Time-varying Model Averaging

Yuying Sun\textsuperscript{1,2}, Xinyu Zhang\textsuperscript{1,2}, Tae-Hwy Lee\textsuperscript{3}, Yongmiao Hong\textsuperscript{4} and Shouyang Wang\textsuperscript{1,2}

\textsuperscript{1}Academy of Mathematics and Systems Science, Chinese Academy of Sciences
\textsuperscript{2}Center for Forecasting Science, Chinese Academy of Sciences
\textsuperscript{3}Department of Economics, University of California
\textsuperscript{4}Department of Economics and Department of Statistics Sciences, Cornell University

SUMMARY

Structural changes often occur in economics and finance due to changes in preferences, technologies, institutional arrangements, policies, crises, etc. Improving the forecast accuracy of economic time series with the evolutionary behavior is a long-standing problem. Model averaging aims at providing an insurance against selecting a poor model. All existing model averaging approaches are designed with constant (non-time-varying) weights. Little attention has been paid to the time-varying model averaging, which is more realistic in economics under structural changes. This paper proposes a novel model averaging estimator which selects the smoothly time-varying weights by minimizing a local jackknife criterion. It is shown that the proposed time-varying jackknife model averaging (TJMA) estimator is asymptotically optimal in the sense of achieving the lowest possible local squared errors in a class of time-varying model averaging estimators, with allowing non-spherical errors. A simulation study and empirical application highlight the merits of the proposed TJMA estimator relative to a variety of popular estimators from constant model averaging and model selection.

KEY WORDS: Asymptotic optimality; Forecast combination; Local stationarity; Non-spherical error; Structural change; Time-varying model averaging.

JEL Classification codes: C2, C13.

1 Introduction

Structural instability is a long-standing problem in time series statistics and econometrics (e.g., Stock & Watson (1996, 2002, 2005), Rossi (2006), and Rossi & Sekhposyan (2011)). Macroeconomic and financial time series, especially over long period, is likely to be affected by structural instability and uncertainty, including changes in preferences, technologies, policies, crises, etc. For example, in macroeconomics, Stock & Watson (1996) find substantial instability in 76 representative US monthly post-war macroeconomic time series. Subsequently, Rossi & Sekhposyan (2011) argue that most predictors for output growth lost their predictive ability in the mid-1970s, and became essentially useless over the last two decades.
In finance, Welch & Goyal (2008) confirm that the predictive regressions of excess stock returns perform poorly in out-of-sample forecast of the U.S. equity premium, and then Rapach & Zhou (2013) argue that model uncertainty and instability seriously impair the forecasting ability of individual predictive regression models. In labor economics, Hansen (2001) finds strong evidence of a structural break in labor productivity between 1992 and 1996, and weaker evidence of a structural break in the 1960s and the early 1980s. Thus, it is crucial to take into account such instability and model uncertainty in economic modelling and forecasting.

An approach to reduce model uncertainty is model averaging, which compromises across the competing models and yields an insurance against selecting a poor model. There has existed a relatively large body of literature on Bayesian model averaging; see Hoeting et al. (1999) for a comprehensive review. In recent years, frequentist model averaging (FMA) has received growing attention in econometrics and statistics (Buckland et al. (1997), Yang (2001), Hjort & Claeskens (2003), Yuan & Yang (2005), Hansen (2007, 2008), Wan et al. (2010), Liu & Okui (2013), Liu (2015), etc). Most of the efforts focus on model averaging weights determination, related inference issues, and asymptotic optimality. Recently, taking heteroscedasticity into consideration, Hansen & Racine (2012) (hereafter HR, 2012) propose a jackknife model averaging (JMA) which selects the model averaging weights by minimizing a cross-validation criterion. The advantage of the JMA estimator lies in that the asymptotic optimality theory is established under heteroskedastic error settings. Subsequently, Zhang et al. (2013) broaden the HR’s scope of the asymptotic optimality of the JMA estimator to encompass models with a non-spherical error covariance structure and lagged dependent variables, thus allowing for dependent data.

However, a potential problem with the aforementioned model averaging approaches is that, one predictive regression may yield the best forecasts in one period while other models outperform the original one in other periods. This implies that the model averaging weights may change over time. There are three reasons for this potential phenomenon. Firstly, a time series dynamics usually suffers from structural instability in economics and finance. For instance, Stock & Watson (2003) point out that a predictor useful in one period does not guarantee its good forecasting performance in other periods. The empirical results in Stock & Watson (2007) suggest that a substantial fraction of forecasting relations are unstable. Secondly, some macroeconomic and financial series are nonstationary, which may follow linear stationary processes in one period and nonlinear processes in other periods. For instance, the original relationship chosen by Phillips (1958) between unemployment and money wage changes was exponential form, while a linear Phillips curve is considered by Cogley & Sargent (2005) in the form of a vector autoregression model with random walk coefficients, where the coefficient drift is interpreted as “a reflection of the process by which policy-makers learn the true model of the economy” (Cogley and Sbordone (2005)).
Thirdly, because of collinearity among predictors, variable selection and model averaging are inherently unstable (Stock & Watson (2012)). Thus, to handle such instability, it may be prudent to use time-varying weights in model averaging instead of constant weights during the whole sample. Furthermore, the underlying economic structure is likely to be affected by technological progress, preference changes and policy switches, crises, etc. It is desirable to use time-varying coefficient models to capture the structural changes in economic time series. To our knowledge, there has been no literature on selecting the optimal time-varying weights in time-varying coefficient model averaging.

This paper fills this gap by proposing a time-varying jackknife model averaging (TJMA) that selects model weights by minimizing a local cross-validation criterion. Our approach complements the existing literature for the constant JMA weights and avoids the difficulty associated with whether structural changes exist in economics and finance. More specifically, we assume that model parameters, as well as the model averaging weights, are smooth unknown functions of time. This setting is consistent with the evidence of types of instability documented in economics, namely smooth structural changes (e.g., Rothman (1998), Grant (2002) and Chen (2015)). To allow the weights in model averaging to change at any time point, we employ the local idea to the squared error loss, leading to a local constant model average estimator. Besides, since Hansen (2001) points out that it might seem more reasonable to allow a structural change to take effect with a period of time rather than to be effective immediately, we follow the spirit of Robinson (1989) and use the local constant method to estimate the time-varying parameters in each candidate model. Furthermore, the candidate models are enlarged from static models to dynamic regression models, which cover more applications in economics and finance.

In this paper, it is shown that the TJMA estimator is asymptotically optimal in the sense of achieving the lowest possible local squared error loss in a class of time-varying model average estimators under three model settings. The first two settings admit a non-diagonal covariance structure in the errors, including heteroscedastic errors in HR (2012), with exogenous regressors. Thus, non-time-varying JMA estimator in HR (2012), under heteroscedastic errors in nested set-up, is a special case of the TJMA estimator in this setting. Our theoretical analysis allows the weights to be continuously changing over time, which avoids restricting the weights to a discrete set in HR (2012). The conditions required for optimality by our method are neither stronger nor weaker than those required by HR’s method. The third model setting that involves lagged dependent variables under i.i.d. errors, for which we prove the asymptotic optimality of the TJMA estimator by allowing the regressors to be locally stationary, which follows the spirits of Ing & Wei (2003) and Vogt (2012).

In simulation study and empirical application, we compare forecast performance of the TJMA estimator with several other estimators, including the Mallow model averaging (MMA) of Hansen (2007), JMA, smoothed Akaike information criterion (SAIC) model averaging...
(Buckland et al. (1997)), smoothed Bayesian information criterion (SBIC) model averaging, nonparametric version of bias-corrected AIC (AICc) model selection (Cai & Tiwari (2000)) and smoothed nonparametric version of AICc (SAICc) model averaging estimators. It is documented that for different structural changes, the TJMA estimator outperforms these alternative estimators under strictly exogenous regressors and ARMA and GARCH-type errors. Additionally, under the dynamic candidate models, the TJMA estimator is still preferred to other estimators.

Compared with the existing model averaging, the proposed approach has a number of appealing features. Firstly, we extend the conventional constant weight model averaging to the time-varying weight model averaging. Specifically, we propose a novel time-varying jackknife model averaging approach, by employing local information at each time point instead of the whole sample, which covers the JMA estimator in HR (2012) as a special case. The TJMA weights selected in this paper are allowed to change smoothly over time, which may be consistent with evolutionary instability of the economic relationship. Secondly, we allow parameters in each candidate model to change smoothly over time. Nonparametric approach is used to estimate parameters, avoiding a potentially misspecified functional form of time-varying coefficients in the parametric approach, e.g., smooth transition regression. Third, we allow regressors to be locally stationary, and thus, time-varying parameter dynamic regression models (e.g., autoregressive models and models with lagged dependent variables) are included in candidate models.

The remainder of this paper is organized as follows. Section 2 introduces the local jackknife criterion and develops the asymptotic optimality theory of the TJMA estimator under heteroscedasticity. In Section 3, a special case of local constant estimators in the time-varying coefficient model is studied. Section 4 develops an asymptotic optimality theory for the TJMA estimator in time-varying coefficient models including lagged values of the dependent variables as regressors. Section 5 presents simulation studies under constant and time-varying coefficient linear regressions. Section 6 examines the empirical forecast performance of stock returns. Section 7 concludes. All mathematical proofs are given in Online Appendix.

2 Model averaging estimator

We consider a general nonlinear model

\[ Y_t = \mu_t + \epsilon_t = f_t(X_t) + \epsilon_t, \quad t = 1, \cdots, T, \quad (1) \]

where \( Y_t \) is a dependent variable, \( X_t = (X_{1t}, X_{2t}, \cdots) \) is countably infinite, \( \epsilon_t \) is an unobservable disturbance with \( E(\epsilon_t | X_t) = 0 \) almost surely (a.s.), \( f_t(\cdot) \) is a unknown smooth
function of time point \( t \), and \( T \) is the sample size. Note that when the functional form of \( f_t(X_t) \) is known up to some finite dimensional parameters, e.g., \( f_t(x) = x' \beta_t \), the conditional mean of \( Y_t \) given \( X_t \) is parametrically specified. A time-varying coefficient regression model for \( f_t(x) = x' \beta_t \) will be treated in Section 3. The linear models are included if assuming \( f_t(\cdot) = f(\cdot) \) is linear. When the functional form of \( f_t(\cdot) \) is unknown, we can estimate \( f_t(\cdot) \) nonparametrically using some well known nonparametric methods, such as the Nadaraya-Watson (NW) estimator and the local linear estimator. Let \( \hat{\mu} \) be a vector, in which the \( \hat{\theta}_t \) estimation based on the delete-one cross-validation is employed. Write \( Y = (Y_1, \cdots, Y_T)' \), \( \mu = (\mu_1, \cdots, \mu_T)' \) and \( X = (X_1', X_2', \cdots, X_T')' \). Furthermore, assume that \( \text{Var}(\epsilon|X) = \Omega \), where \( \epsilon = (\epsilon_1, \cdots, \epsilon_T)' \) and \( \Omega \) is a positive definite symmetric matrix. This setup allows a non-diagonal covariance structure in the errors. Thus, non-spherical error processes, e.g., heteroscedastic and autocorrelated, are included.

### 2.1 Model framework and the jackknife criterion

Consider a sequence of approximating models \( m = 1, \cdots, M_T \), which are allowed to be misspecified. For different models, explanatory variables may be different. Let \( \{\tilde{\mu}^1, \cdots, \tilde{\mu}^{M_T}\} \) be a set of nonparametric estimators of \( \mu \) dependent on these models. Specifically, for the \( m \)-th model, the estimator of \( \mu \) may be written as \( \tilde{\mu}^m = P_m Y \), where \( P_m \) is dependent on \( X \) but not on \( Y \). For each time \( t = 1, \cdots, T \), let \( \mathbf{w} = (\mathbf{w}^1, \cdots, \mathbf{w}^{M_T})' \) be a weight vector, which satisfies

\[
\mathcal{H}_T = \left\{ \mathbf{w} \in [0, 1]^{M_T} : \sum_{m=1}^{M_T} w^m = 1 \right\}.
\]

We obtain \( \tilde{\mu}_t(\mathbf{w}) = \sum_{m=1}^{M_T} w^m \tilde{\mu}^m_t = \sum_{m=1}^{M_T} w^m e_t P_m Y = e_t P(\mathbf{w}) Y \), where \( e_t \) is an \( 1 \times T \) vector, in which the \( t \)-th element is 1 and others are zero, \( \tilde{\mu}^m_t = e_t P_m Y \) and \( P(\mathbf{w}) = \sum_{m=1}^{M_T} w^m P_m \). Then the model average estimator of \( \mu \) can be fitted as \( \tilde{\mu}(\mathbf{w}) = (\tilde{\mu}_1(\mathbf{w}), \cdots, \tilde{\mu}_T(\mathbf{w}))' \).

Denote \( \tilde{\mu}^m_t \) as the time-varying estimator of \( \mu \) in the \( m \)-th model, when jackknife estimation based on the delete-one cross-validation is employed. Write \( \tilde{\mu}^m_t = (\tilde{\mu}^m_1, \cdots, \tilde{\mu}^m_T)' \), where \( \tilde{\mu}^m_t \) is the estimator of \( \mu_t \) obtained with the \( t \)-th observation \((Y_t, X_t)\) removed from the sample in the \( m \)-th candidate model. Thus, we obtain \( \tilde{\mu}^m_t = \tilde{P}_m Y \), where \( \tilde{P}_m \) has zeros on the diagonal and depends only on \( X \). The Jackknife version model averaging estimator of \( \mu_t \), which smooths across the \( M_T \) jackknife estimators, at time point \( t \) is obtained as

\[
\tilde{\mu}_t(\mathbf{w}) = \sum_{m=1}^{M_T} w^m \tilde{\mu}^m_t = e_t \sum_{m=1}^{M_T} w^m \tilde{P}_m Y = e_t \tilde{P}(\mathbf{w}) Y,
\]

where \( \tilde{P}(\mathbf{w}) = \sum_{m=1}^{M_T} w^m \tilde{P}_m \).

Set \( \tilde{\mu}(\mathbf{w}) = (\tilde{\mu}_1(\mathbf{w}), \cdots, \tilde{\mu}_T(\mathbf{w}))' \). Let \( K_t = \text{diag}(k_{t1}, \cdots, k_{tT}) \), where \( k_{st} = k(s/t_H) \).
the kernel $k(\cdot) : [-1, 1] \to \mathbb{R}^+$ is a prespecified symmetric probability density, and $h \equiv h(T)$ is a bandwidth with $h \to 0$ and $Th \to \infty$ as $T \to \infty$. We minimize the local cross-validation (CV) squared error loss criterion,

$$CV_{t,T}(w) = (Y - \tilde{\mu}(w))'K_t(Y - \tilde{\mu}(w)).$$

Then we obtain the time-varying weight vector $\hat{w}_t = \arg\min_{w \in \mathcal{H}_t} CV_{t,T}(w)$, which minimizes $CV_{t,T}(w)$. The TJMA estimator of $\mu$ is $\hat{\mu}(\hat{w}_t)$.

The jackknife criterion (or cross-validation) is widely used in selecting regression models (e.g., Allen (1974), Stone (1974) and Geisser (1975)), and the asymptotic optimality of model selection using CV is established by Li (1987) for homoskedastic regression and by Andrews (1991) for heteroskedastic regression, respectively. In this paper, the CV criterion defined above is locally weighted by $K_t$ at each time point. This CV criterion chooses the optimal weights by generating the smallest CV value over the local sample leaving out the observation $(X_t, Y_t)$ at time $t$. Thus, the time-varying weight vector $\hat{w}_t$ is essentially the constant weight in the neighborhood of any fixed time point $t$, and combines different models to yield the lowest local squared error, defined in the following (2.5).

Note that there are two key differences between our estimator and the JMA estimator proposed by HR (2012) and Zhang et al. (2013). One major difference is that our theoretical analysis allows the model averaging weights to change with time smoothly. This is unlike the method of HR (2012) and Zhang et al. (2013), which restricts the weights to be constant in a discrete set or a continuous set. We extend the constant weights to time-varying weights, that can reflect instability of economic relationship. The other difference is that the models in HR (2012) and Zhang et al. (2013) are composed of linear regressions, while in this paper, parameters are allowed to be smooth unknown functions of time. In this setting, structural changes smoothly occur over time. Theoretically, we obtain the asymptotic optimality of the TJMA from smoothly time-varying coefficient model-sets, which is demonstrated by simulation results that the proposed TJMA estimator outperforms the other existing methods in the presence of smooth structural changes as well as recurrent breaks.

2.2 Asymptotic Optimality

To establish the asymptotic optimality of the TJMA estimator, we consider the following local squared error loss and associated risk criteria:

$$L_{t,T}(w) = (\hat{\mu}(w) - \mu)'K_t(\hat{\mu}(w) - \mu)$$

and

$$R_{t,T}(w) = E(L_{t,T}(w)|X) = \mu' A'(w)K_tA(w)\mu + \text{tr}(P'(w)K_tP(w)\Omega),$$

where $A'(w)$ is the differential of $A(w)$, $\text{tr}$ is the trace of a matrix, and $\Omega$ is the covariance matrix of the error term.
where $\hat{\mu}(w) = \sum_{m=1}^{M_T} w^m \hat{\mu}^m$ is the weighted average of the forecasts of every single model, and $A(w) = I_T - P(w)$.

Let $\tilde{L}_{t,T}(w)$ and $\tilde{R}_{t,T}(w)$ be the local jackknife squared error loss and risk, respectively, which are obtained by replacing $\hat{\mu}(w)$ by $\tilde{\mu}(w)$, $A(w)$ by $\tilde{A}(w)$, and $P(w)$ by $\tilde{P}(w)$. Specifically,

$$\tilde{L}_{t,T}(w) = (\tilde{\mu}(w) - \mu)'K_t(\tilde{\mu}(w) - \mu)$$

and

$$\tilde{R}_{t,T}(w) = \mu'\tilde{A}'(w)K_t\tilde{A}(w)\mu + \text{tr}(\tilde{P}'(w)K_t\tilde{P}(w)\Omega).$$

Denote $\xi_{t,T} = \inf_{w \in G_t} R_{t,T}(w)$ and $\Omega = \Omega - \text{diag}(\Omega_{11}, \cdots, \Omega_{TT})$, where $\Omega_{ii}$ is the i-th diagonal element.

Extending the results of HR (2012) and Zhang et al. (2013), we prove that the TJMA estimator $\hat{\mu}(\hat{w}_t)$ satisfies the optimality (OPT)

$$\text{(OPT)} : \frac{L_{t,T}(\hat{w}_t)}{\inf_{w \in G_t} L_{t,T}(w)} \overset{p}{\to} 1.$$  

This suggests that the local average squared error of the jackknife estimator is asymptotically identical to the local average squared error of the infeasible best possible averaging estimator. This optimality property is almost the same as that in Zhang et al. (2013), except that we allow the weights to change smoothly over time with the local perspective.

To guarantee that the TJMA estimator satisfies the (OPT) property under a model setup that permits smooth-change parameters and a non-diagonal error covariance structure, we impose a set of regularity conditions:

**Assumption 1.** $\{\varepsilon_t\}$ is a sequence of innovations such that $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_T)'$ satisfies $\varepsilon | X \sim N(0, \Omega)$, where $\Omega$ is $T \times T$.

**Assumption 2.** The maximum singular value of $\Omega$ satisfies $\zeta(\Omega) \leq \overline{C} < \infty$, where $\overline{C}$ is a constant.

**Assumption 3.** For $1 \leq m \leq M_T$, the maximum singular value of $P_m$ satisfies

$$\lim_{T \to \infty} \max_{1 \leq m \leq M_T} \zeta(P_m) < \infty, \text{ a.s.},$$

where $\zeta(P_m)$ is the maximum singular value of a matrix $P_m$.

**Assumption 4.** For $1 \leq m \leq M_T$, the maximum singular value of $\tilde{P}_m$ is finite when the sample size increases, i.e., $\lim_{T \to \infty} \max_{1 \leq m \leq M_T} \zeta(\tilde{P}_m) < \infty$, a.s.

**Assumption 5.** The local risk $\tilde{R}_{t,T}(w)$, i.e., the conditional expectation of the local jackknife squared error loss given $X$, satisfies $\sup_{w \in G_t} |\tilde{R}_{t,T}(w)/R_{t,T}(w) - 1| \to 0$, a.s.
Assumption 6. $M_T \xi^{-2G} \sum_{m=1}^{M_T} (R_{t,T}(w^0_m))^G \to 0$, a.s., for some constant $G \geq 1$, where $w^0_m$ is a $M_T \times 1$ weight vector with the $m$th element taking on the value of unity and other elements zeros.

Assumption 6’. $\xi^{-2} \sum_{m=1}^{M_T} (R_{t,T}(w^0_m)) \to 0$, a.s., where $w^0_m$ is a $M_T \times 1$ weight vector with the $m$th element taking on the value of unity and other elements zeros.

Assumption 7. $\sup_{w \in \mathcal{H}_T} |\text{tr}(K_t \tilde{P}(w) \tilde{\Omega})/R_{t,T}(w)| \to 0$, a.s.

Assumption 8. $k : [-1, 1] \to \mathbb{R}^+$ is a symmetric bounded probability density function.

Assumption 9. The bandwidth $h = cT^{-\lambda}$ for $0 < \lambda < 1$ and $0 < c < \infty$.

Assumption 6 is the same as condition (11) in Zhang et al. (2013), which is limited to Gaussian regressions. More importantly, this condition can be removed to obtain the asymptotic optimality of the TJMA estimators for time-varying coefficient regression models in Section 3. Assumption 2 assumes the largest singular values of the correlation matrix $\Omega$ to be finite when the sample size increases, corresponding to condition (12) in Zhang et al. (2013). Assumptions 3-4 correspond to conditions (A.3) and (A.4) of HR (2012), respectively. Both of them are quite mild, because typical estimators satisfy that the maximum singular values of corresponding matrix are bounded. Assumption 5 imposes that the leave-one-out estimator is asymptotically equivalent to the local risk of the regular estimator, uniformly over the class of averaging estimators. This is standard for the application of cross-validation and almost the same as condition (10) in Zhang et al. (2013), except that the continuous time-varying set $\mathcal{H}_T$ is used here instead of the continuous constant set $\mathcal{H}_n$ in Zhang et al. (2013). Assumption 6 is required for the asymptotic optimality of all MMA, JMA and TJMA estimators; see more discussions in Wan et al. (2010) and Zhang et al. (2013). Assumption 6 also requires $M_T \sum_{m=1}^{M_T} (R_{t,T}(w^0_m))^G \to \infty$ at a rate slower than $\xi^{-2G} \to \infty$ with the infinite number $M_T$ of candidate models, as $T \to \infty$. Assumption 6’ is clearly weaker than Assumption 6 when $G$ is set to 1. It is reasonable to assign unity to $G$, because the elements in $\epsilon$ are allowed to be correlated and heteroscedastic. Assumption 7 can be removed when the data generating process (DGP) is a linear regression with time-varying coefficients in Section 3.

Based on Assumption 8 the kernel is symmetric and bounded, which has a compact support $[-1, 1]$ and discounts the observations whose values are far away from the time point that we are interested in. This implies that $\max_{s,t} k_{st} = k_{\max} < \infty$, which is used in mathematical proof. The commonly used kernel function, the Epanechnikov kernel $k(u) = 0.75(1-u^2)I(|u| \leq 1)$ is used in this paper, where $I(\cdot)$ is the indicator function. Assumption 9 implies $h \to 0$ and $Th \to \infty$, which is a standard condition for the bandwidth; see Chen and Hong (2012). Assumption 9 covers the optimal bandwidth $h \propto T^{-1/5}$, which minimizes the integrated mean squared error (MSE) of the nonparametric estimation for $\beta(u)$, $u \in [0, 1]$. 

8
In practice, \( h \) can be chosen via a simple rule-of-thumb approach, namely \( h = 2.34 \sigma_x T^{-1/5} \) for the Epanechnikov kernel, where \( \sigma_x \) is the standard deviation of \( X_t \). Intuitively, when \( \beta_t \) is time-varying, including too many distant past observations reduces the forecast variance but increases its bias, and conversely, a small bandwidth decreases the bias at the cost of increasing the forecast variance.

**Theorem 1.** Suppose Assumptions 1-9 hold, then the TJMA estimator \( \hat{\mu}(\hat{w}_t) \) satisfies the (OPT) asymptotic optimality, i.e.,

\[
\frac{L_{t,T}(\hat{w}_t)}{\inf_{w \in \mathcal{F}_T} L_{t,T}(w)} \xrightarrow{p} 1.
\]

Theorem 1 shows that the local squared error obtained from the time-varying weight vector \( \hat{w}_t \) is asymptotically equivalent to the infeasible optimal weight vector at any time point \( t \). This implies that the TJMA estimator is asymptotically optimal in the class of model averaging estimators obtained from nonlinear models where the weight vector \( w \) is restricted to the set \( \mathcal{F}_T \), which allows the weights to change smoothly.

### 3 Special case: Time-varying coefficient regression

In this section, we focus on local linear estimators on a set of candidate models with specific forms, i.e., time-varying coefficient regressions. This is a special case of the general candidate models in Section 2. Consider the \( m \)-th time-varying parameter regression as follows:

\[
Y_t = X_t^m \beta_t^m + \varepsilon_t^m, \quad t = 1, \ldots, T, \quad m = 1, \ldots, M_T, \tag{9}
\]

where \( X_t^m \) is a \( q_m \times 1 \) vector of explanatory variables, \( \beta_t^m \) is a \( q_m \times 1 \) possibly time-varying parameter vector, \( \varepsilon_t^m \) is an unobservable disturbance, and \( q_m \) is a positive integer and allowed to be infinite. Note that we allow \( E(\varepsilon_t^m | X_t^m) \neq 0 \) in the set of candidate models.

As [Hansen (2001)] suggests, “it may seem unlikely that a structural break could be immediate and might seem more reasonable to allow a structural change to take a period of time to take effect”, we are interested in the \( m \)-th smooth time-varying coefficient model as follows:

\[
Y_t = X_t^m \beta^m \left( \frac{t}{T} \right) + \varepsilon_t^m, \quad t = 1, \ldots, T, \tag{10}
\]

where \( \beta^m : [0, 1] \to \mathbb{R}^{d_m} \) is a \( q_m \)-dimensional vector-valued function on \([0, 1]\). In the neighborhood of each time point, the model is locally stationary yet globally nonstationary.

To capture the evolutional behavior of economic time series, time-varying coefficient models have appeared as a novel tool. For example, the smooth transition regression (STR)
model is proposed by Chan & Tong (1986) and further studied by Lin & Teräsvirta (1994), which allows both the intercept and the slope to change smoothly over time. One of the central issues in time-varying coefficient models is to estimate parameters. If the coefficient function is correctly specified, parametric models for time-varying parameters result in more efficient estimation. However, there is no economic theory to support any concrete functional form for these parameters, and the choice of concrete functional form for time-varying parameters is somewhat arbitrary, probably leading to serious bias with a misspecification. Robinson (1989, 1991) firstly propose the nonparametric time-varying parameter model and it is further studied by Blundell et al. (1998), Cai (2007) and Chen & Hong (2012). One advantage of this nonparametric model is little or restrictive prior information imposed on the functional forms of time-varying parameters, except for the assumption that they evolve over time smoothly. Thus, in terms of the functional form of $\beta^m(t/T)$, we follow the spirit of the nonparametric estimation method proposed in Robinson (1989).

Instead of specifying a parameterization of $\beta^m(t/T)$, which may lead to serious bias, we assume that $\beta^m(\cdot)$ is a smooth time-varying function of the ratio $t/T$. This assumption is based upon a common scaling scheme in the literature (e.g., Robinson (1989)). To reduce the bias and variance of nonparametric estimator of $\beta^m_t$, it is necessary to balance the increase between the sample size $T$ and the amount of local information at a fixed time point $t$. One possible solution suggested in Robinson (1989) and Cai (2007) is to assume a smooth function $\beta(\cdot)$ on an equally spaced grid over $[0,1]$ and consider estimation of $\beta^m(u)$ at fixed points $u \in [0,1]$. We note that the parameter $\beta^m_t$ is dependent on the sample size $T$, so that new information accumulates as $T$ increases. This condition is beneficial to the consistency of parameters (see, Cai (2007), Chen & Hong (2012)).

By assuming that $\beta^m_s$ has a continuous first derivative, $\beta^m_s$ can be approximated at any fixed time point $t \in [0,1]$ as follows:

$$
\beta^m_s = \beta^m_t + O((s-t)/T), \ s \in [t-Th, t+Th].
$$

Define $K_{-t} = \text{diag}(k_{1t}, k_{2t}, \ldots, k_{(t-1)t}, 0, k_{(t+1)t}, \ldots, k_{Tt})$ as the weight for Jackknife estimation, thus for every time point $t$, we obtain the least square estimation $\hat{\mu}^m_t$ and Jackknife estimation $\tilde{\mu}^m_t$ in the $m$th candidate model as follows:

$$
\hat{\mu}^m_t = X^m_t (X^m_t K_t X^m)^{-1} X^m_t K_t Y
$$

and

$$
\tilde{\mu}^m_t = X^m_t (X^m_t K_{-t} X^m)^{-1} X^m_t K_{-t} Y.
$$
Based on the expressions of $\hat{\mu}_t^m$ and $\tilde{\mu}_t^m$, it is easy to obtain that

$$P_m = \begin{bmatrix}
X_1^m (X_m'K_1X_m)^{-1}X_m'K_1 \\
X_2^m (X_m'K_2X_m)^{-1}X_m'K_2 \\
\vdots \\
X_T^m (X_m'K_TX_m)^{-1}X_m'K_T
\end{bmatrix} \quad (14)$$

and

$$\tilde{P}_m = \begin{bmatrix}
X_1^m (X_m'K_{-1}X_m)^{-1}X_m'K_{-1} \\
X_2^m (X_m'K_{-2}X_m)^{-1}X_m'K_{-2} \\
\vdots \\
X_T^m (X_m'K_{-T}X_m)^{-1}X_m'K_{-T}
\end{bmatrix}. \quad (15)$$

Thus, $\tilde{P}_m = D_m(P_m - I_T) + I_T$, where $D_m$ is a diagonal matrix with $i$-th diagonal element $(1 - h_{ii}^m)^{-1}$, where $h_{ii}^m$ is the $(i, i)$ element in $P_m$.

To construct the asymptotic optimality property of $\hat{\mu}(\hat{w})$, the following regularity Assumptions are imposed:

**Assumption 10.** $\text{tr}(P'(w)P(w))\xi_{t,T}^{-1} \xrightarrow{p} 0$.

**Assumption 11.** The local average of $\mu^2_t$ is bounded, i.e., $\frac{1}{T}\mu'K_t\mu = O(1)$, a.s.

**Assumption 12.** There exists some constant $\Lambda$ such that for any $i$, $h_{ii}^m \leq \Lambda q_m T^{-1}$, a.s.

Let $\gamma = \text{rank}(X)$ and $h$ be the bandwidth in candidate models. With some calculation, we have that $\text{tr}(P'(w)P(w)) = O_p(\gamma/h)$, thus Assumption 10 imposes a restriction on the growth rate of the number of regressors, corresponds to condition (22) in Zhang et al. (2013). Assumption 11 is similarly used in Shao (1997) and Wan et al. (2010). Assumption 12 excludes extremely unbalanced designs, which is reasonable and typical for the application of cross-validation; see Li (1987), HR (2012) and Zhang et al. (2013) for more discussions.

**Theorem 2.** Suppose Assumptions 2, 6', 8-12 hold, then $\hat{\mu}(\hat{w}_t)$ satisfies the (OPT) asymptotic optimality.

Theorem 2 shows that the TJMA estimator is asymptotically optimal in the class of weighted averages of time-varying estimators.

4 Asymptotic optimality of the TJMA estimator with lagged dependent variables

In this section, we develop an asymptotic optimality theory for the TJMA estimator in time-varying coefficient models including lagged values of the dependent variable as regressors.
Dynamic regressions are widely used in forecasting macroeconomic variables, including Gross Domestic Product, inflation, income, etc. Thus, it is highly desirable to extend the TIMA estimator in the static regressions to that in the dynamic regression models. Consider DGP as follows

$$Y_t = \sum_{j=1}^{\infty} \beta_{jt} Y_{t-j} + \epsilon_t, \ t = 1, \cdots, T,$$

(16)

where \( \epsilon_t \) is i.i.d. with a mean of zero and variance of \( \sigma^2 \). This the DGP is considered as a special case of the DGP in Section 2.

In spite of (16), the exogenous regressors are allowed to the candidate models with lagged values of \( Y \). The full model (i.e., the model with all regressors) is

$$Y_t = \sum_{i=1}^{r_1} \beta_{it} Y_{t-i} + \sum_{j=1}^{r_2} \beta_{(r_1+j)t} X^*_{ij} + \epsilon^f_t, \ t = 1, \cdots, T,$$

(17)

where \( X^*_{ij} \) is the exogenous variable, \( \epsilon^f_t \) is the innovation in the full model, \( r_1 \) is the maximal lag order in any candidate model, \( r_2 \) is the number of extraneous regressors. Let \( r_1 \) be allowed to increase while \( r_2 \) is fixed when \( T \) increases. Denote \( Y = (Y_1, \cdots, Y_T)' \), \( Y_{LT} = (Y_{t-1}, \cdots, Y_{t-r_1}) \), and let \( Y_L = (Y_{L1}, \cdots, Y_{LT})' \) be \( T \times r_1 \) matrix containing \( T \times r_1 \) observations of lagged dependent regressors, \( X^* = (X^*_{11}, \cdots, X^*_{1T})' \) be a \( T \times r_2 \) matrix including observations of \( r_2 \) exogenous regressors, \( X = (Y_L, X^*) \) be a regressor matrix in the full model with rank \( \gamma = r_1 + r_2 \) and \( \epsilon^f = (\epsilon^f_1, \cdots, \epsilon^f_T)' \). The regressor matrix \( X^m \) of the \( m \)-th candidate model is formed by combining the columns of \( X \). Define \( P \) similarly to \( P_m \) with \( X^m \) replaced by \( X \). Note that \( X^m \) is the regressor matrix in the \( m \)-th candidate model and \( q_m \) is the number of regressors in \( X^m \). Regressors are permitted to be locally stationary, allowing lagged dependent variables. Thus, our framework can cover AR as well as ARX models with time-varying coefficients, and nested as well as nonnested AR process in the lagged component of the model.

**Assumption 13.** \( \{Y_t\} \) is a local stationary process, \( \{X^*_t\} \) is strictly stationary process, and both \( \{Y_t\} \) and \( \{X^*_t\} \) are \( \beta \)-mixing process with mixing coefficients \( \{\beta(j)\} \) satisfying \( \sum_{j=1}^{\infty} j^2 \beta(j)^{\delta/(1+\delta)} < C \) for some constant \( 0 < \delta < 1 \), \( E||Y_t||^4 < \infty \) and \( E||X^*_t||^4 < \infty \).

**Assumption 14.** \( \mu'\mu/T = O_p(1) \), \( T q_m^{-1} h^m_{it} = O_p(1) \), \( \gamma \xi^{*-1}_{t,T} = o_p(1) \), and \( \gamma \mu'P'P\mu \xi^{*2}_{t,T} = o_p(1) \), where \( \xi^*_{t,T} = \inf_{w \in H_t} V_{t,T}(w) \) and \( V_{t,T}(w) = \mu' A'(w) K_t A(w) \mu + \sigma^2 tr(P'(w) K_t P(w)) \).

**Assumption 15.** \( \zeta(T^{-1/2}X^*Y^*) = O_p(1) \), and for all \( t = 1 \cdots T \), \( \sqrt{T}X^*K_t \xi \xrightarrow{d} N(0, \Delta) \), and \( \zeta((T^{-1}X^*M_{t,T}X^*)^{-1}) = O_p(1) \), where \( \Delta \) is positive definite matrix and \( M_{t,T} = I_T - Y_L(Y_L'K_t Y_L)^{-1}Y_L' \).
Assumption 16. \( \{\varepsilon_t\} \) has an identical and independent distribution, and satisfies that with some positive constants \( \alpha_1, \alpha_2 \) and \( \alpha_3 \),

\[
|F_t(d_1) - F_t(d_2)| \leq \alpha_1|d_1 - d_2|^2,
\]

for all \( t \) when \( |d_1 - d_2| \leq \alpha_3 \) and \( F \) is the distribution of \( \varepsilon_t \).

Assumption 17. \( r_1^{α + 3} = O(T) \) for some \( α_4 > 0 \) and \( \sup_t E\varepsilon_t^4 < \infty \).

In Assumption 13, local stationarity is weaker than strict stationarity as assumed in \cite{Robinson1989} and \cite{Cai2007}. Intuitively, it implies that when the standardized time \( \frac{t}{T} \) is in a neighborhood of any point \( \tau \in [0, 1] \), the behavior of time series \( \{Y_t\} \) can be approximated up to a certain high order by a strictly stationary process \( \{Y_t(\tau)\} \), and it holds that \( \|Y_t - Y_t(\tau)\| = O_p \left( h + \frac{1}{T} \right) \), where \( h \) is a bandwidth with \( h \to 0 \) as \( T \to \infty \); see \cite{Dahlhaus1996}, \cite{Dahlhaus1997} and \cite{Vogt2012} for details. Thus, the autocovariance function of \( \{Y_t\} \) for all times \( t \), with \( \frac{t}{T} \) in the neighborhood of \( \tau \), can be approximated by the strictly stationary time series \( Y_t(\tau) \).

Assumption 14 is analogous to Assumptions 10-12, which are used in time-varying coefficient regression models when \( X_t \) is assumed to be strictly stationary. When \( \{X_t, \varepsilon_t\} \) is a stationary and ergodic martingale difference sequence with finite fourth moments and \( T^{-1}X_tX_t' \) converges to a positive definite matrix in probability, Assumption 15 hold.

Assumption 16 is a mild condition easily satisfied by any distribution with a bounded density, which is the same as condition (K.2) of \cite{Ing2003}. This assumption is also used to prove Lemma 1 in Mathematical Appendix. Assumption 17 is a reiteration of assumptions in Lemma 4 in Mathematical Appendix. Assumption 17 can be also replaced by \( r_1^{α + 3} \) and \( \sup_{\infty < i < \infty} E\varepsilon_t^4 < \infty \) for all \( s \).

Next, we need some assumptions about the aforementioned strictly stationary process \( \{Y_t(\tau)\} \).

Assumption 18. For any \( \tau = t/T \) and \( q > 0 \), \( \{Y_t(\tau)\} \) is strictly stationary with \( E|Y_t(\tau)|^q < \infty \) satisfying \( Y_t(\tau) + \sum_{i=1}^{\infty} a_i Y_{t-i}(\tau) = \varepsilon_t, \ t = \ldots, -1, 0, 1, \ldots \ and \ the \ roots \ of \ A(z), \ i.e., \ A(z) = 1 + \sum_{i=1}^{\infty} a_i z^i = 0, \ lies \ outside \ the \ unit \ circle \ |z| = 1, \ \{\varepsilon_t\} \ is \ a \ sequence \ of \ independent \ random \ variables \ with \ zero \ means \ and \ variances \ \sigma^2. \)

Assumption 19. \( \{Y_t(\tau)\} \) is stationary \( \beta-\) mixing process with mixing coefficients \( \{\beta(j)\} \) satisfying \( \sum_{j=1}^{\infty} j^2 \beta(j)^{3/(1+\delta)} < C \) for any \( \tau = t/T \) and some \( 0 < \delta < 1. \)

Assumption 18 is a standard condition for ARMA models; see more discussions in \cite{Ing2003}. The mixing condition in Assumption 19 imposes a restriction on the temporal dependence in \( \{Y_t(\tau)\} \), which is widely used in existing literature (e.g, \cite{Cai2007}, \cite{Chen2007}).
Theorem 3. If Assumptions $8, 9, 13 - 17$ hold, the TJMA estimator $\hat{\mu}(\hat{\mathbf{w}}_t)$ in this section satisfies the (OPT) asymptotic optimality property.

A main contribution of Theorem 3 is extending the asymptotical optimality to dynamic regression models with time-varying coefficients. The assumptions do allow locally stationary process and time-series dependence.

5 Monte carlo simulation

To examine the finite sample performance of the proposed model averaging method, we consider the following DGPs:

DGP 1 (Smooth Structural Changes):

$$Y_t = \mu_t + \epsilon_t = \sum_{j=1}^{\infty} \theta_j F(\tau) X_{tj} + \epsilon_t, t = 1, \ldots, T$$

where $\tau = t/T, F(\tau) = \tau^2, X_{t1} = 1,$ and observations of all other $X_{tj}, j \geq 2$ are generated from the $N(0, 1)$ distribution and are independent. Following HR (2012), $\theta_j = c\sqrt{2} \alpha^{-1/2}$, with $c > 0$ and $\alpha = 1.5$, and the coefficient $c$ is selected to control the population $R^2 = c^2/(1+c^2)$ to vary on a grid from 0.1 to 0.9.

To examine the robustness of the TJMA estimator, we consider three cases for $\{\epsilon_t\}$: (i) $\epsilon_t \sim i.i.d. N(0,1)$; (ii) $\epsilon_t = \epsilon_{t,1} + \epsilon_{t,2}$, $\epsilon_{t,1} \sim N(0, X_{t2}^2)$, $\epsilon_{t,2} = \phi \epsilon_{t-1,2} + u_t$, $u_t \sim i.i.d. N(0, 1)$ and $\phi = 0.5$. This error process is the same as that of Zhang et al. (2013); (iii) $\epsilon_t = \sqrt{h_t} u_t$, $h_t = 0.2 + 0.5X_{t2}^2$, $u_t \sim i.i.d. N(0, 1)$, which follows the error setup in Chen and Hong (2012). Note that $\text{var}(\epsilon_t|X_{t2}) \neq \sigma^2$ under case (iii).

We compare the TJMA estimator with a variety of popular methods, namely (2) nonparametric version of bias-corrected AIC in Cai & Tiwari (2000); (3) smoothed nonparametric version of AICc; (4) JMA of HR (2012); (5) MMA of Hansen (2007); (6) smoothed Akaike information criterion; and (7) smoothed Bayesian information criterion. Specifically, the nonparametric version of the AICc for order selection is $AICc = \log RSS + \frac{T+\text{tr}(S^*)}{T-\text{tr}(S^*)+2}$, where $RSS = \sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2$ based on local linear regression method and $\text{tr}(S^*)$ is the number of parameters involved in the model fitting, which penalizes extra parameters for larger values of $\text{tr}(S^*)$. See the definition of $S^*$ and more discussions in Cai & Tiwari (2000). The SAICc method is the model average estimator with the weights $w_m = \exp(-\frac{1}{2} AICc_m)/\sum_{m=1}^{M} \exp(-\frac{1}{2} AICc_m)$, where AICc is obtained from the second method. The other four estimators are based on linear regressions and constant weights. The AIC and BIC model selection estimators are to minimize $AIC_m = T \ln \hat{\sigma}_m^2 + 2m$ and $BIC_m = T \ln \hat{\sigma}_m^2 + \ln(T)m$, respectively. SAIC, proposed by Buckland et al. (1997), is the least squares model average estimator with the weights $w_m = \exp(-\frac{1}{2} AIC_m)/\sum_{m=1}^{M} \exp(-\frac{1}{2} AIC_m)$, while for
SBIC, the simplified form of Bayesian model averaging, the weights are \( w_m = \exp\left(-\frac{1}{2} \text{BIC}_m\right) / \sum_{m=1}^{M_T} \exp\left(-\frac{1}{2} \text{BIC}_m\right) \).

The number of approximating models is determined by the rule in HR (2012), i.e., \( M_T = \lceil 3T^{1/3} \rceil \) (the nearest integer from \( 3T^{1/3} \)), resulting in \( M_T = 11, 14, 15 \) and 18 for \( T = 50, 75, 100 \) and 200, respectively. The approximating models are \( Y_t = \sum_{j=1}^{m} \beta_j(\tau)X_{tj} + \epsilon_t, \ t = 1, \cdots, T, \ m = 1, \cdots, M_T. \) In JMA, MMA, SAIC and SBIC methods, \( \beta_j(\tau) \) in candidate models are assumed to be constant, and then the candidate models are simplified to \( Y_t = \sum_{j=1}^{m} \beta_jX_{tj} + \epsilon_t, \ t = 1, \cdots, T, \ m = 1, \cdots, M_T. \)

For TJMA and AICc approaches, we use the second order Epanechnikov kernel for the nonparametric estimation, because the Epanechnikov kernel has been shown to be the optimal kernel for density estimation (Epanechnikov (1969)) and robust regression (Lehmann & Casella (2006)), although our experience suggests that the choice of kernel \( k(\cdot) \) has little impact on the performance of our method. To conserve space, we report results based on the normal reference bandwidth \( h = 2.34T^{-1/5} \), which attains the optimal rate for mean squared errors (MSE) (Chen & Hong (2012)). We generate 100 data sets of random sample \( \{Y_t, X_t^{\prime}\}_{t=1}^{T} \). To compare finite performances, we employ the following MSE measure to assess the accuracy of forecasts:

\[
\frac{1}{N} \sum_{n=1}^{N} \|\hat{\mu}(w)^{(n)} - \mu^{(n)}\|^2,
\]

where \( \hat{\mu}(w)^{(n)} \) and \( \mu^{(n)} \) denote the forecast value and true value of the expectation of \( Y \) in the \( n \)-th replication, \( n = 1, \cdots, N = 100. \) To simplify comparisons, the risk (i.e., expected squared error) of all estimators are normalized by the MSE of the infeasible optimal least squares estimator, which is the same as that in HR (2012). To conserve space and simplify discussions, the results of Monte Carlo are reported in graphical forms; see discussions below.

Figures 1-3 report the results of simulations under DGP 1. We find that in most cases, the TJMA estimator delivers the most precise forecasts among those considered, especially when \( R^2 \) is relatively large. Under both conditional heteroscedastic errors and autocorrelated errors, our method displays the best performance in terms of the risk, as is expected. Also, when the sample \( T \) is large enough, the AICc and SAICc estimators can sometimes be marginally similar to the TJMA estimator in case of large \( R^2 \). This happens because the parameters in DGP are changing over time and the candidate models in these three methods are time-varying coefficient models as well. Besides, in most cases, the TJMA estimator is preferred to any of the four estimators based on linear least squares, although occasionally small to moderate reductions in MSE can be made with MMA and JMA estimators with small \( R^2 \) and small \( T \); see \( T = 50 \) for example. Furthermore, it is possible that our results are a bit sensitive to bandwidth selection and how to select the optimal bandwidth to estimate the optimal time-varying weights is left for future study. A possible solution is to consider
the model averaging bandwidths; see Henderson & Parmeter (2016) and Zhu et al. (2017).

Next, a special case in time-varying coefficient models, including lagged values of the dependent variables as regressors, is considered as follows:

**DGP 2** (Dynamic Regression with Smooth Structural Changes):

\[
Y_t = \sum_{j=1}^{\infty} \theta_j f(\tau) Y_{t-j} + \epsilon_t,
\]

where \( \theta_j = 1/\sqrt{2\alpha^j} e^{-\alpha^{-1/2}}, f(\tau) = \tau, \epsilon_t = \frac{c}{\tau} \epsilon_t, c = \sum \theta_j^2 \) and \( \alpha = 1.5 \). \( R^2 \) varies on a grid from 0.1 to 0.9.

To investigate the finite sample performance of DGPs with various structural changes, we consider the three DGPs with case (ii) for \( \{\epsilon_t\} \). For DGPs 3-5, \( \theta_j = c \sqrt{2\alpha^j} e^{-\alpha^{-1/2}} \), with \( c > 0 \) and \( \alpha = 1.5 \), the same as those in DGP 1:

**DGP 3** (Single Structural Break):

\[
Y_t = \sum_{j=1}^{\infty} 0.5 \theta_j X_{tj} I(t \leq 0.3T) + \sum_{j=1}^{\infty} \theta_j X_{tj} I(t > 0.3T) + \epsilon_t.
\]

**DGP 4** (Smooth Transition Regression):

\[
Y_t = \sum_{j=1}^{\infty} \theta_j f(\tau) X_{tj} + \epsilon_t,
\]

where \( f(\tau) = 1.5 - 1.5 \exp(-3(\tau - 0.3)^2) \).

**DGP 5** (Smooth Structural Changes with Periodicity):

\[
Y_t = \sum_{j=1}^{\infty} \theta_j f(\tau) X_{tj} + \epsilon_t,
\]

where \( f(\tau) = \sin(\pi \tau^2) \).

For each of DGPs 2-5, we generate 100 data sets of the random sample \( \{X_t, Y_t\}_{t=1}^{T} \) for each \( T = 50, 75, 100 \) and 200. \( X_{t1} = 1 \) and observations of all other \( X_{tj}, j \geq 2 \) are generated from the \( N(0, 1) \) distribution and are independent. The candidate models for DGPs 2-5 are the same for DGP 1. The results are reported in Figures 4-7.

First, we consider the dynamic regression models with smooth time-varying coefficients (DGP 2). In Figure 4, when the sample \( T \) is large enough, the TJMA estimator yields the smaller risk than other six estimators, which is even more clear for small \( R^2 \).

Second, we consider the deterministic single break (DGP 3), namely, a single break with a given breakpoint and size. In Figure 5, not surprisingly, the TJMA estimator outperforms other estimators when the sample size is larger than 50 for all \( R^2 \), while AICc and SAICc
yield smaller risk than SAIC and SBIC; see, for example, the case with \( T = 200 \) and \( R^2 > 0.4 \).

Third, we consider an alternative with nonmonotonic smooth structural changes (DGP 4). This is a smooth transition regression model, considered in Lin & Teräsvirta (1994), which is further studied by Cai (2007) and Chen (2015). The smooth transition function is a second-order logistic function. In Figure 6, the TJMA estimator dominates all other estimators. We note that in most cases, the AICc estimator is similar to SAICc estimator with large \( T \) and large \( R^2 \), while both of them achieve a higher risk than the TJMA estimator. SAIC achieves a lower risk with small \( R^2 \) and the SBIC yields the least accurate estimator with large \( R^2 \).

Finally, we consider DGP 5 with periodic structural changes, covering long or short period cycles in economics and finance; see Twrdy & Batista (2016) for an example of container throughput forecasting. In Figure 7, the TJMA estimator outperforms all other estimators. The SAICc estimator habitually is the worst performing estimator with the case \( R^2 < 0.3 \), while its performance improves as \( R^2 \) increases and yields the second smallest risk when \( R^2 \geq 0.7 \).

To sum up, the TJMA estimator achieves lower risk than other estimators, under various data generating processes. When the sample size \( T \) increases, even when \( R^2 \) is small, the TJMA is seem to be the best estimator. When \( R^2 \) is large, the smoothed AICc estimator achieves a lower risk than AICc model selection, which is consistent with the findings in the earlier literature, but both of them perform worse than the TJMA estimator with large \( T \) for all \( R^2 \).

6 Empirical application

It is widely accepted that stock return predictability is a crucial yet controversial issue in empirical finance. The conventional wisdom, studied by Campbell (1990) and Cochrane (1996), is that aggregate dividend yields strongly forecast excess stock return, even at longer horizons. Conventional predictive variables are financial ratios, such as dividend-price ratio, earnings-price ratio, and book-to-market ratio (Rozef (1984), Fama & French (1988), Campbell & Shiller (1988), Lewellen (2004)), as well as corporate payout and financing activity (Lamont (1998), Baker & Wurgler (2000)). However, Wang (2003), Welch & Goyal (2008) show that regressions of excess stock returns perform poorly in out-of-sample forecasts of the U.S. equity premium while historical average returns generate superior forecasts, which causes vigorous debates in the literature (Campbell & Thompson (2008)). It is possible that the presence of structural changes leads to a changing predictive relationship. Indeed, Pesaran & Timmermann (2007) find that the size of parameter variation between the break points in models is considerably large, even the coefficients of dividend yield have opposite signs before and after 1991. Chen and Hong (2012) find strong evidence against
stability in both univariate and multivariate predictor regressions for both the postwar and post-oil-shock sample periods. Furthermore, Rapach & Zhou (2013) point out that model uncertainty and instability seriously impair the forecasting ability of individual predictive regression models.

This sensitivity of the empirical results to model parameter estimation highlights the need of smooth-change weights in model averaging. In this section, we compare the performance of stock return forecasts using our approaches and other existing methods. The key distinction between these various methods lies in that we allow the model combining weights to change smooth with time in combining time-varying coefficient predictive models.

We employ Campbell and Thompson’s (2008) popular dataset, which is widely used in Chen and Hong (2012), Jin et al. (2014) and Lu & Su (2015). We consider the following standard predictive regression model:

\[ Y_{t+1} = \alpha_t + \beta_t' X_t + \epsilon_{t+1}, \]

where \( Y_{t+1} = \log([P_{t+1} + D_{t+1}]/P_t) - r_t \), \( P_t \) is the S&P 500 price index, \( D_t \) is the dividend paid on the S&P 500 price index, \( r_t \) is the 3-month treasury bill rates as the risk-free rate, \( X_t \) is a set of predictor variables, i.e., \( X_t = (X_{t1}, \cdots, X_{tN})' \), and \( N \) is the number of predictor variables. Quarterly variables from Welch and Goyal (2008) are available for 1927:01-2005:12, since quarterly stock returns before that time are constructed by interpolation of lower-frequency data which may be not reliable.

Following Welch and Goyal (2008) and Rapach et al. (2010), we consider 14 financial and economic variables, sorted by relevance to \( Y \): default yield spread \( (X_1) \), treasury bill rate \( (X_2) \), net equity expansion \( (X_3) \), term spread \( (X_4) \), log dividend price ratio \( (X_5) \), log earnings price ratio \( (X_6) \), long-term yield \( (X_7) \), book-to-market ratio \( (X_8) \), inflation \( (X_9) \), log dividend yield \( (X_{10}) \), log dividend payout ratio \( (X_{11}) \), stock variance \( (X_{12}) \), long-term return \( (X_{13}) \), default return spread \( (X_{14}) \). For simplicity, we consider the following 14 nested models: \( \{1, X_1\}, \{1, X_1, X_2\}, \cdots, \{1, X_1, \cdots, X_{14}\} \).

The estimation sample starts from 1947Q1 and our estimation is based on \( T_1 = 80, 92, 104, 116, 128, 140, 152, 164, 176, 188, 200, 212 \) and 224. The remaining observations are used for out-of-sample rolling forecast accuracy assessment. Thus, the out-of-sample forecast periods begins from 1967Q1, 1970Q1, 1973Q1, 1976Q1, 1979Q1, 1982Q1, 1985Q1, 1988Q1, 1991Q1, 1994Q1, 1997Q1, 2000Q1 and 2003Q1, respectively, and all ends at 2005Q4. The postwar sample, 1947Q1-2005Q4, and the post-oil-shock subsample, 1976Q1-2005Q4, are widely used in the existing literature, e.g., Welch and Goyal (2008), Chen and Hong (2012), etc. The bandwidth employed in TJMA, AICc and smoothed AICc is set as \( 2.34 T_1^{-0.2} \).
Following Ullah et al. (2017), we use the out-of-sample $\tilde{R}^2$ measure:

$$\tilde{R}^2 = 1 - \frac{\sum_{t=1}^{T-1} (Y_{t+1} - \hat{Y}_{t+1})^2}{\sum_{t=1}^{T-1} (Y_{t+1} - \bar{Y})^2},$$

where $\hat{Y}_{t+1}$ is the prediction of $Y_{t+1}$ based on a given forecast method, and $\bar{Y}$ is the average of the values of $Y$ across the $T_1$ observations. This measure represents the relative difference in squared error predictive risks. The negative value of $\tilde{R}^2$ suggests that $\hat{Y}$ yields a larger sum of squared one-period forecast errors in comparison with the historical average, and vice versa.

Table 1 compares the $\tilde{R}^2$ obtained from the TJMA estimator with those from other estimators. If $\tilde{R}^2$ is larger, the approach used to forecast is better. We find that in most cases, the TJMA estimator is almost always the best estimator among those considered. When the sample used is large enough, e.g., 292 observations, the TJMA estimator has been improved significantly. Our finding supports the argument of Chen and Hong (2012) that instability exists in univariate predictor models for stock return and smooth structural change is a possibility, which explains why the TJMA estimator is more appropriate and more advantageous than JMA and MMA. Besides, JMA estimator produces the second smallest forecast errors in most cases, with the MMA being a close fourth. In most cases, the AICc yields the worst performance. It is possible that the evidence against stability is a bit weak in quarterly data, which is consistent with Chen and Hong (2012).

7 Conclusion

Although structural changes have received considerable attention in time series econometrics for a long time, no work has attempted to focusing on time-varying model averaging from nonlinear candidate models, especially time-varying coefficient models. This paper proposes a frequentist method for model averaging with time-varying model averaging. This method is more appropriate to nonlinear time series than the traditional MMA and JMA approaches. It is shown that our new TJMA estimator is asymptotically optimal in the sense of achieving the lowest possible squared error in a class of model average estimators. In simulation, its finite-sample and forecast performance outperform other existing methods, including the nonparametric version of bias-correct AIC method. An application to predicting stock returns is undertaken that demonstrates the TJMA performs better than other methods for various sizes of sample investigated.

In this paper, a global bandwidth is used for TJMA estimator, which may be severely affected by the existence structural changes. Therefore, it will be highly desirable to use a time-varying bandwidth for each time point, which will be further investigated in future
research.
Figure 1: Finite-sample performance under DGP 1 with case (i)

Notes:
(1) DGP 1. Smooth Structural Changes:
\[ Y_t = \mu_t + \epsilon_t = \sum_{j=1}^{\infty} \theta_j F(\tau) X_{tj} + \epsilon_t, t = 1, \ldots, T, \]
where \( \tau = t/T, F(\tau) = \tau^3, X_{t1} = 1, \) and observations of all other \( X_{tj}, j \geq 2 \) are generated from the \( N(0, 1) \) distribution and are independent; \( \theta_j = c\sqrt{2\alpha j^{-\alpha - 1/2}}, \) with \( c > 0 \) and \( \alpha = 1.5. \)

(2) In each figure, the sample sizes are shown in four panels. The sample varies from \( T = 50, 75, 100 \) and 200.

(3) Three cases for \( \{\epsilon_t\} \): (i) \( \epsilon_t \sim \text{i.i.d.} N(0, 1); \) (ii) \( \epsilon_t = \epsilon_{t,1} + \epsilon_{t,2}, \epsilon_{t,1} \sim N(0, X_{t2}^2), \epsilon_{t,2} = \phi \epsilon_{t-1,2} + \nu_t, \nu_t \sim \text{i.i.d.} N(0, 1) \) and \( \phi = 0.5; \) (iii) \( \epsilon_t = \sqrt{h_t} u_t, h_t = 0.2 + 0.5X_{t2}^2, u_t \sim \text{i.i.d.} N(0, 1). \)

(4) In each panel, the y axis and the x axis display the mean squared error and the population \( R^2, \) respectively. Seven approaches to estimate parameters are shown in these figures, including TJMA, AICc in \cite{Cai & Tiwari 2000}, SAICc, JMA, MMA, SAIC and SBIC estimators.
Figure 2: Finite-sample performance under DGP 1 with case (ii)

Notes:
(1) DGP 1. Smooth Structural Changes:
\[ Y_t = \mu_t + \varepsilon_t = \sum_{j=1}^{\infty} \theta_j F(\tau)X_{tj} + \varepsilon_t, t = 1, \cdots, T, \]
where \( \tau = t/T, F(\tau) = \tau^3, X_{t1} = 1, \) and observations of all other \( X_{tj}, j \geq 2 \) are generated from the \( N(0,1) \) distribution and are independent; \( \theta_j = c\sqrt{2\alpha j^{-\alpha-1/2}}, \) with \( c > 0 \) and \( \alpha = 1.5. \)

(2) In each figure, the sample sizes are shown in four panels. The sample varies from \( T = 50, 75, 100 \) and 200.

(3) Three cases for \( \{\varepsilon_t\} \): (i) \( \varepsilon_t \sim \text{i.i.d.} N(0,1) \); (ii) \( \varepsilon_t = \varepsilon_{t,1} + \varepsilon_{t,2}, \varepsilon_{t,1} \sim N(0,X_{t2}^2), \varepsilon_{t,2} = \phi \varepsilon_{t-1,2} + u_t, u_t \sim \text{i.i.d.} N(0,1) \) and \( \phi = 0.5; \) (iii) \( \varepsilon_t = \sqrt{h_t}u_t, h_t = 0.2 + 0.5X_{t2}^2, u_t \sim \text{i.i.d.} N(0,1). \)

(4) In each panel, the y axis and the x axis display the mean squared error and the population \( R^2, \) respectively. Seven approaches to estimate parameters are shown in these figures, including TJMA, AICc in [Cai & Tiwari (2000)], SAICc, JMA, MMA, SAIC and SBIC estimators.
Notes:
(1) DGP 1. Smooth Structural Changes:

\[ Y_t = \mu_t + \epsilon_t = \sum_{j=1}^{\infty} \theta_j F(\tau) X_{tj} + \epsilon_t, \quad t = 1, \ldots, T, \]

where \( \tau = t/T, F(\tau) = \tau^3, X_{t1} = 1, \) and observations of all other \( X_{tj}, j \geq 2 \) are generated from the \( N(0, 1) \) distribution and are independent; \( \theta_j = c \sqrt{2\alpha j^{-\alpha-1/2}}, \) with \( c > 0 \) and \( \alpha = 1.5. \)

(2) In each figure, the sample sizes are shown in four panels. The sample varies from \( T = 50, 75, 100, \) and \( 200. \)

(3) Three cases for \( \{\epsilon_t\} \): (i) \( \epsilon_t \sim \text{i.i.d.} N(0, 1); \) (ii) \( \epsilon_t = \epsilon_{t,1} + \epsilon_{t,2}, \epsilon_{t,1} \sim N(0, X_{t2}^2), \epsilon_{t,2} = \phi \epsilon_{t-1,2} + u_t, \) \( u_t \sim \text{i.i.d.} N(0, 1) \) and \( \phi = 0.5; \) (iii) \( \epsilon_t = \sqrt{h_t} u_t, \) \( h_t = 0.2 + 0.5 X_{t2}^2, \) \( u_t \sim \text{i.i.d.} N(0, 1). \)

(4) In each panel, the y axis and the x axis display the mean squared error and the population \( R^2, \) respectively. Seven approaches to estimate parameters are shown in these figures, including TJMA, AICc in [Cai & Tiwari (2000)], SAICc, JMA, MMA, SAIC and SBIC estimators.
Notes:
(1) DGP 2. Dynamic Regression with Smooth Structural Changes: We have

\[ Y_t = \sum_{j=1}^{\infty} \theta_j F(\tau)Y_{t-j} + \epsilon_t, \]

where \( \theta_j = 1/\sqrt{2\alpha_j} - \alpha^{-1/2} \), \( c = \sum \theta_j^2 \), \( F(\tau) = \tau \), \( \epsilon_t = \frac{R}{c} \epsilon_t \) with \( R^2 \) varying on a grid from 0.1 to 0.9, \( \epsilon_t \sim \text{i.i.d.} \text{N}(0,1) \), \( \alpha = 1.5 \).

(2) In each figure, the sample sizes are shown in four panels. The sample varies from \( T = 50, 75, 100 \) and 200.

(3) In each panel, the y axis and the x axis display the mean squared error and the population \( R^2 \), respectively. Seven approaches to estimate parameters are shown in these figures, including TJMA, AICc in Cai & Tiwari (2000), SAICc, JMA, MMA, SAIC and SBIC estimators.
Figure 5: Finite-sample performance under DGP 3

Notes:
(1) DGP 3. Single Structural Break: We have

\[ Y_t = \sum_{j=1}^{\infty} 0.5\theta_j X_{tj} I(t \leq 0.3T) + \sum_{j=1}^{\infty} \theta_j X_{tj} I(t > 0.3T) + \epsilon_t, \]

where \(X_{t1} = 1\), observations of all other \(X_{tj}, j \geq 2\) are generated from the \(N(0, 1)\) distribution and are independent; \(\theta_j = c\sqrt{2\alpha j^{-\alpha - 1/2}}\), with \(c > 0\) and \(\alpha = 1.5\).

(2) In each figure, the sample sizes are shown in four panels. The sample varies from \(T = 50, 75, 100\) and \(200\).

(3) In each panel, the y axis and the x axis display the mean squared error and the population \(R^2\), respectively. Seven approaches to estimate parameters are shown in these figures, including TJMA, AICc in Cai & Tiwari (2000), SAICc, JMA, MMA, SAIC and SBIC estimators.
Figure 6: Finite-sample performance under DGP 4

Notes:
(1) DGP 4. Smooth transition regression (STR): We have
\[ Y_t = \sum_{j=1}^{\infty} \theta_j F(\tau) X_{tj} + \epsilon_t, \]
where \( F(\tau) = 1.5 - 1.5 \exp(-3(\tau - 0.3)^2) \), \( X_{t1} = 1 \) and observations of all other \( X_{tj}, j \geq 2 \) are generated from the \( \mathcal{N}(0, 1) \) distribution and are independent; \( \theta_j = c \sqrt{2\alpha} j^{-\alpha-1/2} \), with \( c > 0 \) and \( \alpha = 1.5 \).

(2) In each figure, the sample sizes are shown in four panels. The sample varies from \( T = 50, 75, 100 \) and \( 200 \).

(3) In each panel, the y axis and the x axis display the mean squared error and the population \( R^2 \), respectively. Seven approaches to estimate parameters are shown in these figures, including TJMA, AICc in [Cai & Tiwari (2000)], SAICc, JMA, MMA, SAIC and SBIC estimators.
Figure 7: Finite-sample performance under DGP 5

Notes:
(1) DGP 5. Smooth Structural Changes with Periodicity: We have

$$Y_t = \sum_{j=1}^{\infty} \theta_j F(\tau) X_{tj} + \varepsilon_t,$$

where $F(\tau) = \sin(\pi \tau^2)$, $X_{t1} = 1$, and observations of all other $X_{tj}, j \geq 2$ are generated from the $N(0, 1)$ distribution and are independent; $\theta_j = c \sqrt{2\alpha j^{-\alpha-1/2}}$, with $c > 0$ and $\alpha = 1.5$.

(2) In each figure, the sample sizes are shown in four panels. The sample varies from $T = 50, 75, 100$ and $200$.

(3) In each panel, the y axis and the x axis display the mean squared error and the population $R^2$, respectively. Seven approaches to estimate parameters are shown in these figures, including TJMA, AICc in Cai & Tiwari (2000), SAICc, JMA, MMA, SAIC and SBIC estimators.
Table 1: Out-of-sample $\tilde{R}^2$

<table>
<thead>
<tr>
<th>estimation</th>
<th>prediction</th>
<th>TJMA</th>
<th>AICc</th>
<th>SAICc</th>
<th>JMA</th>
<th>MMA</th>
<th>SAIC</th>
<th>SBIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1947Q1($T_1=80$)</td>
<td>1967Q1</td>
<td>0.1771°</td>
<td>0.0335</td>
<td>0.1018</td>
<td>0.1761°°</td>
<td>0.1657</td>
<td>0.1111</td>
<td>0.1024</td>
</tr>
<tr>
<td>1947Q1($T_1=92$)</td>
<td>1970Q1</td>
<td>0.1242°</td>
<td>-0.0312</td>
<td>0.0549</td>
<td>0.1228°°</td>
<td>0.1124</td>
<td>0.0512</td>
<td>0.0530</td>
</tr>
<tr>
<td>1947Q1($T_1=104$)</td>
<td>1973Q1</td>
<td>0.1212°</td>
<td>-0.0443</td>
<td>0.0770</td>
<td>0.1107°°</td>
<td>0.0977</td>
<td>0.0358</td>
<td>0.0367</td>
</tr>
<tr>
<td>1947Q1($T_1=116$)</td>
<td>1976Q1</td>
<td>0.0372°</td>
<td>-0.1574</td>
<td>-0.0397</td>
<td>0.0025°°</td>
<td>-0.0165</td>
<td>-0.1128</td>
<td>-0.0708</td>
</tr>
<tr>
<td>1947Q1($T_1=128$)</td>
<td>1979Q1</td>
<td>0.0357°</td>
<td>-0.1727</td>
<td>-0.0291</td>
<td>-0.0188°°</td>
<td>-0.0381</td>
<td>-0.1393</td>
<td>-0.1007</td>
</tr>
<tr>
<td>1947Q1($T_1=140$)</td>
<td>1982Q1</td>
<td>-0.1057°</td>
<td>-0.3579</td>
<td>-0.1375°°</td>
<td>-0.1830</td>
<td>-0.2064</td>
<td>-0.3210</td>
<td>-0.2220</td>
</tr>
<tr>
<td>1947Q1($T_1=152$)</td>
<td>1985Q1</td>
<td>-0.1833°</td>
<td>-0.4899</td>
<td>-0.2359°°</td>
<td>-0.2586</td>
<td>-0.2829</td>
<td>-0.4043</td>
<td>-0.2567</td>
</tr>
<tr>
<td>1947Q1($T_1=164$)</td>
<td>1988Q1</td>
<td>-0.2630°°</td>
<td>-0.6292</td>
<td>-0.2522°</td>
<td>-0.3773</td>
<td>-0.4091</td>
<td>-0.5724</td>
<td>-0.3658</td>
</tr>
<tr>
<td>1947Q1($T_1=176$)</td>
<td>1991Q1</td>
<td>-0.2238°°</td>
<td>-0.4310</td>
<td>-0.2242</td>
<td>-0.2194°</td>
<td>-0.2347</td>
<td>-0.2945</td>
<td>-0.3095</td>
</tr>
<tr>
<td>1947Q1($T_1=188$)</td>
<td>1994Q1</td>
<td>-0.2181</td>
<td>-0.4080</td>
<td>-0.1579°</td>
<td>-0.2207</td>
<td>-0.2161°°</td>
<td>-0.2570</td>
<td>-0.3249</td>
</tr>
<tr>
<td>1947Q1($T_1=200$)</td>
<td>1997Q1</td>
<td>-0.0423°°</td>
<td>-0.1979</td>
<td>-0.1005</td>
<td>-0.0365°</td>
<td>-0.0538</td>
<td>-0.0735</td>
<td>-0.0990</td>
</tr>
<tr>
<td>1947Q1($T_1=212$)</td>
<td>2000Q1</td>
<td>0.0125°°</td>
<td>-0.1434</td>
<td>-0.1316</td>
<td>0.0694°</td>
<td>-0.0207</td>
<td>-0.0383</td>
<td>-0.0330</td>
</tr>
<tr>
<td>1947Q1($T_1=224$)</td>
<td>2003Q1</td>
<td>0.1859°°</td>
<td>-0.0714</td>
<td>-0.0560</td>
<td>0.1822</td>
<td>0.2909</td>
<td>0.3013°</td>
<td>-0.0543</td>
</tr>
</tbody>
</table>

Notes: (1) The estimation time begins from 1947Q1.

(2) Seven approaches to estimate parameters are shown in Table 1, corresponding to TJMA, AICc in Cai & Tiwari (2000), Smoothed AICc, JMA, MMA, SAIC and SBIC estimators. The larger the criteria, the better the approach.

(3) The bandwidth used in this empirical application is $2.34T_1^{-0.2}$, as the same as that in simulation.

(4) ° and °° denote the best and the second best forecast among these seven approaches, respectively.
References


Appendix

1 Appendix A.1

Proof of Theorem 1. First we show

\[ \frac{\tilde{L}_{L,T}(\hat{\mathbf{w}}_t)}{\inf_{\mathbf{w} \in \mathcal{H}_T} \tilde{L}_{L,T}(\mathbf{w})} \overset{p}{\to} 1 \]  

(A.1)

with Assumptions 1 - 7.

Note that

\[ CV_{L,T}(\mathbf{w}) = \tilde{L}_{L,T}(\mathbf{w}) + \varepsilon' K_t \varepsilon + 2 \varepsilon' K_t \tilde{A}(\mathbf{w}) \mu - 2 \varepsilon' K_t \tilde{P}(\mathbf{w}) \varepsilon \]

\[ = \tilde{L}_{L,T}(\mathbf{w}) + \varepsilon' K_t \varepsilon + 2 \varepsilon' K_t \tilde{A}(\mathbf{w}) \mu - 2(\varepsilon' K_t \tilde{P}(\mathbf{w}) \varepsilon - \text{tr}(K_t \tilde{P}(\mathbf{w}) \Omega)) - 2 \text{tr}(K_t \tilde{P}(\mathbf{w}) \Omega). \]

With Assumption 7 (A.1) is valid if the following hold: as \( T \to \infty \),

\[ \sup_{\mathbf{w} \in \mathcal{H}_T} \left| \frac{\varepsilon' K_t \tilde{A}(\mathbf{w}) \mu}{R_{L,T}(\mathbf{w})} \right| \overset{p}{\to} 0, \]  

(A.2)

\[ \sup_{\mathbf{w} \in \mathcal{H}_T} \left| \frac{\varepsilon' K_t \tilde{P}(\mathbf{w}) \varepsilon - \text{tr}(K_t \tilde{P}(\mathbf{w}) \Omega)}{R_{L,T}(\mathbf{w})} \right| \overset{p}{\to} 0, \]  

(A.3)

and

\[ \sup_{\mathbf{w} \in \mathcal{H}_T} \left| \frac{\tilde{L}_{L,T}(\mathbf{w})}{R_{L,T}(\mathbf{w})} - 1 \right| \overset{p}{\to} 0. \]  

(A.4)

(A.2)-(A.4) will be verified later.

We next show that

\[ \sup_{\mathbf{w} \in \mathcal{H}_T} \left| \frac{L_{L,T}(\mathbf{w})}{R_{L,T}(\mathbf{w})} - 1 \right| \overset{p}{\to} 0. \]  

(A.5)

It is easy to obtain that

\[ L_{L,T}(\mathbf{w}) - R_{L,T}(\mathbf{w}) \]

\[ = \mu' A'(\mathbf{w}) K_t A(\mathbf{w}) \mu + \varepsilon' P'(\mathbf{w}) K_t P(\mathbf{w}) \varepsilon - 2 \mu' A(\mathbf{w})' K_t P(\mathbf{w}) \varepsilon \]

\[ - [\mu' A'(\mathbf{w}) K_t A(\mathbf{w}) \mu + \text{tr}(P'(\mathbf{w}) K_t P(\mathbf{w}) \Omega)] \]

\[ = \varepsilon' P'(\mathbf{w}) K_t P(\mathbf{w}) \varepsilon - \text{tr}(P(\mathbf{w})' K_t P(\mathbf{w}) \Omega) - 2 \mu' A'(\mathbf{w}) K_t P(\mathbf{w}) \varepsilon \]

\[ = D_1(\mathbf{w}) - D_2(\mathbf{w}), \]
Thus, to prove (A.5), we only need to verify the following two equations:

$$\sup_{\mathbf{w} \in \mathcal{C}_T} \left| \frac{D_1(\mathbf{w})}{R_{t,T}(\mathbf{w})} \right| \xrightarrow{p} 0. \quad \text{(A.6)}$$

and

$$\sup_{\mathbf{w} \in \mathcal{C}_T} \left| \frac{D_2(\mathbf{w})}{R_{t,T}(\mathbf{w})} \right| \xrightarrow{p} 0. \quad \text{(A.7)}$$

Let $\lambda_{\max}(\mathbf{P}'(\mathbf{w}))$ be the maximum eigenvalue of $\mathbf{P}'(\mathbf{w})$. When $\mathbf{X}$ is nonstochastic, based on the Chebyshev’s inequality, Theorem 2 in Whittle (1960) and Assumption [1], we have, for any $\delta > 0$,

$$\Pr \left( \sup_{\mathbf{w} \in \mathcal{C}_T} \left| \mathbf{e}'(\mathbf{w}) \mathbf{K}_t \mathbf{P}(\mathbf{w}) \mathbf{e} - \text{tr}(\mathbf{P}'\mathbf{K}_t \mathbf{P}(\mathbf{w}) \mathbf{\Omega}) \right| > \delta \frac{\mathbf{R}_{t,T}(\mathbf{w})}{\mathbf{R}_{t,T}(\mathbf{w})} \right) \leq \Pr (\sup_{\mathbf{w} \in \mathcal{C}_T} \sum_{k=1}^{M_T} \sum_{m=1}^{M_I} \mathbf{w}_k \mathbf{w}_m | \mathbf{e}' \mathbf{P}' \mathbf{K}_t \mathbf{P}_m \mathbf{e} - \text{tr}(\mathbf{P}' \mathbf{K}_t \mathbf{P}_m \mathbf{\Omega}) | > \delta \frac{\mathbf{R}_{t,T}(\mathbf{w})}{\mathbf{R}_{t,T}(\mathbf{w})} )$$

$$\leq \sum_{k=1}^{M_T} \sum_{m=1}^{M_I} \mathbf{P} ( | \mathbf{e}' \mathbf{P}' \mathbf{K}_t \mathbf{P}_m \mathbf{e} - \text{tr}(\mathbf{P}' \mathbf{K}_t \mathbf{P}_m \mathbf{\Omega}) | > \delta \frac{\mathbf{R}_{t,T}(\mathbf{w})}{\mathbf{R}_{t,T}(\mathbf{w})} )$$

$$\leq \sum_{k=1}^{M_T} \sum_{m=1}^{M_I} \mathbf{E} \left[ \frac{(\mathbf{e}' \mathbf{P}' \mathbf{K}_t \mathbf{P}_m \mathbf{e} - \text{tr}(\mathbf{P}' \mathbf{K}_t \mathbf{P}_m \mathbf{\Omega}))^2}{\delta^2 \mathbf{R}_{t,T}(\mathbf{w})^2} \right]$$

$$\leq C_4 \delta^{-2G} \xi_{t,T}^{-2G} \sum_{k=1}^{M_T} \sum_{m=1}^{M_I} \text{tr} \left( (\mathbf{P}' \mathbf{K}_t \mathbf{P}_m \mathbf{\Omega})^2 \right)^G$$

$$\leq C_4' \delta^{-2G} \xi_{t,T}^{-2G} \sum_{k=1}^{M_T} \sum_{m=1}^{M_I} \text{tr} (\mathbf{P}' \mathbf{K}_t \mathbf{P}_m \mathbf{\Omega})^G$$

$$\leq C_4' \delta^{-2G} \xi_{t,T}^{-2G} M_T \sum_{m=1}^{M_T} (\mathbf{R}_{t,T}(\mathbf{w}_m^0))^G,$$  \quad \text{(A.8)}

where $\mathbf{w}_m^0$ defines the weight with the $m$-th element 1 and others 0, and $C_4$ and $C_4'$ are some
constants. Thus, with Assumption \[6\] on non-stochastic \(X\) is verified.

To prove \[A.7\], we have \(\|\mu' A'_k K_t P_m \|^2 \leq \lambda_{\max}^2 (P'(w)) \) \(K_{\max} \mu' A'_k K_t A_k \mu \leq C \mu' A'_k K_t A_k \mu, 1 \leq k \leq M_T, 1 \leq m \leq M_T. Based on the Chebyshev's inequality, Theorem 2 in Whittle (1960) and Assumption \[4\] for any \(\delta > 0\), we have

\[
P \left\{ \sup_{w \in \mathcal{F}_T} \left| \frac{\mu' A'(w) K_t P_t(w) \epsilon}{R_{t,T}(w)} \right| > \delta \right\} \leq P \left\{ \sup_{w \in \mathcal{F}_T} \sum_{1 \leq k \leq M_T} \sum_{1 \leq m \leq M_T} w_k w_m |\mu'(I_T - P_k)' K_T P_m \epsilon| > \delta \xi_{t,T} \right\} \leq P \left\{ \max_{1 \leq k \leq M_T} \max_{1 \leq m \leq M_T} |\mu'(I_T - P_k)' K_T P_m \epsilon| > \delta \xi_{t,T} \right\} = P \{ |\mu'(I_T - P_1)' K_T P_1 \epsilon| > \delta \xi_{t,T} \} \cup \{ |\mu'(I_T - P_1)' K_T P_2 \epsilon| > \delta \xi_{t,T} \} \cup \cdots \cup \{ |\mu'(I_T - P_M)' K_T P_M \epsilon| > \delta \xi_{t,T} \} \}
\[
\leq \sum_{1 \leq k \leq M_T} \sum_{1 \leq m \leq M_T} E \left[ \frac{|\mu'(I_T - P_k)' K_T P_m \epsilon|^{2G}}{\delta^{2G} \xi_{t,T}^{2G}} \right] \leq C_5 \delta^{-2G} \xi_{t,T}^{-2G} \sum_{1 \leq k \leq M_T} \sum_{1 \leq m \leq M_T} |\mu'(I_T - P_k)' K_T P_m|^{2G} \leq C'_5 \delta^{-2G} \xi_{t,T}^{-2G} \sum_{1 \leq k \leq M_T} \sum_{1 \leq m \leq M_T} (\mu' A'_m K_T A_m \mu)^G \leq C'_5 \delta^{-2G} \xi_{t,T}^{-2G} M_T \sum_{m=1}^{M_T} (R_{t,T}(w_m^0))^G, \tag{A.9}
\]

where \(C_5\) and \(C'_5\) are some constants. With Assumption 6, \([A.7]\) on non-stochastic \(X\) is verified.

Besides, for the case of random \(X\), based on the dominated convergence theorem, Assumption 5 and \([A.2]\), the results in \([A.6]\) and \([A.7]\) are obtained by \([A.8]\) and \([A.9]\), respectively.

Define

\[
V_{t,T}(\hat{w}_t) = \mu' A'(\hat{w}_t) K_t A(\hat{w}_t) \mu + \text{tr}(P'(\hat{w}_t) K_t P(\hat{w}_t) \Omega),
\]

and

\[
\bar{V}_n(\hat{w}_t) = \mu' \tilde{A}'(\tilde{w}_t) K_t \tilde{A}(\tilde{w}_t) \mu + \text{tr}(\tilde{P}'(\tilde{w}_t) K_t \tilde{P}(\tilde{w}_t) \Omega).
\]
Next, we follow the spirit in Zhang et al. (2013) to obtain

\[
\frac{L_{t,T}(\hat{w}_t)}{\inf_{w \in \mathcal{G}_T} L_{t,T}(w)} - 1 = \sup_{w \in \mathcal{G}_T} \left( \frac{L_{t,T}(\hat{w}_t)}{L_{t,T}(w)} - 1 \right)
\]

\[
= \sup_{w \in \mathcal{G}_T} \left( \frac{L_{t,T}(\hat{w}_t)}{L_{t,T}(w)} \frac{R_{t,T}(w)}{R_{t,T}(\hat{w}_t)} V_{t,T}(\hat{w}_t) \frac{L_{t,T}(\hat{w}_t)}{L_{t,T}(w)} - 1 \right)
\]

\[
\leq \sup_{w \in \mathcal{G}_T} \left( \frac{R_{t,T}(w)}{R_{t,T}(\hat{w}_t)} \frac{\tilde{L}_{t,T}(\hat{w}_t)}{\tilde{L}_{t,T}(w)} \frac{\tilde{L}_{t,T}(\hat{w}_t)}{\tilde{L}_{t,T}(w)} - 1 \right).
\]

Thus, with [A.2]-[A.4], Theorem 1 holds.

Next, we only need to verify [A.2]-[A.4]. Denote \( \tilde{\xi}_{t,T} = \inf_{w \in \mathcal{G}_T} \tilde{R}_{t,T}(w) \). If \( \sup_{w \in \mathcal{G}_T} \left| \frac{\tilde{R}_{t,T}(w)}{\tilde{R}_{t,T}(\hat{w}_t)} - 1 \right| \leq 1 \), we have

\[
\tilde{\xi}_{t,T}^{2G} \left[ \sum_{m=1}^{M_T} (\tilde{R}_{t,T}(w^0_m))^G \right]^{-1}
\]

\[
= \left[ \inf_{w \in \mathcal{G}_T} \left( \frac{R_{t,T}(w)}{R_{t,T}(\hat{w}_t)} \right) \right]^{2G} \left[ \sum_{m=1}^{M_T} (R_{t,T}(w^0_m))^G \left( \frac{\tilde{R}_{t,T}(w^0_m)}{R_{t,T}(w^0_m)} \right) \right]^{-1}
\]

\[
\geq \tilde{\xi}_{t,T}^{2G} \left[ \inf_{w \in \mathcal{G}_T} \left( \frac{\tilde{R}_{t,T}(w)}{R_{t,T}(\hat{w}_t)} \right) \right]^{2G} \left[ \max_{1 \leq m \leq M_T} \left( \frac{\tilde{R}_{t,T}(w^0_m)}{R_{t,T}(w^0_m)} \right) \right]^{-G} \left[ \sum_{m=1}^{M_T} (R_{t,T}(w^0_m))^G \right]^{-1}
\]

\[
\geq \tilde{\xi}_{t,T}^{2G} \left[ 1 + \inf_{w \in \mathcal{G}_T} \left( \frac{\tilde{R}_{t,T}(w)}{R_{t,T}(\hat{w}_t)} - 1 \right) \right]^{2G} \left[ \sup_{w \in \mathcal{G}_T} \left( \frac{\tilde{R}_{t,T}(w)}{R_{t,T}(w)} - 1 \right) + 1 \right]^{-G} \left[ \sum_{m=1}^{M_T} (R_{t,T}(w^0_m))^G \right]^{-1}
\]

\[
\geq \tilde{\xi}_{t,T}^{2G} \left[ 1 - \sup_{w \in \mathcal{G}_T} \left| \frac{\tilde{R}_{t,T}(w)}{R_{t,T}(\hat{w}_t)} - 1 \right| + 1 \right]^{-G} \left[ \sum_{m=1}^{M_T} (R_{t,T}(w^0_m))^G \right]^{-1}. \quad (A.10)
\]

Under Assumptions [5] and [6] we obtain the following result from [A.10]:

\[
M_T \tilde{\xi}_{t,T}^{2G} \sum_{m=1}^{M_T} (\tilde{R}_{t,T}(w^0_m))^G \xrightarrow{a.s.} 0. \quad (A.11)
\]

To prove [A.2], by using the Chebyshev inequality, Theorem 2 of Whittle (1960) and
Assumptions 1-2, we have, for any $\delta > 0$:

$$
\Pr \left\{ \sup_{w \in \mathcal{F}_T} \left| \frac{\epsilon' K_t \tilde{A}(w) \mu}{R_{t,T}(w)} \right| > \delta \right\} 
\leq \Pr \left\{ \sup_{w \in \mathcal{F}_T} \sum_{m=1}^{M_T} w^m |\epsilon' K_t \tilde{A}(w_m^0) \mu| > \delta \tilde{\xi}_{t,T} \right\} 
= \Pr \left\{ \max_{1 \leq m \leq M_T} |\epsilon' K_t \tilde{A}(w_m^0) \mu| > \delta \tilde{\xi}_{t,T} \right\} 
= \Pr \left\{ (|\epsilon' K_t \tilde{A}(w_1^0) \mu| > \delta \tilde{\xi}_{t,T}) \cup \cdots \cup (|\epsilon' K_t \tilde{A}(w_{M_T}^0) \mu| > \delta \tilde{\xi}_{t,T}) \right\} 
\leq \sum_{m=1}^{M_T} \Pr \left\{ |\epsilon' K_t \tilde{A}(w_m^0) \mu| > \delta \tilde{\xi}_{t,T} \right\} 
\leq \delta^{-2G} \tilde{\xi}_{t,T}^{-2G} \sum_{m=1}^{M_T} E(\mu' \tilde{A}'(w_m^0) K_t \epsilon)^{2G} 
\leq C_1 \delta^{-2G} \tilde{\xi}_{t,T}^{-2G} \sum_{m=1}^{M_T} [\mu' \tilde{A}'(w_m^0) K_t \Omega K_t \tilde{A}(w_m^0) \mu]^G 
\leq C_1 \delta^{-2G} \tilde{\xi}_{t,T}^{-2G} \delta^G(\Omega) \sum_{m=1}^{M_T} (\text{tr}(K_t \tilde{A}(w_m^0) \mu' \tilde{A}'(w_m^0) K_t)) 
\leq C_1 \delta^{-2G} \tilde{\xi}_{t,T}^{-2G} \delta^G(\Omega) k_{\max} \sum_{m=1}^{M_T} [\mu' \tilde{A}'(w_m^0) K_t \tilde{A}(w_m^0) \mu]^G,
$$

where $C_1$ is a positive constant. Then from (A.11) and Assumptions 2 and 4, $\sup_{w \in \mathcal{F}_T} \left| \frac{\epsilon' K_t \tilde{A}(w) \mu}{R_{t,T}(w)} \right| \xrightarrow{P} 0$.

Next, we verify (A.3).

$$
\Pr \left\{ \sup_{w \in \mathcal{F}_T} \left| \frac{\epsilon' K_t \tilde{P}(w) \epsilon - \text{tr}(K_t \tilde{P}(w) \Omega)}{R_{t,T}(w)} \right| > \delta \right\} 
\leq \sum_{m=1}^{M_T} \Pr \left\{ \sup_{w \in \mathcal{F}_T} \left| \frac{\epsilon' K_t \tilde{P}(w_m^0) \epsilon - \text{tr}(K_t \tilde{P}(w_m^0) \Omega)}{R_{t,T}(w_m^0)} \right| > \delta \right\} 
\leq \delta^{-2G} \tilde{\xi}_{t,T}^{-2G} \sum_{m=1}^{M_T} E[|\epsilon' K_t \tilde{P}(w_m^0) \epsilon - \text{tr}(K_t \tilde{P}(w_m^0) \Omega)|]^{2G} 
\leq C_2 \delta^{-2G} \tilde{\xi}_{t,T}^{-2G} \sum_{m=1}^{M_T} \text{tr} \left( \Omega \tilde{P}(w_m^0) K_t \Omega K_t \tilde{P}(w_m^0) \Omega \right) 
\leq C_2 \delta^{-2G} \tilde{\xi}_{t,T}^{-2G} \zeta(\Omega) \sum_{m=1}^{M_T} \text{tr} \left( \tilde{P}(w_m^0) K_t \tilde{P}(w_m^0) \Omega \right) 
$$
\[ \leq C_2 \delta^{-2G} \tilde{\xi}_{T,T}^{-2G} \zeta(\Omega) K_{\text{max}} \sum_{m=1}^{M_T} \left[ \text{tr} \left( \tilde{P}'(w_m^0) K_{\text{t}} \tilde{P}(w_m^0) \Omega \right) \right], \]

where \( C_2 \) is a positive constant. Then from Assumptions 2 4 and (A.11), we obtain that 
\[ \sup_{w \in \mathcal{G}_T} \left| \frac{\epsilon' \tilde{P}(w) K_{\text{t}} \tilde{A}(w) \mu}{\tilde{R}_{T,T}(w)} \right| \xrightarrow{p} 0. \]

Finally, we will complete the proof of Theorem 1 with verifying (A.4). Note that 
\[ \left| \frac{L_{T,T}(w) - \tilde{R}_{T,T}(w)}{w \in \mathcal{G}_T} \right| = |\epsilon' \tilde{P}'(w) K_{\text{t}} \tilde{P}(w) - \text{tr}(P'(w) K_{\text{t}} P(w) \Omega) - 2 \epsilon' \tilde{P}'(w) K_{\text{t}} \tilde{A}(w) \mu|. \]

We only need to prove the following equations:
\[ \sup_{w \in \mathcal{G}_T} \left| \frac{\epsilon' \tilde{P}'(w) K_{\text{t}} \tilde{A}(w) \mu}{\tilde{R}_{T,T}(w)} \right| \xrightarrow{P} 0 \quad \text{(A.12)} \]

and
\[ \sup_{w \in \mathcal{G}_T} \left| \frac{\epsilon' \tilde{P}'(w) K_{\text{t}} \tilde{P}(w) - \text{tr}(P'(w) K_{\text{t}} P(w) \Omega)}{\tilde{R}_{T,T}(w)} \right| \xrightarrow{P} 0. \quad \text{(A.13)} \]

For (A.12), we have that
\[ \Pr \left( \sup_{w \in \mathcal{G}_T} \left| \frac{\epsilon' \tilde{P}'(w) K_{\text{t}} \tilde{A}(w) \mu}{\tilde{R}_{T,T}(w)} \right| > \delta \right) \leq \Pr \left( \sup_{w \in \mathcal{G}_T} \left| \epsilon' \tilde{P}'(w) K_{\text{t}} \tilde{A}(w) \mu \right| > \delta \tilde{\xi}_{T,T} \right) \leq \sum_{m=1}^{M_T} \Pr \left( \left| \epsilon' \tilde{P}'(w_m^0) K_{\text{t}} \tilde{A}(w_m^0) \mu \right| > \delta \tilde{\xi}_{T,T} \right) \leq \delta^{-2G} \tilde{\xi}_{T,T}^{-2G} \sum_{m=1}^{M_T} E |\epsilon' \tilde{P}'(w_m^0) K_{\text{t}} \tilde{A}(w_m^0) \mu|^{2G} \leq \delta^{-2G} \tilde{\xi}_{T,T}^{-2G} \text{max}_{1 \leq m \leq M_T} \tilde{P}'(w_m^0) \left( \mu' \tilde{A}'(w_m^0) K_{\text{t}} \tilde{A}(w_m^0) \mu \right)^G \leq \delta^{-2G} \tilde{\xi}_{T,T}^{-2G} C K_{\text{max}} \zeta(\Omega) \left( \mu' \tilde{A}'(w_m^0) K_{\text{t}} \tilde{A}(w_m^0) \mu \right)^G. \]

Then from (A.11) and Assumptions 2 4 \( \sup_{w \in \mathcal{G}_T} \left| \frac{\epsilon' \tilde{P}'(w) K_{\text{t}} \tilde{A}(w) \mu}{\tilde{R}_{T,T}(w)} \right| \xrightarrow{P} 0 \). The proof of (A.13) is similar to that of (A.12), and thus the proof is omitted. Thus, (A.3) is completed.
2 Appendix A.2

Proof of Theorem 2. Denote the maximum singular values of matrixes $B_i$, $i = 1, 2$, by $\zeta(B_1)$ and $\zeta(B_2)$. It is acknowledged that for any square matrices $B_1$ and $B_2$ with identical dimensions, the following simple inequalities are obtained: $\zeta(B_1B_2) \leq \zeta(B_1)\zeta(B_2)$ and $\zeta(B_1 + B_2) \leq \zeta(B_1) + \zeta(B_2)$. See more discussions of these inequalities in proof of Theorem 5.2 in Li (1987) and proof of Theorem 2.2 in Zhang et al. (2013).

Let $h^* = \max_{1 \leq m \leq M_T} \max_{1 \leq i \leq n} h_{ii}^m$, and $\tilde{h} = \frac{h^*}{1-h^*}$, then by Assumption 12, we have $h^* = O(\gamma T^{-1})$, $\tilde{h} = O(\gamma T^{-1})$ a.s., and

$$h^* \leq \lambda \max\{q_1, \ldots, q_{M_T}\} T^{-1} \xi_{t,T}^{-1} (\mu'A_m\mu + \text{tr}(P_m K_t P_m \Omega))$$

$$\leq \lambda \gamma T^{-1} \xi_{t,T}^{-1} (\mu'A_m\mu + \text{tr}(P_m K_t P_m \Omega))$$

$$\leq \lambda \gamma \xi_{t,T}^{-1} (\mu'\mu + \zeta(\Omega) q_m) / T.$$ 

Then with Assumption 2 and Assumptions 10-11, we have

$$h^* \rightarrow 0 \text{ and } \tilde{h} \rightarrow 0 \text{ a.s.} \quad (A.14)$$

Let $Q_m$ be a $T \times T$ diagonal matrix with the $(i, i)$-th element $\frac{h_{ii}^m}{1-h_{ii}^m}$. Then it is easy to obtain that $\tilde{P}_m = P_m - Q_mA_m$, and $\tilde{A}_m = A_m + Q_mA_m$. Denote $Q(w) = \sum_{m=1}^{M_T} w^m Q_m$, $T_m = Q_m P_m$ and $T(w) = \sum_{m=1}^{M_T} w^m T_m$. To prove Theorem 2, we only need to verify the following equations:

$$\sup_w \left| \frac{L_{t,T}(w)}{R_{t,T}(w)} - 1 \right| \xrightarrow{p} 0, \quad (A.15)$$

$$\sup_w \left| \frac{\tilde{R}_{t,T}(w)}{R_{t,T}(w)} - 1 \right| \xrightarrow{p} 0, \quad (A.16)$$

$$\sup_w \left| \frac{\tilde{L}_{t,T}(w)}{R_{t,T}(w)} - 1 \right| \xrightarrow{p} 0, \quad (A.17)$$

$$\sup_w \left| \frac{\mu'\tilde{A}(w)K_t \epsilon}{\tilde{R}_{t,T}(w)} \right| \xrightarrow{p} 0, \quad (A.18)$$

$$\sup_w \left| \frac{\epsilon'\tilde{P}(w)K_t \epsilon}{\tilde{R}_{t,T}(w)} \right| \xrightarrow{p} 0. \quad (A.19)$$

Intuitively, (A.15), (A.17) and (A.18) are similar to (A.5), (A.4) and (A.2) in proof of
Theorem 1. To prove (A.15), we have:

\[
\sup_{w \in \mathcal{H}} \left| L_{t,T}(w) - R_{t,T}(w) \right| \leq \sup_{w \in \mathcal{H}} \frac{|\varepsilon'P'(w)K_tP(w)\varepsilon|}{R_{t,T}(w)} + \sup_{w \in \mathcal{H}} \frac{|\mu' A'(w)K_tP(w)\varepsilon|}{R_{t,T}(w)} + \sup_{w \in \mathcal{H}} \frac{\text{tr}(P'(w)K_tP(w)\Omega)}{R_{t,T}(w)} \equiv \Delta_{11} + \Delta_{12} + \Delta_{13}.
\]

We next show that \(\Delta_{11} = o_p(1)\), \(\Delta_{12} = o_p(1)\) and \(\Delta_{13} = o_p(1)\). For \(\Delta_{11}\), we have

\[
|\varepsilon'P'(w)K_tP(w)\varepsilon| \leq k_{\text{max}} \varepsilon'P'(w)P(w)\varepsilon. \quad (A.20)
\]

For \(\Delta_{12}\), we have

\[
\frac{|\mu' A'(w)K_tP(w)\varepsilon|}{R_{t,T}(w)} \leq \left[ k_{\text{max}} \frac{\varepsilon'P'(w)P(w)\varepsilon \mu' A'(w)K_tA(w)\mu}{R_{t,T}(w)} \right]^{rac{1}{2}} \leq \left[ k_{\text{max}} \frac{\varepsilon'P'(w)P(w)\varepsilon}{R_{t,T}(w)} \right]^{rac{1}{2}}. \quad (A.21)
\]

Besides, it is easy to obtain that

\[
\Pr \left\{ \sup_{w \in \mathcal{H}} \frac{\varepsilon'P'(w)P(w)\varepsilon}{R_{t,T}(w)} \geq \delta \right\} \leq \Pr \{ \varepsilon'P'(w)P(w)\varepsilon \geq \xi_{t,T}\delta \} \leq E \{ \varepsilon'P'(w)P(w)\varepsilon \} \xi_{t,T}^{-1}\delta^{-1} = \text{tr}(P'(w)P(w)\Omega)\xi_{t,T}^{-1}\delta^{-1} \leq \zeta(\Omega)\text{tr}(P'(w)P(w))\xi_{t,T}^{-1}\delta^{-1}. \quad (A.22)
\]

With Assumptions 2 and 10 and (A.22), \(\sup_{w \in \mathcal{H}} \frac{\varepsilon'P'(w)P(w)\varepsilon}{R_{t,T}(w)} \to 0\). Combined with (A.20) and (A.21), we have \(\Delta_{11} = o_p(1)\), \(\Delta_{12} = o_p(1)\). For \(\Delta_{13}\), we have:

\[
\text{tr}(P'(w)K_tP(w)\Omega) \leq k_{\text{max}} \text{tr}(P(w)\Omega P'(w)) \leq k_{\text{max}} \zeta(\Omega) \text{tr}(P'(w)P(w)).
\]
Then with Assumptions 2 and 10, we have

\[
\Pr\left(\sup_{w \in \mathcal{C}_T} \frac{\text{tr}(P'(w)K_tP(w)\Omega)}{R_{l,T}(w)} \geq \delta\right) \\
\leq \Pr\left(\sup_{w \in \mathcal{C}_T} \text{tr}(P'(w)K_tP(w)\Omega) \geq \xi_{l,T}\delta\right) \\
\leq \Pr[k_{\max}\zeta(\Omega)\text{tr}(P'(w)P(w))\xi_{l,T}^{-1} \geq \delta] \\
\rightarrow 0,
\]

then $\Delta_{13} = o_p(1)$. Thus, the proof of (A.15) is completed.

To prove (A.16), we have:

\[
\sup_{w \in \mathcal{C}_T} \left| \frac{\tilde{R}_{l,T}(w) - R_{l,T}(w)}{R_{l,T}(w)} \right| \leq \sup_{w \in \mathcal{C}_T} \left| \frac{\mu'\tilde{A}'(w)K_t\tilde{A}(w)\mu - \mu'A'(w)K_tA(w)\mu}{R_{l,T}(w)} \right| \\
+ \sup_{w \in \mathcal{C}_T} \left| \frac{\text{tr}(\tilde{P}(w)K_t\tilde{P}'(w)\Omega)}{R_{l,T}(w)} \right| + \sup_{w \in \mathcal{C}_T} \left| \frac{\text{tr}(P(w)K_tP'(w)\Omega)}{R_{l,T}(w)} \right| \\
\equiv \Delta_{21} + \Delta_{22} + \Delta_{23}.
\]

Because the matrix $P(w)'\Omega P(w)$ is symmetric, $\Delta_{23} = \Delta_{13} = o_p(1)$. Then we only need to verify $\Delta_{21} = o_p(1)$ and $\Delta_{22} = o_p(1)$. For $\Delta_{22},$

\[
\text{tr}(\tilde{P}(w)K_t\tilde{P}'(w)\Omega) \\
\leq \text{tr}(P(w)K_tP'(w)\Omega) + \text{tr}(Q(w)K_tQ'(w)\Omega) + \text{tr}(T(w)K_tT'(w)\Omega) \\
+ 2|\text{tr}(Q(w)K_tP'(w)\Omega)| + 2|\text{tr}(Q(w)K_tT'(w)\Omega)| + 2|\text{tr}(T(w)K_tP'(w)\Omega)|(A.23)
\]

Let $C_t = \sqrt{\text{tr}(K_t^2)} < \infty$ be some constant. Terms on the right side of (A.23) follow that

\[
\text{tr}(Q(w)K_tQ'(w)\Omega) \leq \zeta(Q(w))k_{\max}\text{tr}(Q(w)\Omega) \\
\leq \tilde{h}k_{\max}\text{tr}(Q(w)\Omega) \\
\leq \frac{\tilde{h}}{1 - h^*}\zeta(\Omega)k_{\max} \sum_{m=1}^{M_T} w^m\text{tr}(P'_mP_m) \\
\leq \frac{\tilde{h}}{1 - h^*}\zeta(\Omega)\text{tr}(P(w)'P(w))
\]

and

\[
|\text{tr}(Q(w)K_tP'(w)\Omega)| = |\text{tr}(K_tP'(w)\Omega Q(w))| \\
\leq \sqrt{\text{tr}(K_t^2)\text{tr}(Q'(w)\Omega P(w)P'(w)\Omega Q(w))} \\
\leq C_t\sqrt{\text{tr}(Q^2(w)\Omega P(w)P'(w)\Omega)}
\]
\[ C_t \sqrt{\tilde{h}^2 \text{tr}(\Omega P(w)P'(w)\Omega)} \leq C_t \tilde{h} \sqrt{\text{tr}(\Omega^2 P(w)P'(w))} \leq C_t \tilde{h} \zeta(\Omega) \sqrt{\text{tr}(P(w)P'(w))}. \]

For \( T(w) \), since that

\[
\text{tr}(T(w)T'(w)) = \text{tr}
\left( \sum_{m=1}^{M_T} w_m Q_m P_m \sum_{m=1}^{M_T} w_m P'_m Q'_m \right)
\leq \tilde{h}^2 \sum_{m=1}^{M_T} w_m w_k \text{tr}(Q_m P_m P'_k Q'_k)
\leq \tilde{h}^2 \text{tr}(P(w)P'(w)),
\]

we have

\[
|\text{tr}(T(w)K_t P'(w)\Omega)| \leq \sqrt{\text{tr}(K_t^2) \text{tr}(T'(w)\Omega P(w)P'(w)\Omega T(w))} \leq C_t \tilde{h} \zeta(\Omega) \text{tr}(P(w)P'(w)).
\]

Similarly,

\[
|\text{tr}(T(w)K_t Q'(w)\Omega)| \leq C_t \zeta(\Omega) \tilde{h}^2 \sqrt{\text{tr}(P(w)P'(w))},
\]

and

\[
\text{tr}(T(w)K_t T'(w)\Omega) \leq C_t \zeta(\Omega) \tilde{h}^2 \text{tr}(P(w)P'(w)).
\]

Then with Assumption 10, \( \Pr \left\{ \sup_{w \in \mathcal{H}} \frac{\text{tr}(\tilde{T}(w)K_t \tilde{T}(w)\Omega)}{R_{t,T}(w)} > \delta \right\} \rightarrow 0. \) Thus, \( \Delta_{22} = o_p(1). \)

For \( \Delta_{21} \), we have

\[
\begin{align*}
&\left| \mu' \tilde{A}'(w) K_t \tilde{A}(w) \mu - \mu' A'(w) K_t A(w) \mu \right| / R_{t,T}(w) \\
= &\sum_{k=1}^{M_T} \sum_{m=1}^{M_T} \left\{ w^k w^m \left( \mu' \tilde{A}'_k K_t \tilde{A}_m \mu - \mu' A'_k K_t A_m \mu \right) \right\} / R_{t,T}(w) \\
= &\sum_{k=1}^{M_T} \sum_{m=1}^{M_T} \left\{ w^k w^m \left[ \mu'(A_k + Q_k A_k)' K_t (A_m + Q_m A_m) \mu - \mu' A'_k K_t A_m \mu \right] \right\} / R_{t,T}(w)
\end{align*}
\]
\[ \sum_{k=1}^{M_T} \sum_{m=1}^{M_T} w^k w^m [\mu' A_k Q_k K_t Q_m'] + 2 \sum_{k=1}^{M_T} \sum_{m=1}^{M_T} w^k w^m \mu' A_k Q_k K_t A_m \mu / R_{t,T}(w), \]

Then \( \Delta_{21} = o_p(1) \) and thus (A.16) is proved.

For (A.17),

\[ \bar{L}_{t,T}(w) - \bar{R}_{t,T}(w) = \epsilon' \bar{P}(w) K_t \bar{P}(w) \epsilon - 2 \mu' \bar{A}'(w) K_t \bar{P}(w) \epsilon - \text{tr}\left( \bar{P}(w)' K_t \bar{P}(w) \Omega \right) \equiv \Delta_{31} + \Delta_{32} + \Delta_{33}. \]

Firstly, for \( \Delta_{32} \),

\[ \frac{\mu' \bar{A}'(w) K_t \bar{P}(w) \epsilon}{\bar{R}_{t,T}(w)} \leq \left[ \frac{\epsilon' \bar{P}(w)' K_t \bar{P}(w) \epsilon (\mu' \bar{A}'(w) K_t \bar{A}(w) \mu)}{R^2_{t,T}(w)} \right]^{\frac{1}{2}} \]

\[ \leq \left[ \frac{\epsilon' \bar{P}(w)' K_t \bar{P}(w) \epsilon (\mu' \bar{A}'(w) K_t \bar{A}(w) \mu)}{R^2_{t,T}(w)} \right]^{\frac{1}{2}}. \]
Besides, we can obtain that
\[
\varepsilon' \bar{P}'(w)K_t \bar{Q}(w) \varepsilon \\
\leq \varepsilon' P'(w)K_t P(w) \varepsilon + \varepsilon' Q'(w)K_t Q(w) \varepsilon \\
+ \varepsilon' T'(w)K_t T(w) \varepsilon + 2\varepsilon' P(w)K_t T(w) \varepsilon \\
+ 2|\varepsilon' P(w)K_t Q(w) \varepsilon | + 2|\varepsilon' Q(w)K_t T(w) \varepsilon |. 
\] (A.24)

Terms on the right side of (A.24) follow that:
\[
2|\varepsilon' P(w)K_t T(w) \varepsilon | \leq |\varepsilon' P'(w)K_t P(w) \varepsilon \varepsilon' T'(w)K_t T(w) \varepsilon |^{1/2}, \\
2|\varepsilon' P(w)K_t Q(w) \varepsilon | \leq |\varepsilon' P'(w)K_t P(w) \varepsilon \varepsilon' Q'(w)K_t Q(w) \varepsilon |^{1/2}, \\
2|\varepsilon' Q(w)K_t T(w) \varepsilon | \leq |\varepsilon' T'(w)K_t T(w) \varepsilon \varepsilon' Q(w)K_t Q(w) \varepsilon |^{1/2}.
\]

Furthermore,
\[
\varepsilon' Q'(w)K_t Q(w) \varepsilon = \sum_{k=1}^{M_T} \sum_{m=1}^{M_T} w^k w^m \varepsilon' Q_k K_t Q_m \varepsilon \\
\leq \sum_{m=1}^{M_T} \sum_{m=1}^{M_T} w^k w^m \varepsilon' \varepsilon \tilde{c}(Q_k K_t Q_m) \leq k_{\text{max}} \tilde{h}^2 \varepsilon' \varepsilon.
\]

Similarly, \(\varepsilon' T'(w)K_t T(w) \varepsilon \leq k_{\text{max}} \tilde{h}^2 \varepsilon' \varepsilon\). Combining the proof of \(\Delta_{11}\) in (A.15), we obtain \(\Delta_{31}\) and \(\Delta_{32}\) are both \(o_p(1)\). The proof of \(\Delta_{33}\) is similar to that of (A.16) and is omitted here. Thus, (A.17) is proved.

For (A.18), with Assumption 6' and (A.15), we have
\[
\Pr \left\{ \sup_{w \in \mathcal{C}_T} \left| \frac{\mu' \tilde{A}(w)K_t \varepsilon}{R_{t,T}(w)} \right| > \delta \right\} \\
\leq \Pr \left\{ \sup_{w \in \mathcal{C}_T} \left| \mu' \tilde{A}(w)K_t \varepsilon \right| > \delta \tilde{\xi}_{t,T} \right\} \\
\leq \sum_{m=1}^{M_T} \Pr \left\{ \left| \mu' \tilde{A}'(w^0_m)K_t \varepsilon \right| > \delta \tilde{\xi}_{t,T} \right\} \\
\leq \delta^{-2} \tilde{\xi}_{t,T}^{-2} \sum_{m=1}^{M_T} \mathbb{E} \left( \mu' \tilde{A}'(w^0_m)K_t \varepsilon \right) \\
= \delta^{-2} \tilde{\xi}_{t,T}^{-2} \sum_{m=1}^{M_T} \text{tr} \left( \mu' \tilde{A}'(w^0_m)K_t \Omega K_t \tilde{A}'(w^0_m) \mu \right) \\
\leq k_{\text{max}} \delta^{-2} \tilde{\xi}_{t,T}^{-2} \tilde{\xi}(\Omega) \sum_{m=1}^{M_T} \tilde{R}_{t,T}(w^0_m) \to 0.
\]

45
Thus, Eq (A.18) is proved. And the proof of (A.19) is similar and is omitted here. With
(A.15)-(A.19), Theorem 2 is valid.

3 Appendix A.3

Following Vogt (2012), the process \( \{Y_t\} \) is locally stationary if for each rescaled time point \( \tau \in [0, 1] \) there exists an associated strictly stationary process \( Y_t(\tau) \) with \( \|Y_t - Y_t(\tau)\| = O_p(h + \frac{1}{T}) \). Thus, for every time \( t \), we can replace \( Y_t \), in the neighborhood of \( t \) (i.e., \( [t - Th, t + Th] \)), by a strictly stationary process \( Y_t(\tau) \) with a small cost \( O_p(h + \frac{1}{T}) \), where \( \tau = t/T \). We replace \( Y_t(\tau) \) with \( x_t \) to simplify the note. Denote \( \lambda_{\min}(A) \) and \( \lambda_{\max}(A) \) as

the minimum and maximum eigenvalues of matrix \( A \), respectively.

Before proving Theorem 3, we need to prove the following four Lemmas:

Lemma 1. Under Assumptions 16 and 18, for any \( q > 0 \) and all \( \theta > 0 \),

\[
E\lambda_{\min}^{-q}(\hat{R}_{t,T}(r_1)) = O(r_1^{2+\theta}q), \quad (A.25)
\]

where \( \hat{R}_{t,T}(m) = \frac{1}{(T-m)h}X^mK_1X^m = \frac{1}{T} \sum_{j=m}^{T-1} x_j^{(m)}k_{ij}x_j^{(m)}' \) in the \( m \)-th model, \( 1 \leq p_m \leq r_1 \).

Proof of Lemma 1. Define \( x_j = x_j^{(r_1)} = (x_j, \ldots, x_{j-r_1+1})' \), and

\[
\begin{pmatrix}
1 & a_1 & \cdots & a_{r_1-1} \\
0 & 1 & & \\
& \ddots & \ddots & \\
& & \ddots & a_1 \\
& & & 0 \ldots 0 1
\end{pmatrix}.
\]

Following the spirit of Lemma 1 in Ing and Wei (2003), we consider the following transformation of \( x_j \), \( \phi_j = Ak_j^{1/2}x_j = B_i x_i = \rho_j + \sigma_j \) for any fixed time point \( t \), where \( \rho_j = (\rho_{j1}, \ldots, \rho_{j+r_1-1})' \), and \( \sigma_j = (\sigma_{j1}, \ldots, \sigma_{j_{r_1}})' \) with \( \sigma_{ji}, 1 \leq i \leq r_1 \), being a linear combination of \( \rho_i, 1 \leq j \leq r_1 \). It is easy to obtain the following results:

(F1) \( \phi_j \) is independent of \( \{\sigma_{i_l}, \rho_{i_2}\} \) for \( l_1 \leq j \leq l_2 \leq j - r_1 \),

(F2) \( \lambda_{\min}^{-1}(\sum_{j=r_1}^{T-1} x_jk_{ij}x_j') \leq \lambda_{\max}(B_i'B_i)\lambda_{\min}^{-1}(\sum_{j=r_1}^{T-1} \phi_j\phi_j') \),

(F3) \( \lambda_{\max}(B_i'B_i) = O(1) \) for any fixed time point \( t \).

In view of (F2) and (F3), (A.25) follows from

\[
E\left((T - r_1)^q\lambda_{\min}^{-q}\left(\sum_{j=r_1}^{T-1} \phi_j\phi_j'\right)\right) \leq C\left(r_1^{2+\theta}\right)^q, \quad (A.26)
\]

which is the same as Eq (2.6) in Ing and Wei (2003). With (F1), the proof of (A.26) is similar to Eq (2.6) in Ing and Wei (2003). To save the space, we omit the proof of (A.26). ■
Denote $R_t(r_1) = E_t x_t'$ and $\|C\|^2 = \lambda_{\max}(C'C)$ as the maximum eigenvalue of the matrix $C'C$.

**Lemma 2.** Under Assumptions 18 and 19, and $\sup_{-\infty < t < \infty} E|\epsilon_t|^{2q} < \infty$ for some $q \geq 2$, we have

$$
E\|\tilde{R}_{t,T}(r_1) - R_t(r_1)\|^{q} \leq C \left( \frac{r_1^{q}}{(T-r_1)h} \right)^{q/2}, \tag{A.27}
$$

where $C$ is some positive constant.

**Proof of Lemma 2.** Note that

$$
E\|\tilde{R}_{t,T}(r_1) - R_t(r_1)\|^{q} \leq \frac{r_1^{q}}{r_1^{q}} \sum_{i=1}^{r_1} \sum_{j=1}^{r_1} E|\gamma_{i,j,t} - \gamma_{i,j,t}|^{q}, \tag{A.28}
$$

where $\gamma_{i,j,t}$ and $\gamma_{i,j,t}$ denote the $(i,j)$ components of $\tilde{R}_{t,T}(r_1)$ and $R_t(r_1)$, respectively. With Proposition 1 in Mathematical Appendix of Chen and Hong (2012), we obtain that

$$
E|\gamma_{i,j,t} - \gamma_{i,j,t}| = E\left| \sum_{s=1}^{T} x_{s1} k_{st} x_{sj} - \sum_{s=1}^{T} \sum_{s=1}^{T} E x_{st} k_{st} x_{sj} \right| = O((T-r_1)h^{-1/2}).
$$

Then $E\|\tilde{R}_{t,T}(r_1) - R_t(r_1)\|^{q} \leq C \left( \frac{r_1^{q}}{(T-r_1)h} \right)^{q/2}$. This proof is completed. \[\square\]

**Lemma 3.** Under Assumption 18 if $\sup_{-\infty < t < \infty} E|\epsilon_t|^q < \infty$ for $q \geq 2$, then for $1 \leq p_m \leq r_1$ with $r_1 \leq T - 1$,

$$
E \left\| \frac{1}{\sqrt{(T-r_1)h}} \sum_{j=r_1}^{T-1} k_{jt} x_j^{(m)} \epsilon_{j+1} \right\|^q \leq C(p_m)^{q/2}.
$$

**Proof of Lemma 3.** Following the spirit of Eq (3.8) in Ing and Wei (2003), it is shown that

$$
E \left\| \frac{1}{\sqrt{T-r_1}} \sum_{j=r_1}^{T-1} k_{jt} x_j^{(m)} \epsilon_{j+1} \right\|^q \leq p_m^{q/2} p_m^{-1} \sum_{l=0}^{k-1} E \left\{ ((T-r_1)h)^{-q/2} \left| \sum_{j=r_1}^{T-1} k_{jt} x_{j-l} \epsilon_{j+1} \right| \right\}.
$$

With Assumption 12 and the convexity of $x^{q/2}, x > 0$, there exists some constant $C$ satisfying

$$
E \left( \frac{1}{(T-r_1)h} \sum_{j=r_1}^{T-1} x_j^2 \right)^{q/2} \leq \frac{1}{T-r_1} \sum_{j=r_1}^{T-1} E|x_{j-l}|^q \leq C.
$$
Thus, this Lemma is proved. ■

**Lemma 4.** If Assumptions 13, 16 and 17 hold, \( r_1^{6+\delta} = O(T) \) for some \( \delta > 0 \), and \( \sup_{-\infty < t < \infty} E(|\xi_t|^2 q_1) < \infty \) for some \( q_1 \geq 2 \), then for any \( 0 < q < q_1 \), \( E \left\| \hat{R}_{t,T}^{-1}(r_1) \right\|^q \leq C \), and \( E \left\| \hat{R}_{t,T}^{-1}(r_1) - R_t^{-1}(r_1) \right\|^q \leq C \left( \frac{r_1^6}{|T - r_1|} \right)^{q/4} \).

**Proof of Lemma 4.** With Lemma 1, we obtain that
\[
\left\| \hat{R}_{t,T}^{-1}(r_1) - R_t^{-1}(r_1) \right\|^q \leq C \left( E \left| \hat{R}_{t,T}(r_1) - R_t(r_1) \right|^q \right)^{q/(2^+q_1)} \leq C \left( \frac{r_1^6}{|T - r_1|} \right)^{q/2},
\]
almost surely for large \( T \). Based on the Hölder’s inequality and Lemma 2, we have
\[
E \left\| \hat{R}_{t,T}^{-1}(r_1) - R_t^{-1}(r_1) \right\|^q \leq C (E \left\| \hat{R}_{t,T}(r_1) - R_t(r_1) \right\|^{q_1})^{q/(q_1+\theta)} \leq C \left( \frac{r_1^6}{|T - r_1|} \right)^{q/2},
\]
for sufficiently large \( T \). Set \( 2\theta \leq \delta \), and with the assumption that \( r_1^{6+\delta} = O(T) \), \( E|\hat{R}_{t,T}^{-1}(r_1)| \leq C \) is obtained. Moreover, since the Cauchy-Schwarz equality gives
\[
E \left\| \hat{R}_{t,T}^{-1}(r_1) - R_t^{-1}(r_1) \right\|^{q/2} \leq C (E \left\| \hat{R}_{t,T}^{-1}(r_1) \right\|^q)^{1/2} (E \left\| \hat{R}_{t,T}(r_1) - R_t(r_1) \right\|^q)^{1/2},
\]
Lemma 4 is proved. ■

**Proof of Theorem 3.** The proof of Theorem 3 is similar to the proof of Theorem 3.1 in Zhang et al. (2013). First, substitute \( V_{t,T}(w) \), \( \tilde{V}_{t,T}(w) \), \( \tilde{\xi}_{t,T}^* \), \( \tilde{\xi}_{t,T} \), \( \sigma^2 I_T \) and “in probability”, for \( R_{t,T}(w) \), \( \tilde{R}_{t,T}(w) \), \( \xi_{t,T} \), \( \tilde{\xi}_{t,T} \), \( \Omega \) and “a.s.”, respectively in Theorem 2 and its proof. To prove Theorem 3 we need to verify that
\[
\tilde{\xi}_{t,T}^{-1} \hat{\xi} \xi' \xi = o_p(1), \quad (A.29)
\]
and
\[
\tilde{\xi}_{t,T}^{-1} \xi' P K_t P \xi = o_p(1), \quad (A.30)
\]
and
\[
\sup_{w \in \mathcal{X}} \left| \frac{\mu' \tilde{A}(w) K_t \xi}{\tilde{V}_{t,T}(w)} \right| = o_p(1). \quad (A.31)
\]
Since \( \mu' K_t \xi \) is unrelated to \( w \), we can equally prove
\[
\sup_{w \in \mathcal{X}} \left| \frac{\mu' \tilde{P}(w) K_t \xi}{\tilde{V}_{t,T}(w)} \right| = o_p(1), \quad (A.32)
\]
instead of (A.31). According to (A.17), for proving (A.32), we only need to verify
\[ \xi^{-1}_{t,T} \sup_{w \in \mathcal{W}} \left| \mu' \tilde{P}(w) K_t \xi \right| = o_p(1). \] (A.33)

Considering that \( \{\xi_1, \ldots, \xi_T\} \) is i.i.d and \( \mathbb{E} \xi_1^4 < \infty \), we have that \( \xi' \xi = O_p(T) \). Combined with (A.14) and Assumption 14, (A.29) is proved.

Before the proof of (A.30), we verify some equations first. With Lemma 3 and Assumptions 16-17, we have
\[
T^{-1} \mathbb{E}(\xi' K_t Y_L Y_L' K_t \xi') = \mathbb{E} \left[ \xi' K_t \left( Y_L(\tau) + O_p(h + \frac{1}{T}) \right) \left( Y_L(\tau) + O_p(h + \frac{1}{T}) \right)' K_t \xi' \right] \\
= T^{-1} \mathbb{E} (\xi' K_t Y_L(\tau) Y_L'(\tau) K_t \xi') + O \left( h + \frac{1}{T} \right) \\
= O(r_1) + O \left( h + \frac{1}{T} \right) \\
= O(r_1). \] (A.34)

Thus by the Markov’s inequality,
\[ T^{-1} r_1^{-1} \xi' K_t Y_L Y_L' K_t \xi' = O_p(1). \] (A.35)

Also, by Assumption 15 we have
\[ T^{-1} \xi' K_t X^* X' K_t \xi = O_p(1). \] (A.36)

Thus we have
\[ T^{-1} \gamma^{-1} \xi' K_t X X' K_t \xi = O_p(1). \] (A.37)

With Lemma 4 and Assumptions 16-17 we have
\[ T \mathbb{E} (\zeta((Y_L' K_t Y_L)^{-1}) = T \mathbb{E} \left[ \zeta \left( \left( Y_L'(\tau) K_t Y_L(\tau) + O \left( h + \frac{1}{T} \right) \right)^{-1} \right) \right] = O(1). \] (A.38)

By Markov’s equality, we have
\[ T \zeta((Y_L' K_t Y_L)^{-1}) = O_p(1). \] (A.39)
Let \( J_t = (Y_t'K_tY_t)^{-1}Y_t'K_tX^*(X^*M_tX^*)^{-1/2} \). With [Rao, 1973], it is shown that

\[
(X'K_tX)^{-1} = \begin{pmatrix}
(Y_t'K_tY_t)^{-1} + J_tK_tJ_t' & -J_tK_t(X^*M_tX^*)^{-1/2} \\
-(X^*M_tX^*)^{-1/2}K_tJ_t & (X^*M_tX^*)^{-1}
\end{pmatrix}
\]

\[
= \begin{pmatrix}
(Y_t'K_tY_t)^{-1} 0 \\
0 0
\end{pmatrix} + 2 \begin{pmatrix}
J_tK_tJ_t' & 0 \\
0 & (X^*M_tX^*)^{-1}
\end{pmatrix}
\]

\[
- \begin{pmatrix}
J_tK_tJ_t' \\
(X^*M_tX^*)^{-1/2}K_tJ_t
\end{pmatrix}
\begin{pmatrix}
J_tK_t(X^*M_tX^*)^{-1/2} \\
(X^*M_tX^*)^{-1}
\end{pmatrix}.
\]  \( \text{(A.40)} \)

Thus

\[
\zeta \left( (X'K_tX)^{-1} \right) \leq \zeta \left( (Y_t'K_tY_t)^{-1} \right) + 2 \max \left\{ \zeta J_tK_t, \zeta \left( (X^*M_tX^*)^{-1} \right) \right\}
\]

\[
\leq \zeta \left( (Y_t'K_tY_t)^{-1} \right) + 2 \max \left\{ \zeta \left( (Y_t'K_tY_t)^{-1} \right) \zeta \left( T^{-1}(X^*K_tX^*) \right) \zeta (T^{-1}(X^*M_tX^*)^{-1}), T^{-1}\zeta \left( T^{-1}(X^*M_tX^*)^{-1} \right) \right\}.
\]  \( \text{(A.41)} \)

Combining \( \text{(A.39)} \) and \( \text{(A.41)} \), we have

\[
T\zeta \left( (X'K_tX)^{-1} \right) = O_p(1).
\]  \( \text{(A.42)} \)

Getting back to the proof of \( \text{(A.30)} \), we notice that \( P'K_tP = \{z_{ij}\}_{i,j=1}^{T} \), where

\[
z_{ij} = X_i \sum_{s=1}^{T} (X'K_sX)^{-1}X'_s k_{si}k_{sj}X_s (X'K_sX)^{-1}X'_j.
\]

Notice that for any \((i, j)\) such that \(|i - t| > 2Th\) or \(|j - t| > 2Th\), we obtain that \(z_{ij} = 0\). So let \( K^*_t = \{k^*_{ij}\} \) with

\[
k^*_{ij} = \begin{cases}
1, & i = j, \text{ and } |i - t| < 2Th \\
0, & \text{otherwise.}
\end{cases}
\]  \( \text{(A.43)} \)

Thus we have \( P'K_tP = K^*_tP'K_tPK_t^* \).

Since \( \frac{1}{T}X'K_sX \rightarrow R_s = EX'_sX_s \), we have

\[
z_{ij} \rightarrow \frac{1}{T^2h^2} X_i \left( \sum_{s=1}^{T} R_s^{-1}X'_s k_{si}k_{sj}X_s R_s^{-1} \right) X'_j = z^*_{ij}.
\]

Based on Assumption \[19\], it is easy to obtain that \( R_s = R_t + O(|t - s|/T) \) and \( k_{st} = 0 \) for
$|s - t| > Th$. Thus we have

$$z_{ij}^* = X_i R_t^{-1} \frac{1}{T^2 h^2} \sum_{s=1}^{T} X'_s k_{st} k_{sj} X_s R_t^{-1} X'_j + O \left( \frac{1}{T^2 h} \right)$$

$$\leq k_{max}^2 X_i R_t^{-1} \frac{1}{T^2 h^2} \sum_{s=1}^{T} X'_s k_{st} X_s R_t^{-1} X'_j$$

$$\rightarrow k_{max}^2 X_i R_t^{-1} \frac{1}{Th} R_t^{-1} X'_j$$

$$= \frac{k_{max}^2}{Th} X_i R_t^{-1} X'_j, \text{ if } i, j \in [t - Th, t + Th].$$

and then as $T \rightarrow \infty$,

$$\epsilon' P' K_t P \epsilon \leq \frac{k_{max}^2}{Th} \epsilon' X'R_t^{-1} X'K_t^* \epsilon.$$

For $R_t^{-1}$, since $\frac{1}{Th} X'K_t X \rightarrow R_t$, and combined with (A.41), we have that

$$\frac{1}{h} \zeta(R_t^{-1}) = O_p(1), \quad (A.44)$$

and for $\epsilon' K_t^* XX'K_t^* \epsilon$, notice that $K_t^*$ is a special case of $K_t$ with the bandwidth $h^* = 2h$ and kernel $k$ be uniform. Thus with (A.37) we have

$$\frac{1}{T\gamma} \epsilon' K_t^* XX'K_t^* \epsilon = O_p(1). \quad (A.45)$$

Therefore,

$$\frac{k_{max}^2}{Th} \epsilon' K_t^* XX'K_t^* \epsilon \xi^*_{t,T}$$

$$\leq \frac{k_{max}^2}{Th} \frac{\zeta(R_t^{-1})}{h} \frac{\epsilon' K_t^* XX'K_t^* \epsilon}{T\gamma} \epsilon \xi^*_{t,T}$$

$$\rightarrow 0.$$ 

This completes the proof of (A.30).

For (A.33), we obtain that

$$|\mu' \tilde{A}(w) K_t \epsilon| \leq |\mu' P(w) K_t \epsilon| + |\mu' Q(w) K_t \epsilon| + |\mu' T(w) K_t \epsilon|$$

$$= |\mu' P P(w) K_t \epsilon| + |\mu' Q(w) K_t \epsilon| + |\mu' T(w) K_t \epsilon|$$

$$\leq (\mu' P P(\epsilon' K_t P'(w) P(w) K_t \epsilon)^{1/2} + (\mu' \mu' \epsilon' K_t Q'(w) Q(w) K_t \epsilon)^{1/2}$$

$$+ (\mu' \epsilon' K_t T'(w) T(w) K_t \epsilon)^{1/2}$$
\[
\leq k_{\max} \left( (\mu'P\mu\epsilon'P'(w)P(w)\epsilon)^{1/2} + 2(\mu'\mu\tilde{h}\epsilon'\epsilon)^{1/2} \right)
\]
\[
= k_{\max} \xi^*_t \left( (\gamma\xi^*_t - 2\mu'P\mu)(\gamma^{-1}\epsilon'P\epsilon) \right)^{1/2}
+ 2 \left( (T^{-1}\mu'\mu)(T\gamma^{-1}\tilde{h})(\gamma\xi^*_t - 1)(\xi^*_t - 1\tilde{h}\epsilon'\epsilon) \right)^{1/2}. \tag{A.46}
\]

Since \(\gamma^{-1}\epsilon'P\epsilon = o_p(1)\), combining \(\text{(A.14)}\), \(\text{(A.29)}\) and \(\text{(A.46)}\), we obtain \(\text{(A.33)}\). This completes the proof. 

\[\square\]