Econometric Theory, **42**, 2026, 548–603.
doi:10.1017/S0266466625000064

TIME-VARYING COMPLETE SUBSET AVERAGING IN A DATA-RICH ENVIRONMENT

HAIQI LI
Hunan University

JING ZHANG
University of Jinan

XINGYI CHEN
Xiamen University

YONGMIAO HONG
Chinese Academy of Sciences and University of Chinese Academy of Sciences

This study proposes two novel time-varying model-averaging methods for time-varying parameter regression models. When the number of predictors is small, we propose a novel time-varying complete subset-averaging (TVCSA) procedure, where the optimal time-varying subset size is obtained by minimizing the local leave-h-out cross-validation criterion. The TVCSA method is asymptotically optimal for achieving the lowest possible local mean squared error. When the number of predictors is relatively large, we propose a factor TVCSA method to reduce the computational burden by first reducing the dimension of predictors by extracting a few factors using principal component analysis and then obtaining the TVCSA forecasts from time-varying models with the generated factors. We show that the TVCSA estimator remains asymptotically optimal in the presence of generated factors. Monte Carlo simulation studies have provided favorable evidence for the TVCSA methods relative to the popular model-averaging methods in the literature. Empirical applications to equity premiums and inflation forecasting highlight the practical merits of the proposed methods.

1. INTRODUCTION

The past few decades have witnessed an increase in the use of big data in economics and finance. For example, the popular FRED-MD and FRED-QD data are large

We thank Peter C. B. Phillips (Editor), Robert Taylor (Co-Editor), three referees for their comments and suggestions. Li thanks the support of the National Natural Science Foundation of China (NSFC) (Project No. 72171076). Zhang thanks the support of Shandong Provincial Natural Science Foundation of China for Young Scholars (Project No. ZR2024QG138) and the Higher Education Institution Youth Entrepreneurship Team Plan of Shandong Province (Project No. 2024KJB007). Hong thanks the support of the Basic Scientific Center Project entitled as "Econometric Modeling and Economic Policy Studies" of NSFC (Project No. 71988101). Address correspondence to Jing Zhang, Business School, University of Jinan, Jinan, China, e-mail: jingzhang00@hnu.edu.cn.

© The Author(s), 2025. Published by Cambridge University Press.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

publicly available macroeconomic databases designed for the empirical analysis of big data (McCracken and Ng, 2016, 2021). Many studies have employed tens or hundreds of economic variables from these databases to forecast economic variables of interest, such as inflation rates and oil prices (e.g., Costa et al., 2021; Medeiros et al., 2021). While big data provides more information, it poses a challenge in obtaining accurate estimates and forecasts due to model misspecification, model uncertainty, and parameter instability, which are commonly encountered, especially in a high-dimensional scenario with structural changes.

There are two main approaches to dealing with model uncertainty: model selection and averaging. In the model selection literature, a single optimal model is selected based on a suitable statistical criterion, such as the Akaike information criterion (AIC; Akaike, 1973), the Schwarz-Bayes information criterion (SIC/BIC; Schwarz, 1978), Mallows' C_p (Mallows, 1973), or the focused information criterion (FIC; Claeskens and Hjort, 2003). However, there is no consensus on the specific regression form of this model. It is rather difficult to select a single "best" forecasting model that is capable of yielding accurate predictions under various economic scenarios. As Leung and Barron (2006) pointed out, practitioners may encounter the risk of selecting a poor forecasting model. Thus, as a viable and robust alternative, model averaging combines the forecasts obtained from the set of candidate models under consideration by considering their weighted average (e.g., Hansen, 2008a; Hansen and Racine, 2012; Elliott, Gargano, and Timmermann, 2013; Zhang, Wan, and Zou, 2013; Lu and Su, 2015; Lee and Shin, 2023).[1] In particular, model averaging over a set of candidate models in high-dimensional cases has recently attracted considerable attention (e.g., Ando and Li, 2014; Zhang et al., 2020; Liao et al., 2021).

Parameter instability is another source of inaccurate forecasts, particularly for long-horizon forecasts (e.g., Stock and Watson, 1996; Rapach and Zhou, 2013). This may result from a time-varying relationship between economic variables due to technological progress (Hansen, 2001), business cycles (Boyd, Hu, and Jagannathan, 2005), crisis, and others. For example, Ang and Bekaert (2007) found that the predictive power of dividend yields for excess U.S. stock returns was not robust over different horizons. The sign of the coefficient on dividend yields appears negative at the beginning but becomes positive if the data from the late 1990s are included. Welch and Goyal (2008) investigated the in-sample and out-of-sample performances of linear regressions, and argued that the predictive relations of most models seemed unstable or even spurious. Chen and Hong (2012) showed strong evidence of the instability of return prediction models in post-war and post-oil shock sample periods. Yousuf and Ng (2021) noted that the parameter instability caused by the Great Moderation period reduced the forecasting ability of time-invariant methods.

[1] There are two types of model averaging in the literature: Bayesian model averaging (BMA) and frequentist model averaging (FMA). In this study, we consider the FMA. See Hoeting et al. (1999) for an excellent overview of the BMA. Steel (2020) reviewed both the BMA and FMA and their use in economics.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Although considerable effort has been devoted to addressing parameter instability and model uncertainty in a unified model framework, few attempts have been made in a data-rich environment with structural changes over time. Chan et al. (2012) considered state-space models and Bayesian model averaging, where the coefficients evolve according to a random walk process and the weights are estimated by a rolling window method. However, this assumption might be questionable because the coefficients change in an unbounded manner and might drift to arbitrarily high or low values, which is inconsistent with economic theories. Chen et al. (2018) proposed semiparametric ultra-high-dimensional model averaging for nonlinear time series regression models; however, they failed to consider the time-varying characteristics of the regression parameters. Yousuf and Ng (2021) and Sun et al. (2023) modeled high-dimensional time-varying parameters as smooth functions of the rescaled time, whose functional forms are unknown. This assumption may seem reasonable because structural changes usually require long-term evolution (e.g., Hansen, 2001; Chen and Hong, 2012). Chen and Maung (2023), Sun et al. (2023), and Chen, Hong, and Li (2024) further studied time-varying combination weights to address model uncertainty for high-dimensional data. These papers proposed penalization schemes, including boosting, least absolute shrinkage and selection operator (LASSO), and group smoothly clipped absolute deviation (SCAD). However, it is well known that the estimation error of the optimal weighting vector may be substantial when the number of candidate models is large. Moreover, choosing appropriate tuning parameters for the SCAD and LASSO procedures is crucial.

This study proposes a novel time-varying complete subset-averaging (TVCSA) method in a data-rich environment with smooth structural changes over time. The basic idea is that we first estimate all models composed of a small number (l) of regressors out of the total number (L) of available regressors and then average these time-varying regression estimates over all l dimensional subsets of the regressors. All candidate models used in TVCSA can potentially be misspecified, which is consistent with the fact that we often have little information about the true model in practice. When the number of predictors is relatively large, direct use of the TVCSA method may be computationally intensive and even infeasible, and the traditional estimation methods fail to consistently estimate a large number of parameters. Thus, we propose a two-step factor TVCSA method. First, we reduce the dimension of the predictors by extracting a few factors based on the principal component analysis (PCA). Second, we apply the TVCSA method with the estimated factors. Two local leave-h-out cross-validation (LLHCV) criteria are proposed to obtain the optimal time-varying subset sizes in the TVCSA and factor TVCSA methods, respectively. We show that the proposed TVCSA estimator with the estimated optimal time-varying subset size is asymptotically optimal as in Li (1987); that is, its local mean squared forecast error (MSFE) attains an infeasible lower bound in a class of model-averaging estimators constructed based on time-varying regression models. The TVCSA method typically incurs high computational costs because it requires the evaluation of all possible 2^L

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

models. Following Lee and Shin (2023), we propose employing only a portion of all models within a given subset. Forecast accuracy is insensitive to the model selection methods in the subsets, and the randomly drawn sampling approach is a natural choice because we use a simple arithmetic mean in the model-averaging stage (Elliott, Gargano, and Timmermann, 2015).

This study contributes to the literature on model averaging in several ways. First, we extend and generalize Elliott et al.'s (2013, 2015) complete subset regressions (CSR) and Lee and Shin's (2023) quantile CSR to time-varying CSR, where the time-varying parameters are modeled as unknown smooth functions of the rescaled time t/T. Elliott et al. (2013, 2015) first proposed combining least-squares forecasts based on all the short models within a complete subset indexed by l.[2] They found that an equal-weighting combination of subset regressions outperformed the LASSO or Bayesian model-averaging methods. Lee and Shin (2023) demonstrated the asymptotic optimality of the complete subset quantile regression estimator, following Li (1987). However, all these studies were developed under the assumption of independent and identically distributed (i.i.d.) errors and time-invariant parameters. To the best of our knowledge, no studies consider the CSR in a time-series context, which restricts its applications to macroeconomics and finance.

The proposed TVCSA method employs an equal-weighting scheme. As an important contribution, we demonstrate its superiority over the optimal weighting scheme in terms of efficiency loss when the number of candidate models is relatively large. Intuitively, an equal-weighting scheme avoids weight estimation because the estimation error of the stochastic optimal weights in a data-rich environment can be substantial. As Theorem 5 in Section 3 shows, the efficiency loss of the estimated weights diverges as the number M of candidate models increases. Moreover, the equal-weighting scheme does not require forecasters to determine the order of the initial models. Model sets are naturally determined by subset size l, which can be selected in a data-driven manner. This, to a certain degree, explains the well-known empirical "forecasting combination puzzle" observed by Stock and Watson (1999, 2003, 2004) in the sense that simply equally weighting model averaging outperforms those with combination weights estimated according to some optimality criteria. Similar to Claeskens et al. (2016) and Smith and Wallis (2009), we provide a theoretical explanation of the "forecasting combination puzzle" in a data-rich time-series environment with smooth structural changes.

Second, the subset size l, that is, the number of included regressors in an individual model, is allowed to change over time and increase to infinity as the sample size T increases so that one can fully utilize the information offered by big data. As Elliott et al. (2013) pointed out, the choice of l reflects a trade-off between the bias and variance of the forecasts. In practice, selecting an optimal time-varying subset

[2] In this study, a short model refers to a model with l regressors, where $l < L$ is a fixed and usually small number, and the full model refers to a model with the total L regressors.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

size is desirable to capture the unstable prediction relations. The cross-validation (CV) criterion is widely used to determine the optimal combination weights in the frequentist model-averaging literature (e.g., Hansen and Racine, 2012; Zhang et al., 2013; Cheng and Hansen, 2015; Gao et al., 2016; Liao et al., 2019). However, as Cheng and Hansen (2015) noted, conventional leave-one-out cross-validation (LOOCV) fails to account for serial correlation in time series. Thus, we propose a novel LLHCV method to obtain an optimal time-varying subset size.

Finally, we derive the uniform convergence rate and asymptotic normality for the proposed time-varying subset regression estimator using temporally dependent data. In particular, as a rule rather than an exception, we allow all candidate models to be misspecified. The proposed TVCSA and factor TVCSA estimators are asymptotically optimal in the sense that the TVCSA estimators with the estimated optimal time-varying subset sizes are asymptotically equivalent to the estimator with the infeasible optimal subset size, which extends the asymptotic optimality results of Sun et al. (2021), Lee and Shin (2023), and Sun et al. (2023). Moreover, while Elliott et al. (2013, 2015) and Lee and Shin (2023) did not consider generated regressors, we show that asymptotic optimality still holds in the presence of generated factors, indicating that the generated regressor issue can be ignored in the TVCSA procedure under certain regularity conditions. Our extension is nontrivial and possesses several technical challenges. In particular, to show the asymptotic optimality of the proposed TVCSA method, we require to address projection matrices that involve generated factors. Unlike the time-invariant model averaging, these matrices are highly complex and not necessarily idempotent and even not symmetric.

Extensive simulation results show that in most cases, the proposed TVCSA estimator achieves a lower risk than its competing peers, such as the popular model-averaging estimators, including Hansen and Racine's (2012) jackknife model-averaging (JMA) estimator, Sun et al.'s (2021) time-varying jackknife model-averaging (TVJMA) estimator, Elliott et al.'s (2013) CSR estimator, Buckland, Burnham, and Augustin's (1997) smoothed Akaike information criterion (SAIC), Cai and Tiwari's (2000) smoothed bias-corrected Akaike information criterion (SAICc), and a smoothed Bayesian information criterion (SBIC). Empirical applications to equity premiums and inflation forecasting show that the proposed TVCSA and factor TVCSA methods have superior performance in terms of the MSFE.

The remainder of this paper is organized as follows: Section 2 introduces the model setup and proposes the novel TVCSA method. Section 3 establishes the asymptotic theory of the time-varying complete subset regression estimator, particularly the asymptotic optimality of the TVCSA estimator. Moreover, we provide a theoretical justification for the equal-weighting scheme. Section 4 proposes a factor TVCSA method and derives its asymptotic optimality. Section 5 provides simulation evidence that supports the TVCSA methods. Section 6 presents two empirical applications to equity premiums and inflation forecasting. Finally, Section 7 concludes the paper. Mathematical proofs are collected in the Appendix,

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

and additional simulation and empirical results are provided in the Supplementary Material.

2. MODEL AND ESTIMATION

In this section, we first introduce a time-varying regression framework and the TVCSA estimator. Subsequently, we then describe the choice of time-varying subset size.

2.1. Time-Varying Complete Subset Model-Averaging Estimator

Let $\{y_t, \mathbf{z}_t, \mathbf{x}_t\}_{t=1}^{T}$ be a time-series random sample, where y_t is a scalar dependent variable, $\mathbf{z}_t$ and $\mathbf{x}_t$ are vectors of predictors. We consider the following data-generating process (DGP):

$$y_{t+h} = \mu_t + \varepsilon_{t+h} = \mathbf{z}_t^{\mathsf{T}}\delta_t + \mathbf{x}_t^{\mathsf{T}}\beta_t + \varepsilon_{t+h}, \quad t = 1, \ldots, T, \tag{1}$$

where h is the forecast horizon, ε_{t+h} is an unobservable innovation, δ_t and β_t are time-varying parameter vectors, and T is the sample size. We assume that $E(\varepsilon_{t+h}|\mathbf{z}_t, \mathbf{x}_t) = 0$, and $\mu_t = E(y_{t+h}|\mathbf{z}_t, \mathbf{x}_t)$ denotes the conditional mean of y_{t+h} given $(\mathbf{z}_t, \mathbf{x}_t)$. Suppose that $\mathbf{z}_t$ is a $Q \times 1$ vector of predetermined predictors, which are always included in the forecasting model based on economic theories or empirical evidence, such as the constant and lagged dependent variables. $\mathbf{x}_t$ is an $L_T \times 1$ vector of possibly relevant predictors, which may or may not be included in the model. If the coefficients δ_t and β_t are constant over the entire sample period, that is, $(\delta_t, \beta_t) = (\delta, \beta)$ for all t, then Model (1) reduces to the conventional time-invariant linear models as in Elliott et al. (2015) and Liu and Kuo (2016). Instead of specifying a restrictive parametric model, we let $\delta_t := \delta(t/T)$ and $\beta_t := \beta(t/T)$, where $\delta(\cdot)$ and $\beta(\cdot)$ are unknown smooth functions of the rescaled time t/T rather than original time t. This specification ensures an increasing amount of local information available at any given time t when $T \to \infty$, which is crucial to obtain a consistent nonparametric estimation (e.g., Robinson, 1989; Cai, 2007; Chen and Hong, 2012, 2016). Several approaches have been employed to model time-varying parameters in the literature: a deterministic function of time, a functional-coefficient of economic variables, a latent stationary stochastic process, and a latent nonstationary stochastic process, respectively. Each approach has its own advantages and disadvantages. It is rather challenging to determine which time-varying modeling strategy is most suitable for a particular study in practice. In the present study, our approach of assuming that parameters are deterministic functions of time can avoid the choice of economic variables and the "curse of dimensionality" problem associated with a high-dimensional set of explanatory variables.

Note that the dimension of $\mathbf{x}_t$ is allowed to diverge to infinity with the sample size T. This implies that the degree of model complexity may increase as T increases. An example is the investigation of stock-return predictability with an

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

increasing number of economic and financial predictors being incorporated into the forecasting models (e.g., Rapach, Strauss, and Zhou, 2010; Phan, Sharma, and Narayan, 2015). When $Q + L_T < T$, that is, the number of candidate predictors is less than the sample size, we can estimate δ_t and β_t using a local least-squares method and construct a point forecast of y_{T+h} as $\mathbf{z}_T^{\intercal}\hat{\delta}_T + \mathbf{x}_T^{\intercal}\hat{\beta}_T$. When $Q + L_T > T$, feasible estimation and prediction can be obtained after reducing the dimension of the potential regressors by suitable dimension-reducing methods (e.g., PCA), which we will explore in Section 4. In this study, we will address the model uncertainty issue caused by the potential predictors $\mathbf{x}_t$. For ease of exposition, we confine our attention to the case where $Q = 0$ and propose a time-varying model-averaging scheme. All results could be extended to the case where $Q \neq 0$, which involves fundamentally no new ideas but tedious notations.

Given that we are interested in the h-step-ahead forecast of y_t with L_T predictors, we first consider a set of time-varying parameter models composed of l regressors as follows:

$$y_{t+h} = \mathbf{x}_{t(m,l)}^{\intercal}\beta_{t(m,l)} + \epsilon_{t+h,(m,l)}, \tag{2}$$

where $m = \{1, \ldots, M\}$ is the model index, $l = \{1, \ldots, L_T\}$ is the number of included variables, $\mathbf{x}_{t(m,l)}$ is an $l \times 1$ vector of predictors, $\beta_{t(m,l)}$ is a time-varying parameter vector, and $\epsilon_{t+h,(m,l)} := \mu_t - \mathbf{x}_{t(m,l)}^{\intercal}\beta_{t(m,l)} + \varepsilon_{t+h}$ is a composite error term when we use $l < L_T$ predictors. In a data-rich environment, the set of potential explanatory variables may be large; as a result, we allow L_T to increase to infinity as the sample size T increases. Our asymptotic theory will impose a restriction on the rate at which L_T increases with T. It is worth noting that all candidate models may potentially be misspecified due to omitted variables. In practice, model misspecification is inevitable since any specific model is only an approximation of the DGP.

Various nonparametric methods can be used to estimate the time-varying parameters in Model (2), such as local constant and local polynomial estimations. In this study, we use the local constant estimation. It is well known that local constant estimation suffers from the well-known boundary effect problem, in which the asymptotic bias has a slower convergence rate in the boundary regions than that in the interior region, because there is no symmetric data coverage in the boundary regions. To address this issue, we use Hall and Wehrly's (1991) reflection method to obtain pseudo symmetric data in the boundary regions. Specifically, we obtain the "reflected" data $(y_{-t}, \mathbf{x}_{(-t)(m,l)})$ for $-\lfloor Tb \rfloor \leq t \leq -1$ in the left boundary region and $(y_{2T-t}, \mathbf{x}_{(2T-t)(m,l)})$ for $T+1 \leq t \leq T+\lfloor Tb \rfloor$ in the right boundary region, where $\lfloor Tb \rfloor$ denotes the integer part of Tb, and $b := b(T)$ is a bandwidth controlling the size of the local neighborhood. Moreover, from a forecasting perspective, we only use the one-sided information from historical observations to estimate $\beta_{t(m,l)}$ at any time t. We use the reflection method to obtain the synthesized data at the time t, which consists of the original data $(y_t, x_{t(m,l)})$ and the pseudo data $(y_t, x_{t(m,l)}) = (y_{2t-i}, \mathbf{x}_{(2t-i)(m,l)})$ for $t+1 \leq i \leq t+\lfloor Tb \rfloor$. As a result, the asymptotic bias of the local constant estimator with reflection in the right boundary region has

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

the same order of magnitude and proportionality constant as that in the interior region.

Suppose $\beta_{t(m,l)} = \beta_{(m,l)}(t/T)$, where $\beta_{(m,l)}(\cdot)$ is continuously differentiable on $[0,1]$. By Taylor expansion, for any s in the local neighborhood of the time t, we have

$$\beta_{s(m,l)} = \beta_{t(m,l)} + O((s-t)/T),\ s \in [t - Tb, t + Tb]. \tag{3}$$

A local constant estimator for $\beta_{t(m,l)}$ can then be obtained by minimizing the local weighted sum of the squared residuals as follows:

$$\hat{\beta}_{t(m,l)} = \underset{\beta_{(m,l)} \in \mathbb{R}^l}{\arg\min} \sum_{s=t-\lfloor Tb \rfloor}^{t+\lfloor Tb \rfloor} \left(y_{s+h} - \mathbf{x}_{s(m,l)}^{\intercal} \beta_{(m,l)} \right)^2 K_{st}, \tag{4}$$

where $K_{st} = \frac{1}{b} k\left(\frac{s-t}{Tb}\right)$ and $k(\cdot)$ is the prespecified kernel function. For notational simplicity, let $\mathbf{y} = (y_{1+h}, \ldots, y_{T+h})^{\intercal}$, $\mathbf{x}_{(m,l)} = (\mathbf{x}_{1(m,l)}^{\intercal}, \ldots, \mathbf{x}_{T(m,l)}^{\intercal})^{\intercal}$, and $\mathbf{K}_t = \text{diag}\{K_{1t}, \ldots, K_{Tt}\}$. The solution to the optimization problem (4) is then given by

$$\hat{\beta}_{t(m,l)} = \left(\mathbf{x}_{(m,l)}^{\intercal} \mathbf{K}_t \mathbf{x}_{(m,l)} \right)^{-1} \mathbf{x}_{(m,l)}^{\intercal} \mathbf{K}_t \mathbf{y}, \tag{5}$$

and the corresponding estimator for μ_t in the (m,l)th candidate model is $\hat{\mu}_{t(m,l)} = \mathbf{x}_{t(m,l)}^{\intercal} \hat{\beta}_{t(m,l)}$.

We now combine forecasts from a set of time-varying parameter models with the same number of predictors. To simplify the notations, we drop the subscript T from L_T. Given L predictors, the number of l-variate models is $M = L!/(l!\,(L-l)!)$. For example, when $L = 4$, there are $M_1 = 4$ univariate models, $M_2 = 6$ two-variate models, $M_3 = 4$ three-variate models, and a full model. A complete subset-averaging (CSA) estimator is obtained as a simple averaging of those forecasts within a given complete subset indexed by l:

$$\hat{\mu}_t(l) = \frac{1}{M} \sum_{m=1}^{M} \hat{\mu}_{t(m,l)}. \tag{6}$$

Compared with the existing model-averaging methods with sophisticated optimal weights (e.g., Hansen and Racine, 2012; Sun et al., 2021, 2023; Chen et al., 2024), the CSA method does not require nested or non-nested construction of candidate model sets artificially, nor does it require a prior ordering of the predictors. It may first appear *ad hoc* to assign equal weights to a set of candidate models. One might consider choosing a weight vector based on some risk criteria. However, the number of candidate models M in a complete subset can be very large even when L is of moderate size. When using the estimated optimal weights, the accumulation of weight estimation errors may offset the benefits from model averaging, which is confirmed by Theorem 5 in Section 3.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

2.2. Optimal Time-Varying Subset Size

The key parameter for the CSA method is the subset size parameter l, which controls the degree of model complexity and the number of candidate models in a complete subset. There are three popular methods for selecting the optimal subset size: The first seeks to minimize the sample counterpart of the approximate mean squared error (MSE) (e.g., Elliott et al., 2013); the second is Meligkotsidou et al.'s (2021) likelihood-based method, which is a recursive algorithm with a prespecified error distribution; and the third is Lee and Shin's (2023) LOOCV method, which is often used in the literature to select tuning parameters. To the best of our knowledge, there has been no work on selecting optimal time-varying subset sizes in the TVCSA. In this study, we select the optimal subset size at each time point, which can accommodate the time-varying predictive performance of the candidate models. Specifically, we propose a local LLHCV criterion, a locally weighted sum of the squared leave-h-out prediction residuals. For simplicity, we denote the selection matrix as follows:

$$\phi_t = \begin{cases} (\boldsymbol{I}_{t+h-1}, \mathbf{0}_{(t+h-1)\times(T-h-t+1)}), & 1 \le t \le h, \\ (\mathbf{0}_{(2h-1)\times(t-h)}, \boldsymbol{I}_{2h-1}, \mathbf{0}_{(2h-1)\times(T-h-t+1)}), & h+1 \le t \le T-h, \\ (\mathbf{0}_{(T+h-t)\times(t-h)}, \boldsymbol{I}_{T+h-t}), & T-h+1 \le t \le T, \end{cases} \tag{7}$$

and

$$\pi_t = \begin{cases} (\underbrace{0,\ldots,0}_{t-1}, 1, \underbrace{0,\ldots,0}_{h-1}), & 1 \le t \le h, \\ (\underbrace{0,\ldots,0}_{h-1}, 1, \underbrace{0,\ldots,0}_{h-1}), & h+1 \le t \le T-h, \\ (\underbrace{0,\ldots,0}_{h-1}, 1, \underbrace{0,\ldots,0}_{T-t}), & T-h+1 \le t \le T. \end{cases} \tag{8}$$

For the (m,l)th candidate model, we let $\mathbf{y}_{[t+h]} = \phi_t \mathbf{y}, \mathbf{x}_{[t](m,l)} = \phi_t \mathbf{x}_{(m,l)}$, and $\mu_{[t]} = \phi_t \mu$, where the subscript $[\cdot]$ denotes the operation to select the elements at the corresponding positions. Delete $\{\mathbf{y}_{[t+h]}, \mathbf{x}_{[t](m,l)}\}$ from $\{\mathbf{y}_{t+h}, \mathbf{x}_{t(m,l)}\}$ to obtain the remaining set, denoted as $\{\mathbf{y}_{[-(t+h)]}, \mathbf{x}_{[-t](m,l)}\}$, where the subscript $[-\cdot]$ denotes the operation to delete the elements at the corresponding positions. Define $\mathbf{K}_{[-t]} = \mathrm{diag}(K_{1t}, \ldots, K_{(t-h)t}, K_{(t+h)t}, \ldots, K_{Tt})$ as a $(T-2h+1)\times(T-2h+1)$ weighting matrix. It is straightforward to see that the leave-h-out estimator of $\beta_{t(m,l)}$ in the (m,l)th candidate model is given by

$$\tilde{\beta}_{t(m,l)} = \left(\mathbf{x}_{[-t](m,l)}^{\intercal} \mathbf{K}_{[-t]} \mathbf{x}_{[-t](m,l)}\right)^{-1} \mathbf{x}_{[-t](m,l)}^{\intercal} \mathbf{K}_{[-t]} \mathbf{y}_{[-(t+h)]}. \tag{9}$$

Then, the corresponding leave-h-out estimator of $\mu_{[t]}$ is $\tilde{\tilde{\mu}}_{[t](m,l)} = \mathbf{x}_{[t](m,l)}^{\intercal} \tilde{\beta}_{t(m,l)}$, and the estimator of μ_t at any time t is given by $\tilde{\mu}_{t(m,l)} = \pi_t \tilde{\tilde{\mu}}_{[t](m,l)}$.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Let $\tilde{\mu}(l) = (\tilde{\mu}_1(l), \ldots, \tilde{\mu}_T(l))^\intercal$, where $\tilde{\mu}_t(l) = \frac{1}{M}\sum_{m=1}^{M} \tilde{\mu}_{t(m,l)}$. The LLHCV criterion for the optimal time-varying subset selection can be written as follows:

$$CV_{t,T}(l) = (\mathbf{y} - \tilde{\mu}(l))^\intercal \mathbf{K}_t (\mathbf{y} - \tilde{\mu}(l)). \tag{10}$$

By minimizing (10) with respect to l, the optimal time-varying subset size is obtained as follows:

$$\hat{l}_t = \underset{1 \le l \le L}{\arg\min}\, CV_{t,T}(l). \tag{11}$$

The TVCSA estimator at any given time t is then given by $\hat{\mu}_t(\hat{l}_t)$. Note that the optimal time-varying subset size is associated with the smallest $CV_{t,T}(l)$ value over the local sample, excluding observations around the time of t. Thus, the subset size is constant in the neighborhood of any fixed time t, but changes globally over time.

3. ASYMPTOTIC THEORY

In this section, we derive the convergence rate and asymptotic normality of $\hat{\beta}_{t(m,l)}$. All candidate models used for averaging are potentially misspecified. Therefore, the pseudo-true parameter vector is defined as follows:

$$\beta^*_{t(m,l)} := \underset{\beta_{t(m,l)} \in \mathbb{R}^l}{\arg\min} E\left[\left(y_{t+h} - \mathbf{x}^\intercal_{t(m,l)}\beta_{t(m,l)}\right)^2\right]. \tag{12}$$

We show the uniform convergence rates of estimators $\hat{\beta}_{t(m,l)}$ and $\tilde{\beta}_{t(m,l)}$ to $\beta^*_{t(m,l)}$. Next, we establish the asymptotic optimality of the TVCSA estimator to minimize the local squared error loss and associated conditional risk.

Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and largest eigenvalues of Matrix A, respectively. Throughout the paper, C denotes a generic constant that may vary across different places. First, we impose the following regularity conditions:

Assumption A.1.

(i) $\{\mathbf{x}_{t(m,l)}, \epsilon_{t+h,(m,l)}\}$ is a strictly stationary β-mixing process with the mixing coefficient $\{\beta(j)\}$ satisfying $\sum_{j=1}^{\infty} j^4 \beta(j)^{\delta/(1+\delta)} < C$ for some $0 < \delta < 1$. In addition, $\sup_{i \ge 1} E\left(x_{it}^8\right) < C$, $E\left(\mu_t^4\right) < \infty$, and $E[\epsilon_{t+h,(m,l)}|\mathbf{x}_{t(m,l)}] = 0$.

(ii) The $l \times l$ matrix $A_{(m,l)} := E\left(\mathbf{x}_{t(m,l)}\mathbf{x}^\intercal_{t(m,l)}\right)$ is finite and positive definite. Moreover, there exist positive constants $\underline{C}_{x(m,l)}$ and $\overline{C}_{x(m,l)}$ such that $0 < \underline{C}_{x(m,l)} \le \lambda_{\min}\left(A_{(m,l)}\right) \le \lambda_{\max}\left(A_{(m,l)}\right) \le \overline{C}_{x(m,l)} < \infty$.

(iii) Let $\Gamma_{(m,l)}(j) := Cov(\mathbf{x}_{t(m,l)}\epsilon_{t+h,(m,l)}, \mathbf{x}_{t+j,(m,l)}\epsilon_{t+j+h,(m,l)})$. The $l \times l$ matrix $V_{(m,l)} := \sum_{j=-\infty}^{+\infty} \Gamma_{(m,l)}(j)$ is finite and positive definite. Moreover, there exist positive constants $\underline{C}_{v(m,l)}$ and $\overline{C}_{v(m,l)}$ such that $0 < \underline{C}_{v(m,l)} \le \lambda_{\min}\left(V_{(m,l)}\right) \le \lambda_{\max}\left(V_{(m,l)}\right) \le \overline{C}_{v(m,l)} < \infty$.

(iv) $(\overline{C}_{x(m,l)} + \overline{C}_{v(m,l)})/l = O(\underline{C}^2_{x(m,l)})$.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Assumption A.2.

(i) $\beta_t = \beta(t/T)$, where $\beta(\cdot)$ is twice continuously differentiable on $[0,1]$ except for a finite number of discontinuity points in the interval $[0,1]$. Moreover, $\sup_{t\geq 1}\sup_{|s-t|<Tb}\|\beta_t-\beta_s\|^2 = O(T^{-1}b^{-1})$.
(ii) The kernel function $k(\cdot):[-1,1]\to\mathbb{R}^+$ is a prespecified symmetric bounded probability density function. Further, $k(0)\geq k(u)$ for all $u\in[-1,1]$ and $\nu_0 = \int_{-1}^{1}k^2(u)du<\infty$.
(iii) The bandwidth $b = cT^{-\lambda}$ for $0<\lambda<1$ and $0<c<\infty$. Moreover, for some $\delta>0$, $Tb^{1+4/\delta}\to\infty$ as $T\to\infty$.

Assumption A.3. Let $\underline{C}_x := \min_{1\leq m\leq M}\min_{1\leq l\leq L}\underline{C}_{x(m,l)}$, $\overline{C}_x := \max_{1\leq m\leq M}\max_{1\leq l\leq L}\overline{C}_{x(m,l)}$, and $\overline{C}_v := \max_{1\leq m\leq M}\max_{1\leq l\leq L}\overline{C}_{v(m,l)}$.

(i) $L/\log(Tb) = O(1)$.
(ii) $T\log(Tb)^{L+1}(Tb)^{-0.5L\underline{C}_x^3/(2\overline{C}_x\overline{C}_v)} = o(1)$.

The β-mixing in Assumption A.1(i) is a weak dependence condition in the study of the asymptotic properties of nonparametric time series models, which restricts the strength of dependency among local data. The moment conditions are needed for the application of various inequalities such as Boole and Markov inequalities. Assumptions A.1(ii)–(iii) are similar to Lu and Su's (2015) Assumptions A.2(ii)–(iii). The covariance matrices $A_{(m,l)}$ and $V_{(m,l)}$ are positive definite with bounded eigenvalues, which depend on the dimension of the regressors in an individual model. They are standard in the case of the divergent dimension and model misspecification and can be verified through the eigenvalues of the specific matrices in the (m,l)th model. Note that these bounds $\{\underline{C}_{x(m,l)},\overline{C}_{x(m,l)},\underline{C}_{v(m,l)},\overline{C}_{v(m,l)}\}$ depend on the dimension l of the regressors in the (m,l)th model. It is possible that $\{\underline{C}_{x(m,l)},\underline{C}_{v(m,l)}\}$ converges to zero and $\{\overline{C}_{x(m,l)},\overline{C}_{v(m,l)}\}$ diverges to infinity as $T\to\infty$. Assumption A.1(iv) imposes rate restrictions on these bounds to ensure the $\sqrt{l/Tb}$-consistency of the parameter estimator.

Assumption A.2(i) imposes a weak restriction on the functions of interest (e.g., Chen and Hong, 2012). This condition allows for smooth structural changes and abrupt structural breaks with bounded break sizes. Typical examples include Robinson's (1989) time-varying parameter model and Lin and Teräsvirta's (1994) smooth transition regression model. In Monte Carlo simulation studies of Section 5, we consider several DGPs that satisfy this assumption, including smooth transition changes, periodic smooth changes, smooth structural changes, single structural break, and multiple structural breaks. Assumption A.2(ii) is a standard condition for the kernel function, which implies that $\int_{-1}^{1}k(u)du=1$, $\int_{-1}^{1}uk(u)du=0$, and $\int_{-1}^{1}u^2k(u)du<\infty$. Examples of $k(\cdot)$ include the uniform kernel $k(u)=I(|u|\leq 1)$ and the Epanechnikov kernel $k(u)=0.75(1-u^2)I(|u|\leq 1)$, where $I(\cdot)$ is an indicator function. It could be relaxed to allow kernels with unbounded support at the cost of more tedious proofs (e.g., Hansen, 2008b). Assumption A.2(iii)

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

is commonly imposed in the literature. If $\delta > 1$, then the asymptotic MSE-optimal bandwidth $b \propto T^{-1/5}$ satisfies Assumption A.2(iii), which minimizes the asymptotic mean integrated squared error of a kernel estimator. Assumption A.3(i) implies that the total number of predictors L is allowed to grow at rate $\log(Tb)$ as T increases, which is comparable with the conditions in the model-averaging literature with optimal weighting schemes. For example, Lu and Su (2015) require $\bar{k}^4(\log T)^4/T \to 0$, where $\bar{k}$ is the maximum number of parameters. Assumption A.3(i) is slightly stronger because the number of candidate models $M = L!/[l!\,(L-l)!\,]$ in a complete subset would grow at a factorial rate of L. Assumption A.3(ii) imposes restrictions on the uniform bounds $\underline{C}_x$, $\overline{C}_x$, and $\overline{C}_v$ to obtain the uniform convergence rates for $\hat{\beta}_{t(m,l)}$ and $\tilde{\beta}_{t(m,l)}$.

Theorem 1 establishes the convergence rate and asymptotic normality of the local constant estimator $\hat{\beta}_{t(m,l)}$ in the (m,l)th model when the model is possibly misspecified and the dimension of the regressors diverges to infinity as $T \to \infty$.

THEOREM 1. *Suppose Assumptions A.1–A.3 hold. Let $C_{(m,l)}$ denote an $\ell_{(m,l)} \times l$ matrix such that $\lim_{T\to\infty} C_{(m,l)}C_{(m,l)}^{\intercal} = C_0$, where $\ell_{(m,l)} \in [1,l]$ is a fixed integer, and C_0 is finite and positive definite. Let $\Omega_{(m,l)} = A_{(m,l)}^{-1}\Xi_{(m,l)}A_{(m,l)}^{-1}$, where $\Xi_{(m,l)} = 2v_0V_{(m,l)}$, and $V_{(m,l)}$ and v_0 are defined in Assumptions A.1(iii) and A.2(ii), respectively. Let $\beta''_{t(m,l)}$ be a second-order derivative of $\beta_{t(m,l)}$ with respect to t/T. Then, at any given time t,*

(i) $\left\|\hat{\beta}_{t(m,l)} - \beta^*_{t(m,l)}\right\| = O_p\left(\sqrt{\frac{l}{Tb}}\right)$;

(ii) $\sqrt{Tb}C_{(m,l)}\Omega_{(m,l)}^{-1/2}\left[\hat{\beta}_{t(m,l)} - \beta^*_{t(m,l)} - \frac{b^2}{2}\beta''_{t(m,l)}\mu_2\right] \Rightarrow N(0, C_0)$ *as* $T \to \infty$,

where $\mu_2 = \int_{-1}^{1} u^2k(u)du$.

Theorem 1(i) shows that $\hat{\beta}_{t(m,l)}$ is a consistent estimator of the pseudo-true parameter $\beta^*_{t(m,l)}$. The convergence rate is $\sqrt{l/(Tb)}$, which is determined by the diverging number of parameters l and the effective sample size Tb in the smoothed nonparametric estimation. The matrix $C_{(m,l)}$ in Theorem 1(ii) represents an arbitrary combination of estimators. The asymptotic bias of $\hat{\beta}_{t(m,l)}$ is proportional to b^2. A decrease in b would reduce bias but increase variance. The asymptotically optimal bandwidth is the one that achieves the best trade-off between squared bias and variance, which in practice can be obtained by the rule-of-thumb bandwidth selector or more complicated bandwidth selection methods, such as the CV and plug-in methods.

Since the TVCSA estimator is defined over a complete subset indexed by l, we need to prove the uniform convergence rates of $\hat{\beta}_{t(m,l)}$ and its leave-h-out counterpart $\tilde{\beta}_{t(m,l)}$ over those subsets.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

THEOREM 2. *Suppose Assumptions A.1–A.3 hold. Then,*

$$\text{(i)} \max_{1\le t\le T}\max_{1\le l\le L}\max_{1\le m\le M}\left\|\tilde{\beta}_{t(m,l)}-\beta^*_{t(m,l)}\right\| = O_p\left(\sqrt{\frac{L\log(Tb)}{Tb}}\right);$$

$$\text{(ii)} \max_{1\le t\le T}\max_{1\le l\le L}\max_{1\le m\le M}\left\|\hat{\beta}_{t(m,l)}-\beta^*_{t(m,l)}\right\| = O_p\left(\sqrt{\frac{L\log(Tb)}{Tb}}\right).$$

Theorem 2 indicates that $\hat{\beta}_{t(m,l)}$ and its leave-h-out counterpart $\tilde{\beta}_{t(m,l)}$ are asymptotically equivalent, both of which converge to $\hat{\beta}^*_{t(m,l)}$ uniformly at the rate $\sqrt{L\log(Tb)/Tb}$. In a nonparametric regression setting, the uniform convergence rate relies on the effective sample size in the interval $[t-Tb, t+Tb]$ and the number of predictors L.

In a data-rich environment, forecasting using all possible models in a complete subset of size l is typically too profligate. For example, in the case of $L = 15$ and $l = 6$, there are $M_6 = 5005$ candidate models. To ease the computational burden, one can run only a portion of all regressions. Specifically, let $\overline{M}$ be the maximum number of models to implement the proposed TVCSA method, which are randomly drawn without replacement from the model space $\{1,\dots,M\}$. Define $\tilde{\mu}_t(l) = \frac{1}{\overline{M}}\sum_{m=1}^{\overline{M}} x^{\mathsf{T}}_{t(m,l)}\hat{\beta}_{t(m,l)}$, $\mu^*_t(l) = \hat{\mu}_t(l)|_{\hat{\beta}_{t(m,l)}=\beta^*_{t(m,l)}}$, and $\vec{\mu}^*_t(l) = \lim_{M\to\infty}\mu^*_t(l) < \infty$. Theorem 3 establishes that the predictors $\tilde{\mu}_t(l)$ and $\hat{\mu}_t(l)$ are asymptotically equivalent under suitable regularity conditions.

THEOREM 3. *Suppose Assumptions A.1–A.3 hold. Let* $\overline{M}\to\infty$ *and* $(\log\overline{M})^2\log(TL)/\overline{M}\to 0$ *as* $T\to\infty$. *Moreover,* $\max_{1\le t\le T}\max_{1\le l\le L}\left|\mu^*_t(l)-\vec{\mu}^*_t(l)\right| = o_p(1)$, *and* $\max_{1\le t\le T}\max_{1\le m\le M}\max_{1\le l\le L}\|x_{t(m,l)}\| = O_p(1)$. *Then,*

$$\max_{1\le t\le T}\max_{1\le l\le L}\left|\tilde{\mu}_t(l)-\hat{\mu}_t(l)\right| = o_p(1).$$

Theorem 3 ensures that the approximation error vanishes under a moderate diverging condition on $\overline{M}$ when we replace M with $\overline{M}$. It requires $\overline{M}\to\infty$ at a faster rate than $\log(TL)\to\infty$ as $T\to\infty$. To see this, let $z_{t(m,l)} = x^{\mathsf{T}}_{t(m,l)}\beta^*_{t(m,l)} - E\left[x^{\mathsf{T}}_{t(m,l)}\beta^*_{t(m,l)}\right]$ and $\max_{1\le t\le T}\max_{1\le l\le L}|z_{t(m,l)}| < C_T$. Then, there exist constants C_o and V_o such that

$$\begin{aligned}P\left(\max_{1\le t\le T}\max_{1\le l\le L}\left|\frac{1}{\overline{M}}\sum_{m=1}^{\overline{M}} z_{t(m,l)}\right| > \delta\right) &\le TL\max_{1\le t\le T}\max_{1\le l\le L}P\left(\left|\frac{1}{\overline{M}}\sum_{m=1}^{\overline{M}} z_{t(m,l)}\right| > \delta\right)\\ &\le TL\exp\left\{-\frac{C_o\overline{M}^2\delta^2}{V_o^2\overline{M}+C_T^2+C_T\overline{M}\delta(\log\overline{M})^2}\right\}\\ &= o_p(1),\end{aligned}$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

if $(\log \overline{M})^2 \log(TL)/\overline{M} \to 0$ as $T \to \infty$, where the second inequality follows from the Bernstein inequality. The uniform boundedness assumption on $x_{t(m,l)}$ in the model is mild and can be easily verified in practice. In the literature, there is no theoretical guidance for selecting the specific value of $\overline{M}$. In practice, one can use a screen plot. The TVCSA's forecast performances are insensitive to the choice of $\overline{M}$ when it is moderately large. Hence, like Lee and Shin (2023), we set $\overline{M} = 100$ in the empirical section.

To establish the asymptotic optimality of the proposed TVCSA estimator, we consider the local MSE as follows:

$$\mathcal{L}_{t,T}(l) = \frac{1}{Tb} \left(\hat{\mu}(l) - \mu\right)^{\intercal} \mathbf{K}_t \left(\hat{\mu}(l) - \mu\right). \tag{13}$$

It can be shown that the expectation of the local MSE $\mathcal{L}_{t,T}(l)$ is asymptotically equivalent to the MSFE of y_{T+h} given $\mathbf{x}_T$ up to a constant. To see this, we let

$$\begin{aligned}
\mathcal{R}_{t,T}(l) &= E[y_{T+h} - \hat{y}_{T+h}(l)]^2 \\
&= E[(\mu_T - \hat{\mu}_T(l)) + \varepsilon_{T+h}]^2 \\
&\simeq E[(\mu_t - \hat{\mu}_t(l))^2 + \varepsilon_{t+h}^2] \\
&\simeq E[\mathcal{L}_{t,T}(l)] + E[\varepsilon_{t+h}^2],
\end{aligned}$$

where the approximation in the third line follows from Assumption A.1(i) and the final approximation is because $E[\mathcal{L}_{t,T}(l)] = E[\frac{1}{Tb}\sum_{t=1}^{T} k\left(\frac{s-t}{Tb}\right)(\hat{\mu}_t(l) - \mu_t)^2] \to \int_{-1}^{1} k(u)du E[(\hat{\mu}_t(l) - \mu_t)^2] = E[(\hat{\mu}_t(l) - \mu_t)^2]$. Therefore, $\mathcal{L}_{t,T}(l)$ can be viewed as an approximately unbiased estimator of the MSFE.

Define a $T \times T$ matrix

$$P_{(m,l)} = \begin{bmatrix}
\mathbf{x}_{1(m,l)}^{\intercal} \left(\mathbf{x}_{(m,l)}^{\intercal} \mathbf{K}_1 \mathbf{x}_{(m,l)}\right)^{-1} \mathbf{x}_{(m,l)}^{\intercal} \mathbf{K}_1 \\
\mathbf{x}_{2(m,l)}^{\intercal} \left(\mathbf{x}_{(m,l)}^{\intercal} \mathbf{K}_2 \mathbf{x}_{(m,l)}\right)^{-1} \mathbf{x}_{(m,l)}^{\intercal} \mathbf{K}_2 \\
\vdots \\
\mathbf{x}_{T(m,l)}^{\intercal} \left(\mathbf{x}_{(m,l)}^{\intercal} \mathbf{K}_T \mathbf{x}_{(m,l)}\right)^{-1} \mathbf{x}_{(m,l)}^{\intercal} \mathbf{K}_T
\end{bmatrix},$$

and $P(l) = \frac{1}{M}\sum_{m=1}^{M} P_{(m,l)}$. Then, the estimator $\hat{\mu}_t(l)$ defined in (6) can be rewritten as $\hat{\mu}_t(l) = \mathbf{e}_t P(l)\mathbf{y}$, where $\mathbf{e}_t$ is a $1 \times T$ vector with the tth element taking the value of unity and other elements zeros. Similarly, the leave-h-out estimator $\tilde{\mu}_t(l)$ can be rewritten as $\tilde{\mu}_t(l) = \mathbf{e}_t \tilde{P}(l)\mathbf{y}$, where $\tilde{P}(l) = \frac{1}{M}\sum_{m=1}^{M} \tilde{P}_{(m,l)}$ and $\tilde{P}_{(m,l)}$ has zeros near the diagonal because observations $\{\mathbf{y}_{[t+h]}, \mathbf{x}_{[t](m,l)}\}$ are removed from the sample. Put $\varepsilon = (\varepsilon_{1+h}, \ldots, \varepsilon_{T+h})^{\intercal}$ and $\mathbb{L} = \{1, \ldots, L_T\}$, and define $\Omega = E(\varepsilon\varepsilon^{\intercal}|\mathbf{x})$ and $\xi_{t,T} = Tb \inf_{l \in \mathbb{L}} \mathcal{R}_{t,T}(l)$. Define $\mathcal{R}_{t,T(m,l)} = E[y_{T+h} - \hat{y}_{T+h,(m,l)}]^2$, where $\hat{y}_{T+h,(m,l)} = \mathbf{x}_{T(m,l)}^{\intercal}\hat{\beta}_{T(m,l)}$. Let $\zeta(A)$ denote the maximum singular value of Matrix A, and $[A]_{tt}$ be the (t,t)th element of Matrix A. We now impose the following regularity conditions.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Assumption A.4.

(i) $\zeta(\Omega) \leq C < \infty$.
(ii) $\overline{\lim}_{T\to\infty} \max_{1\leq l\leq L} \max_{1\leq m\leq M} \zeta(P_{(m,l)}) < C$, and $\overline{\lim}_{T\to\infty} \max_{1\leq l\leq L} \max_{1\leq m\leq M} \zeta(\tilde{P}_{(m,l)}) < C$, almost surely (a.s.).

Assumption A.5. For any given time t,

(i) $\xi_{t,T}^{-1} \sup_{l\in\mathbb{L}} \mathrm{tr}\,(P^{\intercal}(l)P(l)) = o_p(1)$;
(ii) $\xi_{t,T}^{-2} \sum_{m=1}^{M} \mathcal{R}_{t,T(m,l)} \to 0$, a.s., as $T \to \infty$;
(iii) the local average of μ_t^2 satisfies $(Tb)^{-1}\mu^{\intercal}\mathbf{K}_t\mu = O(1)$, a.s., as $T \to \infty$;
(iv) $p_{tt}^* = \max_{1\leq l\leq L} \max_{1\leq m\leq M} \max_{1\leq t\leq T} [P_{(m,l)}]_{tt} = O(T^{-1}b^{-1})$ and $b^{-1}\xi_{t,T}^{-1} \to 0$ as $T \to \infty$.

Assumption A.4(i) is similar to Condition (12) in Zhang et al. (2013). It requires the largest singular value of the error covariance matrix Ω to be bounded as $T \to \infty$. The first part of Assumption A.4(ii) is rather mild because typical estimators satisfy the condition that the maximum singular values of the projection matrices are bounded (e.g., the OLS estimator and the kernel smoothing estimator), which is similar to Condition (A.3) in Hansen and Racine (2012). The second part of A.4(ii) imposes a similar condition on the leave-h-out estimators.

Assumption A.5(i) is an analog of Condition (22) in Zhang et al. (2013) and Assumption 10 in Sun et al. (2021), implying that bias dominates risk. Therefore, the conditional risk measure $\mathcal{R}_{t,T}(l)$ is equivalent to $\mathcal{L}_{t,T}(l)$ when data are strictly stationary and weakly temporally dependent. To ensure that Assumption A.5(i) holds, we restrict the number of regressors $L \to \infty$ at a slower rate than that of the infimum risk $\xi_{t,T} \to \infty$. As Hansen (2007) noted, the condition $\xi_{t,T} \to \infty$ implies that the candidate models cannot include the true model (i.e., all candidate models are misspecified) and $\xi_{t,T}$ diverges quickly enough, because $\xi_{t,T}$ does not increase with the sample size T if there exist correctly specified models. In practice, it is reasonable to assume that all candidate models are the approximations to the unknown true DGP. Similar to Kabaila (2002) and Wan, Zhang, and Zou (2010), we provide an example under which Assumption 5(i) holds. Specifically, let $y_{t+h} = \mu_t + \varepsilon_{t+h} = f(x_t) + \varepsilon_{t+h}$, where $f(x_t) = \sum_{l=1}^{\infty} \theta_l \cos((l-1)x_t)/l$ and $x_t = 2\pi(t-1)/T$. Here, the random errors ε_{t+h} are i.i.d. $N(0,1)$. Suppose $\theta_l = T^{-1/3}l^{-7/12}$, one can show that Assumption A.5(i) is satisfied if $L = o(\log T)$ similar to Wan et al. (2010). We note that when the number L of predictors is finite, Assumption 5(i) holds trivially. Especially, because the number M of candidate models and the number l of variables in the candidate models depend on L, we do not need to control their divergence rates.

Assumption A.5(ii) imposes a restriction on the growth rate of the infimum risk $\xi_{t,T}$ and maximum risk of an individual model. This assumption is similar to the convergence condition in Sun et al.'s (2021) Assumption 6′, Zhang et al.'s (2013) Condition 13 and Wan et al.'s (2010) Condition 8. Let $\eta_{t,T} = \max_{1\leq m\leq M} \mathcal{R}_{t,T(m,l)}$. This assumption implies that $M\eta_{t,T}\xi_{t,T}^{-2} \to 0$, that is, $\xi_{t,T}^2 \to \infty$ at a faster rate than $M\eta_{t,T} \to \infty$, as $T \to \infty$. In this study, the number of candidate models is

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

$M = L!/[l!(L-l)!] < 2^L$ in a complete subset, which is determined by the total number of regressors $L = O(\log Tb)$ from Assumption 3(i). Like Ando and Li (2014) and Sun et al. (2021), given $\xi_{t,T} \to \infty$ with the rate $T^{1-\delta}$ for $\delta < 1/2$, Assumption 5(ii) holds if $\eta_{t,T} = O(Tb)$. Assumption 5(ii) also implies that the risk of each individual model cannot explode too quickly. This is crucial for model-averaging methods with non-nested models. Like Wan et al. (2010), this condition can be ensured by removing the uninformative variables that have poor forecasting abilities at the outset prior to model averaging. In practice, when there are many candidate variables in a complete subset, one could apply variable screening methods to remove uninformative variables before model averaging (e.g., Yuan and Yang, 2005; Elliott et al., 2013).

We note that Assumptions A.5(i)–(ii) are stronger than those of Hansen (2007) and Hansen and Racine (2012). However, they only consider discrete weights and require that the candidate models be sequentially nested. The nested candidate models depend on the ordering of regressors; however, there is no clear justification for the ordering. We consider a general weight set and allow the candidate models to be non-nested.

Assumption A.5(iii) implies that the local average of μ_t^2 is bounded, which is the same as Assumption 11 of Sun et al. (2021). In Assumption A.5(iv), p_{tt}^* is defined based on index sets of $t \in \{1, \ldots, T\}$, $m \in \{1, \ldots, M\}$, and $l \in \{1, \ldots, L\}$. It requires the weight $p_{tt}^* \to 0$ for all candidate models, which excludes extremely unbalanced designs in which only one observation remains relevant asymptotically. This condition is typical of the application of CV in model selection and averaging literature (e.g., Li, 1987; Hansen and Racine, 2012; Zhang et al., 2013; Sun et al., 2021).

Theorem 4 establishes the asymptotic optimality of the proposed TVCSA estimator.

THEOREM 4. *Suppose Assumptions A.1–A.5 hold. Then, at any given time t,*

$$\frac{\mathcal{L}_{t,T}(\hat{l}_t)}{\inf_{l \in \mathbb{L}} \mathcal{L}_{t,T}(l)} \xrightarrow{p} 1, \text{ as } T \to \infty.$$

Theorem 4 implies that the TVCSA-selected subset size $\hat{l}_t$ is asymptotically optimal in the sense that it is asymptotically equivalent to the infeasible best subset size for minimizing $\mathcal{L}_{t,T}$. This finding implies that the TVCSA estimator is asymptotically optimal in a class of model-averaging estimators constructed from a complete subset of time-varying parameter regression models. This asymptotic optimality is nearly the same as that of Lee and Shin (2023); however, they restrict the subset size to a constant value over the entire sample period. Sun et al. (2021) established asymptotic optimality for optimal time-varying weights in an optimal weighting scenario. However, their results are conditional on the original set of estimators chosen by the researcher. Our asymptotic optimality is unconditional because the proposed TVCSA method considers all complete subsets once the total

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

number of predictors L is given. We would like to emphasize that our optimality criterion is the one for (out-of-sample) forecast accuracy performance as measured by the squared forecast error loss, not the criterion for weight estimation efficiency.

The proposed TVCSA uses an equal-weighting combination scheme to combine forecasts within a complete subset. We do not attempt to use the optimal weighting schemes due to the well-known stylized fact that a simple combination of point forecasts often outperforms sophisticated weighting combinations, which is known as the "forecast combination puzzle" (e.g., Stock and Watson, 1999, 2003, 2004). This is likely due to the increase in variance in the estimation of optimal weights, especially when the number of candidate models M is rather large (e.g., Claeskens et al., 2016).

As an important theoretical contribution, this study provides a formal theoretical explanation for the "forecast combination puzzle" in a nonparametric framework by comparing the efficiency loss of model-averaging estimators using estimated optimal weights and equal weights, respectively. Put $F_{t,T}(\mathbf{w}_t) := E[\mathcal{L}_{t,T}(\mathbf{w}_t)]$, where $\mathcal{L}_{t,T}(\mathbf{w}_t)$ is a locally weighted loss function. An example used by Sun et al. (2021) is the local quadratic loss $\mathcal{L}_{t,T}(\mathbf{w}_t) = (\hat{\mu}(\mathbf{w}_t) - \mu)^T K_t(\hat{\mu}(\mathbf{w}_t) - \mu)$. The infeasible optimal weight vector is defined as follows:

$$\mathbf{w}_t^* := \underset{\mathbf{w}_t \in \mathcal{H}^{M-1}}{\arg\min}\ F_{t,T}(\mathbf{w}_t),$$

where $\mathcal{H}^{M-1} = \{\mathbf{w}_t \in [0,1]^M : \sum_{m=1}^M w_t^m = 1\}$. The estimated optimal weight vector $\widehat{\mathbf{w}}_t$ is obtained by minimizing $\mathcal{L}_{t,T}(\mathbf{w}_t)$. Denote $\hat{\eta}_t := \widehat{\mathbf{w}}_t - \mathbf{w}_t^*$ as the estimation errors, and let $\overline{\mathbf{w}} := M^{-1}\mathbf{1}_M$ be an equal-weighting vector, where $\mathbf{1}_M$ represents an M-dimensional unit vector. We impose the following regularity conditions.

Assumption A.6. For any given time t,

(i) $F_{t,T}(\mathbf{w})$ is twice differentiable with respect to $\mathbf{w}$ in $\mathcal{H}^{M-1}$, and $\sup_{\mathbf{w}_t \in \mathcal{H}^{M-1}} \|\nabla_2 F_{t,T}(\mathbf{w}_t)\| \leq C < \infty$ uniformly in M;

(ii) $E(\hat{\eta}_t) = 0, \underline{\sigma}_\eta^2 = \min_{1\leq t\leq T} \min_{1\leq m\leq M} Var(\hat{\eta}_t^m) > 0$, and $\overline{\sigma}_\eta^2 = \max_{1\leq t\leq T} \max_{1\leq m\leq M} Var(\hat{\eta}_t^m) < \infty$, where $\hat{\eta}_t^m$ is the mth element of $\hat{\eta}_t$.

Assumption A.6(i) implies that the eigenvalues of the matrix $\nabla_2 F_{t,T}(\mathbf{w}_t)$ are absolutely summable, that is, $\sum_{m=1}^\infty |\lambda_m| < \infty$ since $\|\nabla_2 F_{t,T}(\mathbf{w}_t)\| = (\sum_{m=1}^M \lambda_m^2)^{1/2}$, where the λ_m's are eigenvalues of $\nabla_2 F_{t,T}(\mathbf{w}_t)$. This condition holds for conventional loss functions, such as the quadratic loss function (e.g., Sun et al., 2021). Assumption A.6(ii) imposes a condition on the weights to exclude the nonstochastic equal-weighting case. Theorem 5 establishes a general framework for comparing the efficiency loss between using equal weights and estimated optimal weights.

THEOREM 5. *Suppose Assumption A.6 holds. Let* $\underline{\lambda}_{\min} := \inf_t \liminf_M \inf_{\mathbf{w}_t} \lambda_{\min}\left(\nabla_2 F_{t,T}(\mathbf{w}_t)\right)$ *and* $\overline{\lambda}_{\max} := \sup_t \limsup_M \sup_{\mathbf{w}_t} \lambda_{\max}\left(\nabla_2 F_{t,T}(\mathbf{w}_t)\right)$. *Then,*

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

(i) $\max_{1\le t\le T} |F_{t,T}(\overline{\boldsymbol{w}}) - F_{t,T}(\boldsymbol{w}_t^*)| \le \frac{1}{2}\overline{\lambda}_{\max}(1 - M^{-1})$;

(ii) $\min_{1\le t\le T} E|F_{t,T}(\widehat{\boldsymbol{w}}_t) - F_{t,T}(\boldsymbol{w}_t^*)| \ge \frac{1}{2}\underline{\lambda}_{\min} M\underline{\sigma}_\eta^2$.

Theorem 5 shows that the difference between the expected loss using the equal weights and that using the infeasible optimal weights is uniformly smaller than the upper bound $\frac{1}{2}\overline{\lambda}_{\max}$ as M increases. The upper bound $\frac{1}{2}\overline{\lambda}_{\max}$ depends on the functional form of the loss function. For example, we take the local quadratic loss function in this study: $F_{t,T}(\mathbf{w}_t) = E[\mathbf{w}_t^\intercal(\hat{\mu} - \mu)K_t(\hat{\mu} - \mu)^\intercal \mathbf{w}_t]$. In contrast, the lower bound of the difference between the expected loss using the infeasible optimal weights and that using the estimated optimal weights may diverge to infinity if $M\underline{\sigma}_\eta^2 \to \infty$. To ensure the convergence of the difference, one should guarantee that the convergence rate of the variance $\underline{\sigma}_\eta^2$ is faster than the divergence rate of the number of candidate models M. This is difficult to satisfy in a data-rich environment because M is often rather large. Therefore, the model averaging with an equal-weighting scheme is expected to outperform the one with estimated optimal weights in situations where M is relatively large.

For the special case of the local quadratic loss function, we can strictly establish the optimality of the TVCSA estimator relative to the model-averaging estimator using a sophisticated weighting scheme, as stated in Corollary 1.

COROLLARY 1. *Let* $\Sigma_t := E(e(\boldsymbol{w}_t)\boldsymbol{K}_t e(\boldsymbol{w}_t)^\intercal)$ *such that* $\max_{1\le t\le T} \lambda_{\max}(\Sigma_t) < \infty$. *Then, at any given time t,*

$$|F_{t,T}(\overline{\boldsymbol{w}}) - F_{t,T}(\boldsymbol{w}_t^*)| \to 0 \text{ as } M \to \infty.$$

Corollary 1 shows that under the special case of the local quadratic loss function, the proposed TVCSA estimator that uses equal weights is asymptotically equivalent to the estimator using the infeasible optimal weights in terms of minimizing the expected loss as $M \to \infty$. Intuitively, the efficiency loss caused by using the estimated optimal weights $\hat{\mathbf{w}}_t$ diverges to infinity as $M \to \infty$ unless the variance σ_η^2 of the estimated weight converges to zero uniformly, which results in an equal-weighting scheme. Thus, Corollary 1 implies that there is essentially no theoretical gain from model averaging with the use of estimated optimal weights as the number of candidate models M increases to infinity.

4. EXTENSION TO HIGH-DIMENSIONAL PREDICTORS

It is well known that when the number of predictors is relatively large, nonparametric smoothing suffers from the so-called "curse of dimensionality" problem. The direct use of nonparametric methods may lead to poor estimation and prediction results because of collinearity or numerical instability problems (Chen et al., 2018). If the number of predictors exceeds the sample size, these problems become

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

worse because of the lack of model identifiability. Thus, dimension reduction methods are often used before model-averaging forecasting.

As summarized by Ng (2013), two classes of dimension reduction strategies are typically used in forecasting scenarios: a sparse class that assumes that the predictor set is sparse, including Tibshirani's (1996) LASSO and Fan and Li's (2001) SCAD; and a dense class that assumes that the predictors have a low rank representation, including Bai and Ng's (2002) PCA and Kelly and Pruitt's (2015) partial least squares. The sparsity assumption is quite restrictive in forecasting models where the number of potentially relevant predictors is large because regularization methods only select strong signals and ignore weak predictors.

In this study, we extract a few common factors to reduce the dimension of the predictor set prior to model averaging, which can summarize the information contained in a large set of predictors. Specifically, suppose x_{it} has a factor structure as follows:

$$x_{it} = \lambda_i^{\mathsf{T}} F_t + e_{it},\ i = 1, 2, \ldots, L, \tag{14}$$

where F_t is an $R \times 1$ vector of unobserved factors, λ_i is a vector of factor loadings associated with F_t, product $\lambda_i^{\mathsf{T}} F_t$ is the common component, and e_{it} is the idiosyncratic component with zero mean. Here, for simplicity, the number of factors R is assumed to be known. In practice, we can obtain a consistent estimate for R using information criteria (e.g., Bai and Ng, 2002).

Once we obtain the common factor F_t, we construct the predictive regression as follows:

$$y_{t+h} = F_t^{\mathsf{T}} \alpha_t + \varepsilon_{t+h},\ t = 1, 2, \ldots, T, \tag{15}$$

where $\alpha_t := \alpha(t/T)$ is the $R \times 1$ time-varying regression coefficient vector and is twice differentiable with bounded derivatives. Since F_t is unobservable, we consider a two-step procedure: First, we estimate F_t via (14) and denote it as $\widehat{F}_t$; second, we regress y_{t+h} on $\widehat{F}_t$ to obtain the coefficient estimate $\hat{\alpha}_t$, from which a forecast $\hat{\mu}_t = \widehat{F}_t^{\mathsf{T}} \hat{\alpha}_t$ can be formed.

Specifically, in the first step, we estimate the factors $\widehat{F}_t$ using the commonly used PCA approach. As shown by Bai and Ng (2002), the estimated factors are consistent under some mild conditions. Therefore, the factor approach could extract most of the useful information from the original data using only a few factors and achieve dimension reduction before the second step of model averaging. In the second step, we apply the proposed TVCSA method as in Section 2, with the only exception that the original potential predictors $\mathbf{x}_t$ are now replaced by the estimated factors $\widehat{F}_t$. Specifically, the (m, r)th candidate forecasting model can be formulated as follows:

$$y_{t+h} = F_{t(m,r)}^{\mathsf{T}} \alpha_{t(m,r)} + \epsilon_{t+h,(m,r)}, \tag{16}$$

where $F_{t(m,r)}$ is an $r \times 1$ vector of factors, and $\epsilon_{t+h,(m,r)}$ is an unobservable error with zero mean. Since $F_{t(m,r)}$ is latent, we replace it with an $r \times 1$ vector of "generated

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

factor" and consider the following feasible forecasting model:

$$y_{t+h} = \widehat{F}_{t(m,r)}^{\intercal}\alpha_{t(m,r)} + \epsilon_{t+h,(m,r)}. \tag{17}$$

We obtain a consistent estimator of $\alpha_{t(m,l)}$ by solving the local least-squares minimization problem as follows:

$$\hat{\alpha}_{t(m,r)} = \underset{\alpha_{(m,r)}\in\mathbb{R}^r}{\arg\min} \sum_{s=t-\lfloor Tb\rfloor}^{t+\lfloor Tb\rfloor} \left(y_{s+h} - \widehat{F}_{s(m,r)}^{\intercal}\alpha_{(m,r)}\right)^2 K_{st}. \tag{18}$$

The solution to the above minimization problem is given by

$$\hat{\alpha}_{t(m,r)} = \left(\widehat{F}_{(m,r)}^{\intercal}\mathbf{K}_t\widehat{F}_{(m,r)}\right)^{-1}\widehat{F}_{(m,r)}^{\intercal}\mathbf{K}_t\mathbf{y}, \tag{19}$$

where $\widehat{F}_{(m,r)} = (\widehat{F}_{1(m,r)},\ldots,\widehat{F}_{T(m,r)})^{\intercal}$. $\hat{\alpha}_{t(m,r)}$ is called a "two-step" estimator. Given r, the factor TVCSA estimator for the conditional mean μ_t is defined as follows:

$$\hat{\mu}_t(r) = \frac{1}{M_r}\sum_{m=1}^{M_r}\hat{\mu}_{t(m,r)}, \tag{20}$$

where $\hat{\mu}_{t(m,r)} = \widehat{F}_{t(m,r)}^{\intercal}\hat{\alpha}_{t(m,r)}$ and $M_r = R!/[r!\,(R-r)!]$. We then obtain the optimal time-varying subset size r_t by minimizing the following LLHCV criterion:

$$\hat{r}_t = \underset{1\le r\le R}{\arg\min}\ \mathrm{CV}_{t,T}(r) = (\mathbf{y}-\tilde{\mu}(r))^{\intercal}\mathbf{K}_t(\mathbf{y}-\tilde{\mu}(r)), \tag{21}$$

where $\tilde{\mu}(r) = (\tilde{\mu}_1(r),\ldots,\tilde{\mu}_T(r))^{\intercal}$ is the leave-h-out estimator for μ. The final factor TVCSA estimator of μ_t at any time t is given by $\hat{\mu}_t(\hat{r}_t)$.

To investigate the asymptotic properties of the proposed TVCSA method using the estimated factors, we impose the following regularity conditions.

Assumption A.7.

(i) $\{F_t\}$, $\{\lambda_i\}$, and $\{e_{it}\}$ are jointly independent. Serial dependence within each group is allowed.

(ii) $E\|F_t\|^4 \le C$, and $T^{-1}\sum_{t=1}^{T}F_tF_t^{\intercal} \xrightarrow{p} \Sigma_F$ as $T\to\infty$, where Σ_F is an $R\times R$ positive definite matrix .

(iii) λ_i is either deterministic such that $||\lambda_i|| < C$ or stochastic such that $E\|\lambda_i\|^4 \le C$. In both cases, $L^{-1}\sum_{i=1}^{L}\lambda_i\lambda_i^{\intercal} \xrightarrow{p} \Sigma_\lambda$ as $L\to\infty$, where Σ_λ is an $R\times R$ positive definite matrix.

(iv) The errors e_{it} satisfy that $E(e_{it}) = 0$ and $E|e_{it}|^8 \le C$. For every (i,j), $E|L^{-1/2}\sum_{i=1}^{L}[e_{is}e_{jt} - E(e_{is}e_{jt})]|^4 \le C$. Let $\sigma_{ij,st} := E(e_{is}e_{jt})$. $|\sigma_{ij,st}| \le \bar{\sigma}_{ij}$ for all (s,t), $|\sigma_{ij,st}| \le \tau_{st}$ for all (i,j), $L^{-1}\sum_{i=1}^{L}\sum_{j=1}^{L}\bar{\sigma}_{ij} \le C$, $T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}\tau_{st} \le C$, and $L^{-1}T^{-1}\sum_{i=1}^{L}\sum_{j=1}^{L}\sum_{s=1}^{T}\sum_{t=1}^{T}|\sigma_{ij,st}| \le C$.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Assumption A.8.

(i) $\frac{1}{Tb}\sum_{s=t-\lfloor Tb\rfloor}^{t+\lfloor Tb\rfloor} K_{st}F_{s(m,r)}F_{s(m,r)}^{\intercal} \xrightarrow{p} \Sigma_{F_t} > 0$ as $T \to \infty$.

(ii) $\frac{1}{\sqrt{Tb}}\sum_{s=t-\lfloor Tb\rfloor}^{t+\lfloor Tb\rfloor} K_{st}F_{s(m,r)}\epsilon_{s+h,(m,r)} \xrightarrow{d} N(0,\Sigma_{F_t,\epsilon})$ as $T \to \infty$, where $\Sigma_{F_t,\epsilon}$ is a positive definite matrix.

(iii) $\max_{1\le r\le R}\max_{1\le m\le M_r}\max_{1\le t\le T} E\|F_{t(m,r)}\| \le C$.

(iv) $T/(L^2b) \to 0$ and $Lb \to \infty$ as $T, L \to \infty$.

Assumption A.7 is standard in factor analysis, which is used to show that the estimated factors $\widehat{F}_t$ are consistent estimators for a rotation of the true factors F_t. Assumption A.7(i) is similar to but weaker than Assumption R(ii) in Cheng and Hansen (2015). Assumption A.7(ii) is a high-level regularity condition, which can be shown by imposing more primitive conditions. Assumption A.7(iii) ensures that there exist R factors that have nontrivial contributions to the variance of $\mathbf{x}_t$. Assumption A.7(iv) allows for weak temporal and cross-section dependence and heteroskedasticity in the idiosyncratic component. Assumptions A.8(i)–(ii) are local versions of Bai and Ng's (2006) Assumption E, which ensure that the time-varying parameters can be identified. Assumption A.8(ii) is a high-level regularity condition, which can be shown by imposing more primitive conditions. Assumption A.8(iii) requires the factors to be uniformly bounded. Assumption A.8(iv) imposes a restriction on the relative growth rates of T, L, and bandwidth b, which implies that the estimated factors do not generate a nontrivial asymptotic bias on the asymptotic distribution of the estimator in the second step.

First, we derive the convergence rate and limiting distribution for the local constant estimate $\hat{\alpha}_{t(m,r)}$ in the (m,r)th model. Let $\tilde{V}_{(m,r)}$ be the $r \times r$ diagonal matrix consisting of r eigenvalues selected from the largest R eigenvalues of $\mathbf{x}\mathbf{x}^{\intercal}/(TL)$, where $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$. Let $\Lambda_{(m,r)} = (\lambda_1, \ldots, \lambda_L)^{\intercal}$ be the factor-loading matrix and $\Phi = \tilde{V}_{(m,r)}^{-1}(\widehat{F}_{(m,r)}^{\intercal}\widehat{F}_{(m,r)}/T)(\Lambda_{(m,r)}^{\intercal}\Lambda_{(m,r)}/L)$ be the $r \times r$ rotation matrix.

THEOREM 6. *Suppose Assumptions A.1(i), A.2(ii)–(iii), A.7, and A.8 hold. Then, at any given time t,*

$$\sqrt{Tb}\left(\hat{\alpha}_{t(m,r)} - \Phi^{-1}\alpha_{t(m,r)} - b^2\varpi_{t(m,r)}\right) \xrightarrow{d} N\left(0, \Phi_0^{\intercal-1}\Sigma_{F_t}^{-1}\Sigma_{F_t,\epsilon}\Sigma_{F_t}^{-1}\Phi_0^{-1}\right) \text{ as } T \to \infty,$$

where $\varpi_{t(m,r)} = \mu_2\alpha''_{t(m,r)}E(\Phi_0F_{t(m,r)}F_{t(m,r)}^{\intercal}\Phi_0^{\intercal})^{-1}E(\Phi_0F_{t(m,r)}F_{t(m,r)}^{\intercal})$, $\mu_2 = \int_{-1}^{1}u^2k(u)du$, *and* $\Phi_0 = \operatorname*{plim}_{T\to\infty}\Phi$.

Theorem 6 implies that the local constant estimator $\hat{\alpha}_{t(m,r)}$ is $\sqrt{Tb}$ consistent, although it converges to a transformation of $\alpha_{t(m,r)}$ because the true factors F_t can only be identified up to an invertible rotation matrix. Using the estimated factors as regressors does not affect the consistency of the time-varying parameter estimates.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

We now study the asymptotic optimality property of the proposed TVCSA estimator in the presence of estimated factors. Denote $P_{(m,r)} = F_{(m,r)}(F_{(m,r)}^{\intercal}\mathbf{K}_t F_{(m,r)})^{-1}F_{(m,r)}^{\intercal}\mathbf{K}_t$, $\widehat{P}_{(m,r)} = P_{(m,r)}|_{F_{(m,r)}=\widehat{F}_{(m,r)}}$ and $\widetilde{P}_{(m,r)} = \widehat{F}_{[-t](m,r)}(\widehat{F}_{[-t](m,r)}^{\intercal}\mathbf{K}_{[-t]}\widehat{F}_{[-t](m,r)})^{-1}\widehat{F}_{[-t](m,r)}^{\intercal}\mathbf{K}_{[-t]}$. We impose the following regularity conditions.

Assumption A.9.

(i) For certain positive constants c_1 and c_2, $0 < c_1 \leq \lambda_{\min}(\frac{1}{Tb}F_{\Phi(m,r)}^{\intercal}\mathbf{K}_t F_{\Phi(m,r)}) \leq \lambda_{\max}(\frac{1}{Tb}F_{\Phi(m,r)}^{\intercal}\mathbf{K}_t F_{\Phi(m,r)}) \leq c_2 < \infty$ with probability approaching 1 as $T \to \infty$, where $F_{\Phi(m,r)} := \Phi F_{(m,r)}$ is a rotation of $F_{(m,r)}$.

(ii) For $1 \leq r \leq R$ and $1 \leq m \leq M_r$, $\overline{\lim_{T\to\infty}}\, \zeta(P_{(m,r)}) < \infty$, $\overline{\lim_{T\to\infty}}\, \zeta(\widetilde{P}_{(m,r)}) < \infty$, and $\overline{\lim_{T\to\infty}}\, \zeta(\widehat{P}_{(m,r)}) < \infty$.

(iii) $R\xi_{t,T}^{-1} = o_p(1)$ and $R\mu^{\intercal}\mu\xi_{t,T}^{-2} = o_p(1)$, where $\xi_{t,T} = Tb\inf_{1\leq r\leq R}E(\mathcal{L}_{t,T}(r))$.

(iv) For any given time t, $p_{tt}^* = \max_{1\leq r\leq R}\max_{1\leq m\leq M_r}\max_{1\leq t\leq T}[P_{(m,r)}]_{tt} = O_p(T^{-1}b^{-1})$.

Assumption A.9(i) is similar to Conditions (C.2)–(C.3) in Zhang et al. (2020). Assumption A.9(ii) is similar to Assumption A.4(ii) in Section 3. Assumption A.9(iii) requires $\mathrm{tr}(P^{\intercal}(r)P(r))\mu^{\intercal}\mu \to \infty$ at a slower rate than $\xi_{t,T}^2 \to \infty$. This condition imposes restrictions on the growth rate of the number of factors R. Assumption A.9(iv) is similar to Condition (C.9) of Sun et al. (2023), which is related to CV methods.

THEOREM 7. *Suppose Assumptions A.1(i), A.2(ii)–(iii), A.5(iii), and A.7–A.9 hold. Then, at any given time t,*

$$\frac{\mathcal{L}_{t,T}(\hat{r}_t)}{\inf_{1\leq r\leq R}\mathcal{L}_{t,T}(r)} \xrightarrow{p} 1 \text{ as } T \to \infty.$$

Theorem 7 establishes the asymptotic optimality of the proposed TVCSA method in the presence of estimated factors. Therefore, the generated factors do not affect the asymptotic optimality of the TVCSA method. This result extends Theorem 4 to high-dimensional cases where the number of regressors may exceed the sample size.

5. MONTE CARLO SIMULATIONS

In this section, we use Monte Carlo simulations to examine the finite-sample performance of the proposed TVCSA method. The DGPs take the following form:

$$y_{t+1} = \mu_t + \varepsilon_{t+1} = \sum_{i=1}^{1000}\beta_i F(\tau)x_{it} + \varepsilon_{t+1},\ t = 1,\ldots,T,$$

where $\tau = t/T$. Following Zhang et al. (2013), we set $\beta_i = c\sqrt{2\alpha}i^{-\alpha-1/2}$ with $c > 0$ and $\alpha = 0.5$. The parameter c is selected to render the population coefficient

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

of determination $R^2 = c^2/(1+c^2)$ to vary on a grid. We set $x_{1t} = 1$ to be the intercept, which is always included in each candidate model. We consider two typical settings for the remaining regressors: (i) $\{x_{it}, i \geq 2\}$ are generated from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$, where the diagonal elements of Σ are 1 and the off-diagonal elements are ρ. Thus, the regressors are positively correlated when $\rho > 0$, and mutually independent of each other when $\rho = 0$; (ii) $x_{it} = 0.5x_{i(t-1)} + v_{it}$, where $\{v_{it}, i \geq 2\} \sim N(\mathbf{0}, \Sigma)$. This design allows the regressors to have both temporal and cross-sectional dependencies. We consider two types of non-Gaussian error processes[3]: (i) Heavy-tailed process: $\varepsilon_{t+1} \sim i.i.d.t(3)$; (ii) GARCH(1,1) process: $\varepsilon_{t+1} = \sigma_{t+1}u_{t+1}$, $\sigma^2_{t+1} = 0.5 + 0.1\sigma^2_t + 0.8\varepsilon^2_t$, and $u_{t+1} \sim i.i.d.N(0,1)$. For a robustness check, we consider six DGPs with various time-varying coefficients $F(\tau)$ as follows:

DGP 1 (Smooth transition changes): $F(\tau) = \tau^3$.

DGP 2 (Periodic smooth changes): $F(\tau) = \sin(\pi\tau^2)$.

DGP 3 (Single structural break): $F(\tau) = 0.1 + I(\tau > 0.2)$.

DGP 4 (Smooth structural changes): $F(\tau) = 1.5 - 1.5\exp(-3(\tau - 0.2)^2)$.

DGP 5 (Multiple structural breaks): $F(\tau) = 0.1 + 0.5I(0.2 < \tau < 0.5) + I(\tau \geq 0.5)$.

The coefficient function $F(\tau)$ is designed to change smoothly as the sample size T increases in DGPs 1, 2, and 4, whereas $F(\tau)$ changes abruptly at certain points in DGPs 3 and 5. Following one referee's suggestion, we include the time-invariant DGP 6 to show the costs of complicating the model:

DGP 6 (No structural change): $F(\tau) = 0.5$.

To implement the kernel regression estimator, we use the Epanechnikov kernel, which is shown to be the optimal kernel function for minimizing the asymptotic MSE. The bandwidth b is determined by a rule-of-thumb bandwidth selector for the Epanechnikov kernel, that is, $b = 2.34sT^{-1/5}$ with $s = 1/\sqrt{12}$ (e.g., Chen and Hong, 2012). The scale factor $1/\sqrt{12}$ can be regarded as the standard deviation of the standard uniform distribution $U(0,1)$, which can be regarded as the limiting distribution of grid points t/T, $t = 1, \ldots, T$, as $T \to \infty$. We consider two sample sizes: $T = 50$ and 150. The number of candidate regressors is set to $L = 4\lfloor \ln(Tb) \rfloor$ where $\lfloor \cdot \rfloor$ denotes the integer part of a real number, resulting in $L = 8$ and 12 for $T = 50$ and 150, respectively. Moreover, computing forecasts based on all models is rather demanding, even when $L = 8$. Hence, we randomly draw $\overline{M}$ models without replacement from the model space if $M = L!/[l!(L-l)!] > \overline{M}$; otherwise, we use the complete subset. In this study, we set the maximum number of sub-models in each complete subset to $\overline{M} = 100$ (e.g., Lee and Shin, 2023).

[3]The results under Gaussian and conditionally homoscedastic errors are available upon request.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

We compare (1) the proposed TVCSA estimator with several popular model-averaging estimators in the literature, including (2) Sun et al.'s (2021) TVJMA estimator,[4] (3) Cai and Tiwari's (2000) SAICc, (4) Elliott et al.'s (2013) CSR estimator, (5) Hansen and Racine's (2012) JMA estimator, (6) Buckland et al.'s (1997) SAIC, and (7) a smoothed Bayesian information criterion (SBIC). Both the TVJMA and SAICc estimators are based on local constant estimators with time-varying combination weights, whereas the other estimators are based on conventional linear regressions with time-invariant weights. The TVJMA and JMA adopt optimal weights by minimizing the LOOCV criteria. The CSR method uses equal-weighting combinations. We consider two special cases: $l = 1$ and $l = 2$, where the former corresponds to the univariate forecast combinations of Rapach et al. (2010) and the latter is shown to be almost the best by Elliott et al. (2013). The SAICc, SAIC, and SBIC methods use weights that are inversely proportional to certain information criteria, namely $\text{AICc} = \ln \text{RSS} + (T + \text{tr}(S))/(T - (\text{tr}(S) + 2))$, $\text{AIC} = T\ln(\hat{\sigma}^2) + 2\text{tr}(S)$, and $\text{BIC} = T\ln(\hat{\sigma}^2) + \ln(T)\text{tr}(S)$, where $\text{RSS} = \sum_{t=1}^{T}(y_t - \hat{y}_t)^2$, $\hat{\sigma}^2 = \text{RSS}/T$, and $\text{tr}(S)$ denotes the number of parameters to be estimated in the model. They are referred to as "smoothed" averaging due to the exponential weighting form $w_m = \exp(-0.5\text{I}_\text{m})/\sum_{m=1}^{M}\exp(-0.5\text{I}_\text{m})$, where I_m is an information criterion for the mth model. Except for the CSA and the proposed TVCSA, the set of candidate models used for all other alternative weighting methods is generated in an encompassing manner; for example, $\{x_{1t}\}$, $\{x_{1t}, x_{2t}\}, \ldots, \{x_{1t}, x_{2t}, \ldots, x_{12t}\}$.

We vary $R^2 = \{0.1, 0.2, \ldots, 0.9\}$ and $\rho = \{0, 0.1, 0.2, \ldots, 0.9\}$ to investigate the performance of the various model-averaging estimators. To assess forecast accuracy, we follow Hansen and Racine (2012) and Sun et al. (2023) and compute the risk (expected squared error) defined as follows:

$$J^{-1}\sum_{j=1}^{J}||\hat{\mu}_T^{(j)}(w) - \mu_T^{(j)}||^2,$$

where $J = 1000$ is the number of Monte Carlo replications, $\hat{\mu}_T^{(j)}(w)$ is the predicted value by a specific model-averaging method in the jth replication, and $\mu_T^{(j)}$ is the true value of the conditional expectation $E(y_{T+1}|x_T)$.

Figures 1–4 report forecast results of various model-averaging estimators under DGP 1 when the sample size $T = 50$ and 150. The results are displayed over $R^2 = \{0.1, 0.2, \ldots, 0.9\}$ in Figures 1 and 2 and over $\rho = \{0, 0.1, 0.2, \ldots, 0.9\}$ in Figures 3 and 4.[5] The four panels in each group display the results under various regressor and error settings. Overall, the proposed TVCSA method outperforms its peers in terms of the quadratic risk under nearly all DGP designs. First,

[4]We would like to thank Youngki Shin and Yuying Sun for sharing their codes.

[5]We note that forecast results under DGPs 2–6 are similar to those under DGP 1. To save space, we only report the results under DGP 1 in the main text and put the results under DGPs 4–5 in the Supplementary Material. The results under DGPs 2, 3, and 6 are available upon request.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

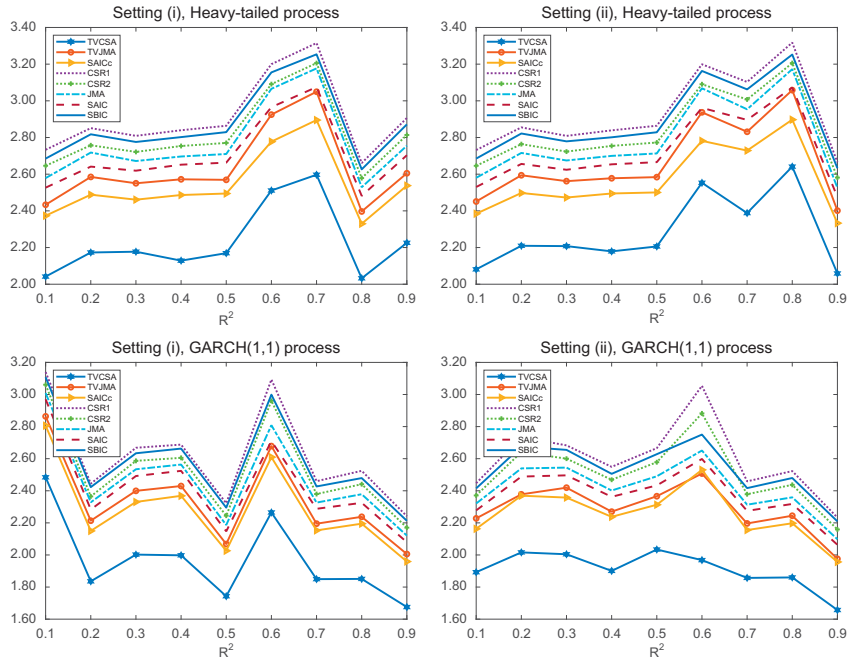


FIGURE 1. Forecast accuracy over various R^2 under DGP 1 (Smooth transition change) with $T = 50$. *Note*: (1) Setting (i): $\{x_{it}\}$ is generated from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$, where the diagonal elements of Σ are 1 and the off-diagonal elements are ρ; and (ii) $x_{it} = 0.5x_{i(t-1)} + v_{it}$, where $\{v_{it}, i \geq 2\} \sim N(\mathbf{0}, \Sigma)$. (2) Heavy-tailed process: $\varepsilon_{t+1} \sim i.i.d.t(3)$; GARCH(1,1) process: $\varepsilon_{t+1} = \sigma_{t+1}u_{t+1}$, $\sigma^2_{t+1} = 0.5 + 0.1\sigma^2_t + 0.8\varepsilon^2_t$, and $u_{t+1} \sim i.i.d.N(0, 1)$. (3) The y-axis shows the quadratic risk of various model-averaging methods, including the proposed TVCSA, Sun et al.'s (2021) TVJMA, Cai and Tiwari's (2000) SAICc, Elliott et al.'s (2013) CSR with $l = 1$ and 2, Hansen and Racine's (2012) JMA, and averaging estimators based on the AIC and BIC criteria. (4) The x-axis represents the population coefficient of determination R^2. We vary $R^2 = \{0.1, 0.2, \ldots, 0.9\}$ while fixing ρ to 0.5.

we note that the performance of CSR1 (i.e., $l = 1$) seems to be the worst. This method corresponds to a simple equal-weighting combination of all possible univariate models. CSR2 ($l = 2$) performs slightly better than CSR1. One possible explanation is that the combination of the two-variate models has a lower bias than that of the individual univariate models, which is consistent with the findings of Elliott et al. (2013). Second, the time-varying model-averaging methods (TVCSA, TVJMA, and SAICc) display better performance than the time-invariant model-averaging methods (CSR1, CSR2, JMA, SAIC, and SBIC) under DGPs 1–5, as is expected. This happens because these time-invariant methods ignore possible structural changes in parameters under these DGPs. Moreover, the performances of the TVJMA and SAICc estimators are indistinguishable in most cases, and a gain in forecast accuracy can be achieved from the TVCSA. Under

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

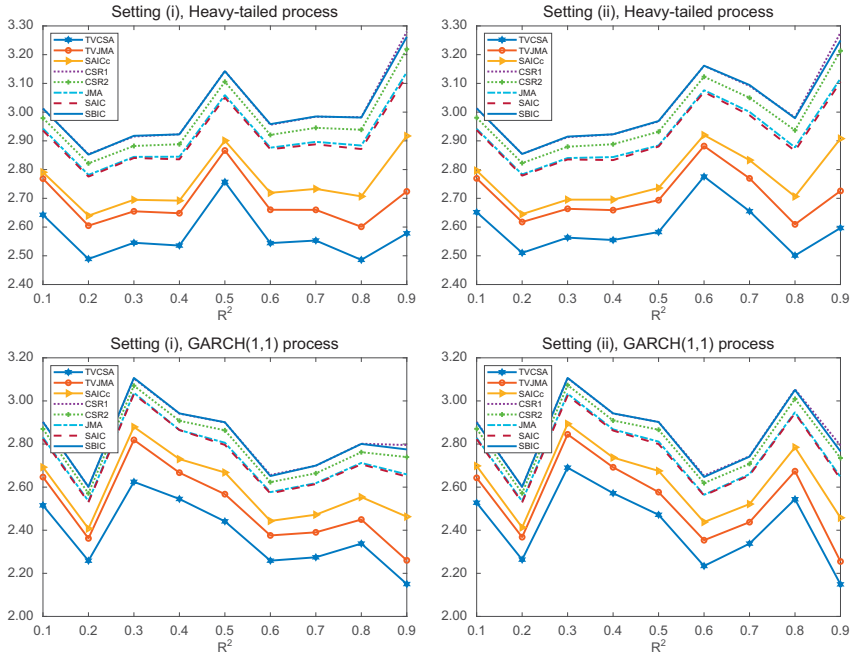


FIGURE 2. Forecast accuracy over various R^2 under DGP 1 (Smooth transition change) with $T = 150$. *Note*: (1) Setting (i): $\{x_{it}\}$ is generated from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$, where the diagonal elements of Σ are 1 and the off-diagonal elements are ρ; and (ii) $x_{it} = 0.5x_{i(t-1)} + v_{it}$, where $\{v_{it}, i \geq 2\} \sim N(\mathbf{0}, \Sigma)$. (2) Heavy-tailed process: $\varepsilon_{t+1} \sim i.i.d.t(3)$; GARCH(1,1) process: $\varepsilon_{t+1} = \sigma_{t+1}u_{t+1}$, $\sigma^2_{t+1} = 0.5 + 0.1\sigma^2_t + 0.8\varepsilon^2_t$, and $u_{t+1} \sim i.i.d.N(0,1)$. (3) The y-axis shows the quadratic risk of various model-averaging methods, including the proposed TVCSA, Sun et al.'s (2021) TVJMA, Cai and Tiwari's (2000) SAICc, Elliott et al.'s (2013) CSR with $l = 1$ and 2, Hansen and Racine's (2012) JMA, and averaging estimators based on the AIC and BIC criteria. (4) The x-axis represents the population coefficient of determination R^2. We vary $R^2 = \{0.1, 0.2, \ldots, 0.9\}$ while fixing ρ to 0.5.

DGP 6, the gap between the forecast performance of time-varying coefficient model-averaging methods and fixed coefficient model-averaging methods narrows slightly. However, the time-invariant model-averaging methods do not perform particularly well in this setup. This can be explained as follows: Although the smoothed nonparametric time-varying weighting model-averaging estimator is not as efficient as the fixed-weighting model-averaging estimator under DGP 6, both of them are consistent, and their out-of-sample MSEs converge to the same population MSE at the same convergence rate.

We also investigate the effects of various degrees of smoothing using bandwidths $\tilde{b} = cb$ with $c = 0.8$ and 1.2, where $b = \frac{2.34}{\sqrt{12}}T^{-1/5}$.[6] It is well known that the

[6]The results using bandwidths $\tilde{b} = 0.8b$ and $1.2b$ are available upon request.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

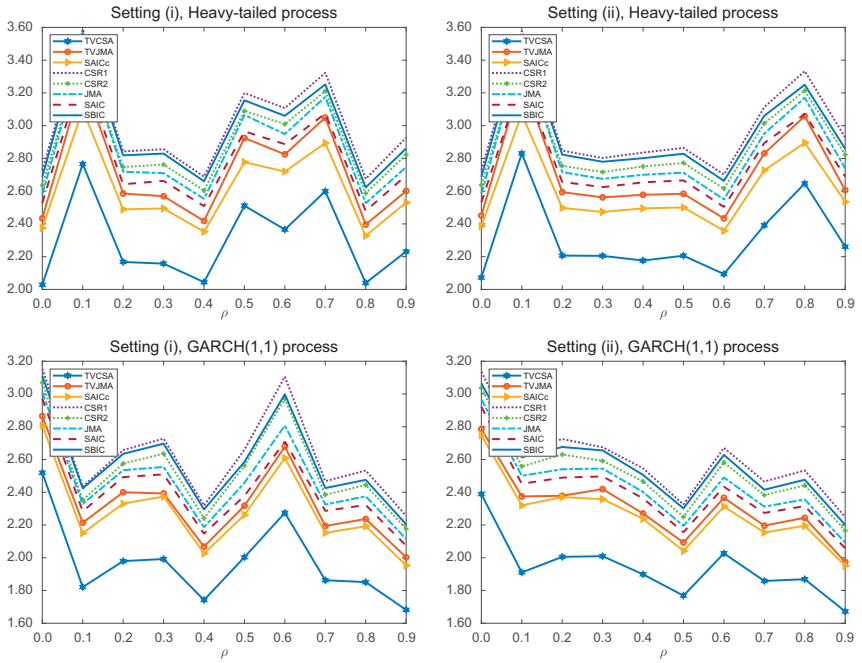


FIGURE 3. Forecast accuracy over various ρ under DGP 1 (Smooth transition change) with $T = 50$. *Note*: (1) Setting (i): $\{x_{it}\}$ is generated from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$, where the diagonal elements of Σ are 1 and the off-diagonal elements are ρ; and (ii) $x_{it} = 0.5x_{i(t-1)} + v_{it}$, where $\{v_{it}, i \geq 2\} \sim N(\mathbf{0}, \Sigma)$. (2) Heavy-tailed process: $\varepsilon_{t+1} \sim i.i.d.t(3)$; GARCH(1,1) process: $\varepsilon_{t+1} = \sigma_{t+1}u_{t+1}$, $\sigma^2_{t+1} = 0.5 + 0.1\sigma^2_t + 0.8\varepsilon^2_t$, and $u_{t+1} \sim i.i.d.N(0, 1)$. (3) The y-axis shows the quadratic risk of various model-averaging methods, including the proposed TVCSA, Sun et al.'s (2021) TVJMA, Cai and Tiwari's (2000) SAICc, Elliott et al.'s (2013) CSR with $l = 1$ and 2, Hansen and Racine's (2012) JMA, and averaging estimators based on the AIC and BIC criteria. (4) The x-axis controls the correlation of regressors. We vary $\rho = \{0.0, 0.1, 0.2, \ldots, 0.9\}$ while fixing R^2 to 0.5.

choice of bandwidth reflects a trade-off between estimation bias and variance: a smaller (larger) bandwidth would reduce (increase) the bias but increase (reduce) the variance. We note that using various bandwidths mildly impacts the forecast accuracy; however, TVCSA still yields the best performance. Following one referee's suggestion, the regressors could be persistent in economic and financial applications. Therefore, we provide a Monte Carlo simulation study under Setting (iii), in which the regressors are generated by the random walk processes, that is, $x_{it} = x_{i(t-1)} + v_{it}$. Figures S9 and S10 in the Supplementary Material report the simulation results under DGPs 1–6 with Setting (iii). Evidently, the TVCSA outperforms the competing methods.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

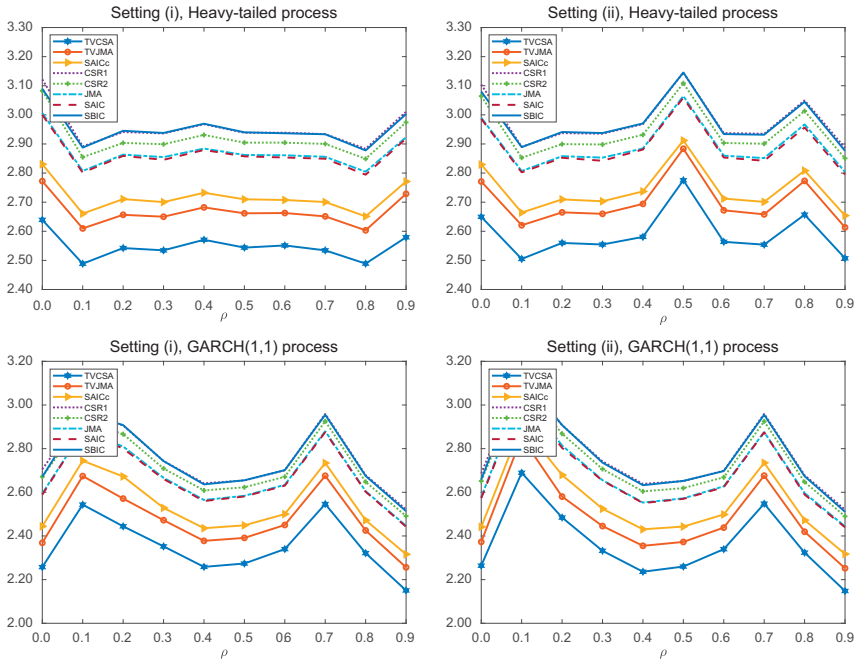


FIGURE 4. Forecast accuracy over various ρ under DGP 1 (Smooth transition change) with $T = 150$. *Note*: (1) Setting (i): $\{x_{it}\}$ is generated from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$, where the diagonal elements of Σ are 1 and the off-diagonal elements are ρ; and (ii) $x_{it} = 0.5x_{i(t-1)} + v_{it}$, where $\{v_{it}, i \geq 2\} \sim N(\mathbf{0}, \Sigma)$. (2) Heavy-tailed process: $\varepsilon_{t+1} \sim i.i.d.t(3)$; GARCH(1,1) process: $\varepsilon_{t+1} = \sigma_{t+1}u_{t+1}$, $\sigma^2_{t+1} = 0.5 + 0.1\sigma^2_t + 0.8\varepsilon^2_t$, and $u_{t+1} \sim i.i.d.N(0,1)$. (3) The y-axis shows the quadratic risk of various model-averaging methods, including the proposed TVCSA, Sun et al.'s (2021) TVJMA, Cai and Tiwari's (2000) SAICc, Elliott et al.'s (2013) CSR with $l = 1$ and 2, Hansen and Racine's (2012) JMA, and averaging estimators based on the AIC and BIC criteria. (4) The x-axis controls the correlation of regressors. We vary $\rho = \{0.0, 0.1, 0.2, \ldots, 0.9\}$ while fixing R^2 to 0.5.

In summary, the proposed TVCSA method performs quite well compared to competing methods for a range of DGPs, population R^2, and the correlation of regressors ρ, implying great application prospects in economics and finance.

6. EMPIRICAL APPLICATIONS

In this section, we use the proposed TVCSA method to predict equity premiums and inflation rates, which are among the most prominent variables in empirical studies in macroeconomics and finance. The two variables exhibit different time-series characteristics. Specifically, the equity premium series appears stationary except for a few outliers, and the inflation rate series is sensitive to the effects of policy changes.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

6.1. Equity Premium Forecasting

This subsection provides an empirical application to equity premium forecasting. A vast body of literature exists on this topic and many types of predictors have been considered, including macroeconomic (Chava, Gallmeyer, and Park, 2015), sentiment (Hirshleifer, Hou, and Teoh, 2009), and commodities (Driesprong, Jacobsen, and Maat, 2008). However, no consensus has been reached on whether they can improve the accuracy of stock forecasts. Welch and Goyal (2008) showed that prominent predictors from earlier academic literature failed to outperform the historical average in terms of both in-sample and out-of-sample predictability. Their study was conservative because they ignored prior information and assumed that the underlying model was well-specified. Campbell and Thompson (2008) argued that most of these variables can yield better out-of-sample forecasting performance using restricted regressions. Recent studies have attempted to combine forecasts from possible predictive models to address model uncertainty and parameter instability, indicating that forecasting combinations generate better out-of-sample forecasts than the historical average (e.g., Rapach et al., 2010; Meligkotsidou et al., 2014; Sun et al., 2021).

We compare the performance of U.S. equity premium forecasts using the proposed TVCSA method against competing alternatives, including Sun et al.'s (2021) TVJMA, Cai and Tiwari's (2000) SAICc, Elliott et al.'s (2013) CSR, Hansen and Racine's (2012) JMA, and Buckland et al.'s (1997) SAIC, and SBIC. We also consider the proposed TVCSA with a fixed number of regressors, including the simple averaging of time-varying univariate regressions (SA) and a full model with time-varying parameters (FULL). The predictive regression model is formulated as follows:

$$y_{t+1} = \alpha_t + x_t^{\mathsf{T}} \beta_t + \varepsilon_{t+1},\ t = 1, 2, \ldots, T_1, \tag{22}$$

where y_{t+1} is the equity premium calculated as the difference between the continuously compounded S&P 500 returns (including dividends) and the 3-month Treasury-Bill rate. x_t is a set of predictors related to stock characteristics, interest rates, and macroeconomic indicators. We consider 13 financial and economic variables frequently explored in the literature: log dividend price ratio (dp), log dividend yield (dy), log earnings price ratio (ep), book-to-market ratio (bm), stock variance (svar), net equity expansion (ntis), treasury-bill rate (tbl), long-term return (ltr), term spread (tms), default yield spread (dfy), default return spread (dfr), inflation (infl), and investment capital ratio (ik). Model (22) may be unbalanced because the regressand y_{t+1} is more likely to be a random noise, whereas the regressors may be rather persistent. Unbalancing issues often arise in the context of predictive regression and can be attenuated by considering a nonlinear transformation of the regressors (Phillips, 2015). Time-varying parameter regression models can be regarded as special nonlinear transformations if we let the nonlinear transformation be $f_t(x_t) = x_t^{\mathsf{T}} \beta_t$. Here, the regressors involving long memory or near-unit root behavior violate Assumption A.1. Following one referee's suggestion, we conduct

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

a simulation experiment to evaluate the performance of the model-averaging estimators in this setting and find that the proposed TVCSA estimator performs satisfactorily (see Section 5 for a detailed discussion). It remains a challenge to examine model averaging with various degrees of persistence in the predictors, which is beyond the scope of this study and left for future research.

Quarterly data are obtained from Welch and Goyal (2008)[7] and cover the period from 1947Q1 to 2022Q4 with $T = 304$. We use a sub-sample size T_1 to calculate a one-period-ahead forecast of the equity premium, and the remaining $T - T_1$ observations to evaluate the prediction accuracy recursively. The estimation sample sizes are varied over the $T_1 = \{56, 68, 80, 92, 116, 140\}$ period for robustness check. For time-varying model-averaging methods, the model parameters are estimated using the local constant method with reflection, where the kernel is set to the Epanechnikov kernel and the bandwidth $h = 2.34/\sqrt{12}T_1^{-1/5}$. Note that the performance of model-averaging methods using unequal weights depends on the order of candidate regressors. Thus, we order these predictors according to their absolute values of correlation with the dependent variable and construct 13 nested candidate models: $\{1, x_1\}, \{1, x_1, x_2\}, \ldots, \{1, x_1, \ldots, x_{13}\}$. However, such an ordering is not required for the TVCSA method. We then compare the forecasting performances of the various methods using the MSFE measure as follows:

$$\text{MSFE} = \frac{1}{T - T_1} \sum_{t=T_1}^{T-1} (\hat{y}_{t+1} - y_{t+1})^2,$$

where $\hat{y}_{t+1}$ is the one-period-ahead forecast of the equity premium using a specific method. We also conduct Giacomini and White's (2006) pairwise tests of the equal conditional predictive ability (CPA) of the various forecasting methods.

To reduce the computational burden, we set the maximum number of candidate models for the TVCSA method. Figure 5 displays the out-of-sample MSFEs of TVCSA under various values of $\overline{M}$. It is straightforward to see that the forecast performances of the TVCSA are not sensitive to the choice of $\overline{M}$ when it is moderately large. Hence, we set $\overline{M} = 100$ as in Lee and Shin (2023).

Table 1 presents the MSFEs and the forecast standard errors. First, we observe that the TVCSA method with a time-varying subset size delivers the smallest MSFE among all methods under study. When the subset size is set to one, the CSA-based method is equivalent to Rapach and Zhou's (2013) simple averaging (i.e., CSR1 in the time-invariant setting). Our findings provide favorable evidence for instability in both the regression coefficient and the optimal subset size. Many episodes occur during such long estimation periods; therefore, the relationships among economic variables may change. For example, stock market disasters (e.g., Black Monday in 1987, the 2008 global financial crisis, and the sovereign debt crisis in 2011) have profound impacts on both stock returns and predictors. By accounting for possible structural changes, the TVCSA method yields a better

[7]The data are available at Amit Goyal's website.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

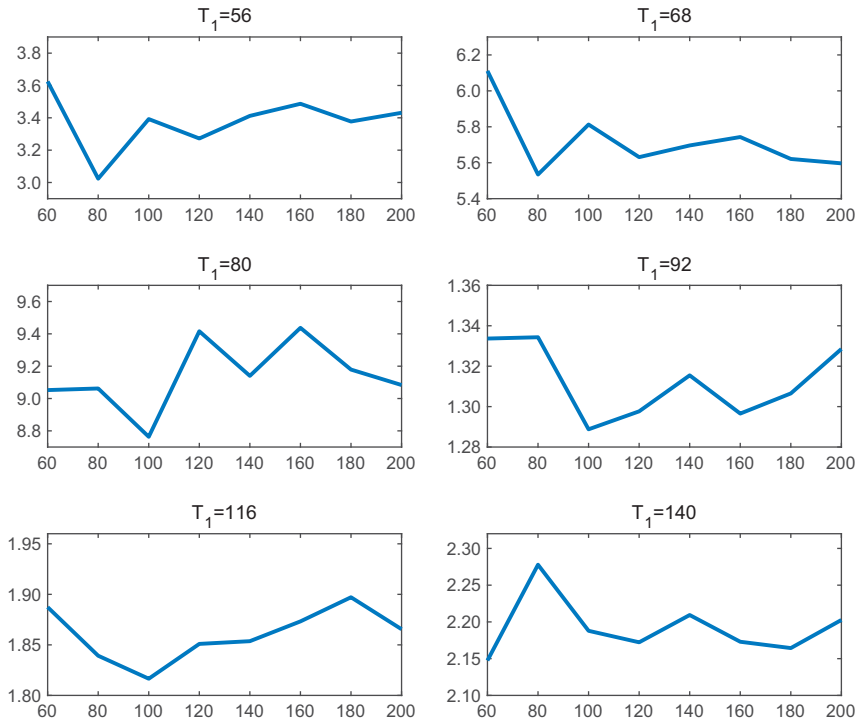


FIGURE 5. MSFEs over the maximum number of models in the forecast combination. The *y*-axis displays the MSFE values, and the *x*-axis shows the maximum number of models. For comparative purposes, the MSFEs are multiplied by 10,000 when $T_1 = 56, 68, 80$, and 1,000 when $T_1 = 92, 116, 140$.

performance than CSR1 and SA. Moreover, compared to the full model (FULL), TVCSA achieves considerable improvement. One possible explanation is the well-known bias-variance trade-off; hence, models with either very few or very many regressors yield the largest squared bias or variance, which is consistent with the findings of Elliott et al. (2013). Second, TVJMA performs best among the model-averaging methods with unequal weights, which sheds light on the importance of time-varying combination weights. Table S1 in the Supplementary Material shows the results of Giacomini and White's (2006) CPA tests for various model-averaging methods. As expected, the CPA tests between TVCSA and the competing methods are significant in nearly all cases, implying that the TVCSA exhibits a significantly better performance.

6.2. Inflation Forecasting

Inflation forecasting plays a crucial role in the formulation and implementation of monetary policies because of its impact on inflation expectations. Since Stock

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Table 1. Out-of-sample MSFEs for equity premium forecasts.

T_1	TVCSA	TVJMA	SAICc	CSR1	CSR2	JMA	SAIC	SBIC	SA	FULL
56	0.0352	0.2074	0.4132	0.6565	0.6118	0.4767	0.4245	0.6131	0.3976	3.2922
	(0.14)	(0.61)	(0.78)	(1.22)	(1.12)	(0.89)	(0.76)	(1.11)	(3.19)	(41.64)
68	0.0577	0.2300	0.4248	0.6558	0.6134	0.4932	0.4449	0.6045	0.4146	1.2509
	(0.17)	(0.61)	(0.76)	(1.20)	(1.10)	(0.91)	(0.77)	(1.06)	(2.92)	(25.05)
80	0.0913	0.2458	0.4563	0.6830	0.6429	0.5212	0.4701	0.6228	0.4146	1.2509
	(0.23)	(0.59)	(0.81)	(1.23)	(1.13)	(0.91)	(0.81)	(1.08)	(2.74)	(20.22)
92	0.1302	0.2521	0.4894	0.7044	0.6708	0.5669	0.5065	0.6530	0.5050	1.7096
	(0.28)	(0.58)	(0.85)	(1.26)	(1.18)	(0.96)	(0.88)	(1.09)	(2.55)	(13.28)
116	0.1841	0.2631	0.4907	0.6358	0.6177	0.5614	0.5250	0.6179	0.7641	1.1362
	(0.34)	(0.51)	(0.85)	(1.08)	(1.06)	(0.94)	(0.88)	(1.07)	(2.01)	(12.22)
140	0.2261	0.2894	0.5363	0.6688	0.6520	0.5975	0.6017	0.6323	0.9699	1.3364
	(0.38)	(0.55)	(0.93)	(1.14)	(1.12)	(1.01)	(1.02)	(1.08)	(1.80)	(10.74)

Note: We perform one-period-ahead rolling-window prediction using fixed T_1 data. For comparative purposes, the MSFEs are multiplied by 100. The numbers in parentheses are the forecast standard errors.

and Watson (1996), studies have considered structural instability and model uncertainty in macroeconomic time series relations. Stock and Watson (2007) discussed the merits of a time-varying univariate model for inflation. Elliott et al. (2015) showed that the CSR approach yields superior point forecasts for U.S. inflation, even with many weak predictors. Garcia, Medeiros, and Vasconcelos (2017) confirmed the superiority of CSR in a data-rich environment. However, as Medeiros et al. (2021) noted, the relationships among macroeconomic indicators may be highly nonlinear. Moreover, inflation dynamics are sensitive to monetary policies and structural breaks, highlighting the necessity for time-varying combination forecasts. This subsection aims to forecast U.S. inflation using the proposed TVCSA method.

The data are obtained from the FRED-QD database.[8] It consists of the 248 macroeconomic series collected from 14 categories, including output and income, labor market, housing, money and credit, interest rates, exchange rates, and stock markets, which form the basis for studies that use big data. The data are recorded at a quarterly frequency, and all variables are transformed to achieve stationarity before analysis (see the FRED-QD Appendix for details). Our sample begins with $1959Q1$ and ends with $2022Q4$, with $T = 255$ observations. Data from $1959Q1$ to $2006Q4$ ($T_1 = 188$) are used as the training set, and those from $2007Q1 - 2022Q4$ are used for forecasting. The number of available regressors is too large compared

[8]Data are available on Michael W. McCracken's website: https://research.stlouisfed.org/econ/mccracken/fred-databases/.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

with the sample size, which may lead to the notorious "curse of dimensionality." In addition, similarities within exogenous variables pose a significant challenge in forecasting inflation. Therefore, we compare the proposed factor TVCSA method with its competing peers (TVJMA, SAICc, CSR1, CSR2, JMA, SAIC, and SBIC). For the 246 series, only 208 variables with all observations are used to estimate the factors. The predictive regression model with the estimated factors is as follows:

$$\pi_{t+h} = \alpha_{0t} + \widehat{F}_t^{\intercal}\alpha_t + \varepsilon_{t+h},\ t = 1, \ldots, T_1, \tag{23}$$

where $\pi_{t+h} = \frac{400}{h}\ln(P_{t+h}/P_t)$ is the annualized inflation in quarter $t+h$, P_t is the quarterly consumer price index, and $\widehat{F}_t = (\widehat{F}_{1t}, \ldots, \widehat{F}_{Rt})^{\intercal}$ is an R-dimensional vector of the estimated factors. The information criterion IC_{p2} in Bai and Ng (2002) suggest the presence of $R = 10$ factors. As a robustness check, we also consider the cases of $R = 20$ and 30. We evaluate combination forecasts over various horizons $h = \{1, 2, 3, 4\}$ using the following out-of-sample MSFE:

$$\text{MSFE} = \frac{1}{T - T_1 - h + 1}\sum_{t=T_1}^{T-h}(\hat{\pi}_{t+h} - \pi_{t+h})^2.$$

Table 2 reports the MSFEs and the forecast standard errors of various model-averaging methods. Several findings are summarized as follows: First, when the number of regressors is small, the full model with a time-varying parameter performs best for short horizons (e.g., $R = 10$ and $h = 1$), but it deteriorates when the forecast horizon h becomes longer. Second, the performance of a simple average for all univariate time-varying models (SA) is relatively stable over the number of factors R and the forecast horizon h. Significantly, it outperforms most model-averaging methods when R and h are small. This finding is consistent with the well-known "forecast combination puzzle." Third, the MSFE results of model-averaging methods, including the proposed factor TVCSA and its competing peers, decrease rapidly as the forecast horizon h increases, and the proposed factor TVCSA estimator yields the smallest forecast errors, especially for long horizons and high-dimensional regressors. Compared with the SA and Full methods corresponding to the settings $r = 1$ and $r = R$, respectively, the factor TVCSA with time-varying subset size r_t achieves an attractive bias-variance trade-off of the combined forecasts.

Table S2 in the Supplementary Material provides the relative MSFEs and p-values of Giacomini and White's (2006) CPA tests for various model-averaging methods in terms of inflation forecast. As Table S2 in the Supplementary Material shows, the factor TVCSA performs equally well as the TVJMA, SAICc, and CSR1 when the forecast horizon $h = 1$. However, the factor TVCSA method significantly outperforms its competing peers as the forecast horizon h increases. Thus, it is crucial to capture the structural changes in forecast combinations by either the time-varying model sets as we do in the TVCSA estimator or time-varying weights for a given set of candidate models as in TVJMA or SAICc.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Table 2. Out-of-sample MSFEs for inflation forecasts.

R	h	TVCSA	TVJMA	SAICc	CSR1	CSR2	JMA	SAIC	SBIC	SA	FULL
10	1	4.8711	7.1127	6.7272	13.9288	14.3639	6.6293	7.1567	7.1577	2.0675	1.6700
		(12.27)	(19.41)	(18.24)	(19.65)	(18.56)	(15.44)	(15.45)	(15.54)	(0.57)	(3.22)
	2	2.5158	3.9477	4.9021	11.0426	11.6670	4.7916	5.6338	5.5381	2.0575	1.4022
		(4.97)	(8.04)	(10.91)	(15.92)	(15.25)	(8.96)	(11.09)	(10.55)	(0.46)	(2.91)
	3	1.4025	2.9398	3.4974	9.4171	10.1052	3.3709	4.1596	4.3441	2.0444	1.6603
		(2.95)	(6.74)	(7.38)	(13.81)	(13.34)	(6.48)	(8.99)	(8.97)	(0.45)	(1.81)
	4	0.8142	2.2347	2.5877	8.3213	9.0680	2.5606	3.1253	3.4271	2.0414	1.8341
		(1.01)	(5.8)	(4.97)	(12.29)	(11.85)	(4.55)	(6.48)	(7.19)	(0.43)	(1.53)
20	1	2.5520	6.1798	5.6312	13.9089	14.1570	6.3665	6.7410	7.1081	2.0854	2.4351
		(3.86)	(16.41)	(14.24)	(20.14)	(19.61)	(13.49)	(13.19)	(13.86)	(0.48)	(3.91)
	2	1.0715	3.6019	4.0902	10.9281	11.2475	4.4749	5.2509	5.3525	2.0813	2.0198
		(1.94)	(7.37)	(8.81)	(16.39)	(15.98)	(8.28)	(10.73)	(9.81)	(0.44)	(3.49)
	3	0.4882	2.4774	2.7059	9.2707	9.6970	3.0032	3.7410	4.1232	2.0686	1.7642
		(0.69)	(5.98)	(6.03)	(14.17)	(13.74)	(6.18)	(8.61)	(8.62)	(0.42)	(2.80)
	4	0.3418	1.8961	1.8119	8.1136	8.5916	2.1327	2.6854	2.9738	2.0592	1.7325
		(0.53)	(5.21)	(3.46)	(12.53)	(12.57)	(3.83)	(5.81)	(6.17)	(0.40)	(1.65)
30	1	1.2033	3.3399	4.7903	13.7445	13.7857	6.1350	6.1805	7.1081	2.1072	3.9175
		(2.63)	(5.98)	(11.83)	(20.13)	(19.57)	(12.66)	(11.94)	(13.86)	(0.46)	(10.04)
	2	0.2333	2.8846	3.4931	10.7354	10.9196	4.1270	3.9223	5.3524	2.0937	1.8608
		(0.43)	(6.37)	(7.44)	(16.29)	(15.99)	(8.00)	(8.7)	(9.81)	(0.45)	(5.31)
	3	0.1581	2.0608	2.2175	9.0799	9.2453	2.7739	2.4962	4.1232	2.0795	1.1895
		(0.32)	(5.33)	(5.12)	(14.02)	(13.59)	(5.89)	(6.69)	(8.62)	(0.43)	(3.32)
	4	0.1029	1.4744	1.3820	7.9549	8.3577	1.9771	1.7024	2.9738	2.0726	1.5707
		(0.25)	(4.67)	(2.48)	(12.45)	(12.74)	(3.74)	(3.27)	(6.17)	(0.42)	(2.68)

Note: The numbers in parentheses are forecast standard errors.

7. CONCLUSION

This study contributes to the time-varying model-averaging literature by proposing a TVCSA method that accounts for both model uncertainty and parameter instability. This method estimates a set of candidate models with time-varying parameters and averages them over complete subsets using an equal-weighting scheme. All the models can be misspecified, which is more realistic in economics and finance. The optimal time-varying subset size is selected by minimizing the LLHCV criterion. Theoretically, we derive the asymptotic normality and uniform convergence rates of the time-varying complete subset regression estimator under smooth structural changes and model misspecification. We further demonstrate

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

that the proposed TVCSA estimator, which assigns equal weights to all candidate models with the same number of regressors, achieves the asymptotic optimality by attaining the lowest local mean squared error loss. Finally, we propose a factor TVCSA method and establish the asymptotic optimality when the number of predictors is relatively large. When the total number of candidate models is large, we recommend randomly drawing a finite number of candidate models from each given subset and establish their forecasting equivalence. Monte Carlo simulations show that the TVCSA method yields better forecasting performance than existing model-averaging methods with estimated optimal weights. Two empirical applications of the TVCSA methods to equity premiums and inflation forecasting illustrate its advantages.

APPENDIX

In this appendix, we provide useful lemmas and technical proofs of the theorems presented in the main text. Lemmas A.1, A.2, and A.3 are used to show Theorems 1, 4, and 6, respectively. The proofs of Lemmas A.1, A.2, and A.3 are placed in the Supplementary Material. Unless otherwise specified, all convergences are considered under the condition $T \to \infty$.

LEMMA A.1. *Suppose Assumptions A.1 and A.2(ii)–(iii) hold. Thus, we obtain the following:*

(i) $\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal} \xrightarrow{p} A_{(m,l)} := E(x_{t(m,l)} x_{t(m,l)}^{\intercal})$;

(ii) $TbVar\left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} \epsilon_{s+h,(m,l)}\right) \xrightarrow{p} \Xi_{(m,l)}$;

(iii) $\sqrt{Tb}\left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} \epsilon_{s+h,(m,l)}\right) \Rightarrow N(0, \Xi_{(m,l)})$.

LEMMA A.2. *$\tilde{P}_{(m,l)} = \tilde{D}_{(m,l)}(P_{(m,l)} - I_T) + I_T$, where I_T is an identity matrix, $P_{(m,l)}$ and $\tilde{P}_{(m,l)}$ are defined as follows:*

$$P_{(m,l)} = \begin{bmatrix} x_{1(m,l)}^{\intercal}\left(x_{(m,l)}^{\intercal}\boldsymbol{K}_1 x_{(m,l)}\right)^{-1} x_{(m,l)}^{\intercal}\boldsymbol{K}_1 \\ x_{2(m,l)}^{\intercal}\left(x_{(m,l)}^{\intercal}\boldsymbol{K}_2 x_{(m,l)}\right)^{-1} x_{(m,l)}^{\intercal}\boldsymbol{K}_2 \\ \vdots \\ x_{T(m,l)}^{\intercal}\left(x_{(m,l)}^{\intercal}\boldsymbol{K}_T x_{(m,l)}\right)^{-1} x_{(m,l)}^{\intercal}\boldsymbol{K}_T \end{bmatrix},$$

$$\tilde{P}_{(m,l)} = \begin{bmatrix} x_{1(m,l)}^{\intercal}\left(x_{(m,l)}^{\intercal}\boldsymbol{K}_{-1,h} x_{(m,l)}\right)^{-1} x_{(m,l)}^{\intercal}\boldsymbol{K}_{-1,h} \\ x_{2(m,l)}^{\intercal}\left(x_{(m,l)}^{\intercal}\boldsymbol{K}_{-2,h} x_{(m,l)}\right)^{-1} x_{(m,l)}^{\intercal}\boldsymbol{K}_{-2,h} \\ \vdots \\ x_{T(m,l)}^{\intercal}\left(x_{(m,l)}^{\intercal}\boldsymbol{K}_{-T,h} x_{(m,l)}\right)^{-1} x_{(m,l)}^{\intercal}\boldsymbol{K}_{-T,h} \end{bmatrix},$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

where $\boldsymbol{K}_{-t,h} = \text{diag}\{K_{1t}, \ldots, K_{(t-h)t}, 0, 0, \ldots, 0, K_{(t+h)t}, \ldots, K_{Tt}\}$ *are the weighting matrix for leave-h-out estimation.* $\tilde{D}_{(m,l)}$ *is a matrix with the ith row element* $(0, \ldots, 0, d_{i,i-h}^{(m,l)}, d_{i,i-h+1}^{(m,l)}, \ldots, d_{i,i+h-1}^{(m,l)}, d_{i,i+h}^{(m,l)}, 0, \ldots, 0)$, *where* $d_{i,j}^{(m,l)} = 1 + p_{i,j}^{(m,l)}$, *and* $p_{i,j}^{(m,l)}$ *is the* (i,j)*th element of the matrix* $P_{(m,l)}$.

LEMMA A.3. *Suppose Assumptions A.2(ii)–(iii), A.7, and A.8 hold. Then,*

(i) $\frac{1}{Tb}\sum_{s=1}^{T} K_{st}||\widehat{F}_s - \Phi F_s||^2 = O_p(\max\{T^{-1}b^{-1}, L^{-1}b^{-1}\})$, *as* $T, L \to \infty$;
(ii) $\frac{1}{Tb}\sum_{s=1}^{T} K_{st}F_s(\widehat{F}_s - \Phi F_s)^{\intercal} = O_p(\max\{T^{-1}b^{-1}, L^{-1}b^{-1}\})$, *as* $T, L \to \infty$.

Proof of Theorem 1.

(i) Recall the definitions of $\beta_{t(m,l)}^* = \arg\min_{\beta_{t(m,l)} \in \mathbb{R}^l} E[(y_{t+h} - x_{t(m,l)}^{\intercal}\beta_{t(m,l)})^2]$ and

$$\hat{\beta}_{t(m,l)} = \underset{\beta_{(m,l)} \in \mathbb{R}^l}{\arg\min}\ Q_t(\beta_{(m,l)}) := \sum_{s=1}^{T} K_{st}(y_{s+h} - x_{s(m,l)}^{\intercal}\beta_{(m,l)})^2.$$

Let $a_T = \sqrt{l/(Tb)}$ and $v_{(m,l)} \in \mathbb{R}^l$ such that $\|v_{(m,l)}\| = C$. To demonstrate Theorem 1(i), it is sufficient to show that for any given $\epsilon > 0$,

$$\Pr\left\{\inf_{\|v_{(m,l)}\|=C} Q_t(\beta_{t(m,l)}^* + a_T v_{(m,l)}) > Q_t(\beta_{t(m,l)}^*)\right\} \geq 1 - \epsilon.$$

Note that

$$\begin{aligned}
&Q_t(\beta_{t(m,l)}^* + a_T v_{(m,l)}) - Q_t(\beta_{t(m,l)}^*) \\
&= \sum_{s=1}^{T} K_{st}(y_{s+h} - x_{s(m,l)}^{\intercal}(\beta_{t(m,l)}^* + a_T v_{(m,l)}))^2 - \sum_{s=1}^{T} K_{st}(y_{s+h} - x_{s(m,l)}^{\intercal}\beta_{t(m,l)}^*)^2 \\
&= -2a_T \sum_{s=1}^{T} K_{st}(y_{s+h} - x_{s(m,l)}^{\intercal}\beta_{t(m,l)}^*)x_{s(m,l)}v_{(m,l)} + a_T^2 v_{(m,l)}^{\intercal} \sum_{s=1}^{T} K_{st}x_{s(m,l)}^{\intercal}x_{s(m,l)}v_{(m,l)} \\
&= -2a_T \sum_{s=1}^{T} K_{st}(y_{s+h} - x_{s(m,l)}^{\intercal}\beta_{s(m,l)}^*)x_{s(m,l)}v_{(m,l)} \\
&\quad + 2a_T \sum_{s=1}^{T} K_{st}x_{s(m,l)}^{\intercal}(\beta_{s(m,l)}^* - \beta_{t(m,l)}^*)x_{s(m,l)}v_{(m,l)} \\
&\quad + a_T^2 v_{(m,l)}^{\intercal} \sum_{s=1}^{T} K_{st}x_{s(m,l)}^{\intercal}x_{s(m,l)}v_{(m,l)} \\
&:= I + II + III.
\end{aligned}$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

For I, we have $E(I) = 0$ by the first-order condition of the optimization. By Lemma A.1(i)–(ii),

$$\begin{aligned} Var(I) &= 4a_T^2 v_{(m,l)}^{\intercal} E\left(\sum_{s=1}^{T}\sum_{\iota=1}^{T} K_{st}K_{\iota t} x_{s(m,l)} x_{s(m,l)}^{\intercal} \epsilon_{s+h,(m,l)}\epsilon_{\iota+h,(m,l)}\right) v_{(m,l)} \\ &= 4a_T^2 (Tb) v_{(m,l)}^{\intercal} \Xi_{(m,l)} v_{(m,l)} \\ &\le 4a_T^2 (Tb) \lambda_{\max}\left(\Xi_{(m,l)}\right) \|v_{(m,l)}\|^2, \end{aligned}$$

For II, by Assumptions A1(ii) and A2(i), we have

$$\begin{aligned} &E\left(2a_T \sum_{s=1}^{T} K_{st} x_{s(m,l)}^{\intercal} (\beta^*_{s(m,l)} - \beta^*_{t(m,l)}) x_{s(m,l)} v_{(m,l)}\right)^2 \\ &\le 4a_T^2 \sum_{s=1}^{T} (\beta^*_{s(m,l)} - \beta^*_{t(m,l)})^{\intercal} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal} (\beta^*_{s(m,l)} - \beta^*_{t(m,l)}) \\ &\quad \cdot \sum_{s=1}^{T} v_{(m,l)}^{\intercal} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal} v_{(m,l)} \\ &\le 4a_T^2 (Tb)^2 \sup_{|s-t|<Tb} \left\|\beta^*_{s(m,l)} - \beta^*_{t(m,l)}\right\|^2 \lambda_{\max}\left(A_{(m,l)}\right) \|v_{(m,l)}\|^2 \\ &\le 4a_T^2 (Tb) \overline{C}_x \|v_{(m,l)}\|^2. \end{aligned}$$

For III, by Assumption A1(ii), we have

$$III \ge a_T^2 (Tb) \underline{C}_x \|v_{(m,l)}\|^2.$$

Hence,

$$\begin{aligned} &Q_t(\beta^*_{t(m,l)} + a_T v_{(m,l)}) - Q_t(\beta^*_{t(m,l)}) \\ &\quad \ge a_T^2 (Tb) \underline{C}_x \|v_{(m,l)}\|^2 - 2a_T \sqrt{Tb}\, \overline{C}_v^{1/2} \|v_{(m,l)}\| - 2a_T \sqrt{Tb}\, \overline{C}_x^{1/2} \|v_{(m,l)}\|, \end{aligned}$$

which is positive when $\|v_{(m,l)}\|$ is sufficiently large. This implies that a local minimum exists in the space $\{\beta^*_{t(m,l)} + a_T v_{(m,l)} : \|v_{(m,l)}\| \le C\}$, such that $\left\|\hat{\beta}_{t(m,l)} - \beta^*_{t(m,l)}\right\| = o_p(a_T)$.

(ii) Note that

$$\begin{aligned} \hat{\beta}_{t(m,l)} &= \left(\sum_{s=1}^{T} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal}\right)^{-1} \left(\sum_{s=1}^{T} K_{st} x_{s(m,l)} y_{s+h}\right) \\ &= \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal}\right)^{-1} \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} y_{s+h}\right) \\ &= \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal}\right)^{-1} \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} (x_{s(m,l)}^{\intercal} \beta_{s(m,l)} + \epsilon_{s+h,(m,l)})\right) \end{aligned}$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

$$
\begin{aligned}
&= \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal}\right)^{-1} \\
&\times \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal}\left(\beta_{t(m,l)} + \beta_{t(m,l)}^{'}\left(\frac{s-t}{T}\right) + \beta_{t(m,l)}^{''}\left(\frac{s-t}{T}\right)^2 + o(b^2)\right)\right) \\
&+ \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal}\right)^{-1}\left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} \epsilon_{s+h,(m,l)}\right) \\
&= \beta_{t(m,l)} + \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal}\right)^{-1} \\
&\times \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal}\left(\beta_{t(m,l)}^{'}\left(\frac{s-t}{T}\right) + \beta_{t(m,l)}^{''}\left(\frac{s-t}{T}\right)^2 + o(b^2)\right)\right) \\
&+ \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal}\right)^{-1}\left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} \epsilon_{s+h,(m,l)}\right).
\end{aligned}
$$

Therefore, by Lemma A.1, we have

$$
\begin{aligned}
&\sqrt{Tb}(\hat{\beta}_{t(m,l)} - \beta_{t(m,l)}^{*} - \mu_2 h^2 \beta_{t(m,l)}^{''}) + o(b^2) \\
&= \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st} x_{s(m,l)} x_{s(m,l)}^{\intercal}\right)^{-1}\left(\frac{1}{\sqrt{Tb}}\sum_{s=1}^{T} K_{st} x_{s(m,l)} \epsilon_{s+h,(m,l)}\right) \\
&\Rightarrow N(0, A_{(m,l)}^{-1} \Xi_{(m,l)} A_{(m,l)}^{-1}).
\end{aligned}
$$

Proof of Theorem 2.

(i) Let $\delta_T := C\left(\sqrt{(Tb)^{-1} L \log(Tb)}\right)$ and $\mathcal{Q}_{t(m,l)}(\beta_{t(m,l)}) := E[(y_{t+h} - x_{t(m,l)}^{\intercal}\beta_{t(m,l)})^{\intercal} K_{st}(y_{t+h} - x_{t(m,l)}^{\intercal}\beta_{t(m,l)})]$. Denote

$$
\mathcal{D}(\delta_T) := \inf_{1\le m\le M} \inf_{\left\|\beta_{t(m,l)} - \beta_{t(m,l)}^{*}\right\| > \delta_T} \left[\mathcal{Q}_{t(m,l)}(\beta_{t(m,l)}) - \mathcal{Q}_{t(m,l)}(\beta_{t(m,l)}^{*})\right],
$$

$$
\mathcal{S}_{t(m,l)}(\delta_T) := \left\{\beta_{t(m,l)} : \left\|\beta_{t(m,l)} - \beta_{t(m,l)}^{*}\right\| > \delta_T, \left\|\beta_{t(m,l)} - \beta_{t(m,l)}^{*}\right\| = o(1)\right\}.
$$

For any $\beta_{t(m,l)} \in \mathcal{S}_{t(m,l)}(\delta_T)$, we have

$$
\begin{aligned}
&\mathcal{Q}_{t(m,l)}(\beta_{t(m,l)}) - \mathcal{Q}_{t(m,l)}(\beta_{t(m,l)}^{*}) \\
&= E\left[(y_{t+h} - x_{t(m,l)}^{\intercal}\beta_{t(m,l)})^{\intercal} K_{st}(y_{t+h} - x_{t(m,l)}^{\intercal}\beta_{t(m,l)})\right] \\
&\quad - E\left[(y_{t+h} - x_{t(m,l)}^{\intercal}\beta_{t(m,l)}^{*})^{\intercal} K_{st}(y_{t+h} - x_{t(m,l)}^{\intercal}\beta_{t(m,l)}^{*})\right]
\end{aligned}
$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

$$
\begin{aligned}
&= E(\beta_{t(m,l)}^{\mathsf{T}} x_{t(m,l)} K_{st} x_{t(m,l)}^{\mathsf{T}} \beta_{t(m,l)}) - 2E(\beta_{t(m,l)}^{\mathsf{T}} x_{t(m,l)} K_{st} y_{t+h}) \\
&\quad + 2E(\beta_{t(m,l)}^{*\mathsf{T}} x_{t(m,l)} K_{st} y_{t+h}) - E(\beta_{t(m,l)}^{*\mathsf{T}} x_{t(m,l)} K_{st} x_{t(m,l)}^{\mathsf{T}} \beta_{t(m,l)}^{*}) \\
&= (\beta_{t(m,l)} - \beta_{t(m,l)}^{*})^{\mathsf{T}} E(x_{t(m,l)} K_{st} x_{t(m,l)}^{\mathsf{T}})(\beta_{t(m,l)} - \beta_{t(m,l)}^{*}) \\
&\quad + 2E\left[(\beta_{t(m,l)}^{*} - \beta_{t(m,l)})^{\mathsf{T}} x_{t(m,l)} \epsilon_{t+h,(m,l)}\right] \\
&\geq \underline{C}_x \left\| \hat{\beta}_{t(m,l)} - \beta_{t(m,l)} \right\|^2 \\
&\geq \underline{C}_x \delta_T^2.
\end{aligned}
$$

By Boole inequality, we can obtain

$$
\begin{aligned}
&\Pr\Big(\max_{1\leq t\leq T} \max_{1\leq l\leq L} \max_{1\leq m\leq M} \|\tilde{\beta}_{t(m,l)} - \beta_{t(m,l)}^{*}\| \geq \delta_T \Big) \\
&\leq TLM \max_{1\leq t\leq T} \max_{1\leq l\leq L} \max_{1\leq m\leq M} \Pr(\|\tilde{\beta}_{t(m,l)} - \beta_{t(m,l)}^{*}\| \geq \delta_T) \\
&\leq TLM \max_{1\leq t\leq T} \max_{1\leq l\leq L} \max_{1\leq m\leq M} \Pr(\mathcal{Q}_{t(m,l)}(\beta_{t(m,l)}) - \mathcal{Q}_{t(m,l)}(\beta_{t(m,l)}^{*}) \geq \mathcal{D}(\delta_T)) \\
&\simeq TLM \max_{1\leq t\leq T} \max_{1\leq l\leq L} \max_{1\leq m\leq M} \\
&\quad \times \Pr(Tb(\beta_{t(m,l)} - \beta_{t(m,l)}^{*})^{\mathsf{T}} E(x_{t(m,l)} K_{st} x_{t(m,l)}^{\mathsf{T}})(\beta_{t(m,l)} - \beta_{t(m,l)}^{*}) \geq Tb\mathcal{D}(\delta_T)).
\end{aligned}
$$

Recalling the results of Theorem 1(ii), we have

$$
\sqrt{Tb} C_{(m,l)} \Omega_{t(m,l)}^{-1/2} [\hat{\beta}_{t(m,l)} - \beta_{t(m,l)}^{*} - \mu_2 b^2 \beta_{t(m,l)}^{*''}] \Rightarrow N(0, C_0).
$$

Let $\dot{\beta}_{t(m,l)} := \hat{\beta}_{t(m,l)} - \beta_{t(m,l)}^{*} - \mu_2 b^2 \beta_{t(m,l)}^{*''}$ and $\ddot{\beta}_{t(m,l)} := \sqrt{Tb}\dot{\beta}_{t(m,l)}^{\mathsf{T}} (C_{(m,l)} C_{(m,l)}^{\mathsf{T}})^{-1/2}$ $C_{(m,l)} \Omega_{t(m,l)}^{-1/2}$. We have

$$
\begin{aligned}
&Tb(\hat{\beta}_{t(m,l)} - \beta_{t(m,l)}^{*})^{\mathsf{T}} E(x_{t(m,l)} K_{st} x_{t(m,l)}^{\mathsf{T}})(\hat{\beta}_{t(m,l)} - \beta_{t(m,l)}^{*}) \\
&\quad = Tb(\dot{\beta}_{t(m,l)} + \mu_2 b^2 \beta_{t(m,l)}^{*''})^{\mathsf{T}} E(x_{t(m,l)} K_{st} x_{t(m,l)}^{\mathsf{T}})(\dot{\beta}_{t(m,l)} + \mu_2 b^2 \beta_{t(m,l)}^{*''}) \\
&\quad \leq 2\sqrt{Tb}\dot{\beta}_{t(m,l)}^{\mathsf{T}} E(x_{t(m,l)} K_{st} x_{t(m,l)}^{\mathsf{T}}) \sqrt{Tb}\dot{\beta}_{t(m,l)} \\
&\qquad + 2Tb\mu_2 b^2 \beta_{t(m,l)}^{*''\mathsf{T}} E(x_{t(m,l)} K_{st} x_{t(m,l)}^{\mathsf{T}}) \mu_2 b^2 \beta_{t(m,l)}^{*''} \\
&\quad \leq 2\ddot{\beta}_{t(m,l)}^{\mathsf{T}} \left[\left((C_{(m,l)} C_{(m,l)}^{\mathsf{T}})^{-1/2} C_{(m,l)} \Omega_{(m,l)}^{-1/2} \right)^{+} \right]^{\mathsf{T}} \\
&\qquad \times E(x_{t(m,l)} K_{st} x_{t(m,l)}^{\mathsf{T}}) \left[\left((C_{(m,l)} C_{(m,l)}^{\mathsf{T}})^{-1/2} C_{(m,l)} \Omega_{(m,l)}^{-1/2} \right)^{+} \right] \ddot{\beta}_{t(m,l)} \\
&\qquad + 2Tb^5 \mu_2^2 \beta_{t(m,l)}^{*''\mathsf{T}} E(x_{t(m,l)} K_{st} x_{t(m,l)}^{\mathsf{T}}) \beta_{t(m,l)}^{*''} \\
&\quad \leq 2\overline{C}_x \ddot{\beta}_{t(m,l)}^{\mathsf{T}} \left[\left((C_{(m,l)} C_{(m,l)}^{\mathsf{T}})^{-1/2} C_{(m,l)} \Omega_{(m,l)}^{-1/2} \right)^{+} \right]^{\mathsf{T}} \\
&\qquad \times \left[\left((C_{(m,l)} C_{(m,l)}^{\mathsf{T}})^{-1/2} C_{(m,l)} \Omega_{(m,l)}^{-1/2} \right)^{+} \right] \ddot{\beta}_{t(m,l)} + 2Tb^5 \mu_2^2 \overline{C}_x \left\| \beta_{t(m,l)}^{*''} \right\|^2
\end{aligned}
$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

$$
\begin{aligned}
&= 2\overline{C}_x \ddot{\beta}_{t(m,l)}^{\mathsf{T}} \left\{ (C_{(m,l)} C_{(m,l)}^{\mathsf{T}})^{-1/2} (C_{(m,l)} \Omega_{(m,l)}^{-1} C_{(m,l)}^{\mathsf{T}}) (C_{(m,l)} C_{(m,l)}^{\mathsf{T}})^{-1/2} \right\}^{-1} \ddot{\beta}_{t(m,l)} \\
&\quad + 2Tb^5 \mu_2^2 \overline{C}_x \left\| \beta_{t(m,l)}^{*''} \right\|^2 \\
&= 2\overline{C}_x \ddot{\beta}_{t(m,l)}^{\mathsf{T}} \left\{ (C_{(m,l)} C_{(m,l)}^{\mathsf{T}})^{1/2} (C_{(m,l)} \Omega_{(m,l)}^{-1} C_{(m,l)}^{\mathsf{T}})^{-1} (C_{(m,l)} C_{(m,l)}^{\mathsf{T}})^{1/2} \right\} \ddot{\beta}_{t(m,l)} \\
&\quad + 2Tb^5 \mu_2^2 \overline{C}_x \left\| \beta_{t(m,l)}^{*''} \right\|^2 \\
&= 2\overline{C}_x \|\Omega\| \left\| \ddot{\beta}_{t(m,l)} \right\|^2 + 2Tb^5 \mu_2^2 \overline{C}_x \left\| \beta_{t(m,l)}^{*''} \right\|^2 \\
&\leq \frac{2\overline{C}_x \overline{C}_v}{\underline{C}_x^2} \left\| \ddot{\beta}_{t(m,l)} \right\|^2 + 2Tb^5 \mu_2^2 \overline{C}_x \left\| \beta_{t(m,l)}^{*''} \right\|^2.
\end{aligned}
$$

Let $C_{xv} := (\overline{C}_x \overline{C}_v)/\underline{C}_x^2$, $\bar{l} := \max_{1 \leq m \leq M} \max_{1 \leq l \leq L} l_{(m,l)}$, and assume $\lambda_{\max}(\beta_{t(m,l)}^{*''}) \leq \overline{C}_\beta < \infty$. Therefore,

$$
\begin{aligned}
&\Pr\Big(\max_{1 \leq t \leq T} \max_{1 \leq l \leq L} \max_{1 \leq m \leq M} \|\tilde{\beta}_{t(m,l)} - \beta_{t(m,l)}^*\| \leq \delta_T \Big) \\
&\quad \simeq TLM \max_{1 \leq t \leq T} \max_{1 \leq l \leq L} \max_{1 \leq m \leq M} \\
&\qquad \times \Pr(Tb(\beta_{t(m,l)} - \beta_{t(m,l)}^*)^{\mathsf{T}} E(x_{t(m,l)} K_{st} x_{t(m,l)}^{\mathsf{T}})(\beta_{t(m,l)} - \beta_{t(m,l)}^*) \geq Th\mathcal{D}(\delta_T)) \\
&\quad = TLM \max_{1 \leq t \leq T} \max_{1 \leq l \leq L} \max_{1 \leq m \leq M} \Pr\left(2C_{xv} \left\| \ddot{\beta}_{t(m,l)} \right\|^2 + 2Tb^5 \mu_2^2 \overline{C}_x \overline{C}_\beta^2 \geq Tb\mathcal{D}(\delta_T) \right) \\
&\quad \leq TLM \max_{1 \leq t \leq T} \max_{1 \leq l \leq L} \max_{1 \leq m \leq M} \Pr\left(\left\| \ddot{\beta}_{t(m,l)} \right\|^2 \geq \frac{Tb\mathcal{D}(\delta_T)}{2C_{xv}} - \frac{Tb^5 \mu_2^2 \underline{C}_x^2 \overline{C}_\beta^2}{\overline{C}_v} \right) \\
&\quad \leq \limsup_{T \to \infty} TLM \Pr\left(\chi^2(l_{(m,l)}) \geq \frac{Tb\mathcal{D}(\delta_T)}{2C_{xv}} - \frac{Tb^5 \mu_2^2 \underline{C}_x^2 \overline{C}_\beta^2}{\overline{C}_v} \right) \\
&\quad \leq \limsup_{T \to \infty} TLM \Pr\left(\chi^2(\bar{l}) \geq \bar{l} + \left(\frac{Tb\underline{C}_x \delta_T^2}{2C_{xv}} - \frac{Tb^5 \mu_2^2 \underline{C}_x^2 \overline{C}_\beta^2}{\overline{C}_v} - \bar{l} \right) \right) \\
&\quad \leq \limsup_{T \to \infty} TLM \exp\left(-\frac{1}{2} \left(\frac{Tb\underline{C}_x \delta_T^2}{2C_{xv}} - \frac{Tb^5 \mu_2^2 \underline{C}_x^2 \overline{C}_\beta^2}{\overline{C}_v} - \bar{l} \right) \right) \\
&\qquad\qquad \times \exp\left(\frac{\bar{l}}{2} \log\left(1 + \left(\frac{Tb\underline{C}_x \delta_T^2}{2C_{xv}} - \frac{Tb^5 \mu_2^2 \underline{C}_x^2 \overline{C}_\beta^2}{\overline{C}_v} - \bar{l} \right) \frac{1}{\bar{l}} \right) \right)
\end{aligned}
$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

$$\leq \limsup_{T\to\infty} TLM \exp\left\{-\frac{1}{2}\left(\frac{Tb\underline{C}_x\delta_T^2}{2C_{xv}} - \frac{Tb^5\mu_2^2\underline{C}_x^2\overline{C}_\beta^2}{\overline{C}_v} - \bar{l}\right)\right.$$
$$\left.\times\left(1 - \frac{\log\left(\frac{Tb\underline{C}_x\delta_T^2}{2C_{xv}\bar{l}} - \frac{Tb^5\mu_2^2\underline{C}_x^2\overline{C}_\beta^2}{\overline{C}_v\bar{l}}\right)}{\frac{Tb\underline{C}_x\delta_T^2}{2C_{xv}\bar{l}} - \frac{Tb^5\mu_2^2\underline{C}_x^2\overline{C}_\beta^2}{\overline{C}_v\bar{l}} - 1}\right)\right\}.$$

Since $\left[\log\left(\frac{Tb\underline{C}_x\delta_T^2}{2C_{xv}\bar{l}} - \frac{Tb^5\mu_2^2\underline{C}_x^2\overline{C}_\beta^2}{\overline{C}_v\bar{l}}\right)\right] \times \left(\frac{Tb\underline{C}_x\delta_T^2}{2C_{xv}\bar{l}} - \frac{Tb^5\mu_2^2\underline{C}_x^2\overline{C}_\beta^2}{\overline{C}_v\bar{l}} - 1\right)^{-1} = o(1)$ for sufficiently large C, the leading term becomes

$$TLM \exp\left\{-\frac{1}{2}\left(\frac{Th\underline{C}_x\delta_T^2}{2C_{xv}}\right)\right\} = TLM \exp\left\{-\frac{1}{2}\left(\frac{C^2\log(Tb)L\underline{C}_x}{2C_{xv}}\right)\right\}$$
$$= TLM(Tb)^{-\frac{1}{2}\times\frac{C^2L\underline{C}_x}{2C_{xv}}}$$
$$\leq TLL!\,(Tb)^{-\frac{1}{2}\times\frac{C^2L\underline{C}_x}{2C_{xv}}}$$
$$\leq cT\log(Tb)^{L+1}(Tb)^{-\frac{1}{2}\frac{C^2L\underline{C}_x}{2C_{xv}}} = o(1),$$

where the last line holds under Assumption A.3. Hence, we obtain Theorem 2(i). The proof of (ii) is similar and is omitted.

Proof of Theorem 3.

We demonstrate that the difference between $\tilde{\mu}_t(l) := \frac{1}{M}\sum_{m=1}^{\overline{M}} x_{t(m,l)}^{\intercal}\hat{\beta}_{t(m,l)}$ and $\hat{\mu}_t(l) := \frac{1}{M}\sum_{m=1}^{M} x_{t(m,l)}^{\intercal}\hat{\beta}_{t(m,l)}$ is asymptotically negligible. Using the triangular inequality, we obtain

$$\max_{1\leq t\leq T}\max_{1\leq l\leq L}\left\|\hat{\mu}_{t(l)} - \tilde{\mu}_{t(l)}\right\|$$
$$= \max_{1\leq t\leq T}\max_{1\leq l\leq L}\left\|\frac{1}{M}\sum_{m=1}^{M} x_{t(m,l)}^{\intercal}\hat{\beta}_{t(m,l)} - \frac{1}{\overline{M}}\sum_{m'=1}^{\overline{M}} x_{t(m',k)}^{\intercal}\hat{\beta}_{t(m',k)}\right\|$$
$$= \max_{1\leq t\leq T}\max_{1\leq l\leq L}\left\|\frac{1}{M}\sum_{m=1}^{M}\left(x_{t(m,l)}^{\intercal}\hat{\beta}_{t(m,l)} - \mu_t^*(l)\right) - \frac{1}{\overline{M}}\sum_{m'=1}^{\overline{M}}\left(x_{t(m',k)}^{\intercal}\hat{\beta}_{t(m',k)} - \mu_t^*(l)\right)\right\|$$
$$\leq \max_{1\leq t\leq T}\max_{1\leq l\leq L}\left\|\frac{1}{M}\sum_{m=1}^{M}\left(x_{t(m,l)}^{\intercal}\hat{\beta}_{t(m,l)} - \mu_t^*(l)\right)\right\|$$
$$+ \max_{1\leq t\leq T}\max_{1\leq l\leq L}\left\|\frac{1}{\overline{M}}\sum_{m'=1}^{\overline{M}}\left(x_{t(m',k)}^{\intercal}\hat{\beta}_{t(m',k)} - \mu_t^*(l)\right)\right\| := \Delta_1 + \Delta_2.$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

For $\mathcal{I}_1$, we have

$$\begin{aligned}\Delta_1 &\le \max_{1\le t\le T}\max_{1\le l\le L}\left\|\frac{1}{M}\sum_{m=1}^{M}\left(x_{t(m,l)}^{\mathsf{T}}\beta_{t(m,l)}^{*}-\mu_t^{*}(l)\right)\right\| \\ &\quad+\max_{1\le t\le T}\max_{1\le l\le L}\left\|\frac{1}{M}\sum_{m=1}^{M}x_{t(m,l)}^{\mathsf{T}}\left(\hat{\beta}_{t(m,l)}-\beta_{t(m,l)}^{*}\right)\right\| \\ &\le \max_{1\le t\le T}\max_{1\le l\le L}\left\|\frac{1}{M}\sum_{m=1}^{M}\left(x_{t(m,l)}^{\mathsf{T}}\beta_{t(m,l)}^{*}-\mu_t^{*}(l)\right)\right\| \\ &\quad+\left(\max_{1\le t\le T}\max_{1\le l\le L}\max_{1\le m\le M}\left\|x_{t(m,l)}\right\|\right)\left(\max_{1\le t\le T}\max_{1\le l\le L}\max_{1\le m\le M}\left\|\hat{\beta}_{t(m,l)}-\beta_{t(m,l)}^{*}\right\|\right) \\ &= o_p(1)+O_p(1)\times o_p(1)=o_p(1).\end{aligned}$$

Similarly, we can show $\Delta_2 = o_p(1)$. This completes the proof of Theorem 3.

Proof of Theorem 4.

Let $\tilde{\mathcal{L}}_{t,T}(l) = \frac{1}{Tb}(\tilde{\mu}(l)-\mu)^{\mathsf{T}}K_t(\tilde{\mu}(l)-\mu)$, and $\tilde{\mathcal{R}}_{t,T}(l) = E[\tilde{\mathcal{L}}_{t,T}(l)]$. We conduct the following decomposition as in Zhang et al. (2013):

$$\begin{aligned}\frac{\mathcal{L}_{t,T}(\hat{l})}{\inf_{l\in\mathbb{L}}\mathcal{L}_{t,T}(l)} &= \sup_{l\in\mathbb{L}}\left(\frac{\mathcal{L}_{t,T}(\hat{l})}{\mathcal{L}_{t,T}(l)}\right) \\ &\le \sup_{l\in\mathbb{L}}\left(\frac{\mathcal{L}_{t,T}(l)}{\mathcal{R}_{t,T}(l)}\right)\sup_{l\in\mathbb{L}}\left(\frac{\mathcal{R}_{t,T}(l)}{\mathcal{L}_{t,T}(l)}\right)\sup_{l\in\mathbb{L}}\left(\frac{\tilde{\mathcal{R}}_{t,T}(l)}{\mathcal{R}_{t,T}(l)}\right)\sup_{l\in\mathbb{L}}\left(\frac{\mathcal{R}_{t,T}(l)}{\tilde{\mathcal{R}}_{t,T}(l)}\right) \\ &\quad\times\sup_{l\in\mathbb{L}}\left(\frac{\tilde{\mathcal{R}}_{t,T}(l)}{\tilde{\mathcal{L}}_{t,T}(l)}\right)\sup_{l\in\mathbb{L}}\left(\frac{\tilde{\mathcal{L}}_{t,T}(l)}{\tilde{\mathcal{R}}_{t,T}(l)}\right)\frac{\tilde{\mathcal{L}}_{t,T}(\hat{l})}{\inf_{l\in\mathbb{L}}\tilde{\mathcal{L}}_{t,T}(l)},\end{aligned}$$

where $\mathbb{L} := \{1,\dots,L\}$. Let $A_{(m,l)} = I_T - P_{(m,l)}$, $\tilde{A}_{(m,l)} = I_T - \tilde{P}_{(m,l)}$, $H_{(m,l)} = I_T - D_{(m,l)}$ and $T_{(m,l)} = H_{(m,l)}P_{(m,l)}$. Then we have the following equations:

$$\begin{aligned}\tilde{P}_{(m,l)} &= P_{(m,l)} - H_{(m,l)}A_{(m,l)}, \\ \tilde{A}_{(m,l)} &= A_{(m,l)} + H_{(m,l)}A_{(m,l)}.\end{aligned}$$

Further, we let $A(l) = \frac{1}{M}\sum_{m=1}^{M}A_{(m,l)}$, $\tilde{A}(l) = \frac{1}{M}\sum_{m=1}^{M}\tilde{A}_{(m,l)}$, $H(l) = \frac{1}{M}\sum_{m=1}^{M}H_{(m,l)}$, and $T(l) = \frac{1}{M}\sum_{m=1}^{M}T_{(m,l)}$.

Note that,

$$\begin{aligned}CV_{t,T}(l) &= (\mathbf{y}-\tilde{\mu}(l))^{\mathsf{T}}\mathbf{K}_t(\mathbf{y}-\tilde{\mu}(l)) \\ &= (\varepsilon+\mu-\tilde{\mu}(l))^{\mathsf{T}}\mathbf{K}_t(\varepsilon+\mu-\tilde{\mu}(l)) \\ &= Tb\tilde{\mathcal{L}}_{t,T}(l)+\varepsilon^{\mathsf{T}}\mathbf{K}_t\varepsilon+2\varepsilon^{\mathsf{T}}\mathbf{K}_t(\mu-\tilde{\mu}(l)) \\ &= Tb\tilde{\mathcal{L}}_{t,T}(l)+\varepsilon^{\mathsf{T}}\mathbf{K}_t\varepsilon+2\varepsilon^{\mathsf{T}}\mathbf{K}_t\tilde{A}(l)\mu-2\varepsilon^{\mathsf{T}}\mathbf{K}_t\tilde{P}(l)\varepsilon.\end{aligned}$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Hence, to prove Theorem 4, it suffices to show that

$$\sup_{l\in\mathbb{L}}\left|\frac{\mathcal{L}_{t,T}(l)}{\mathcal{R}_{t,T}(l)}-1\right|\xrightarrow{p}0, \tag{A.1}$$

$$\sup_{l\in\mathbb{L}}\left|\frac{\tilde{\mathcal{R}}_{t,T}(l)}{\mathcal{R}_{t,T}(l)}-1\right|\xrightarrow{p}0, \tag{A.2}$$

$$\sup_{l\in\mathbb{L}}\left|\frac{\tilde{\mathcal{R}}_{t,T}(l)}{\tilde{\mathcal{L}}_{t,T}(l)}-1\right|\xrightarrow{p}0, \tag{A.3}$$

$$\sup_{l\in\mathbb{L}}\left|\frac{\varepsilon^{\intercal}\mathbf{K}_t\tilde{A}(l)\mu}{\tilde{\mathcal{R}}_{t,T}(l)}\right|\xrightarrow{p}0, \tag{A.4}$$

$$\sup_{l\in\mathbb{L}}\left|\frac{\varepsilon^{\intercal}\mathbf{K}_t\tilde{P}(l)\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right|\xrightarrow{p}0. \tag{A.5}$$

We first show (A.1). Using the triangle inequality, we have

$$\begin{aligned}
&\mathcal{L}_{t,T}(l)-\mathcal{R}_{t,T}(l)\\
&\quad=\mu^{\intercal}A(l)^{\intercal}\mathbf{K}_tA(l)\mu+\varepsilon^{\intercal}P(l)^{\intercal}\mathbf{K}_tP(l)\varepsilon-2\mu^{\intercal}A(l)^{\intercal}\mathbf{K}_tP(l)\varepsilon\\
&\qquad-\left[\mu^{\intercal}A(l)^{\intercal}\mathbf{K}_tA(l)\mu+\operatorname{tr}\left(P(l)^{\intercal}\mathbf{K}_tP(l)\Omega\right)\right]\\
&\quad=\varepsilon^{\intercal}P(l)^{\intercal}\mathbf{K}_tP(l)\varepsilon-\operatorname{tr}\left(P(l)^{\intercal}\mathbf{K}_tP(l)\Omega\right)-2\mu^{\intercal}A(l)^{\intercal}\mathbf{K}_tP(l)\varepsilon
\end{aligned}$$

and

$$\begin{aligned}
\sup_{l\in\mathbb{L}}\left|\frac{\mathcal{L}_{t,T}(l)-\mathcal{R}_{t,T}(l)}{\mathcal{R}_{t,T}(l)}\right|&\le\sup_{l\in\mathbb{L}}\left|\frac{\varepsilon^{\intercal}P(l)^{\intercal}\mathbf{K}_tP(l)\varepsilon}{\mathcal{R}_{t,T}(l)}\right|+\sup_{l\in\mathbb{L}}\left|\frac{\operatorname{tr}(P(l)^{\intercal}\mathbf{K}_tP(l)\Omega)}{\mathcal{R}_{t,T}(l)}\right|\\
&\quad+2\sup_{l\in\mathbb{L}}\left|\frac{\mu^{\intercal}A(l)^{\intercal}\mathbf{K}_tP(l)\varepsilon}{\mathcal{R}_{t,T}(l)}\right|\\
&:=\Lambda_{11}+\Lambda_{12}+\Lambda_{13}.
\end{aligned}$$

We denote $\overline{K}$ as the maximum value of $\mathbf{K}_t$, and $\zeta(\cdot)$ as the maximum singular values of a matrix. For Λ_{12}, using Assumptions A.4(i) and A.5(i), for any $\delta>0$,

$$\begin{aligned}
\Pr\left(\sup_{l\in\mathbb{L}}\left|\frac{\operatorname{tr}(P(l)^{\intercal}\mathbf{K}_tP(l)\Omega)}{\mathcal{R}_{t,T}(l)}\right|\ge\delta\right)&\le\Pr\left(\overline{K}\sup_{l\in\mathbb{L}}\operatorname{tr}\left(P(l)^{\intercal}P(l)\Omega\right)\xi_{t,T}^{-1}\ge\delta\right)\\
&\le\Pr\left(\overline{K}\zeta(\Omega)\sup_{l\in\mathbb{L}}\operatorname{tr}\left(P(l)^{\intercal}P(l)\right)\xi_{t,T}^{-1}\ge\delta\right)\\
&\le\Pr\left(\overline{K}\zeta(\Omega)\sup_{l\in\mathbb{L}}P(l)^{\intercal}P(l)\xi_{t,T}^{-1}\ge\delta\right)\to0.
\end{aligned}$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

For Λ_{11}, we have

$$\sup_{l\in\mathbb{L}}\left|\frac{\varepsilon^\intercal P(l)^\intercal K_t P(l)\varepsilon}{\mathcal{R}_{t,T}(l)}\right| \le \sup_{l\in\mathbb{L}}\left|\overline{K}\frac{\varepsilon^\intercal P(l)^\intercal P(l)\varepsilon}{\mathcal{R}_{t,T}(l)}\right|.$$

For Λ_{13}, using the triangle inequality, we have

$$\begin{aligned}\left|\frac{\mu^\intercal A(l)^\intercal \mathbf{K}_t P(l)\varepsilon}{\mathcal{R}_{t,T}(l)}\right| &\le \left[\overline{K}\frac{\mu^\intercal A(l)^\intercal \mathbf{K}_t A(l)\mu}{\mathcal{R}_{t,T}(l)}\frac{\varepsilon^\intercal P(l)^\intercal P(l)\varepsilon}{\mathcal{R}_{t,T}(l)}\right]^{1/2}\\ &\le \left[\overline{K}\frac{\varepsilon^\intercal P(l)^\intercal P(l)\varepsilon}{\mathcal{R}_{t,T}(l)}\right]^{1/2}.\end{aligned}$$

Based on Markov inequality and Assumption A.5(i), for any $\delta > 0$,

$$\begin{aligned}\Pr\left(\sup_{l\in\mathbb{L}}\frac{\varepsilon^\intercal P(l)^\intercal P(l)\varepsilon}{\mathcal{R}_{t,T}(l)} \ge \delta\right) &\le \Pr\left(\varepsilon^\intercal P(l)^\intercal P(l)\varepsilon \ge \delta\xi_{t,T}\right)\\ &\le E\left[\varepsilon^\intercal P(l)^\intercal P(l)\varepsilon\right]\delta^{-1}\xi_{t,T}^{-1}\\ &= \operatorname{tr}\left(P(l)^\intercal P(l)\Omega\right)\delta^{-1}\xi_{t,T}^{-1}\\ &\le \zeta(\Omega)\operatorname{tr}\left(P(l)^\intercal P(l)\right)\delta^{-1}\xi_{t,T}^{-1} \to 0.\end{aligned}$$

Hence, we have $\Lambda_{11} = o_p(1)$ and $\Lambda_{13} = o_p(1)$. This completes the proof of Equation (A.1).

We next show (A.2). By the triangle inequality,

$$\begin{aligned}\sup_{l\in\mathbb{L}}\left|\frac{\tilde{\mathcal{R}}_{t,T}(l)}{\mathcal{R}_{t,T}(l)} - 1\right| &= \sup_{l\in\mathbb{L}}\frac{\left|\tilde{\mathcal{R}}_{t,T}(l) - \mathcal{R}_{t,T}(l)\right|}{\mathcal{R}_{t,T}(l)}\\ &\le \sup_{l\in\mathbb{L}}\frac{\left|\mu^\intercal A(l)^\intercal \mathbf{K}_t A(l)\mu - \mu^\intercal \tilde{A}(l)^\intercal \mathbf{K}_t \tilde{A}(l)\mu\right|}{\mathcal{R}_{t,T}(l)}\\ &\quad + \sup_{l\in\mathbb{L}}\frac{\operatorname{tr}(\tilde{P}(l)\mathbf{K}_t\tilde{P}(l)^\intercal\Omega)}{\mathcal{R}_{t,T}(l)} + \sup_{l\in\mathbb{L}}\frac{\operatorname{tr}(P(l)\mathbf{K}_t P(l)^\intercal\Omega)}{\mathcal{R}_{t,T}(l)}\\ &:= \Lambda_{21} + \Lambda_{22} + \Lambda_{23}.\end{aligned}$$

The proof of $\Lambda_{23} = o_p(1)$ is similar to that of $\Lambda_{12} = o_p(1)$ and is omitted. We only need to show $\Lambda_{21} \to 0$ and $\Lambda_{22} \to 0$.

For Λ_{21}, we decompose it into

$$\begin{aligned}\Lambda_{21} &= \sup_{l\in\mathbb{L}}\frac{\left|\frac{1}{M^2}\sum_{m=1}^M\sum_{s=1}^M\left(\mu^\intercal\tilde{A}_{(m,l)}^\intercal\mathbf{K}_t\tilde{A}_{(s,l)}\mu - \mu^\intercal A_{(m,l)}^\intercal\mathbf{K}_t A_{(s,l)}\mu\right)\right|}{\mathcal{R}_{t,T}(l)}\\ &= \sup_{l\in\mathbb{L}}\frac{\left|\frac{1}{M^2}\sum_{m=1}^M\sum_{s=1}^M\left(\mu^\intercal(A_{(m,l)}+H_{(m,l)}A_{(m,l)})^\intercal\mathbf{K}_t(A_{(s,l)}+H_{(s,l)}A_{(s,l)})\mu - \mu^\intercal A_{(m,l)}^\intercal\mathbf{K}_t A_{(s,l)}\mu\right)\right|}{\mathcal{R}_{t,T}(l)}\\ &= \sup_{l\in\mathbb{L}}\frac{\left|\frac{1}{M^2}\sum_{m=1}^M\sum_{s=1}^M\left(\mu^\intercal A_{(m,l)}^\intercal H_{(m,l)}^\intercal\mathbf{K}_t H_{(s,l)}A_{(s,l)}\mu + 2\frac{1}{M^2}\sum_{m=1}^M\sum_{s=1}^M\mu^\intercal A_{(m,l)}^\intercal H_{(m,l)}^\intercal\mathbf{K}_t A_{(s,l)}\mu\right)\right|}{\mathcal{R}_{t,T}(l)}\\ &:= \Lambda_{211} + \Lambda_{212}.\end{aligned}$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

We use $\zeta(AB) \leq \zeta(A)\zeta(B)$ and $\zeta(A+B) \leq \zeta(A)+\zeta(B)$ for square matrices A and B with identical dimensions.[9] Let $\tilde{p} = p^*_{tt}/(1-p^*_{tt})$. By Assumptions A.5(iii) and (iv), we obtain

$$\begin{aligned}
\Lambda_{211} &= \sup_{l\in\mathbb{L}} \frac{\left|\frac{1}{M^2}\sum_{m=1}^{M}\sum_{s=1}^{M}\mu^{\intercal}A_{(m,l)}^{\intercal}H_{(m,l)}^{\intercal}\mathbf{K}_t H_{(s,l)}A_{(s,l)}\mu\right|}{\mathcal{R}_{t,T}(l)} \\
&\leq \frac{1}{2}\frac{\overline{K}}{M^2}\sum_{m=1}^{M}\sum_{s=1}^{M}\mu^{\intercal}\left(A_{(m,l)}^{\intercal}H_{(m,l)}^{\intercal}H_{(s,l)}A_{(s,l)} + A_{(s,l)}^{\intercal}H_{(s,l)}^{\intercal}H_{(m,l)}A_{(m,l)}\right)\mu/\mathcal{R}_{t,T}(l) \\
&\leq \frac{1}{2}\overline{K}\xi_{t,T}^{-1}\max_{1\leq m\leq M}\max_{1\leq s\leq M}\mu^{\intercal}\zeta(A_{(m,l)}^{\intercal}H_{(m,l)}^{\intercal}H_{(s,l)}A_{(s,l)})\mu \\
&\leq \overline{K}\xi_{t,T}^{-1}\tilde{p}^2\mu^{\intercal}\mu \to 0,
\end{aligned}$$

and

$$\begin{aligned}
\Lambda_{212} &= \sup_{l\in\mathbb{L}} \frac{\left|\frac{2}{M^2}\sum_{m=1}^{M}\sum_{s=1}^{M}\mu^{\intercal}A_{(m,l)}^{\intercal}H_{(m,l)}^{\intercal}\mathbf{K}_t A_{(s,l)}\mu\right|}{\mathcal{R}_{t,T}(l)} \\
&\leq \frac{\left(\frac{2}{M^2}\sum_{m=1}^{M}\sum_{s=1}^{M}\mu^{\intercal}A_{(m,l)}^{\intercal}H_{(m,l)}^{\intercal}\mathbf{K}_t H_{(s,l)}A_{(s,l)}\mu\mu^{\intercal}A(l)^{\intercal}\mathbf{K}_t A(l)\mu\right)^{1/2}}{\mathcal{R}_{t,T}(l)} \\
&\leq \left[\frac{\left(\frac{2}{M^2}\sum_{m=1}^{M}\sum_{s=1}^{M}\mu^{\intercal}A_{(m,l)}^{\intercal}H_{(m,l)}^{\intercal}\mathbf{K}_t H_{(s,l)}A_{(s,l)}\mu\right)}{\mathcal{R}_{t,T}(l)}\right]^{1/2} \to 0.
\end{aligned}$$

For Λ_{22}, we can obtain that

$$\begin{aligned}
\mathrm{tr}(\tilde{P}(l)\mathbf{K}_t\tilde{P}(l)^{\intercal}\Omega) &= \mathrm{tr}\left[(P(l)-H(l)+T(l))\mathbf{K}_t(P(l)-H(l)+T(l))^{\intercal}\Omega\right] \\
&\leq \mathrm{tr}(P(l)\mathbf{K}_t P(l)^{\intercal}\Omega) + \mathrm{tr}(H(l)\mathbf{K}_t H(l)^{\intercal}\Omega) + \mathrm{tr}(T(l)\mathbf{K}_t T(l)^{\intercal}\Omega) \\
&\quad + 2\left|\mathrm{tr}(H(l)\mathbf{K}_t T(l)^{\intercal}\Omega)\right| + 2\left|\mathrm{tr}(H(l)\mathbf{K}_t P(l)^{\intercal}\Omega)\right| \\
&\quad + 2\left|\mathrm{tr}(T(l)\mathbf{K}_t P(l)^{\intercal}\Omega)\right|.
\end{aligned}$$

Given that $\mathrm{rank}(P_{(m,l)}) \leq T$, $p^*_{tt} = O(T^{-1}b^{-1})$ a.s. and $\tilde{p} = O(T^{-1}b^{-1})$ a.s., we have

$$\begin{aligned}
\left|\mathrm{tr}(H(l)\mathbf{K}_t H(l)^{\intercal}\Omega)\right| &\leq \overline{K}\left|\mathrm{tr}(H(l)H(l)^{\intercal}\Omega)\right| \\
&\leq \overline{K}\zeta(H(l))\left|\mathrm{tr}(H(l)\Omega)\right| \\
&\leq \tilde{p}\overline{K}\left|\mathrm{tr}(H(l)\Omega)\right| \\
&\leq \frac{\tilde{p}}{1-p^*_{tt}}\zeta(\Omega)\overline{K}\max_{1\leq m\leq M}\mathrm{rank}(P_{(m,l)})\max_{1\leq m\leq M}\zeta(P_{(m,l)}) \\
&\leq O(b^{-1}\zeta(\Omega))\overline{K}\max_{1\leq m\leq M}\zeta(P_{(m,l)}),
\end{aligned}$$

[9] See Li (1987) for the proofs.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

$$
\begin{aligned}
\left|\operatorname{tr}(T(l)K_tT(l)^\intercal\Omega)\right| &\le \zeta(\Omega)\tilde{p}^2\overline{K}\operatorname{rank}(P(l)) \max_{1\le m\le M}\zeta^2(P_{(m,l)}) \\
&= O(T^{-1}b^{-2}\zeta(\Omega))\overline{K}\max_{1\le m\le M}\zeta^2(P_{(m,l)}), \\
\left|\operatorname{tr}(T(l)K_tH(l)^\intercal\Omega)\right| &\le \zeta(\Omega)\tilde{p}^2\overline{K}\operatorname{rank}(P(l)) \max_{1\le m\le M}\zeta(P_{(m,l)}) \\
&= O(T^{-1}b^{-2}\zeta(\Omega))\overline{K}\max_{1\le m\le M}\zeta(P_{(m,l)}), \\
\left|\operatorname{tr}(H(l)K_tP(l)^\intercal\Omega)\right| &\le \operatorname{rank}(P(l))\zeta(H(l)\mathbf{K}_tP(l)^\intercal\Omega) \\
&\le \tilde{p}\overline{K}\operatorname{rank}(P(l))\zeta(\Omega)\max_{1\le m\le M}\zeta(P_{(m,l)}) \\
&= O(b^{-1}\zeta(\Omega))\overline{K}\max_{1\le m\le M}\zeta(P_{(m,l)}), \\
\left|\operatorname{tr}(T(l)K_tP(l)^\intercal\Omega)\right| &\le \operatorname{rank}(P(l))\zeta(T(l)\mathbf{K}_tP(l)^\intercal\Omega) \\
&\le \operatorname{rank}(P(l))\overline{K}\zeta\left(\frac{1}{M}\sum_{m=1}^{M}P_{(m,l)}H_{(m,l)}\right)\zeta(P(l)^\intercal)\zeta(\Omega) \\
&\le \tilde{p}\overline{K}\zeta(\Omega)\max_{1\le m\le M}\zeta^2(P_{(m,l)}) \\
&= O(b^{-1}\zeta(\Omega))\overline{K}\max_{1\le m\le M}\zeta^2(P_{(m,l)}).
\end{aligned}
$$

These results, together with Assumptions A.4 and A.5(iv), imply that $\Lambda_{22} = o_p(1)$.

To prove (A.3), we use arguments similar to the proof of (A.1):

$$
\begin{aligned}
\sup_{l\in\mathbb{L}}\left|\frac{\tilde{\mathcal{L}}_{t,T}(l)-\tilde{\mathcal{R}}_{t,T}(l)}{\tilde{\mathcal{R}}_{t,T}(l)}\right| &\le \sup_{l\in\mathbb{L}}\left|\frac{\varepsilon^\intercal\tilde{P}(l)^\intercal\mathbf{K}_t\tilde{P}(l)\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right| + \sup_{l\in\mathbb{L}}\left|\frac{\operatorname{tr}(\tilde{P}(l)^\intercal\mathbf{K}_t\tilde{P}(l)\Omega)}{\tilde{\mathcal{R}}_{t,T}(l)}\right| \\
&\quad + 2\sup_{l\in\mathbb{L}}\left|\frac{\mu^\intercal\tilde{A}(l)^\intercal\mathbf{K}_t\tilde{P}(l)\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right| \\
&:= \Lambda_{31}+\Lambda_{32}+\Lambda_{33}.
\end{aligned}
$$

For Λ_{33}, using the triangle inequality, we have

$$
\begin{aligned}
\left|\frac{\mu^\intercal\tilde{A}(l)^\intercal\mathbf{K}_t\tilde{P}(l)\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right| &\le \left[\frac{\mu^\intercal\tilde{A}(l)^\intercal\mathbf{K}_t\tilde{A}(l)\mu}{\tilde{\mathcal{R}}_{t,T}(l)}\,\frac{\varepsilon^\intercal\tilde{P}(l)^\intercal\mathbf{K}_t\tilde{P}(l)\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right]^{1/2} \\
&\le \left[\frac{\varepsilon^\intercal\tilde{P}(l)^\intercal\mathbf{K}_t\tilde{P}(l)\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right]^{1/2}.
\end{aligned}
$$

Combining with the definition of $\tilde{P}(l)$ and the triangle inequality, we have

$$
\begin{aligned}
\left|\varepsilon^\intercal\tilde{P}(l)^\intercal\mathbf{K}_t\tilde{P}(l)\varepsilon\right| &\le \left|\varepsilon^\intercal P(l)^\intercal\mathbf{K}_tP(l)\varepsilon\right| + \left|\varepsilon^\intercal H(l)^\intercal\mathbf{K}_tH(l)\varepsilon\right| + \left|\varepsilon^\intercal T(l)^\intercal\mathbf{K}_tT(l)\varepsilon\right| \\
&\quad + 2\left|\varepsilon^\intercal P(l)^\intercal\mathbf{K}_tT(l)\varepsilon\right| + 2\left|\varepsilon^\intercal P(l)^\intercal\mathbf{K}_tH(l)\varepsilon\right| + 2\left|\varepsilon^\intercal H(l)^\intercal\mathbf{K}_tT(l)\varepsilon\right|,
\end{aligned}
$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

where

$$\begin{aligned}
&\left|\varepsilon^{\intercal}P(l)^{\intercal}\mathbf{K}_tT(l)\varepsilon\right| \le \left[\varepsilon^{\intercal}P(l)^{\intercal}\mathbf{K}_tP(l)\varepsilon\varepsilon^{\intercal}T(l)^{\intercal}\mathbf{K}_tT(l)\varepsilon\right]^{1/2},\\
&\left|\varepsilon^{\intercal}P(l)^{\intercal}\mathbf{K}_tH(l)\varepsilon\right| \le \left[\varepsilon^{\intercal}P(l)^{\intercal}\mathbf{K}_tP(l)\varepsilon\varepsilon^{\intercal}H(l)^{\intercal}\mathbf{K}_tH(l)\varepsilon\right]^{1/2},\\
&\left|\varepsilon^{\intercal}H(l)^{\intercal}\mathbf{K}_tT(l)\varepsilon\right| \le \left[\varepsilon^{\intercal}H(l)^{\intercal}\mathbf{K}_tH(l)\varepsilon\varepsilon^{\intercal}T(l)^{\intercal}\mathbf{K}_tT(l)\varepsilon\right]^{1/2}.
\end{aligned}$$

Based on Markov inequality, we have that for any given $\delta > 0$,

$$\begin{aligned}
\Pr\left(\sup_{l\in\mathbb{L}}\frac{\varepsilon^{\intercal}H(l)^{\intercal}\mathbf{K}_tH(l)\varepsilon}{\mathcal{R}_{t,T}(l)} \ge \delta\right) &\le \overline{K}\Pr\left(\varepsilon^{\intercal}H(l)^{\intercal}H(l)\varepsilon \ge \delta\xi_{t,T}\right)\\
&\le \overline{K}E\left[\varepsilon^{\intercal}H(l)^{\intercal}H(l)\varepsilon\right]\delta^{-1}\xi_{t,T}^{-1}\\
&\le \overline{K}\zeta(\Omega)\tilde{p}^2\delta^{-1}\xi_{t,T}^{-1} \to 0,\\
\Pr\left(\sup_{l\in\mathbb{L}}\frac{\varepsilon^{\intercal}T(l)^{\intercal}\mathbf{K}_tT(l)\varepsilon}{\mathcal{R}_{t,T}(l)} \ge \delta\right) &\le \overline{K}\Pr\left(\varepsilon^{\intercal}T(l)^{\intercal}T(l)\varepsilon \ge \delta\xi_{t,T}\right)\\
&\le \overline{K}E\left[\varepsilon^{\intercal}P(l)^{\intercal}H(l)^{\intercal}H(l)P(l)\varepsilon\right]\delta^{-1}\xi_{t,T}^{-1}\\
&\le \overline{K}\zeta(\Omega)\max_{1\le l\le L}\max_{1\le m\le M}\zeta(P_{(m,l)})\tilde{p}^2\delta^{-1}\xi_{t,T}^{-1} \to 0.
\end{aligned}$$

Then we obtain $\Lambda_{33} = \Lambda_{31} = o_p(1)$. The proof of $\Lambda_{32} = o_p(1)$ is similar to that of $\Lambda_{22} = o_p(1)$. This completes the proof of Equation (A.3).

Now, we show (A.4). From Chebyshev's inequality and Assumption A.5(ii), we obtain

$$\begin{aligned}
\Pr\left(\sup_{l\in\mathbb{L}}\left|\frac{\mu^{\intercal}\tilde{A}(l)\mathbf{K}_t\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right| \ge \delta\right) &\le \frac{1}{M}\sum_{m=1}^{M}\Pr\left\{\left|\mu^{\intercal}\tilde{A}_{(m,l)}\mathbf{K}_t\varepsilon\right| \ge \delta\xi_{t,T}\right\}\\
&\le \delta^{-2}\xi_{t,T}^{-2}\frac{1}{M}\sum_{m=1}^{M}E\left(\mu^{\intercal}\tilde{A}_{(m,l)}\mathbf{K}_t\varepsilon\right)^2\\
&\le \delta^{-2}\xi_{t,T}^{-2}\frac{1}{M}\sum_{m=1}^{M}\mathrm{tr}\left(\mu^{\intercal}\tilde{A}_{(m,l)}\mathbf{K}_t\Omega K_t\tilde{A}_{(m,l)}^{\intercal}\mu\right)\\
&\le \overline{K}^2\zeta(\Omega)\delta^{-2}\xi_{t,T}^{-2}\frac{1}{M}\sum_{m=1}^{M}\mathcal{R}_{t,T(m,l)} \to 0,
\end{aligned}$$

where $\mathcal{R}_{t,T(m,l)}$ denotes the local risk of the (m,l)th model.

Finally, we show (A.5). Note that

$$\begin{aligned}
\sup_{l\in\mathbb{L}}\left|\frac{\varepsilon^{\intercal}\mathbf{K}_t\tilde{P}(l)\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right| &\le \sup_{l\in\mathbb{L}}\left|\frac{\varepsilon^{\intercal}\mathbf{K}_tP(l)\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right| + \sup_{l\in\mathbb{L}}\left|\frac{\varepsilon^{\intercal}\mathbf{K}_tH(l)A(l)\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right|\\
&:= \Lambda_{51} + \Lambda_{22}.
\end{aligned}$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Based on Assumptions A.4 and A.5, we can show

$$\begin{aligned}\Pr\left(\sup_{l\in\mathbb{L}}\left|\frac{\varepsilon^{\intercal}\mathbf{K}_t P(l)\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right|\geq\delta\right) &\leq \frac{1}{M}\sum_{m=1}^{M}\Pr\left(\sup_{l\in\mathbb{L}}\left|\frac{\varepsilon^{\intercal}K_t P_{(m,l)}\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right|\geq\delta\right)\\ &\leq \delta^{-2}\xi_{t,T}^{-2}\frac{1}{M}\sum_{m=1}^{M}E\left[\varepsilon^{\intercal}\mathbf{K}_t P_{(m,l)}\varepsilon\right]^2\\ &\leq \delta^{-2}\xi_{t,T}^{-2}\frac{1}{M}\sum_{m=1}^{M}\operatorname{tr}\left(P_{(m,l)}\mathbf{K}_t P_{(m,l)}\Omega\right)\\ &\leq \overline{K}\delta^{-2}\xi_{t,T}^{-2}\zeta(\Omega)\frac{1}{M}\sum_{m=1}^{M}\operatorname{tr}\left(P_{(m,l)}P_{(m,l)}\right)\to 0,\end{aligned}$$

and

$$\begin{aligned}\Pr\left(\sup_{l\in\mathbb{L}}\left|\frac{\varepsilon^{\intercal}\mathbf{K}_t H(l)A(l)\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right|\right) &\leq \frac{1}{M}\sum_{m=1}^{M}\Pr\left(\left|\frac{\varepsilon^{\intercal}\mathbf{K}_t H_{(m,l)}A_{(m,l)}\varepsilon}{\tilde{\mathcal{R}}_{t,T}(l)}\right|\geq\delta\right)\\ &\leq \frac{1}{M}\sum_{m=1}^{M}\Pr\left(\left|\varepsilon^{\intercal}\mathbf{K}_t H_{(m,l)}A_{(m,l)}\varepsilon\right|\geq\delta\xi_{t,T}\right)\\ &\leq \overline{K}\frac{1}{M}\sum_{m=1}^{M}\Pr\left(\left|\zeta(H_{(m,l)}A_{(m,l)})\varepsilon^{\intercal}\varepsilon\right|\geq\delta\xi_{t,T}\right)\\ &\leq \tilde{p}\overline{K}\Pr\left(\varepsilon^{\intercal}\varepsilon\geq\delta\xi_{t,T}\right)\\ &\leq \overline{K}\zeta(\Omega)\delta^{-1}b^{-1}\xi_{t,T}^{-1}\to 0.\end{aligned}$$

This yields that $\Lambda_{51}=o_p(1)$ and $\Lambda_{52}=o_p(1)$. Therefore, the proof of Theorem 4 is completed.

Proof of Theorem 5.

(i) The second-order Taylor expansion of $F_{t,T}(\overline{\mathbf{w}})$ at $\mathbf{w}_t^*$ is considered:

$$\begin{aligned}F_{t,T}(\overline{\mathbf{w}}) &= F_{t,T}(\mathbf{w}_t^*)+\nabla_1 F_{t,T}^{\intercal}(\mathbf{w}_t^*)(\overline{\mathbf{w}}-\mathbf{w}_t^*)+\frac{1}{2}(\overline{\mathbf{w}}-\mathbf{w}_t^*)^{\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})(\overline{\mathbf{w}}-\mathbf{w}_t^*)\\ &= F_{t,T}(\mathbf{w}_t^*)-\tilde{\lambda}\mathbf{1}_M^{\intercal}(\overline{\mathbf{w}}-\mathbf{w}_t^*)+\frac{1}{2}(\overline{\mathbf{w}}-\mathbf{w}_t^*)^{\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})(\overline{\mathbf{w}}-\mathbf{w}_t^*)\\ &= F_{t,T}(\mathbf{w}_t^*)+\frac{1}{2}(\overline{\mathbf{w}}-\mathbf{w}_t^*)^{\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})(\overline{\mathbf{w}}-\mathbf{w}_t^*),\end{aligned}$$

where $\widetilde{\mathbf{w}}_t$ lies between $\mathbf{w}_t^*$ and $\overline{\mathbf{w}}$, and $\tilde{\lambda}$ is the Lagrange multiplier from the optimization problem $\mathbf{w}_t^* := \arg\min_{\mathbf{w}_t\in\mathcal{W}} F_{t,T}(\mathbf{w}_t)$ with constraint $\mathcal{W}=\{\mathbf{w}_t\in[0,1]^M : \sum_{m=1}^{M} w^m = 1\}$. The last equality follows from the normalization of the weights. For the second term on the

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

right-hand side of the last equation, we have

$$\begin{aligned}
&\max_{1\le t\le T}\left|F_{t,T}(\overline{\mathbf{w}})-F_{t,T}(\mathbf{w}_t^*)\right| \\
&\quad=\max_{1\le t\le T}\left|\frac{1}{2}(\overline{\mathbf{w}}-\mathbf{w}_t^*)^{\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})(\overline{\mathbf{w}}-\mathbf{w}_t^*)\right| \\
&\quad\le\max_{1\le t\le T}\limsup_{M}\sup_{\mathbf{w}_t}\lambda_{\max}\left(\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})\right)\frac{1}{2}\left|(\overline{\mathbf{w}}-\mathbf{w}_t^*)^{\intercal}(\overline{\mathbf{w}}-\mathbf{w}_t^*)\right| \\
&\quad=\frac{1}{2}\overline{\lambda}_{\max}\left|M^{-2}\mathbf{1}_M^{\intercal}\mathbf{1}_M-2M^{-1}\mathbf{1}_M^{\intercal}\mathbf{w}_t^*+\mathbf{w}_t^{*\intercal}\mathbf{w}_t^*\right| \\
&\quad=\frac{1}{2}\overline{\lambda}_{\max}\left|\mathbf{w}_t^{*\intercal}\mathbf{w}_t^*-M^{-1}\right| \\
&\quad\le\frac{1}{2}\overline{\lambda}_{\max}(1-M^{-1}).
\end{aligned}$$

(ii) Let $\hat{\eta}_t:=\widehat{\mathbf{w}}_t-\mathbf{w}_t^*$. As in the proof of Theorem 5(i), we have

$$F_{t,T}(\widehat{\mathbf{w}}_t)=F_{t,T}(\mathbf{w}_t^*)+\frac{1}{2}\hat{\eta}_t^{\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}}_t)\hat{\eta}_t.$$

Thus, Theorem 5(ii) follows from

$$\begin{aligned}
\min_{1\le t\le T}E\left|F_{t,T}(\widehat{\mathbf{w}}_t)-F_{t,T}(\mathbf{w}_t^*)\right| &= \min_{1\le t\le T}E\left|\frac{1}{2}\hat{\eta}_t^{\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}}_t)\hat{\eta}_t\right| \\
&\ge \min_{1\le t\le T}\liminf_{M}\inf_{\mathbf{w}_t}\lambda_{\min}\left(\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})\right)\frac{1}{2}E\left\|\hat{\eta}_t\right\|^2 \\
&= \frac{1}{2}\underline{\lambda}_{\min}E\left\|\hat{\eta}_t\right\|^2\ge\frac{1}{2}\underline{\lambda}_{\min}M\underline{\sigma}_{\eta}^2.
\end{aligned}$$

Proof of Corollary 1.

According to Theorem 5(i), the difference between $F_{t,T}(\overline{\mathbf{w}})$ and $F_{t,T}(\mathbf{w}_t^*)$ becomes

$$\begin{aligned}
&F_{t,T}(\overline{\mathbf{w}})-F_{t,T}(\mathbf{w}_t^*) \\
&\quad=\frac{1}{2}\left(\overline{\mathbf{w}}^{\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})\overline{\mathbf{w}}-2\overline{\mathbf{w}}^{\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})\mathbf{w}_t^*+\mathbf{w}_t^{*\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})\mathbf{w}_t^*\right),
\end{aligned}$$

and

$$\overline{\mathbf{w}}^{\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})\overline{\mathbf{w}}-2\overline{\mathbf{w}}^{\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})\mathbf{w}_t^*\le M^{-1}\overline{\lambda}_{max}\to 0 \text{ as } M\to\infty.$$

Now, we show $\mathbf{w}_t^{*\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})\mathbf{w}_t^*=o_p(1)$. For the local quadratic loss function $F_{t,T}(\mathbf{w}_t)=E(\mathbf{w}_t^{\intercal}e\mathbf{K}_t e^{\intercal}\mathbf{w}_t)$, we obtain a closed-form solution $\mathbf{w}_t^*=(\mathbf{1}_M^{\intercal}\Sigma_t^{-1}\mathbf{1}_M)^{-1}\Sigma_t^{-1}\mathbf{1}_M$. Therefore,

$$\begin{aligned}
\mathbf{w}_t^{*\intercal}\nabla_2 F_{t,T}(\widetilde{\mathbf{w}})\mathbf{w}_t^* &= \mathbf{w}_t^{*\intercal}\Sigma_t\mathbf{w}_t^*=(\mathbf{1}_M^{\intercal}\Sigma_t^{-1}\mathbf{1}_M)^{-1}\le\overline{\lambda}_{max}(\mathbf{1}_M^{\intercal}\mathbf{1}_M)^{-1} \\
&= M^{-1}\overline{\lambda}_{max}\to 0.
\end{aligned}$$

Therefore, we establish Corollary 1.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Proof of Theorem 6.

Let $F_{t(m,r)} = \Phi^{-1}(\Phi F_{t(m,r)} - \widehat{F}_{t(m,r)}) + \Phi^{-1}\widehat{F}_{t(m,r)}$, where the first term on the right is $o_p(1)$ because $\hat{w}_{t(m,r)} := \Phi F_{t(m,r)} - \hat{F}_{t(m,r)} = O_p\left(\max\left\{T^{-1}, L^{-1/2}\right\}\right) = o_p(1)$. Therefore, we can obtain $\hat{F}_{t(m,r)} = \Phi F_{t(m,r)} + o_p(1)$. Then, we have

$$
\begin{aligned}
\hat{\alpha}_{t(m,r)} &= \left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}\widehat{F}_{s(m,r)}^{\intercal}\right)^{-1}\left(\sum_{s=1}^{T} K_{st}\hat{F}_{s(m,r)}y_{s+h}\right) \\
&= \left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}\widehat{F}_{s(m,r)}^{\intercal}\right)^{-1}\left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}F_{s(m,r)}^{\intercal}\alpha_{s(m,r)}\right) \\
&\quad + \left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}\widehat{F}_{s(m,r)}^{\intercal}\right)^{-1}\left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}\epsilon_{s+h,(m,r)}\right) \\
&= \left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}\widehat{F}_{s(m,r)}^{\intercal}\right)^{-1}\left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}\widehat{F}_{s(m,r)}^{\intercal}\Phi^{-1}\alpha_{s(m,r)}\right) \\
&\quad + \left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}\widehat{F}_{s(m,r)}^{\intercal}\right)^{-1}\left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}\Phi^{-1}\hat{w}_{s(m,r)}\alpha_{s(m,r)}\right) \\
&\quad + \left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}\widehat{F}_{s(m,r)}^{\intercal}\right)^{-1}\left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}\epsilon_{s+h,(m,r)}\right) := I + II + III.
\end{aligned}
$$

For the first term I, by Taylor expansion of $\alpha_{s(m,r)}$, we have

$$
\begin{aligned}
I &= \left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}\widehat{F}_{s(m,r)}^{\intercal}\right)^{-1}\left(\sum_{s=1}^{T} K_{st}\widehat{F}_{s(m,r)}\widehat{F}_{s(m,r)}^{\intercal}\Phi^{-1}\alpha_{s(m,r)}\right) \\
&= \Phi^{-1}\alpha_{t(m,r)} + \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st}\Phi F_{s(m,r)}F_{s(m,r)}^{\intercal}\Phi^{\intercal}\right)^{-1} \\
&\quad \times \left(\frac{1}{T}\sum_{s=1}^{T} K_{st}\Phi F_{s(m,r)}F_{s(m,r)}^{\intercal}\alpha'_{t(m,r)}\left(\frac{s-t}{Tb}\right)\right) \\
&\quad + \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st}\Phi F_{s(m,r)}F_{s(m,r)}^{\intercal}\Phi^{\intercal}\right)^{-1} \\
&\quad \times \left(\frac{1}{Tb}\sum_{s=1}^{T} K_{st}\Phi F_{s(m,r)}F_{s(m,r)}^{\intercal}\left(\alpha''_{t(m,r)}\left(\frac{s-t}{T}\right)^2 + o(b^2)\right)\right) \\
&\xrightarrow{p} \Phi^{-1}\alpha_{t(m,r)} + b^2\mu_2\alpha''_{t(m,r)}\left[E(\Phi_0 F_{t(m,r)}F_{t(m,r)}^{\intercal}\Phi_0^{\intercal})\right]^{-1}E(\Phi_0 F_{t(m,r)}F_{t(m,r)}^{\intercal}) \\
&:= \Phi^{-1}\alpha_{t(m,r)} + b^2\varpi_{t(m,r)},
\end{aligned}
$$

where $\varpi_{t(m,r)}$ is implicitly defined.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

For the second term II, note that,

$$\begin{aligned}
&\frac{1}{\sqrt{Tb}}\sum_{s=1}^{T}K_{st}\widehat{F}_{s(m,r)}\hat{w}_{s(m,r)}\\
&\quad=\frac{1}{\sqrt{Tb}}\sum_{s=1}^{T}K_{st}\widehat{F}_{s(m,r)}(\Phi F_{s(m,r)}-\widehat{F}_{s(m,r)})^{\intercal}\\
&\quad=\frac{1}{\sqrt{Tb}}\sum_{s=1}^{T}K_{st}\left(\Phi F_{s(m,r)}-(\Phi F_{s(m,r)}-\widehat{F}_{s(m,r)})\right)(\Phi F_{s(m,r)}-\widehat{F}_{s(m,r)})^{\intercal}\\
&\quad=\frac{1}{\sqrt{Tb}}\sum_{s=1}^{T}K_{st}\Phi F_{s(m,r)}(\Phi F_{s(m,r)}-\widehat{F}_{s(m,r)})^{\intercal}\\
&\qquad-\frac{1}{\sqrt{Tb}}\sum_{s=1}^{T}K_{st}\left(\Phi F_{s(m,r)}-\widehat{F}_{s(m,r)}\right)(\Phi F_{s(m,r)}-\widehat{F}_{s(m,r)})^{\intercal}\\
&\quad:=II_1+II_2.
\end{aligned}$$

It follows that $II_1+II_2=o_p(1)$ when $T/(L^2b)\to 0$ by Lemma A.3(i)–(ii).

For the third term III, using the fact that $\widehat{F}_{s(m,r)}=\Phi F_{s(m,r)}+o_p(1)$, we have

$$\begin{aligned}
III&=\left(\sum_{s=1}^{T}K_{st}\widehat{F}_{s(m,r)}\widehat{F}_{s(m,r)}^{\intercal}\right)^{-1}\left(\sum_{s=1}^{T}K_{st}\widehat{F}_{s(m,r)}\epsilon_{s+h,(m,r)}\right)\\
&=\left(\sum_{s=1}^{T}K_{st}\Phi F_{s(m,r)}F_{s(m,r)}^{\intercal}\Phi^{\intercal}\right)^{-1}\left(\sum_{s=1}^{T}K_{st}\Phi F_{s(m,r)}\epsilon_{s+h,(m,r)}\right)+o_p(1).
\end{aligned}$$

Combing these results, we obtain,

$$\begin{aligned}
&\sqrt{Tb}\left(\hat{\alpha}_{t(m,r)}-\Phi^{-1}\alpha_{t(m,r)}-b^2\varpi_t\right)\\
&\quad=\left(\frac{1}{Tb}\sum_{s=1}^{T}K_{st}\Phi F_{s(m,r)}F_{s(m,r)}^{\intercal}\Phi^{\intercal}\right)^{-1}\left(\frac{1}{\sqrt{Tb}}\sum_{s=1}^{T}K_{st}\Phi F_{s(m,r)}\epsilon_{s+h,(m,r)}\right)+o_p(1)\\
&\quad\xrightarrow{d}\left(\Phi_0\Sigma_{F_t}\Phi_0^{\intercal}\right)^{-1}\Phi_0 N\left(0,\Sigma_{F_t,\varepsilon}\right)\\
&\quad=N\left(0,\Phi_0^{\intercal-1}\Sigma_{F_t}^{-1}\Sigma_{F_t,\varepsilon}\Sigma_{F_t}^{-1}\Phi_0^{-1}\right).
\end{aligned}$$

This completes the proof of Theorem 7.

Proof of Theorem 7.

Rewrite $\tilde{P}_{(m,r)}=\tilde{D}_{(m,r)}(\hat{P}_{(m,r)}-I)+I$ as $\tilde{P}_{(m,r)}=\hat{P}_{(m,r)}+T_{(m,r)}-Q_{(m,r)}$, where $Q_{(m,r)}=\tilde{D}_{(m,r)}-I$ and $T_{(m,r)}=Q_{(m,r)}\hat{P}_{(m,r)}$. Let $\tilde{P}(r)=\frac{1}{M}\sum_{m=1}^{M}\tilde{P}_{(m,r)}=\hat{P}(r)+T(r)-Q(r)$, where $\hat{P}(r)=\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{(m,r)}$, $T(r)=\frac{1}{M}\sum_{m=1}^{M}T_{(m,r)}$, and $Q(r)=$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

$\frac{1}{M}\sum_{m=1}^{M} Q_{(m,r)}$. Following the proof of Theorem 4, it suffices to show that,

$$\sup_{l\in\mathbb{L}}\left|\frac{\mathcal{L}_{t,T}(l)}{\mathcal{R}_{t,T}(l)}-1\right|\xrightarrow{p}0, \tag{B.1}$$

$$\sup_{l\in\mathbb{L}}\left|\frac{\tilde{\mathcal{R}}_{t,T}(l)}{\mathcal{R}_{t,T}(l)}-1\right|\xrightarrow{p}0, \tag{B.2}$$

$$\sup_{l\in\mathbb{L}}\left|\frac{\tilde{\mathcal{R}}_{t,T}(l)}{\tilde{\mathcal{L}}_{t,T}(l)}-1\right|\xrightarrow{p}0, \tag{B.3}$$

$$\sup_{r\in\mathcal{R}}\left|\frac{\mu^{\intercal}\tilde{P}^{\intercal}(r)\mathbf{K}_t\boldsymbol{\varepsilon}}{R_{t,T}(r)}\right|\xrightarrow{p}0, \tag{B.4}$$

$$\sup_{r\in\mathcal{R}}\left|\frac{\boldsymbol{\varepsilon}^{\intercal}\tilde{P}^{\intercal}(r)\mathbf{K}_t\boldsymbol{\varepsilon}}{R_{t,T}(r)}\right|\xrightarrow{p}0, \tag{B.5}$$

where $\mathcal{R}=\{1,\ldots,R\}$. The proofs of (B.1)–(B.3) are similar to those of (A.1)–(A.3) and are omitted.

We then show (B.4). By the triangle inequality, we have

$$\begin{aligned}\frac{|\mu^{\intercal}\tilde{P}^{\intercal}(r)\mathbf{K}_t\boldsymbol{\varepsilon}|}{R_{t,T}(r)} &\le \frac{|\mu^{\intercal}\hat{P}^{\intercal}(r)\mathbf{K}_t\boldsymbol{\varepsilon}|}{R_{t,T}(r)}+\frac{|\mu^{\intercal}T(r)^{\intercal}\mathbf{K}_t\boldsymbol{\varepsilon}|}{R_{t,T}(r)}+\frac{|\mu^{\intercal}Q(r)^{\intercal}\mathbf{K}_t\boldsymbol{\varepsilon}|}{R_{t,T}(r)}\\ &:= I+II+III.\end{aligned}$$

For the first term I, using the trace inequality that $tr(AB)\le tr(A)\lambda_{\max}(B)$, we have

$$\begin{aligned}\frac{|\mu^{\intercal}\hat{P}^{\intercal}(r)\mathbf{K}_t\boldsymbol{\varepsilon}|}{R_{t,T}(r)} &\le \xi_{t,T}^{-1}|\boldsymbol{\varepsilon}^{\intercal}\mathbf{K}_t\hat{P}(r)\mu\mu^{\intercal}\hat{P}^{\intercal}(r)\mathbf{K}_t\boldsymbol{\varepsilon}|^{1/2}\\ &\le \mathrm{tr}(\boldsymbol{\varepsilon}^{\intercal}\mathbf{K}_t\hat{P}(r)\hat{P}^{\intercal}(r)\mathbf{K}_t\boldsymbol{\varepsilon})^{1/2}\lambda_{\max}^{1/2}(\xi_{t,T}^{-2}\mu\mu^{\intercal})=o_p(1).\end{aligned}$$

To see this, we let u_{ij} be the (i,j)th element of the matrix $\mathbf{K}_t\hat{P}(r)\hat{P}^{\intercal}(r)\mathbf{K}_t$. From Assumption A.8(i), we obtain

$$\begin{aligned}u_{ij} &= K_{it}K_{jt}\widehat{F}_{i(m,r)}^{\intercal}\left(\widehat{F}_{(m,r)}^{\intercal}K_i\widehat{F}_{(m,r)}\right)^{-1}\\ &\quad\times\sum_{s=1}^{T}\widehat{F}_{s(m,r)}K_{is}K_{js}\widehat{F}_{s(m,r)}^{\intercal}\left(\widehat{F}_{(m,r)}^{\intercal}K_j\widehat{F}_{(m,r)}\right)^{-1}\widehat{F}_{j(m,r)}\\ &= \frac{K_{it}K_{jt}}{T^2b^2}\widehat{F}_{i(m,r)}^{\intercal}\widehat{\Sigma}_{i(m,r)}^{-1}\sum_{s=1}^{T-l}\widehat{F}_{s(m,r)}K_{is}K_{js}\widehat{F}_{s(m,r)}^{\intercal}\widehat{\Sigma}_{j(m,r)}^{-1}\widehat{F}_{j(m,r)}\\ &\le \frac{\overline{K}K_{it}K_{jt}}{Tb}\widehat{F}_{i(m,r)}^{\intercal}\widehat{\Sigma}_{i(m,r)}^{-1}\frac{1}{Tb}\sum_{s=1}^{T}\widehat{F}_{s(m,r)}K_{is}\widehat{F}_{s(m,r)}^{\intercal}\widehat{\Sigma}_{j(m,r)}^{-1}\widehat{F}_{j(m,r)}\end{aligned}$$

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

$$\xrightarrow{p} \frac{\overline{K}K_{it}K_{jt}}{Tb}\widehat{F}_{i(m,r)}^{\mathsf{T}}\Sigma_{i(m,r)}^{-1}\Sigma_{i(m,r)}\Sigma_{j(m,r)}^{-1}\widehat{F}_{j(m,r)}(1+o_p(1))$$

$$\leq \lambda_{\max}(\Sigma_{j(m,r)}^{-1})\frac{\overline{K}K_{it}K_{jt}}{Tb}\widehat{F}_{i(m,r)}^{\mathsf{T}}\widehat{F}_{j(m,r)},$$

where $\widehat{\Sigma}_{i(m,r)} = \widehat{F}_{(m,r)}^{\mathsf{T}}K_i\widehat{F}_{(m,r)}/(Tb)$. Thus,

$$\begin{aligned}\mathrm{tr}(\boldsymbol{\varepsilon}^{\mathsf{T}}\mathbf{K}_t\hat{P}(r)\hat{P}^{\mathsf{T}}(r)\mathbf{K}_t\boldsymbol{\varepsilon})^{1/2} &\leq \left[\overline{K}\lambda_{\max}(\Sigma_{j(m,r)}^{-1})T^{-1}b^{-1}(\boldsymbol{\varepsilon}^{\mathsf{T}}\mathbf{K}_t\widehat{F}_{(m,r)}^{\mathsf{T}}\widehat{F}_{(m,r)}\mathbf{K}_t\boldsymbol{\varepsilon})\right]^{1/2}\\ &= O_p(r^{1/2}).\end{aligned}$$

Then, by Assumption A.9(iii), we have $I = o_p(1)$ holds true. The second term $II = o_p(1)$ can be expressed in a similar manner. For the third term III, by Assumption A.9(iv), we have

$$\frac{|\mu^{\mathsf{T}}Q(r)^{\mathsf{T}}\mathbf{K}_t\boldsymbol{\varepsilon}|}{R_{t,T}(r)} \leq \xi_{t,T}^{-1}Tbp_{tt}^{*}\mathrm{tr}(T^{-1}b^{-1}\mu^{\mathsf{T}}\mathbf{K}_t\varepsilon) = o_p(1).$$

We finally show (B.5). By the triangle inequality, we have

$$\begin{aligned}\sup_{r\in\mathcal{R}}\left|\frac{\boldsymbol{\varepsilon}^{\mathsf{T}}\tilde{P}^{\mathsf{T}}(r)\mathbf{K}_t\boldsymbol{\varepsilon}}{R_{t,T}(r)}\right| &\leq \sup_{r\in\mathcal{R}}\frac{|\boldsymbol{\varepsilon}^{\mathsf{T}}\hat{P}(r)\mathbf{K}_t\boldsymbol{\varepsilon}|}{R_{T,t}(r)} + \sup_{r\in\mathcal{R}}\frac{|\boldsymbol{\varepsilon}^{\mathsf{T}}(T(r)-Q(r))^{\mathsf{T}}\mathbf{K}_t\boldsymbol{\varepsilon}|}{R_{T,t}(r)}\\ &\leq o_p(1)+\xi_{t,T}^{-1}\sup_{r\in\mathcal{R}}\frac{1}{M}\sum_{m=1}^{M}|\boldsymbol{\varepsilon}^{\mathsf{T}}(T_{(m,r)}-Q_{(m,r)})^{\mathsf{T}}\mathbf{K}_t\boldsymbol{\varepsilon}|\\ &\leq o_p(1)+C\xi_{t,T}^{-1}p_{tt}^{*}\mathrm{tr}(\boldsymbol{\varepsilon}^{\mathsf{T}}\mathbf{K}_t\boldsymbol{\varepsilon})\\ &= o_p(1)+C(r\xi_{t,T}^{-1})(r^{-1}Tbp_{tt}^{*})(T^{-1}b^{-1}\boldsymbol{\varepsilon}^{\mathsf{T}}\mathbf{K}_t\boldsymbol{\varepsilon}) = o_p(1),\end{aligned}$$

where the last equality holds by using Assumption A.9 and the fact that $T^{-1}b^{-1}\boldsymbol{\varepsilon}^{\mathsf{T}}\mathbf{K}_t\boldsymbol{\varepsilon} = O_p(1)$. Therefore, we complete the proof of Theorem 7.

SUPPLEMENTARY MATERIAL

Li, H., J. Zhang, X. Chen and Y. Hong (2025): Supplement to "Time-Varying Complete Subset Averaging in a Data-Rich Environment," Econometric Theory Supplementary Material. To view, please visit https://doi.org/10.1017/S0266466625000064.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B., & Csaki, F. (Eds.), *Second international symposium on information theory* (pp. 267–281). Akademiai Kiado.

Ando, T., & Li, K. C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109, 254–265.

Ang, A., & Bekaert, G. (2007). Stock return predictability: Is it there? *The Review of Financial Studies*, 20(3), 651–707.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1), 191–221.
Bai, J., & Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4), 1133–1150.
Boyd, J. H., Hu, J., & Jagannathan, R. (2005). The stock market's reaction to unemployment news: Why bad news is usually good for stocks. *The Journal of Finance*, 60(2), 649–672.
Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, 53(2), 603–618.
Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics*, 136(1), 163–188.
Cai, Z., & Tiwari, R. C. (2000). Application of a local linear autoregressive model to BOD time series. *Environmetrics*, 11(3), 341–350.
Campbell, J. Y., & Thompson, S. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4), 1509–1531.
Chan, J. C., Koop, G., Leon-Gonzalez, R., & Strachan, R. W. (2012). Time varying dimension models. *Journal of Business & Economic Statistics*, 30(3), 358–367.
Chava, S., Gallmeyer, M., & Park, H. (2015). Credit conditions and stock return predictability. *Journal of Monetary Economics*, 74, 117–132.
Chen, B., & Hong, Y. (2012). Testing for smooth structural changes in time series models via nonparametric regression. *Econometrica*, 80(3), 1157–1183.
Chen, B., & Hong, Y. (2016). Detecting for smooth structural changes in GARCH models. *Econometric Theory*, 32(3), 740–791.
Chen, B., & Maung, K. (2023). Time-varying forecast combination for high-dimensional data. *Journal of Econometrics*, 237(2), 105418.
Chen, J., Li, D., Linton, O., & Lu, Z. (2018). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association*, 113(522), 919–932.
Chen, Q., Hong, Y., & Li, H. (2024). Time-varying forecast combination for factor-augmented regressions with smooth structural changes. *Journal of Econometrics*, 240(1), 105693.
Cheng, X., & Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics*, 186(2), 280–293.
Claeskens, G., & Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98(464), 900–916.
Claeskens, G., Magnus, J., Vasnev, A., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32, 754–762.
Costa, A., Ferreira, P., Gaglianone, W., Guillén, O., Issler, J., & Lin, Y. (2021). Machine learning and oil price point and density forecasting. *Energy Economics*, 102, 105494.
Driesprong, G., Jacobsen, B., & Maat, B. (2008). Striking oil: Another puzzle? *Journal of Financial Economics*, 89(2), 307–327.
Elliott, G., Gargano, A., & Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, 177(2), 357–373.
Elliott, G., Gargano, A., & Timmermann, A. (2015). Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics and Control*, 54, 86–110.
Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.
Gao, Y., Zhang, X., Wang, S., & Zou, G. (2016). Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics*, 192(1), 139–151.
Garcia, M. G., Medeiros, M. C., & Vasconcelos, G. F. (2017). Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting*, 33(3), 679–693.
Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6), 1545–1578.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Hall, P., & Wehrly, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association*, 86(415), 665–672.

Hansen, B. E. (2001). The new econometrics of structural change: Dating breaks in US labour productivity. *Journal of Economic Perspectives*, 15(4), 117–128.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75(4), 1175–1189.

Hansen, B. E. (2008a). Least-squares forecast averaging. *Journal of Econometrics*, 146(2), 342–350.

Hansen, B. E. (2008b). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(3), 726–748.

Hansen, B. E., & Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1), 38–46.

Hirshleifer, D., Hou, K., & Teoh, S. H. (2009). Accruals, cash flows, and aggregate stock returns. *Journal of Financial Economics*, 91(3), 389–406.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4), 382–417.

Kabaila, P. (2002). On variable selection in linear regression. *Econometric Theory*, 18(4), 913–925.

Kelly, B., & Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 186(2), 294–316.

Lee, J. H., & Shin, Y. (2023). Complete subset averaging for quantile regressions. *Econometric Theory*, 39(1), 146–188.

Leung, G., & Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8), 3396–3410.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15(8), 958–975.

Liao, J., Zong, X., Zhang, X., & Zou, G. (2019). Model averaging based on leave-subject-out cross-validation for vector autoregressions. *Journal of Econometrics*, 209(1), 35–60.

Liao, J., Zou, G., Gao, Y., & Zhang, X. (2021). Model averaging prediction for time series models with a diverging number of parameters. *Journal of Econometrics*, 223(1), 190–221.

Lin, C.-F. J., & Teräsvirta, T. (1994). Testing the constancy of regression parameters against continuous structural change. *Journal of Econometrics*, 62(2), 211–228.

Liu, C.-A., & Kuo, B.-S. (2016). Model averaging in predictive regressions. *The Econometrics Journal*, 19(2), 203–231.

Lu, X., & Su, L. (2015). Jackknife model averaging for quantile regressions. *Journal of Econometrics*, 188(1), 40–58.

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

McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4), 574–589.

McCracken, M. W., & Ng, S. (2021). FRED-QD: A quarterly database for macroeconomic research. *Review, Federal Reserve Bank of St. Louis*, 103(1), 1–44.

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1), 98–119.

Meligkotsidou, L., Panopoulou, E., Vrontos, I. D., & Vrontos, S. D. (2014). A quantile regression approach to equity premium prediction. *Journal of Forecasting*, 33(7), 558–576.

Meligkotsidou, L., Panopoulou, E., Vrontos, I. D., & Vrontos, S. D. (2021). Out-of-sample equity premium prediction: A complete subset quantile regression approach. *The European Journal of Finance*, 27(1-2), 110–135.

Ng, S. (2013). Variable selection in predictive regressions. In Elliott, G., & Timmermann, A. (Eds.), *Handbook of economic forecasting* (volume 2, pp. 752–789). Elsevier.

Phan, D. H. B., Sharma, S. S., & Narayan, P. K. (2015). Stock return forecasting: Some new evidence. *International Review of Financial Analysis*, 40, 38–51.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press

Phillips, P. C. B. (2015). Pitfalls and possibilities in predictive regression. *Journal of Financial Econometrics*, 13(3), 521–555.

Rapach, D., & Zhou, G. (2013). Forecasting stock returns. In Elliott, G., & Timmermann, A. (Eds.), *Handbook of economic forecasting* (volume 2, pp. 328–383). Elsevier.

Rapach, D. E., Strauss, J. K., & Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, 23(2), 821–862.

Robinson, P. M. (1989). Nonparametric estimation of time-varying parameters. In Hackl, P. (Ed.), *Statistical analysis and forecasting of economic structural change* (volume 19, pp. 253–264). Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Smith, J., & Wallis, K. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3), 331–355.

Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, 58(3), 644–719.

Stock, J. H., & Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, 14(1), 11–30.

Stock, J. H., & Watson, M. W. (1999). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In Engle, R., & White, H. (Eds.), *Cointegraion, causality, and forecasting: A festschrift in honour of Clive W.J. Granger*, pages 1–44. Oxford University Press.

Stock, J. H., & Watson, M. W. (2003). How did leading indicator forecasts perform during the 2001 recession? *FRB Richmond Economic Quarterly*, 89(3), 71–90.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6), 405–430.

Stock, J. H., & Watson, M. W. (2007). Why has US inflation become harder to forecast? *Journal of Money, Credit and Banking*, 39, 3–33.

Sun, Y., Hong, Y., Lee, T.-H., Wang, S., & Zhang, X. (2021). Time-varying model averaging. *Journal of Econometrics*, 222(2), 974–992.

Sun, Y., Hong, Y., Wang, S., & Zhang, X. (2023). Penalized time-varying model averaging. *Journal of Econometrics*, 235(2), 1355–1377.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267–288.

Wan, A. T., Zhang, X., & Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, 156(2), 277–283.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4), 1455–1508.

Yousuf, K., & Ng, S. (2021). Boosting high dimensional predictive regressions with time varying parameters. *Journal of Econometrics*, 224(1), 60–87.

Yuan, Z., & Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100(472), 1202–1214.

Zhang, X., Wan, A. T., & Zou, G. (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics*, 174(2), 82–94.

Zhang, X., Zou, G., Liang, H., & Carroll, R. J. (2020). Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association*, 115(530), 972–984.

https://doi.org/10.1017/S0266466625000064 Published online by Cambridge University Press