w \in \mathcal{W}} \frac{L_n(w)}{R_n(w)} &= \inf_{w \in \mathcal{W}} \left(\frac{L_n(w)}{R_n(w)} - 1 \right) + 1 \\ &\geq - \sup_{w \in \mathcal{W}} \left| \frac{L_n(w)}{R_n(w)} - 1 \right| + 1 \xrightarrow{p} 1. \end{aligned} \tag{23}$$

From (20), (23) and $\vartheta_n \to 0$, we have

$$\begin{aligned} \inf_{w \in \mathcal{W}} \frac{|L_n(w) - \vartheta_n|}{R_n(w)} \geq \inf_{w \in \mathcal{W}} \frac{L_n(w) - \vartheta_n}{R_n(w)} &\geq \inf_{w \in \mathcal{W}} \frac{L_n(w)}{R_n(w)} - \frac{\vartheta_n}{\inf_{w \in \mathcal{W}} R_n(w)} \\ &\geq - \sup_{w \in \mathcal{W}} \left| \frac{L_n(w)}{R_n(w)} - 1 \right| + 1 - \frac{\vartheta_n}{\inf_{w \in \mathcal{W}} R_n(w)} \\ &\xrightarrow{p} 1. \end{aligned} \tag{24}$$

Now, by the definition of w^*, (18), (20), (22)–(24), and $\vartheta_n \to 0$, we have, for any $\delta > 0$,

$$
\begin{aligned}
&\Pr\left\{\left|\frac{\inf_{w\in\mathcal{W}} L_n(w)}{L_n(w^*)} - 1\right| > \delta\right\} = \Pr\left\{\frac{L_n(w^*) - \inf_{w\in\mathcal{W}} L_n(w)}{L_n(w^*)} > \delta\right\} \\
&= \Pr\left\{\frac{\inf_{w\in\mathcal{W}}\left(L_n(w) + a_n(w)\right) - a_n(w^*) - \inf_{w\in\mathcal{W}} L_n(w)}{L_n(w^*)} > \delta\right\} \\
&\leq \Pr\left\{\frac{L_n(w(n)) + a_n(w(n)) - a_n(w^*) - L_n(w(n)) + \vartheta_n}{L_n(w^*)} > \delta\right\} \\
&\leq \Pr\left\{\frac{|a_n(w(n))|}{L_n(w^*)} + \frac{|a_n(w^*)|}{L_n(w^*)} + \frac{\vartheta_n}{L_n(w^*)} > \delta\right\} \\
&\leq \Pr\left\{\frac{|a_n(w(n))|}{\inf_{w\in\mathcal{W}} L_n(w)} + \frac{|a_n(w^*)|}{L_n(w^*)} + \frac{\vartheta_n}{L_n(w^*)} > \delta\right\} \\
&= \Pr\left\{\frac{|a_n(w(n))|}{L_n(w(n)) - \vartheta_n} + \frac{|a_n(w^*)|}{L_n(w^*)} + \frac{\vartheta_n}{L_n(w^*)} > \delta\right\} \\
&\leq \Pr\left\{\sup_{w\in\mathcal{W}} \frac{|a_n(w)|}{L_n(w) - \vartheta_n} + \sup_{w\in\mathcal{W}} \frac{|a_n(w)|}{L_n(w)} + \sup_{w\in\mathcal{W}} \frac{\vartheta_n}{L_n(w)} > \delta\right\} \\
&\leq \Pr\left\{\sup_{w\in\mathcal{W}} \frac{|a_n(w)|}{R_n(w)} \sup_{w\in\mathcal{W}} \frac{R_n(w)}{|L_n(w) - \vartheta_n|} + \sup_{w\in\mathcal{W}} \frac{|a_n(w)|}{R_n(w)} \sup_{w\in\mathcal{W}} \frac{R_n(w)}{L_n(w)}\right. \\
&\quad \left. + \sup_{w\in\mathcal{W}} \frac{\vartheta_n}{R_n(w)} \sup_{w\in\mathcal{W}} \frac{R_n(w)}{L_n(w)} > \delta\right\} \\
&= \Pr\left\{\sup_{w\in\mathcal{W}} \frac{|a_n(w)|}{R_n(w)} \left[\inf_{w\in\mathcal{W}} \frac{|L_n(w) - \vartheta_n|}{R_n(w)}\right]^{-1} + \sup_{w\in\mathcal{W}} \frac{|a_n(w)|}{R_n(w)} \left[\inf_{w\in\mathcal{W}} \frac{L_n(w)}{R_n(w)}\right]^{-1}\right. \\
&\quad \left. + \frac{\vartheta_n}{\inf_{w\in\mathcal{W}} R_n(w)} \left[\inf_{w\in\mathcal{W}} \frac{L_n(w)}{R_n(w)}\right]^{-1} > \delta\right\} \\
&\to 0. \qquad (25)
\end{aligned}
$$

Therefore, $\inf_{w\in\mathcal{W}} L_n(w)/L_n(w^*) \xrightarrow{p} 1$, which implies (21). □

Proof of Theorem 1. First, from the fact that $X_{(m)}(\gamma)$ is of full column rank, we have $tr\,\hat{P}(w) = tr\,P^*(w) \leq 2\sum_{m=1}^{M} w_m k_m$. Let $\hat{A}(w) = I_n - \hat{P}(w)$, so that

$$
\begin{aligned}
\mathcal{L}_n(w) &= \|Y - \hat{\mu}(w)\|^2 \Big(1 + 2\frac{tr\,\hat{P}(w)}{n}\Big) \\
&= L_n(w) + \|e\|^2 + 2\mu'(\hat{A}(w) - A^*(w))e + 2\mu' A^*(w)e \\
&\quad + 2\big(\sigma^2 tr\,P^*(w) - e' P^*(w)e\big) + 2e'\big(P^*(w) - \hat{P}(w)\big)e \\
&\quad + 2tr\,P^*(w)\big(\|A^*(w)Y\|^2/n - \sigma^2\big) \\
&\quad + 2tr\,P^*(w)\big(\|\hat{A}(w)Y\|^2 - \|A^*(w)Y\|^2\big)/n.
\end{aligned}
$$

Since $\|e\|^2$ is unrelated to w and Condition (20) with $\mathcal{W} = \mathcal{H}_n$ is implied by Condition (7), according to Lemma 1, Theorem 1 is valid if

$$\sup_{w\in\mathcal{H}_n} |\mu' A^*(w)e|/R_n^*(w) \xrightarrow{p} 0, \tag{26}$$

$$\sup_{w\in\mathcal{H}_n} |e' P^*(w)e - \sigma^2 tr P^*(w)|/R_n^*(w) \xrightarrow{p} 0, \tag{27}$$

$$\sup_{w\in\mathcal{H}_n} |L_n^*(w)/R_n^*(w) - 1| \xrightarrow{p} 0, \tag{28}$$

$$\sup_{w\in\mathcal{H}_n} |tr P^*(w)(\|A^*(w)Y\|^2/n - \sigma^2)|/R_n^*(w) \xrightarrow{p} 0, \tag{29}$$

$$\sup_{w\in\mathcal{H}_n} \left|\mu'\big(P^*(w) - \hat{P}(w)\big)e\right|/R_n^*(w) \xrightarrow{p} 0, \tag{30}$$

$$\sup_{w\in\mathcal{H}_n} \left|e'\big(P^*(w) - \hat{P}(w)\big)e\right|/R_n^*(w) \xrightarrow{p} 0, \tag{31}$$

$$\sup_{w\in\mathcal{H}_n} |L_n(w) - L_n^*(w)|/R_n^*(w) \xrightarrow{p} 0, \tag{32}$$

and

$$\sup_{w\in\mathcal{H}_n} \left|tr P^*(w)\big(\|A^*(w)Y\|^2 - \|\hat{A}(w)Y\|^2\big)\right|/nR_n^*(w) \xrightarrow{p} 0. \tag{33}$$

(26)–(28) can been shown by following the proof of Theorem 1$'$ of Wan et al. (2010). Therefore, we only need to verify (29)–(33). First, we prove (29). Note that

$$\begin{aligned}
&\sup_{w\in\mathcal{H}_n} |tr P^*(w)\big(\|A^*(w)Y\|^2/n - \sigma^2\big)|/R_n^*(w)\\
&= \sup_{w\in\mathcal{H}_n} \left\{\frac{tr P^*(w)}{nR_n^*(w)}\left|\|\mu - P^*(w)Y\|^2 + \|e\|^2 + 2\mu' A^*(w)e - 2e' P^*(w)e - n\sigma^2\right|\right\}\\
&\le \sup_{w\in\mathcal{H}_n} \frac{L_n^*(w)}{R_n^*(w)} \sup_{w\in\mathcal{H}_n} \frac{tr P^*(w)}{n} + \sup_{w\in\mathcal{H}_n} \frac{2|\mu' A^*(w)e|}{R_n^*(w)} \sup_{w\in\mathcal{H}_n} \frac{tr P^*(w)}{n}\\
&\quad + \frac{|\|e\|^2 - n\sigma^2|}{\sqrt{n}} \sup_{w\in\mathcal{H}_n} \frac{1}{R_n^*(w)} \sup_{w\in\mathcal{H}_n} \frac{tr P^*(w)}{\sqrt{n}}\\
&\quad + \sup_{w\in\mathcal{H}_n} \frac{2|e' P^*(w)e - \sigma^2 tr P^*(w)|}{R_n^*(w)} \sup_{w\in\mathcal{H}_n} \frac{tr P^*(w)}{n}\\
&\quad + 2\sigma^2 \sup_{w\in\mathcal{H}_n} \frac{1}{R_n^*(w)} \sup_{w\in\mathcal{H}_n} \frac{tr^2 P^*(w)}{n}.
\end{aligned}$$

By the central limit theorem, we have $|\|e\|^2 - n\sigma^2|/\sqrt{n} = O_p(1)$. In addition, it follows from (7) and (9) that

$$\sup_{w\in\mathcal{H}_n} \frac{1}{R_n^*(w)} = o_p(1), \quad \sup_{w\in\mathcal{H}_n} tr^2 P^*(w)/n = O(1) \quad \text{and} \quad \sup_{w\in\mathcal{H}_n} tr P^*(w)/n = o(1).$$

Together with (26)–(28), (29) is obtained.

To prove (30), we observe that

$$\begin{aligned}
&\sup_{w\in\mathcal{H}_n}\left|\mu'\big(P^*(w)-\hat{P}(w)\big)e\right|/R_n^*(w)\\
&\quad\le\frac{1}{\xi_n^*}\sup_{w\in\mathcal{H}_n}\left[\|\mu\|^2e'\big(P^*(w)-\hat{P}(w)\big)^2e\right]^{1/2}\\
&\quad\le\frac{1}{\xi_n^*}\frac{\|\mu\|}{\sqrt{n}}\frac{\|e\|}{\sqrt{n}}\,n\max_{1\le m\le M}\lambda_{\max}(P^*_{(m)}-\hat{P}_{(m)}).
\end{aligned}$$

By Conditions (8) and (10), (30) is verified.

Note that

$$L_n(w)=\|e\|^2+\|\hat{A}(w)\mu\|^2+\|\hat{A}(w)e\|^2-2e'\hat{A}(w)\mu-2e'\hat{A}(w)e+2\mu'\hat{A}^2(w)e,$$

so

$$\begin{aligned}
&\sup_{w\in\mathcal{H}_n}|L_n(w)-L_n^*(w)|/R_n^*(w)\xrightarrow{p}0\Leftrightarrow\\
&\quad\sup_{w\in\mathcal{H}_n}\Big|2\mu'\big(P^*(w)-\hat{P}(w)\big)\mu+2\mu'\big(P^*(w)-\hat{P}(w)\big)e\\
&\quad-\mu'\big(P^*(w)+\hat{P}(w)\big)\big(P^*(w)-\hat{P}(w)\big)\mu\\
&\quad-e'\big(P^*(w)+\hat{P}(w)\big)\big(P^*(w)-\hat{P}(w)\big)e\\
&\quad-2\mu'P^*(w)\big(P^*(w)-\hat{P}(w)\big)e\\
&\quad-2\mu'\big(P^*(w)-\hat{P}(w)\big)\hat{P}(w)e\Big|/R_n^*(w)\xrightarrow{p}0.
\end{aligned}$$

Thus, if

$$\sup_{w\in\mathcal{H}_n}\left|\mu'\big(P^*(w)+\hat{P}(w)\big)\big(P^*(w)-\hat{P}(w)\big)\mu\right|/R_n^*(w)\xrightarrow{p}0,\tag{34}$$

$$\sup_{w\in\mathcal{H}_n}\left|e'\big(P^*(w)+\hat{P}(w)\big)\big(P^*(w)-\hat{P}(w)\big)e\right|/R_n^*(w)\xrightarrow{p}0,\tag{35}$$

$$\sup_{w\in\mathcal{H}_n}\left|\mu'P^*(w)\big(P^*(w)-\hat{P}(w)\big)e\right|/R_n^*(w)\xrightarrow{p}0,\tag{36}$$

$$\sup_{w\in\mathcal{H}_n}\left|\mu'\big(P^*(w)-\hat{P}(w)\big)\hat{P}(w)e\right|/R_n^*(w)\xrightarrow{p}0,\tag{37}$$

and

$$\sup_{w\in\mathcal{H}_n}\left|\mu'\big(P^*(w)-\hat{P}(w)\big)\mu\right|/R_n^*(w)\xrightarrow{p}0,\tag{38}$$

then (32) is valid. From Condition (8) and the following result

$$
\begin{aligned}
&\sup_{w\in\mathcal{H}_n}\big|e'\big(P^*(w)+\hat{P}(w)\big)\big(P^*(w)-\hat{P}(w)\big)e\big|/R_n^*(w)\\
&\le \frac{1}{2\xi_n^*}\sup_{w\in\mathcal{H}_n}\big|e'\big[\big(P^*(w)+\hat{P}(w)\big)\big(P^*(w)-\hat{P}(w)\big)\\
&\quad+\big(P^*(w)-\hat{P}(w)\big)\big(P^*(w)+\hat{P}(w)\big)\big]e\big|\\
&\le \frac{\|e\|^2}{2\xi_n^*}\sup_{w\in\mathcal{H}_n}\lambda_{\max}\big[\big(P^*(w)+\hat{P}(w)\big)\big(P^*(w)-\hat{P}(w)\big)\\
&\quad+\big(P^*(w)-\hat{P}(w)\big)\big(P^*(w)+\hat{P}(w)\big)\big]\\
&\le \frac{\|e\|^2}{\xi_n^*}\sup_{w\in\mathcal{H}_n}\big[\lambda_{\max}\big(P^*(w)+\hat{P}(w)\big)\lambda_{\max}\big(P^*(w)-\hat{P}(w)\big)\big]\\
&\le \frac{\|e\|^2}{\xi_n^*}\sup_{w\in\mathcal{H}_n}\big[\lambda_{\max}\big(P^*(w)\big)+\lambda_{\max}\big(\hat{P}(w)\big)\big]\sum_{m=1}^{M}w_m\lambda_{\max}(P^*_{(m)}-\hat{P}_{(m)})\\
&\le \frac{2}{\xi_n^*}\frac{\|e\|^2}{n}n\max_{1\le m\le M}\lambda_{\max}(P^*_{(m)}-\hat{P}_{(m)}),
\end{aligned}
$$

we obtain (35). Similarly, (31), (34) and (38) can be verified. On the other hand, analogous to the proof of (30), one can obtain (36) and (37).

Further, it can be shown that

$$
\begin{aligned}
&\sup_{w\in\mathcal{H}_n}\big|tr\,P^*(w)\big(\|A^*(w)Y\|^2-\|\hat{A}(w)Y\|^2\big)\big|/nR_n^*(w)\\
&\le \sup_{w\in\mathcal{H}_n}\frac{tr\,P^*(w)}{n}\sup_{w\in\mathcal{H}_n}\frac{|\|A^*(w)Y\|^2-\|\hat{A}(w)Y\|^2|}{R_n^*(w)}\\
&\le a_1\sup_{w\in\mathcal{H}_n}\frac{|\|A^*(w)Y\|^2-\|\hat{A}(w)Y\|^2|}{R_n^*(w)},
\end{aligned}
$$

where the last step is from Condition (9). Observe that

$$
\begin{aligned}
&|\|A^*(w)Y\|^2-\|\hat{A}(w)Y\|^2|\\
&=|2\mu'(\hat{P}(w)-P^*(w))\mu+\mu'(P^*(w)+\hat{P}(w))(P^*(w)-\hat{P}(w))\mu\\
&\quad+2e'(\hat{P}(w)-P^*(w))e+e'(P^*(w)+\hat{P}(w))(P^*(w)-\hat{P}(w))e\\
&\quad+4\mu'(\hat{P}(w)-P^*(w))e+2\mu'P^*(w)(P^*(w)-\hat{P}(w))e\\
&\quad+2\mu'(P^*(w)-\hat{P}(w))\hat{P}(w)e|,
\end{aligned}
$$

so from (30), (31) and (34)–(38), we have

$$\sup_{w\in\mathcal{H}_n} \frac{|\|A^*(w)Y\|^2 - \|\hat{A}(w)Y\|^2|}{R_n^*(w)} \xrightarrow{p} 0.$$

Thus, we obtain (33). This completes the proof of Theorem 1. □

The following lemma is used in the proof of Theorem 2.

Lemma 2 *For any $\hat{\gamma}_{(m)}$ and $\gamma^*_{(m)} \in \Gamma$ and any random variable Y, if Assumptions (a.3) and (a.4) are satisfied, and*

$$|E(Y|z_i = \gamma, \hat{\gamma}_{(m)})| \leq \bar{E}, \tag{39}$$

where $\bar{E}$ is a finite constant, then

$$E\big(Y|I(z_i \leq \gamma^*_{(m)}) - I(z_i \leq \hat{\gamma}_{(m)})|\big) = O(n^{-\rho}). \tag{40}$$

Proof The proof is similar to that of Lemma A.1 in Hansen (2000).

$$\begin{aligned}\frac{\partial E(Y\mathrm{I}(z_i \leq \gamma)|\hat{\gamma}_{(m)})}{\partial\gamma} &= \int_{-\infty}^{+\infty} y\frac{\partial\int_{-\infty}^{\gamma} f(y,z|\hat{\gamma}_{(m)})\,dz}{\partial\gamma}dy\\ &= \int_{-\infty}^{+\infty} yf(y,\gamma|\hat{\gamma}_{(m)})dy\\ &= \int_{-\infty}^{+\infty} yf_1(y|\gamma,\hat{\gamma}_{(m)})f_2(\gamma|\hat{\gamma}_{(m)})dy\\ &= f_2(\gamma|\hat{\gamma}_{(m)})E(Y\big|z_i = \gamma,\hat{\gamma}_{(m)}),\end{aligned}$$

where f, f_1 and f_2 are density functions. Let $C = \bar{f}_2\bar{E}$. By Lagrange's mean value theorem, there exists a $\tilde{\gamma}_{(m)}$ between $\gamma^*_{(m)}$ and $\hat{\gamma}_{(m)}$ such that

$$\begin{aligned}&E(Y\mathrm{I}(z_i \leq \hat{\gamma}_{(m)})|\hat{\gamma}_{(m)}) - E(Y\mathrm{I}(z_i \leq \gamma^*_{(m)})|\hat{\gamma}_{(m)})\\ &\quad= f_2(\tilde{\gamma}_{(m)}|\hat{\gamma}_{(m)})E(Y\big|z_i = \tilde{\gamma}_{(m)},\hat{\gamma}_{(m)})(\hat{\gamma}_{(m)} - \gamma^*_{(m)})\\ &\quad\leq C|\hat{\gamma}_{(m)} - \gamma^*_{(m)}|.\end{aligned} \tag{41}$$

Define $f_3(\gamma)$ as the density of $\hat{\gamma}_{(m)}$. By (41) and Assumptions (a.3) and (a.4), we have

$$\begin{aligned}&E\big(Y|\mathrm{I}(z_i \leq \gamma^*_{(m)}) - \mathrm{I}(z_i \leq \hat{\gamma}_{(m)})|\big)\\ &\quad= \int_{\underline{\gamma}}^{\bar{\gamma}} E\big(Y|\mathrm{I}(z_i \leq \gamma^*_{(m)}) - \mathrm{I}(z_i \leq \hat{\gamma}_{(m)})|\big|\hat{\gamma}_{(m)}\big)f_3(\hat{\gamma}_{(m)})d\hat{\gamma}_{(m)}\\ &\quad= \int_{\underline{\gamma}}^{\gamma^*_{(m)}} E\big(Y\big(\mathrm{I}(z_i \leq \gamma^*_{(m)}) - \mathrm{I}(z_i \leq \hat{\gamma}_{(m)})\big)\big|\hat{\gamma}_{(m)}\big)f_3(\hat{\gamma}_{(m)})d\hat{\gamma}_{(m)}\end{aligned}$$

$$
\begin{aligned}
&+\int_{\gamma_{(m)}^*}^{\bar{\gamma}} E\big(Y\big(\mathrm{I}(z_i \le \hat{\gamma}_{(m)}) - \mathrm{I}(z_i \le \gamma_{(m)}^*)\big)\big|\hat{\gamma}_{(m)}\big) f_3(\hat{\gamma}_{(m)}) d\hat{\gamma}_{(m)} \\
&\le \int_{\underline{\gamma}}^{\bar{\gamma}} C|\hat{\gamma}_{(m)} - \gamma_{(m)}^*| f_3(\hat{\gamma}_{(m)}) d\hat{\gamma}_{(m)} = O(n^{-\rho}).
\end{aligned}
$$

The proof of Lemma 2 is completed. □

Proof of Theorem 2. Note that $\mu' A^*(w)e = \mu' e - \mu' P^*(w)e$. From the proof of Theorem 1 and the fact that $\mu' e$ is unrelated to w, Theorem 2 is valid if

$$\sup_{w \in \mathcal{H}_n} |e'P^*(w)e - \sigma^2 tr P^*(w)|/Q_n^*(w) \xrightarrow{p} 0, \tag{42}$$

$$\sup_{w \in \mathcal{H}_n} |\mu' P^*(w)e|/Q_n^*(w) \xrightarrow{p} 0, \tag{43}$$

$$\sup_{w \in \mathcal{H}_n} |L_n^*(w)/Q_n^*(w) - 1| \xrightarrow{p} 0, \tag{44}$$

$$\sup_{w \in \mathcal{H}_n} |tr P^*(w)(\|A^*(w)Y\|^2/n - \sigma^2)|/Q_n^*(w) \xrightarrow{p} 0, \tag{45}$$

$$\sup_{w \in \mathcal{H}_n} \big|\mu'\big(P^*(w) - \hat{P}(w)\big)e\big|/Q_n^*(w) \xrightarrow{p} 0, \tag{46}$$

$$\sup_{w \in \mathcal{H}_n} \big|e'\big(P^*(w) - \hat{P}(w)\big)e\big|/Q_n^*(w) \xrightarrow{p} 0, \tag{47}$$

$$\sup_{w \in \mathcal{H}_n} |L_n(w) - L_n^*(w)|/Q_n^*(w) \xrightarrow{p} 0, \tag{48}$$

and

$$\sup_{w \in \mathcal{H}_n} \big|tr P^*(w)\big(\|A^*(w)Y\|^2 - \|\hat{A}(w)Y\|^2\big)\big|/nQ_n^*(w) \xrightarrow{p} 0. \tag{49}$$

Because x_i contains the lag values of y_i, the proofs of (42)–(44) are different from those of (26)–(28).

According to Theorem 3.35 of White (1984), Assumption (a.1) implies that $x_{(m)i}x'_{(m)i}\mathrm{I}(z_i \le \gamma_{(m)}^*)$ is stationary and ergodic. Further, Assumption (a.2) ensures $E|x_{(m)ij}x_{(m)ik}\mathrm{I}(z_i \le \gamma_{(m)}^*)| < \infty$. By Theorem 3.34 of White (1984), we have

$$\frac{X_{(m)}^{*\prime} X_{(m)}^*}{n} \xrightarrow{p} \begin{pmatrix} E(x_{(m)i}x'_{(m)i}\mathrm{I}(z_i \le \gamma_{(m)}^*)) & 0 \\ 0 & E(x_{(m)i}x'_{(m)i}\mathrm{I}(z_i > \gamma_{(m)}^*)) \end{pmatrix} \equiv V_{(m)}, \tag{50}$$

where $V_{(m)}$ is an invertible matrix. From Assumptions (a.1) and (a.2), $x_i\mathrm{I}(z_i \le \gamma)e_i$ is a square integrable stationary martingale difference sequence. Therefore, by the central limit theorem for martingale difference sequence, we obtain $\frac{1}{\sqrt{n}}X_{(m)}^{*\prime}e \xrightarrow{d} N(0, \sigma^2 V_{(m)})$. Thus, $\frac{1}{\sqrt{n}}X_{(m)}^{*\prime}e = O_p(1)$. Together with the fact that k_{M^*} and M are bounded, it can be shown that

$$e'P^*_{(m)}e = \frac{1}{\sqrt{n}}e'X^*_{(m)}\left(\frac{X^{*\prime}_{(m)}X^*_{(m)}}{n}\right)^{-1}\frac{1}{\sqrt{n}}X^{*\prime}_{(m)}e = O_p(1) \tag{51}$$

and

$$tr P^*(w) = \sum_{m=1}^{M} w_m tr P^*_{(m)} \leq 2\sum_{m=1}^{M} w_m k_m \leq 2k_{M^*} < \infty. \tag{52}$$

From Condition (12), we have

$$\sup_{w\in\mathcal{H}_n} |e'P^*(w)e - \sigma^2 tr P^*(w)|/Q^*_n(w) \leq \zeta_n^{*-1} \max_{1\leq m\leq M} |e'P^*_{(m)}e| + 2\zeta_n^{*-1}\sigma^2 k_{M^*} \xrightarrow{p} 0. \tag{53}$$

Consequently, (42) is verified.

Under (51) and Condition (10), it can be shown that

$$\begin{aligned}|\mu'P^*(w)e| &= |e'P^*(w)\mu\mu'P^*(w)e|^{\frac{1}{2}} \leq \|\mu\||e'P^{*2}(w)e|^{\frac{1}{2}} \\ &\leq \|\mu\|\lambda^{1/2}_{\max}\big(P^*(w)\big)|e'P^*(w)e|^{1/2} = O_p(\sqrt{n}).\end{aligned} \tag{54}$$

Hence, (43) is valid by Condition (12).

For (44), similar to (54), it can be shown that

$$e'P^{*2}(w)e = O_p(1) \tag{55}$$

and

$$|\mu'P^{*2}(w)e| = O_p(\sqrt{n}). \tag{56}$$

In addition,

$$tr P^{*2}(w) \leq \lambda_{\max}\big(P^*(w)\big) tr P^*(w) \leq 2k_{M^*}. \tag{57}$$

Thus,

$$\begin{aligned}|L^*_n(w) - Q^*_n(w)| &= \big|\|P^*(w)e\|^2 - 2\mu'A^*(w)P^*(w)e - \sigma^2 tr P^{*2}(w)\big| \\ &\leq \|P^*(w)e\|^2 + 2|\mu'P^*(w)e| + 2|\mu'P^{*2}(w)e| + 2\sigma^2 k_{M^*} \\ &= O_p(\sqrt{n}).\end{aligned}$$

Hence, (44) holds by Condition (12).

The proof of (45) is similar to that of (29). From the proofs of (30)–(33), if

$$n\zeta_n^{*-1} \max_{1\leq m\leq M} \lambda_{\max}(P^*_{(m)} - \hat{P}_{(m)}) \xrightarrow{p} 0, \tag{58}$$

then (46)–(49) will hold. In the following, we will verify (58).

By Lemma 2, for the mth candidate model,

$$E|x_{(m)ij}x_{(m)ik}\big(\mathrm{I}(z_i \le \gamma^*_{(m)}) - \mathrm{I}(z_i \le \hat{\gamma}_{(m)})\big)| = O(n^{-\rho})$$

uniformly in i. Hence,

$$\frac{X^{*\prime}_{(m)}X^*_{(m)}}{n} - \frac{\hat{X}'_{(m)}\hat{X}_{(m)}}{n} = O_p(n^{-\rho}), \tag{59}$$

and

$$\frac{(X^*_{(m)} - \hat{X}_{(m)})'(X^*_{(m)} - \hat{X}_{(m)})}{n} = O_p(n^{-\rho}). \tag{60}$$

From (50) and (59), it follows that

$$\frac{\hat{X}'_{(m)}\hat{X}_{(m)}}{n} \xrightarrow{p} V_{(m)}. \tag{61}$$

Thus, by (50), (59) and (61), we obtain

$$\left(\frac{X^{*\prime}_{(m)}X^*_{(m)}}{n}\right)^{-1} - \left(\frac{\hat{X}'_{(m)}\hat{X}_{(m)}}{n}\right)^{-1} = O_p(n^{-\rho}). \tag{62}$$

Note that

$$\begin{aligned} P^*_{(m)} - \hat{P}_{(m)} = {} & X^*_{(m)}[(X^{*\prime}_{(m)}X^*_{(m)})^{-1} - (\hat{X}'_{(m)}\hat{X}_{(m)})^{-1}]X^{*\prime}_{(m)} \\ & -(\hat{X}_{(m)} - X^*_{(m)})(\hat{X}'_{(m)}\hat{X}_{(m)})^{-1}(\hat{X}_{(m)} - X^*_{(m)})' \\ & -(\hat{X}_{(m)} - X^*_{(m)})(\hat{X}'_{(m)}\hat{X}_{(m)})^{-1}X^{*\prime}_{(m)} \\ & -X^*_{(m)}(\hat{X}'_{(m)}\hat{X}_{(m)})^{-1}(\hat{X}_{(m)} - X^*_{(m)})' \\ \equiv {} & \Delta P_{(m)1} + \Delta P_{(m)2} + \Delta P_{(m)3} + \Delta P_{(m)4}. \end{aligned} \tag{63}$$

By using (60)–(62), we have

$$\begin{aligned} \lambda_{\max}(\Delta P_{(m)1}) &\le \lambda_{\max}\left[\left(\frac{X^{*\prime}_{(m)}X^*_{(m)}}{n}\right)^{-1} - \left(\frac{\hat{X}'_{(m)}\hat{X}_{(m)}}{n}\right)^{-1}\right]\lambda_{\max}\left(\frac{X^{*\prime}_{(m)}X^*_{(m)}}{n}\right) \\ &= O_p(n^{-\rho}), \\ \lambda_{\max}(\Delta P_{(m)2}) &\le \lambda_{\max}\left[\left(\frac{\hat{X}'_{(m)}\hat{X}_{(m)}}{n}\right)^{-1}\right]\lambda_{\max}\left(\frac{(\hat{X}_{(m)} - X^*_{(m)})'(\hat{X}_{(m)} - X^*_{(m)})}{n}\right) \\ &= O_p(n^{-\rho}), \end{aligned}$$

and

$$\begin{aligned}
&\lambda_{\max}(\Delta P_{(m)3}) = \lambda_{\max}(\Delta P_{(m)4}) \\
&\quad = \lambda_{\max}^{1/2}\big((\hat{X}_{(m)} - X^*_{(m)})(\hat{X}'_{(m)}\hat{X}_{(m)})^{-1}X^{*\prime}_{(m)}X^*_{(m)}(\hat{X}'_{(m)}\hat{X}_{(m)})^{-1}(\hat{X}_{(m)} - X^*_{(m)})'\big) \\
&\quad \le \lambda_{\max}\left[\left(\frac{\hat{X}'_{(m)}\hat{X}_{(m)}}{n}\right)^{-1}\right]\lambda_{\max}^{1/2}\left(\frac{X^{*\prime}_{(m)}X^*_{(m)}}{n}\right) \\
&\qquad \lambda_{\max}^{1/2}\left(\frac{(\hat{X}_{(m)} - X^*_{(m)})'(\hat{X}_{(m)} - X^*_{(m)})}{n}\right) \\
&\quad = O_p(n^{-\rho/2}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lambda_{\max}(P^*_{(m)} - \hat{P}_{(m)}) &\le \lambda_{\max}(\Delta P_{(m)1}) + \lambda_{\max}(\Delta P_{(m)2}) \\
&\quad + \lambda_{\max}(\Delta P_{(m)3}) + \lambda_{\max}(\Delta P_{(m)4}) \\
&= O_p(n^{-\rho/2}).
\end{aligned}$$

Thus, (58) holds under Condition (12). The proof of Theorem 2 is completed. □

Proof of Theorem 3. Let $A(w) = I_n - P(w)$. From Lemma 1, we need only to verify that

$$\sup_{w\in\widetilde{\mathcal{H}}_n} |\mu' A(w)e|/\widetilde{R}_n(w) \xrightarrow{p} 0, \tag{64}$$

$$\sup_{w\in\widetilde{\mathcal{H}}_n} |e'P(w)e - \sigma^2 tr P(w)|/\widetilde{R}_n(w) \xrightarrow{p} 0, \tag{65}$$

$$\sup_{w\in\widetilde{\mathcal{H}}_n} |\widetilde{L}_n(w)/\widetilde{R}_n(w) - 1| \xrightarrow{p} 0, \tag{66}$$

and

$$\sup_{w\in\widetilde{\mathcal{H}}_n} |tr P(w)(\|A(w)Y\|^2/n - \sigma^2)|/\widetilde{R}_n(w) \xrightarrow{p} 0. \tag{67}$$

We obtain (64)–(66) by following the proof of Theorem 1′ of Wan et al. (2010), while (67) is valid from the proof of (29). □

References

Buckland, S. T., Burnham, K. P., Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, *53*, 603–618.

Caner, M., Hansen, B. E. (2001). Threshold autoregression with a unit root. *Econometrica*, *69*, 1555–1596.

Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics*, *21*, 520–533.

Cheng, T. C. F., Ing, C. K., Yu, S. H. (2014). Inverse moment bounds for sample autocovariance matrices based on detrended time series and their applications. *Linear Algebra & Its Applications*, *473*, 180–201.

Cheng, T. C. F., Ing, C. K., Yu, S. H. (2015). Toward optimal model averaging in regression models with time series errors. *Journal of Econometrics*, *189*, 321–334.

Craven, P., Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, *31*, 377–403.

Cuaresma, J. C., Doppelhofer, G. (2007). Nonlinearities in cross-country growth regressions: A Bayesian averaging of thresholds (BAT) approach. *Journal of Macroeconomics*, *29*, 541–554.

Delgado, M. A., Hidalgo, J. (2000). Nonparametric inference on structural breaks. *Journal of Econometrics*, *96*, 113–144.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica*, *68*, 575–603.

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

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics*, *146*, 342–350.

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

Hansen, B. E., Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, *167*, 38–46.

Hjort, N. L., Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, *98*, 879–899.

Kapetanios, G. (2001). Model selection in threshold models. *Journal of Time Series Analysis*, *22*, 733–754.

Koo, B., Seo, M. H. (2015). Structural-break models under mis-specification: Implications for forecasting. *Social Science Electronic Publishing*, *188*, 166–181.

Li, K. C. (1987). Asymptotic optimality for C_p, C_l, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, *15*, 958–975.

Liang, H., Zou, G., Wan, A. T. K., Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association*, *106*, 1053–1066.

Liu, Q., Okui, R. (2013). Heteroskedasticity-robust C_p model averaging. *Econometrics Journal*, *16*, 463–472.

Shen, X., Huang, H. C. (2006). Optimal model assessment, selection and combination. *Journal of the American Statistical Association*, *101*, 554–568.

Tong, H. (1983). *Threshold models in nonlinear time series analysis: Lecture notes in statistics* (Vol. 21). Berlin: Springer.

Tong, H. (1990). *Non-linear time series: A dynamical system approach*. Oxford: Oxford University Press.

Tong, H., Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society-Series B*, *42*, 245–292.

Wan, A. T. K., Zhang, X., Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, *156*, 277–283.

White, H. (1984). *Asymptotic theory for econometricians*. Orlando, Florida: Academic Press.

Xu, G., Wang, S., Huang, J. (2013). Focused information criterion and model averaging based on weighted composite quantile regression. *Scandinavian Journal of Statistics*, *41*, 365–381.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, *96*, 574–588.

Yang, Y. (2004). Combining forecasting procedures: Some theoretical resutls. *Econometric Theory*, *20*, 176–222.

Zhang, X., Wan, A. T. K., Zou, G. (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics*, *174*, 82–94.