ARTICLE IN PRESS

Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Time-varying model averaging☆

Yuying Sun[a,b,c], Yongmiao Hong[d,e,f], Tae-Hwy Lee[g], Shouyang Wang[a,b,c], Xinyu Zhang[a,b,*]

[a] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China*
[b] *Center for Forecasting Science, Chinese Academy of Sciences, China*
[c] *School of Economics and Management, University of Chinese Academy of Sciences, China*
[d] *Department of Economics, Cornell University, United States of America*
[e] *Department of Statistics and Data Science, Cornell University, United States of America*
[f] *MOE Key Laboratory of Econometrics, Xiamen University, China*
[g] *Department of Economics, University of California, Riverside, United States of America*

ARTICLE INFO

Article history:
Received 5 December 2017
Received in revised form 15 May 2019
Accepted 11 February 2020
Available online xxxx

JEL classification:
C52
C53

Keywords:
Asymptotic optimality
Forecast combination
Local stationarity
Model averaging
Structural change
Time-varying model averaging

ABSTRACT

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

1. Introduction

Structural instability is a long-standing problem in time series econometrics (e.g., Stock and Watson (1996, 2002, 2005), Rossi (2006) and Rossi and Sekhposyan (2011)). Macroeconomic and financial time series, especially over a long

☆ We thank two referees, an associate editor, Oliver Linton (co-editor), Cheng Hsiao, Qingfeng Liu, Zudi Lu, Whitney Newey, Aman Ullah, Alan Wan, and workshop participants at Shanghai University of Finance and Economics, UC Riverside, 2nd International Conference on Econometrics and Statistics at Hong Kong, 4th Guangzhou Econometrics Workshop, 2nd Annual Forum for Chinese Econometricians at Beijing, 4th Annual International Conference on Quantitative Finance and Econometrics at Shenzhen, 12th Annual Society for Financial Econometrics Conference at Shanghai, and 2019 KER International Conference at Sogang University, Korea for their comments and suggestions. All remaining errors are solely ours. Sun thanks support from National Natural Science Foundation of China (NNSFC) Grant (Nos. 71703156 and 71973116) and Fujian Provincial Key Laboratory of Statistics, Xiamen University (No. 201601). Hong and Wang thank support from NNSFC Grant (No. 71988101) entitled as Econometric Modeling and Economic Policy Studies. Zhang thanks support from NNSFC Grant (Nos. 71925007 and 11688101), the Youth Innovation Promotion Association of the Chinese Academy of Sciences, the Beijing Academy of Artificial Intelligence, and a joint grant from Academy for Multidisciplinary Studies, Capital Normal University.

* Corresponding author at: Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China.
E-mail address: xinyu@amss.ac.cn (X. Zhang).

https://doi.org/10.1016/j.jeconom.2020.02.006

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

An approach to reducing the adverse impact of model instability and uncertainty is model averaging, which compromises across the competing models and yields an insurance against selecting a poor model. There has existed a relatively large literature on Bayesian model averaging; see Hoeting et al. (1999) for a comprehensive review. In recent years, frequentist model averaging has received growing attention in econometrics and statistics (e.g., Buckland et al. (1997), Yang (2001), Hjort and Claeskens (2003), Yuan and Yang (2005), Hansen (2007, 2008), Wan et al. (2010), Liu and Okui (2013), Liu (2015) and Chen et al. (2018)). Most of the works focus on model averaging weights determination, related inference, and asymptotic optimality. Recently, Hansen and Racine (2012) have proposed a jackknife model averaging (JMA) which selects model averaging weights by minimizing a cross-validation criterion. The advantage of the JMA estimator mainly lies in the asymptotic optimality theory established under heteroskedastic error settings. Zhang et al. (2013) broaden Hansen and Racine's (2012) scope of asymptotic optimality of the JMA estimator to encompass models with a non-spherical error covariance structure and lagged dependent variables, thus allowing for dependent data and dynamic regression models.

However, a potential problem with the aforementioned model averaging approaches is that, one predictive regression model may yield the best forecast in one period but can be dominated by other models in another period. This implies that optimal model averaging weights should change over time. There are various reasons for adopting this potentially useful time-varying approach. First, a time series model may suffer from structural instability in economics and finance. Therefore, as Stock and Watson (2003) point out, a predictor useful in one period does not guarantee its forecasting performance in other periods. The empirical results in Stock and Watson (2007) suggest that a substantial fraction of forecasting relations are unstable. Second, macroeconomic and financial series may follow different dynamics in different time periods. For example, they may have state-dependent dynamic structures. Third, because of possible collinearity among predictors, variable selection and model selection are inherently unstable (Stock and Watson, 2012). Thus, to handle such instability, it may be better to use time-varying weights instead of constant weights in model averaging. Furthermore, since the underlying economic structure is likely to be affected by technological progress, preference changes, policy switches, crises, and so on, it is desirable to use time-varying parameter models to capture structural changes. To our knowledge, there has been no work on selecting optimal time-varying weights in model averaging where each model itself may also have time-varying parameters.

The present paper fills this gap by proposing a time-varying jackknife model averaging (TVJMA) estimator that selects model averaging weights by minimizing a local cross-validation criterion. Our approach complements the existing literature on constant JMA weights and avoids the difficulty associated with whether structural changes exist. Specifically, we assume that model parameters, as well as model averaging weights, are smooth unknown functions of time. This approach is consistent with the evidence of types of instability documented in economics, namely smooth structural changes (e.g., Rothman (1998), Grant (2002) and Chen and Hong (2012), Chen (2015)). Hansen (2001) points out that it might seem more reasonable to allow a structural change to take effect with a period of time rather than to be effective immediately. To allow the weights in model averaging to change over time, we employ the local smoothing idea to the squared error loss, leading to a local constant model averaging estimator. Moreover, we follow the spirit of Robinson (1989) and use a local constant method to estimate the time-varying parameters in each candidate model. Furthermore, we extend the candidate models from static regressions to dynamic regressions, which cover more applications in economics and finance.

In this paper, we show that the proposed TVJMA estimator is asymptotically optimal in the sense of achieving the lowest possible local squared error loss in a class of time-varying model averaging estimators, under three model settings (i.e., Sections 2, 3, and 4). The first two settings (i.e., Sections 2 and 3) admit a non-diagonal covariance structure for regression errors, including heteroscedastic errors as in Hansen and Racine (2012), with exogenous regressors. As a result, we include the non-time-varying JMA estimator in Hansen and Racine (2012) as a special case of our TVJMA estimator, under heteroscedastic errors in a nested set-up. Our theoretical analysis allows the model averaging weights to be continuously changing over time, which avoids restricting the weights to a discrete set as in Hansen and Racine (2012). The conditions required for optimality of our TVJMA estimator are neither stronger nor weaker than those required by Hansen and Racine (2012). The third model setting we consider involves lagged dependent variables with i.i.d. regression errors, where we prove the asymptotic optimality of the TVJMA estimator by allowing the regressors to be locally stationary, in the sense of Ing and Wei (2003) and Vogt (2012).

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

Compared with the existing model averaging literature, our proposed approach has a number of appealing features. First, we extend conventional constant weight model averaging to time-varying weight model averaging. In particular, we propose a novel time-varying jackknife model averaging approach by exploring local information at each time point instead of over the whole sample period. The TVJMA weights selected by our method are allowed to change smoothly over time, which is consistent with evolutionary instability of economic relationships. Our result includes the constant JMA estimator in Hansen and Racine (2012) as a special case. Second, we also allow parameters in each candidate model to change smoothly over time. A nonparametric approach is used to estimate the time-varying model parameters, avoiding a potentially misspecified functional form of time-varying parameters by any parametric approach (e.g., time-varying smooth transition regression). Third, we allow regressors to be locally stationary (Dahlhaus, 1996, 1997; Vogt, 2012), and as a result, time-varying parameter dynamic regression models (e.g., time-varying parameter models with lagged dependent variables) can be included as candidate models.

The remainder of this paper is organized as follows. Section 2 introduces the local jackknife criterion and develops the asymptotic optimality theory of the proposed TVJMA estimator for a general nonlinear model with heteroscedasticity. In Section 3, we consider a special class of local constant TVJMA estimators for a time-varying parameter model. Section 4 develops an asymptotic optimality theory of the TVJMA estimator for a time-varying parameter regression model with lagged dependent variables. Section 5 presents a simulation study under constant and time-varying parameter linear regressions respectively. Section 6 examines the empirical forecast performance of the TVJMA estimator for S&P 500 stock returns. Section 7 concludes. Throughout, all convergences occur when the sample size $T \to \infty$. All mathematical proofs are given in an Online Supplementary Material.

2. Model averaging estimator

We consider a general nonlinear data generating process (DGP)

$$Y_t = \mu_t + \varepsilon_t = f_t(\mathbf{X}_t) + \varepsilon_t, \ \ t = 1, \ldots, T, \tag{1}$$

where Y_t is a dependent variable, $\mathbf{X}_t = (X_{1t}, X_{2t}, \ldots)$ is possibly countably infinite, ε_t is an unobservable disturbance with $\mathbb{E}(\varepsilon_t|\mathbf{X}_t) = 0$ almost surely (a.s.), $f_t(\mathbf{x})$ is an unknown smooth function of time t, and T is the sample size. Note that when the functional form of $f_t(\cdot)$ is known up to some finite dimensional parameters, e.g., $f_t(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_t$, the conditional mean of Y_t given $\mathbf{X}_t$ is parametrically specified, where parameter $\boldsymbol{\beta}_t$ is possibly time-varying. A time-varying parameter regression with $f_t(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_t$ will be considered in Section 3. The conventional constant parameter linear models are included as a special case if we assume that $f_t(\cdot) = f(\cdot)$ is linear. When the functional form of $f_t(\cdot)$ is unknown, we can estimate $f_t(\cdot)$ using nonparametric methods, such as the Nadaraya–Watson estimator or the local linear estimator. For notational simplicity, we let $\mathbf{Y} = (Y_1, \ldots, Y_T)'$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_T)'$ and $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_T')'$. Furthermore, we assume that $\mathbb{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$ where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)'$ so that $\boldsymbol{\mu} = \mathbb{E}(\mathbf{Y}|\mathbf{X})$. We denote $\mathrm{var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is a positive definite symmetric matrix. This setup allows a non-diagonal covariance structure for regression errors. Therefore, heteroscedastic and autocorrelated errors are allowed.

2.1. Model framework and jackknife criterion

Consider a sequence of candidate models indexed by $m = 1, \ldots, M_T$, which are allowed to be misspecified for the underlying DGP. The number of models, M_T, may depend on the sample size T. For different models, explanatory variables may be different. Let $\{\widehat{\boldsymbol{\mu}}^1, \ldots, \widehat{\boldsymbol{\mu}}^{M_T}\}$ be a set of nonparametric estimators of $\boldsymbol{\mu}$. Specifically, for the mth model, the estimator of $\boldsymbol{\mu}$ may be written as $\widehat{\boldsymbol{\mu}}^m = \mathbf{P}_m\mathbf{Y}$, where $\mathbf{P}_m$ is a $T \times T$ matrix, which depends on both $\mathbf{K}_t$ and $\mathbf{X}$ but not on $\mathbf{Y}$, $\mathbf{K}_t = \mathrm{diag}\{k_{1t}, \ldots, k_{Tt}\}$, $k_{st} = k(\frac{s-t}{Th})$, the kernel $k(\cdot) : [-1, 1] \to \mathbb{R}^+$ is a prespecified symmetric probability density, and $h \equiv h(T)$ is a bandwidth which depends on the sample size T such that $h \to 0$ and $Th \to \infty$ as $T \to \infty$. For instance, $\mathbf{P}_m$ is defined in (18) when a local constant estimator is used, and so $\widehat{\boldsymbol{\mu}}^m$ is a local estimator for the conditional mean. For each time $t = 1, \ldots, T$, let $\mathbf{w} = (w^1, \ldots, w^{M_T})'$ be a weight vector which satisfies

$$\mathcal{H}_T = \left\{ \mathbf{w} \in [0, 1]^{M_T} : \sum_{m=1}^{M_T} w^m = 1 \right\}. \tag{2}$$

Given $\mathbf{w}$, an averaging estimator at any time point t for the conditional mean is

$$\widehat{\mu}_t(\mathbf{w}) \equiv \sum_{m=1}^{M_T} w^m \widehat{\mu}_t^m = \sum_{m=1}^{M_T} w^m \mathbf{e}_t \mathbf{P}_m \mathbf{Y} = \mathbf{e}_t \mathbf{P}(\mathbf{w}) \mathbf{Y}, \tag{3}$$

where $\mathbf{e}_t$ is a $1 \times T$ vector, in which the tth element is 1 and all others are zero, $\widehat{\mu}_t^m = \mathbf{e}_t \mathbf{P}_m \mathbf{Y}$ and $\mathbf{P}(\mathbf{w}) = \sum_{m=1}^{M_T} w^m \mathbf{P}_m$. Then the model averaging estimator of $\boldsymbol{\mu}$ can be fitted as $\widehat{\boldsymbol{\mu}}(\mathbf{w}) = (\widehat{\mu}_1(\mathbf{w}), \ldots, \widehat{\mu}_T(\mathbf{w}))'$.

Denote $\widetilde{\boldsymbol{\mu}}^m = (\widetilde{\mu}_1^m, \ldots, \widetilde{\mu}_T^m)'$ as the jackknife estimator of $\boldsymbol{\mu}$ for the mth model, where $\widetilde{\mu}_t^m$ is the estimator $\widehat{\mu}_t^m$ obtained with the tth observation $(Y_t, \mathbf{X}_t)$ removed from the sample, the so-called "leave-one-out" estimator. Then, we obtain $\widetilde{\boldsymbol{\mu}}^m = \widetilde{\mathbf{P}}_m \mathbf{Y}$, where $\widetilde{\mathbf{P}}_m$ has zeros on the diagonal and depends on $\mathbf{K}_t$ and $\mathbf{X}$; see (19) for an example in a special setup. The jackknife model averaging estimator of μ_t, which smooths across the M_T jackknife estimators at time point t, is obtained as

$$\widetilde{\mu}_t(\mathbf{w}) = \sum_{m=1}^{M_T} w^m \widetilde{\mu}_t^m = \mathbf{e}_t \sum_{m=1}^{M_T} w^m \widetilde{\mathbf{P}}_m \mathbf{Y} = \mathbf{e}_t \widetilde{\mathbf{P}}(\mathbf{w})\mathbf{Y}, \tag{4}$$

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

Set $\widetilde{\boldsymbol{\mu}}(\mathbf{w}) = (\widetilde{\mu}_1(\mathbf{w}), \ldots, \widetilde{\mu}_T(\mathbf{w}))'$. We shall minimize the local cross-validation (CV) squared error criterion,

$$\mathrm{CV}_{t,T}(\mathbf{w}) = (\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w}))' \mathbf{K}_t (\mathbf{Y} - \widetilde{\boldsymbol{\mu}}(\mathbf{w})). \tag{5}$$

We obtain the optimal time-varying weight vector $\widehat{\mathbf{w}}_t = \operatorname{argmin}_{\mathbf{w} \in \mathcal{H}_T} \mathrm{CV}_{t,T}(\mathbf{w})$, which minimizes $\mathrm{CV}_{t,T}(\mathbf{w})$. The TVJMA estimator of μ_t for any given time point t is $\widehat{\mu}_t(\widehat{\mathbf{w}}_t)$.

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

Note that there are two key differences between our TVJMA estimator and the JMA estimators proposed by Hansen and Racine (2012) and Zhang et al. (2013). One major difference is that we allow the model averaging weights to change with time smoothly. In contrast, Hansen and Racine (2012) and Zhang et al. (2013) restrict the weights to be constant in a discrete set or a continuous set. We extend constant wights to time-varying weights, which can accommodate time-varying predictive power of candidate models. Another difference is that the models in Hansen and Racine (2012) and Zhang et al. (2013) are linear regressions, while in the present paper, $f_t(\cdot)$ can be nonlinear, and parameters in each candidate model are allowed to be unknown smooth functions of time. Theoretically, we establish the asymptotic optimality of the TVJMA estimator based on a set of smoothly time-varying parameter models. Simulation studies show that the proposed TVJMA estimator outperforms the existing model averaging methods in the presence of smooth structural changes as well as recurrent breaks.

2.2. Asymptotic optimality

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

$$L_{t,T}(\mathbf{w}) = (\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu})' \mathbf{K}_t (\widehat{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}), \tag{6}$$

and

$$R_{t,T}(\mathbf{w}) = \mathbb{E}(L_{t,T}(\mathbf{w})|\mathbf{X}) = \boldsymbol{\mu}' \mathbf{A}'(\mathbf{w}) \mathbf{K}_t \mathbf{A}(\mathbf{w}) \boldsymbol{\mu} + \mathrm{tr}(\mathbf{P}'(\mathbf{w}) \mathbf{K}_t \mathbf{P}(\mathbf{w}) \boldsymbol{\Omega}), \tag{7}$$

where $\widehat{\boldsymbol{\mu}}(\mathbf{w}) = \sum_{m=1}^{M_T} w^m \widehat{\boldsymbol{\mu}}^m$ is the weighted average of the forecasts of M_T models, and $\mathbf{A}(\mathbf{w}) = \mathbf{I}_T - \mathbf{P}(\mathbf{w})$.

Denote $\widetilde{\mathbf{A}}(\mathbf{w}) = \mathbf{I}_T - \widetilde{\mathbf{P}}(\mathbf{w})$. Let $\widetilde{L}_{t,T}(\mathbf{w})$ and $\widetilde{R}_{t,T}(\mathbf{w})$ be the local jackknife squared error loss and risk, which are obtained by replacing $\widehat{\boldsymbol{\mu}}(\mathbf{w})$ by $\widetilde{\boldsymbol{\mu}}(\mathbf{w})$, $\mathbf{A}(\mathbf{w})$ by $\widetilde{\mathbf{A}}(\mathbf{w})$, and $\mathbf{P}(\mathbf{w})$ by $\widetilde{\mathbf{P}}(\mathbf{w})$, respectively. Specifically,

$$\widetilde{L}_{t,T}(\mathbf{w}) = (\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu})' \mathbf{K}_t (\widetilde{\boldsymbol{\mu}}(\mathbf{w}) - \boldsymbol{\mu}) \tag{8}$$

and

$$\widetilde{R}_{t,T}(\mathbf{w}) = \boldsymbol{\mu}' \widetilde{\mathbf{A}}'(\mathbf{w}) \mathbf{K}_t \widetilde{\mathbf{A}}(\mathbf{w}) \boldsymbol{\mu} + \mathrm{tr}(\widetilde{\mathbf{P}}'(\mathbf{w}) \mathbf{K}_t \widetilde{\mathbf{P}}(\mathbf{w}) \boldsymbol{\Omega}). \tag{9}$$

Let

$$\xi_{t,T} = \inf_{\mathbf{w} \in \mathcal{H}_T} R_{t,T}(\mathbf{w}) \tag{10}$$

and

$$\widetilde{\boldsymbol{\Omega}} = \boldsymbol{\Omega} - \mathrm{diag}(\Omega_{11}, \ldots, \Omega_{TT}), \tag{11}$$

where Ω_{tt} is the tth diagonal element of $\boldsymbol{\Omega}$, and $\zeta(\mathbf{A})$ denotes the maximum singular value of a matrix $\mathbf{A}$.

Extending the results of Hansen and Racine (2012) and Zhang et al. (2013), we prove that for any given time point t, the TVJMA estimator satisfies the following optimality (OPT) property

$$(OPT): \quad \frac{L_{t,T}(\widehat{\mathbf{w}}_t)}{\inf_{\mathbf{w}\in\mathcal{H}_T} L_{t,T}(\mathbf{w})} \xrightarrow{p} 1, \text{ as } T \to \infty.$$

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

To guarantee that the TVJMA estimator satisfies the OPT property under a DGP that allows smooth-changing parameters and a non-diagonal error covariance structure, we impose a set of regularity conditions:

Assumption 1. $\{\varepsilon_t\}$ is a sequence of innovations such that $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)'$ satisfies $\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is a $T \times T$ symmetric positive-definite matrix.

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

Assumption 3. For $1 \le m \le M_T$, the maximum singular value of $\mathbf{P}_m$ satisfies $\overline{\lim}_{T\to\infty} \max_{1\le m\le M_T} \zeta(\mathbf{P}_m) < \infty$ a.s.

Assumption 4. For $1 \le m \le M_T$, the maximum singular value of $\widetilde{\mathbf{P}}_m$ is finite when the sample size $T \to \infty$, i.e., $\overline{\lim}_{T\to\infty} \max_{1\le m\le M_T} \zeta(\widetilde{\mathbf{P}}_m) < \infty$ a.s.

Assumption 5. For any given time point t, the local risk $\widetilde{R}_{t,T}(\mathbf{w})$, i.e., the conditional expectation of the local jackknife squared error criterion given $\mathbf{X}$, satisfies $\sup_{\mathbf{w}\in\mathcal{H}_T} |\widetilde{R}_{t,T}(\mathbf{w})/R_{t,T}(\mathbf{w}) - 1| \to 0$ a.s. as $T \to \infty$.

Assumption 6. For any given time point t, $M_T \xi_{t,T}^{-2G} \sum_{m=1}^{M_T} R_{t,T}^G(\mathbf{w}_m^0) \to 0$ a.s., for some constant $G \ge 1$, where $\mathbf{w}_m^0$ is an $M_T \times 1$ weight vector with the mth element taking the value of unity and other elements zeros.

Assumption 6′. For any given time point t, $\xi_{t,T}^{-2} \sum_{m=1}^{M_T} R_{t,T}(\mathbf{w}_m^0) \to 0$ a.s., where $\mathbf{w}_m^0$ is an $M_T \times 1$ weight vector with the mth element taking the value of unity and other elements zeros.

Assumption 7. For any given time point t, $\sup_{\mathbf{w}\in\mathcal{H}_T} |\mathrm{tr}(\mathbf{K}_t \widetilde{\mathbf{P}}(\mathbf{w})\widetilde{\boldsymbol{\Omega}})/\widetilde{R}_{t,T}(\mathbf{w})| \to 0$ a.s. as $T \to \infty$.

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

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

Assumption 1 is the same as condition (11) in Zhang et al. (2013), which is limited to Gaussian regressions. This condition can be removed to obtain the asymptotic optimality of the TVJMA estimator for a time-varying parameter regression in Section 3. Assumption 2 ensures the largest singular values of the error covariance matrix $\boldsymbol{\Omega}$ to be finite when the sample size $T \to \infty$, corresponding to condition (12) in Zhang et al. (2013). Assumptions 3 and 4 correspond to conditions (A.3) and (A.4) of Hansen and Racine (2012), respectively. Both of them are rather mild, because typical estimators satisfy the regularity conditions that the maximum singular values of the corresponding matrixes are bounded.

Assumption 5 imposes the condition that the leave-one-out estimator is asymptotically equivalent to the local risk of the regular estimator $\widehat{\boldsymbol{\mu}}(\mathbf{w})$, uniformly over the class of averaging estimators. This is a standard condition for the application of cross-validation and almost the same as condition (10) in Zhang et al. (2013), except that the continuous time-varying set $\mathcal{H}_T$ is used here instead of the continuous constant set $\mathcal{H}_n$ in Zhang et al. (2013). In Section 3, we will consider time-varying parameter regressions as candidate models, where Assumption 5 is ensured by more primitive conditions; see (A.18) in Supplementary Material.

Assumption 6 requires $M_T \sum_{m=1}^{M_T} R_{t,T}^G(\mathbf{w}_m^0) \to \infty$ at a rate slower than $\xi_{t,T}^{2G} \to \infty$ as $T \to \infty$. Assumption 6′ is weaker than Assumption 6, when G is set to 1. To gain further insight into Assumptions 6 and 6′, we define $\eta_{t,T} = \max_{1\le m\le M_T} R_{t,T}(\mathbf{w}_m^0)$. Then, we obtain more primitive conditions for Assumptions 6 and 6′ that $M_T^2 \xi_{t,T}^{-2G} \eta_{t,T}^G \to 0$ a.s. and $M_T \xi_{t,T}^{-2} \eta_{t,T} \to 0$ a.s., respectively. These conditions restrict the rates of $M_T \to \infty$, $\xi_{t,T} \to \infty$ and $\eta_{t,T} \to \infty$; in particular they require that the infimum risk $\xi_{t,T}$ explode quickly enough and the maximum risk of an individual model do not explode very quickly. Note that $\xi_{t,T} \to \infty$ is obviously necessary for Assumptions 6 and 6′ to hold, which is pointed out by Hansen (2007) that this is no finite approximating model for which the bias is zero in linear regression as well as nonparametric regression. Like Ando and Li (2014), we consider a case with $\xi_{t,T} \sim T^{1-\widetilde{\delta}}$ for $\widetilde{\delta} < 1/2$. From Assumptions 2, 3 and 8, we can obtain $\eta_{t,T} = O_p(Th)$. Given $\xi_{t,T} \to \infty$ with the rate $T^{1-\widetilde{\delta}}$, $M_T \to \infty$ with a slower rate than $T^{G-2\widetilde{\delta}G} h^{-G}$ and $\eta_{t,T} = O_p(Th)$, Assumptions 6 and 6′ hold. Assumption 6 is required for the asymptotic optimality of all MMA and JMA estimators; see more discussions in Wan et al. (2010) and Zhang et al. (2013).

Assumption 7 restricts the correlation strength among unobservable disturbances and can be removed when disturbances are not correlated. Under the set-up of linear DGP, Assumption 7 can be simplified to $\sup_{\mathbf{w}\in\mathcal{H}_T} |\mathrm{tr}(\widetilde{\mathbf{P}}(\mathbf{w})\widetilde{\boldsymbol{\Omega}})/\widetilde{R}_{t,T}(\mathbf{w})|$

$\to 0$ a.s. as $T \to \infty$, which is the same as condition (14) in Zhang et al. (2013). If all candidate models are linear regressions with constant parameters, it can be shown that $\sup_{\mathbf{w}\in\mathcal{H}_T} |\mathrm{tr}(\widetilde{\mathbf{P}}(\mathbf{w})\widetilde{\boldsymbol{\Omega}})/\widetilde{R}_{t,T}(\mathbf{w})| \leq \xi_{t,T}^{-1}\gamma \max_{1\leq m\leq M_T} \zeta(\widetilde{\mathbf{P}}_m\widetilde{\boldsymbol{\Omega}})$, where γ is the number of regressors. It follows that condition (14) boils down to condition (22) in Zhang et al. (2013), which assumes that the growth rate of the number of regressors in the largest model must be slower than the rate at which $\xi_{t,T} \to \infty$. In this paper, under the linear regression setting with time-varying parameters in Section 3, we can establish the asymptotic optimality without Assumption 7.

In Assumption 8, the kernel is symmetric and bounded, and has a compact support $[-1, 1]$. It usually discounts the observations whose values are far away from the time point of interest. This implies that $k_{\max} \equiv \max_{s,t} k_{st} < \infty$, which is used in our proof. A commonly used kernel function, the Epanechnikov kernel $k(u) = 0.75(1-u^2)I(|u| \leq 1)$, is employed in this paper, where $I(\cdot)$ is the indicator function. Assumption 9 implies $h \to 0$ and $Th \to \infty$ as $T \to \infty$, which is a standard condition for the bandwidth; see Chen and Hong (2012). Assumption 9 includes the optimal bandwidth $h \propto T^{-1/5}$, which minimizes the integrated mean squared error (MSE) of a smoothed nonparametric estimator; see more discussions in Cai (2007) and Chen and Hong (2012).

We now state the main result of this section.

Theorem 1. *Suppose Assumptions 1–9 hold. Then for any given time point t, the TVJMA estimator satisfies the asymptotic optimality (OPT) property, i.e.,*

$$\frac{L_{t,T}(\widehat{\mathbf{w}}_t)}{\inf_{\mathbf{w}\in\mathcal{H}_T} L_{t,T}(\mathbf{w})} \xrightarrow{p} 1.$$

Theorem 1 shows that the local squared error loss obtained from the time-varying combination weight vector $\widehat{\mathbf{w}}_t$ is asymptotically equivalent to the infeasible optimal combination weight vector at any time point t. This implies that the TVJMA estimator is asymptotically optimal in the class of time-varying model averaging estimators based on possibly nonlinear models where the weight vector $\mathbf{w}$ is restricted to the set $\mathcal{H}_T$, which allows the combination weights to change smoothly over time.

3. Time-varying parameter regression

In this section, we focus on a set of candidate models with a specific form, i.e., time-varying parameter linear regressions. This is a special case of the general candidate models in Section 2. Consider the mth time-varying parameter regression model

$$Y_t = \mathbf{X}_t^m \boldsymbol{\beta}_t^m + \varepsilon_t^m, \ t = 1, \ldots, T, \ m = 1, \ldots, M_T, \tag{12}$$

where $\mathbf{X}_t^m$ is a $1 \times q_m$ vector of explanatory variables, $\boldsymbol{\beta}_t^m$ is a $q_m \times 1$ possibly time-varying parameter vector, ε_t^m is an unobservable disturbance, and q_m is a positive integer that may be infinite. Note that we allow $\mathbb{E}(\varepsilon_t^m|\mathbf{X}_t^m) \neq 0$ in the set of candidate models, which arises when the mth model is misspecified for $\mathbb{E}(Y_t|\mathbf{X}_t^m)$.

As Hansen (2001) points out, "it may seem unlikely that a structural break could be immediate and might seem more reasonable to allow a structural change to take a period of time to take effect". We are thus interested in the following mth smooth time-varying parameter model:

$$Y_t = \mathbf{X}_t^m \boldsymbol{\beta}^m\left(\frac{t}{T}\right) + \varepsilon_t^m, t = 1, \ldots, T, \tag{13}$$

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

Various smooth time-varying parameter models have been considered to capture the evolutionary behavior of economic time series. For example, a smooth transition regression (STR) model is proposed by Chan and Tong (1986) and further studied by Lin and Teräsvirta (1994), which allows both the intercept and the slope to change smoothly over time. If the parameter function is correctly specified, parametric models for time-varying parameters can be consistently estimated with the root-T convergence rate. However, there is no economic theory to justify any concrete functional form assumption for these time-varying parameters, and the choice of a particular functional form for time-varying parameters is somewhat arbitrary, probably leading to serious misspecification. Robinson (1989, 1991) considers a nonparametric time-varying parameter model and it is further studied by Blundell et al. (1998), Cai (2007) and Chen and Hong (2012). One advantage of the nonparametric approach is that little or restrictive prior information is required for the functional forms of time-varying parameters, except for the regularity assumption that they evolve over time smoothly. In the present context, for the time-varying parameter $\boldsymbol{\beta}^m(t/T)$, we follow the spirit of the smoothed nonparametric estimation in Robinson (1989).

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

For any s in a neighborhood of a fixed time point t, $\boldsymbol{\beta}_s^m$ follows a Taylor expansion:

$$\boldsymbol{\beta}_s^m \approx \boldsymbol{\beta}_t^m, \; s \in [t - Th, t + Th]. \tag{14}$$

Define $\mathbf{K}_{-t} = \mathrm{diag}\{k_{1t}, k_{2t}, \ldots, k_{(t-1)t}, 0, k_{(t+1)t}, \ldots, k_{Tt}\}$ as the weights for Jackknife estimation and $\mathbf{X}^m$ as a $T \times q_m$ matrix with $\mathbf{X}_t^m$ as its tth row. Thus for every time point t, we obtain a local constant estimator $\widehat{\boldsymbol{\beta}}_t^m$ for $\boldsymbol{\beta}_t^m$, and so a local least square estimator $\widehat{\mu}_t^m$ and a Jackknife estimator $\widetilde{\mu}_t^m$ for the mth candidate model respectively:

$$\widehat{\boldsymbol{\beta}}_t^m = (\mathbf{X}^{m'}\mathbf{K}_t\mathbf{X}^m)^{-1}\mathbf{X}^{m'}\mathbf{K}_t\mathbf{Y}, \tag{15}$$

$$\widehat{\mu}_t^m = \mathbf{X}_t^m(\mathbf{X}^{m'}\mathbf{K}_t\mathbf{X}^m)^{-1}\mathbf{X}^{m'}\mathbf{K}_t\mathbf{Y} \tag{16}$$

and

$$\widetilde{\mu}_t^m = \mathbf{X}_t^m(\mathbf{X}^{m'}\mathbf{K}_{-t}\mathbf{X}^m)^{-1}\mathbf{X}^{m'}\mathbf{K}_{-t}\mathbf{Y}. \tag{17}$$

Based on the expressions of $\widehat{\mu}_t^m$ and $\widetilde{\mu}_t^m$, it is straightforward to obtain

$$\mathbf{P}_m = \begin{bmatrix} \mathbf{X}_1^m(\mathbf{X}^{m'}\mathbf{K}_1\mathbf{X}^m)^{-1}\mathbf{X}^{m'}\mathbf{K}_1 \\ \mathbf{X}_2^m(\mathbf{X}^{m'}\mathbf{K}_2\mathbf{X}^m)^{-1}\mathbf{X}^{m'}\mathbf{K}_2 \\ \cdots \\ \mathbf{X}_T^m(\mathbf{X}^{m'}\mathbf{K}_T\mathbf{X}^m)^{-1}\mathbf{X}^{m'}\mathbf{K}_T \end{bmatrix} \tag{18}$$

and

$$\widetilde{\mathbf{P}}_m = \begin{bmatrix} \mathbf{X}_1^m(\mathbf{X}^{m'}\mathbf{K}_{-1}\mathbf{X}^m)^{-1}\mathbf{X}^{m'}\mathbf{K}_{-1} \\ \mathbf{X}_2^m(\mathbf{X}^{m'}\mathbf{K}_{-2}\mathbf{X}^m)^{-1}\mathbf{X}^{m'}\mathbf{K}_{-2} \\ \cdots \\ \mathbf{X}_T^m(\mathbf{X}^{m'}\mathbf{K}_{-T}\mathbf{X}^m)^{-1}\mathbf{X}^{m'}\mathbf{K}_{-T} \end{bmatrix}. \tag{19}$$

Thus, $\widetilde{\mathbf{P}}_m = \mathbf{D}_m(\mathbf{P}_m - \mathbf{I}_T) + \mathbf{I}_T$, where $\mathbf{D}_m$ is a diagonal matrix with the tth diagonal element $(1 - h_{tt}^m)^{-1}$, and h_{tt}^m is the (t, t) element in $\mathbf{P}_m$.

To establish the asymptotic optimality property of $\widehat{\boldsymbol{\mu}}(\widehat{\mathbf{w}})$, we impose the following regularity conditions:

Assumption 10. For any given time point t, $\sup_{\mathbf{w}\in\mathcal{H}_T} \mathrm{tr}(\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w}))\xi_{t,T}^{-1} = o_p(1)$.

Assumption 11. For any given time point t, the local average of μ_t^2 is bounded, i.e., $\frac{1}{Th}\boldsymbol{\mu}'\mathbf{K}_t\boldsymbol{\mu} = O(1)$ a.s. as $T \to \infty$.

Assumption 12. For any given time point t, $h^* = O(T^{-1}h^{-1})$ and $h^{-1}\xi_{t,T}^{-1} \to 0$ a.s. as $T \to \infty$, where $\xi_{t,T}$ is defined in (10) and $h^* = \max_{1\le m\le M_T} \max_{1\le t\le T} h_{tt}^m$.

As pointed out by a referee, Assumption 10 implies that the bias part dominates the risk since $\mathrm{tr}(\mathbf{P}'(\mathbf{w})\mathbf{P}(\mathbf{w}))$ is related to the variance part of the risk. Typically, the risk is minimized by equating its bias part and its variance part. One way to make Assumption 10 hold is to restrict the set for weights. Another way is to restrict the number of candidate models or the number of variables in candidate models. Assumption 10 is the price for allowing a dependent and non-normal random error ε_t. When Assumption 1 (normal errors) is imposed or it is assumed that $\boldsymbol{\varepsilon}$ is a vector of independent variables as in the existing literature on JMA (e.g., Hansen and Racine (2012) and Ando and Li (2014)), Assumption 10 is no longer needed. Since Theorem 1 has considered normal errors, Theorem 2′ will consider the situations where $\boldsymbol{\varepsilon}$ is a vector of independent variables without using Assumption 10.

Given $\mathbf{K}_t$, we have $\frac{1}{Th}\boldsymbol{\mu}'\mathbf{K}_t\boldsymbol{\mu} = \frac{1}{Th}(\mu_1, \ldots, \mu_T)'\mathbf{K}_t(\mu_1, \ldots, \mu_T) = \frac{1}{Th}\sum_{s=1}^T k_{st}\mu_s^2 \xrightarrow{a.s.} \mathbb{E}\mu_t^2$, as $T \to \infty$. Thus, Assumption 11 implies that the local average of μ_t^2 is bounded. This is similar to condition (11) in Wan et al. (2010) and condition (23) in Zhang et al. (2013), which concern the average of μ_t^2 over the whole sample period. Finally, the first part of Assumption 12 is rather mild, which corresponds to condition (C.2) in Zhang (2015) and equation (5.2) in Andrews (1991). The second part of Assumption 12 excludes extremely unbalanced designs. This condition is reasonable and typical for the application of cross-validation; see Li (1987), Hansen and Racine (2012) and Zhang et al. (2013) for more discussions.

Theorem 2. *Suppose Assumptions 2, 3, 6′ and 8–12 hold. Then for any given time point t, the TVJMA estimator satisfies the asymptotic optimality (OPT) property.*

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

Next, we establish the asymptotic optimality (OPT) result without Assumption 10. Theorem 2′ addresses the asymptotic optimality of the TVJMA estimator.

Theorem 2′. *Suppose $\boldsymbol{\varepsilon}$ is a vector of independent variables and Assumptions 2, 3, 6′, 8–9 and 11–12 hold. Then for any given time point t, the TVJMA estimator satisfies the asymptotic optimality (OPT) property.*

Finally, we consider asymptotic properties of the time-varying parameter averaging estimator. Suppose the DGP is a linear time-varying parameter regression, i.e., $Y_t = \mathbf{X}_t\boldsymbol{\beta}_t + \varepsilon_t$, where $\mathbf{X}_t$ is a $1 \times q$ vector of explanatory variables, $\boldsymbol{\beta}_t \equiv \boldsymbol{\beta}(t/T)$ is a $q \times 1$ smooth time-varying parameter vector, and $\boldsymbol{\beta} : [0, 1] \to \mathbb{R}^q$ is an unknown smooth function except for a finite number of points on [0, 1]. Here, q is a fixed integer, and ε_t is an unobservable disturbance with $\mathbb{E}(\varepsilon_t|\mathbf{X}_t) = 0$ almost surely. A model including only all regressors with nonzero parameters is called a true model; see Zhang (2015). Any candidate model omitting regressors with nonzero parameters is called an under-fitted model; see more discussions in Zhang (2015), Zhang and Liu (2019). It is not required that the true model be one of the candidate models. However, at least one candidate model should not be under-fitted. This implies that one candidate model must include all these regressors with nonzero parameters and may have some redundant regressors as well. From (15), the time-varying model averaging estimator of parameter $\boldsymbol{\beta}_t$ is $\widehat{\boldsymbol{\beta}}_t(\mathbf{w}) = \sum_{m=1}^{M_T} w^m \boldsymbol{\Pi}'_m \widehat{\boldsymbol{\beta}}_t^m$, where $\boldsymbol{\Pi}_m = (\mathbf{I}_{q_m}, \mathbf{0}_{q_m\times(q-q_m)})$ (i.e., a column permutation thereof) and the maximum number of columns of $\mathbf{X}^m$ in all candidate models (i.e., $\max_{1\le m\le M_T} q_m$) is bounded.

Next, we impose the following regularity conditions:

Assumption 13. For each $j = 1, \ldots, q$, the jth element of $\boldsymbol{\beta}(\cdot)$ is continuously differentiable over the unit interval [0, 1].

Assumption 14. For any given time point t, $\boldsymbol{\Psi}_{t,T} \equiv T^{-1}h^{-1}\sum_{s=1}^{T} k_{st}\mathbf{X}_s'\mathbf{X}_s \xrightarrow{p} \boldsymbol{\Psi}$ as $T \to \infty$, where $\boldsymbol{\Psi}$ is a $q\times q$ symmetric, bounded and positive definite matrix, and $T^{-1/2}h^{-1/2}\sum_{s=1}^{T} k_{st}\ \mathbf{X}_t\varepsilon_t = O_p(1)$.

Assumption 13 places a smoothness condition on parameters, which is commonly imposed in the literature; see Robinson (1989, 1991). Assumption 14 can be obtained from Proposition A.1 in Chen and Hong (2012) and Lemma 3 in Cai (2007). The following theorem shows that the TVJMA parameter estimator $\widehat{\boldsymbol{\beta}}_t(\widehat{\mathbf{w}}_t)$ is $\sqrt{Th}$-consistent under these regularity assumptions.

Theorem 3. *Suppose Assumptions 3, 8 and 12–14 hold, and $h = cT^{-\lambda}$ for $\frac{1}{5} \le \lambda < 1$, where $0 < c < \infty$. Then for any given time point t in the interior region $t \in [Th, T - Th]$, $\sqrt{Th}(\widehat{\boldsymbol{\beta}}_t(\widehat{\mathbf{w}}_t) - \boldsymbol{\beta}_t) = O_p(1)$ as $T \to \infty$.*

A similar result holds for the boundary regions $[1, Th] \cup [T - Th, T]$ if we assume $h = cT^{-\lambda}$ for $\frac{1}{3} \le \lambda < 1$, where $0 < c < \infty$. This happens because the local constant estimator suffers from the well-known boundary effect problem in smoothed nonparametric estimation. As shown in Cai (2007), the convergence rate of the asymptotic bias with the local constant estimator is h^2 in the interior region, but only h in the boundary regions.

4. Asymptotic optimality of TVJMA with lagged dependent variables

In this section, we develop an asymptotic optimality theory for the TVJMA estimator based on time-varying parameter regression models that include lagged dependent variables as regressors. Dynamic regressions are widely used in macroeconomic forecasts. It is highly desirable to extend the TVJMA estimator from static regressions to dynamic regressions. Consider the following DGP

$$Y_t = \sum_{j=1}^{\infty} \beta_{jt}Y_{t-j} + \varepsilon_t,\ \ t = 1, \ldots, T, \tag{20}$$

where ε_t is i.i.d. with mean zero and variance σ^2. This is a special case of the DGP in Section 2.

More generally, exogenous regressors can be added to the candidate models with finitely many lagged dependent variables. This yields an augmented regression model

$$Y_t = \sum_{j=1}^{r_1} \beta_{jt}Y_{t-j} + \sum_{j=1}^{r_2} \beta_{(r_1+j)t}X_{tj}^* + \varepsilon_t^f,\ \ t = 1, \ldots, T, \tag{21}$$

where X_{tj}^* is an exogenous variable, ε_t^f is the innovation, r_1 is the maximal lag order, and r_2 is the number of exogenous regressors. Let r_1 be allowed to increase and r_2 be fixed when T increases. Denote $\mathbf{Y} = (Y_1, \ldots, Y_T)'$, $\mathbf{Y}_{Lt} = (Y_{t-1}, \ldots, Y_{t-r_1})$, and let $\mathbf{Y}_L = (\mathbf{Y}_{L1}', \ldots, \mathbf{Y}_{LT}')'$ be a $T \times r_1$ matrix containing T observations of r_1 lagged dependent regressors, $\mathbf{X}^* = (\mathbf{X}_1^{*'}, \ldots, \mathbf{X}_T^{*'})'$ with $\mathbf{X}_t^* = (X_{t1}^*, \ldots, X_{tr_2}^*)$ be a $T\times r_2$ matrix containing observations of r_2 exogenous regressors, $\mathbf{X} = (\mathbf{Y}_L, \mathbf{X}^*)$ be a $T \times \gamma$ matrix with rank $\gamma = r_1 + r_2$, and $\boldsymbol{\varepsilon}^f = (\varepsilon_1^f, \ldots, \varepsilon_T^f)'$. The regressor matrix $\mathbf{X}^m$ of the mth candidate model is formed by combining the columns of $\mathbf{X}$. Define $\mathbf{P}$ in a similar way to $\mathbf{P}_m$ with $\mathbf{X}^m$ replaced by $\mathbf{X}$. Note that $\mathbf{X}^m$ is the regressor matrix in the mth candidate model and q_m is the number of regressors in $\mathbf{X}^m$. Regressors are allowed to be locally stationary (Dahlhaus, 1996, 1997). Thus, our framework covers AR as well as ARX models with time-varying parameters. For each candidate model, time-varying parameters are estimated by a local constant method, which is the same as (15) in Section 3.

We impose the following regularity conditions:

Assumption 15. $\{Y_t\}$ is a locally stationary process, $\{\mathbf{X}_t^*\}$ is a strictly stationary process, and both $\{Y_t\}$ and $\{\mathbf{X}_t^*\}$ are β-mixing processes with mixing coefficients $\{\beta(j)\}$ satisfying $\sum_{j=1}^{\infty} j^2\beta(j)^{\delta/(1+\delta)} < C < \infty$, $\sup_t \mathbb{E}\|Y_t\|^4 < C$ and $\mathbb{E}\|\mathbf{X}_t^*\|^4 < \infty$ for some constant $0 < \delta < 1$ and $C > 0$.

Assumption 16. $Tq_m^{-1}h_{tt}^m = O_p(1)$, $t = 1, \ldots, T$, $m = 1, \ldots, M_T$, and for any given time point t, $T^{-1}h^{-1}\boldsymbol{\mu}'\mathbf{K}_t\boldsymbol{\mu} = O_p(1)$, $\gamma\xi_{t,T}^{*-1} = o_p(1)$, and $\gamma\boldsymbol{\mu}'\boldsymbol{\mu}\xi_{t,T}^{*-2} = o_p(1)$, where $\xi_{t,T}^* = \inf_{\mathbf{w}\in\mathcal{H}_T} V_{t,T}(\mathbf{w})$ and $V_{t,T}(\mathbf{w}) = \boldsymbol{\mu}'\mathbf{A}'(\mathbf{w})\mathbf{K}_t\mathbf{A}(\mathbf{w})\boldsymbol{\mu} + \sigma^2\mathrm{tr}(\mathbf{P}'(\mathbf{w})\mathbf{K}_t\mathbf{P}(\mathbf{w}))$.

Assumption 17. For any given time point t, $\zeta(T^{-1}h^{-1}\mathbf{X}^{*'}\mathbf{K}_t\mathbf{X}^*) = O_p(1)$, $\mathbf{X}^{*'}\mathbf{K}_t\boldsymbol{\varepsilon}/\sqrt{Th} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Delta})$, and $\zeta((T^{-1}\mathbf{X}^{*'}\mathbf{M}_t\mathbf{X}^*)^{-1}) = O_p(1)$, where $\boldsymbol{\Delta}$ is a symmetric, bounded and positive definite matrix, and $\mathbf{M}_t \equiv \mathbf{K}_t - \mathbf{K}_t\mathbf{Y}_L(\mathbf{Y}_L'\mathbf{K}_t\mathbf{Y}_L)^{-1}\mathbf{Y}_L'$.

Assumption 18. The innovation process $\{\varepsilon_t\}$ is an i.i.d. sequence with mean 0 and variance σ^2, and satisfies that with some positive constants α_1, α_2 and α_3,

$$|F_t(d_1) - F_t(d_2)| \le \alpha_1|d_1 - d_2|^{\alpha_2},$$

for all t when $|d_1 - d_2| \le \alpha_3$, where $F(\cdot)$ is the distribution function of ε_t.

Assumption 19. $r_1^{6+\alpha_4} = O(T)$ for some $\alpha_4 > 0$ and $\sup_t \mathbb{E}\varepsilon_t^4 < \infty$.

In Assumption 15, local stationarity is weaker than strict stationarity. Intuitively, local stationarity implies that when the standardized time t/T is in a neighborhood of any fixed point $\tau \in [0, 1]$, the behavior of time series $\{Y_t\}$ can be approximated up to a certain high order by a strictly stationary process $\{Y_t(\tau)\}$, and it holds that $\|Y_t - Y_t(\tau)\| = O_p(h + 1/T)$, where h is a bandwidth such that $h \to 0$ as $T \to \infty$; see Dahlhaus (1996, 1997) and Vogt (2012) for details. Thus, the autocovariance function of $\{Y_t\}$ for all times t, with t/T in the neighborhood of τ, can be approximated arbitrarily well by that of the strictly stationary time series $\{Y_t(\tau)\}$.

Assumption 16 is analogous to Assumptions 10–12, which are used for time-varying parameter regression models when $\mathbf{X}_t$ is assumed to be strictly stationary. The first part of Assumption 16 is a counterpart of Assumption 12 and excludes extremely unbalanced designs. The second part of Assumption 16 concerns the local average behavior of μ_t^2 for any given time point t. Like in Shao (1997) and Wan et al. (2010), if $\{Y_t, \mathbf{X}_t\}$ is a strictly stationary process, this is the average behavior of μ_t^2 over the whole sample period. By $\boldsymbol{\mu}'\boldsymbol{\mu}/T = O_p(1)$ and Assumption 3, a sufficient condition of the fourth part of Assumption 16 is $\gamma T\xi_{t,T}^{*-2} = o_p(1)$. By comparing the expression of $V_{t,T}(\mathbf{w})$ with the risk $R_{t,T}(\mathbf{w})$ defined in (7), we can view $V_{t,T}(\mathbf{w})$ as a kind of risk as well, which may be called as a pseudo-risk. Hence, the third and fourth parts of Assumption 16 impose a restriction on the relationship among the number of regressors γ, the sample size T, and the infimum pseudo-risk $\xi_{t,T}^*$. Similar assumptions are used in Zhang et al. (2013), Liu and Okui (2013) and Ando and Li (2014).

When $\{\mathbf{X}_t^{*'}\varepsilon_t\}$ is a stationary ergodic martingale difference sequence with finite fourth moments and $T^{-1}\mathbf{X}^{*'}\mathbf{X}^*$ converges to a symmetric positive definite matrix in probability, the first part of Assumption 17 holds. The second part of Assumption 17 can be ensured by more primitive conditions; see more discussions in equation (A.7) in Cai (2007). Here, $\mathbf{X}^* = (\mathbf{X}_1^{*'}, \ldots, \mathbf{X}_T^{*'})'$, with $\mathbf{X}_t^* = (X_{t1}^*, \ldots, X_{tr_2}^*)$, is a $T \times r_2$ matrix containing observations of r_2 exogenous regressors. In this paper, we assume that r_2 is fixed when T increases. It is conceivable that we could allow r_2 to increase with T at the cost of more tedious proof and other assumptions. Assumption 18 is a mild condition which is the same as condition (K.2) of Ing and Wei (2003). It holds for any distribution with a bounded probability density. This assumption is also used to prove Lemma 1 in the Supplementary Material. Assumption 19 is a reiteration of assumptions in Lemma 4 in the Supplementary Material. It can be replaced by the conditions of $r_1^{2+\alpha_4} = O(T)$ and $\sup_{-\infty<t<\infty}\mathbb{E}|\varepsilon_t|^S < \infty$ for all $S = 1, 2, \ldots$.

Next, we impose conditions on the strictly stationary process $\{Y_t(\tau)\}$ indexed by $\tau \in [0, 1]$.

Assumption 20. For any $\tau \in [0, 1]$ and $q > 0$, $\{Y_t(\tau)\}$ is strictly stationary with $\mathbb{E}|Y_t(\tau)|^q < \infty$ and $Y_t(\tau) + \sum_{j=1}^{\infty} a_jY_{t-j}(\tau) = \varepsilon_t, t = \ldots, -1, 0, 1, \ldots$, where the roots of $A(z) = 1 + \sum_{j=1}^{\infty} a_jz^j = 0$ lie outside the unit circle $|z| = 1$, and $\{\varepsilon_t\}$ is a sequence of independent random variables with mean 0 and variance σ^2.

Assumption 21. For any $\tau \in [0, 1]$, $\{Y_t(\tau)\}$ is a stationary $\beta-$mixing process with mixing coefficients $\{\beta(j)\}$ satisfying $\sum_{j=1}^{\infty} j^2\beta(j)^{\delta/(1+\delta)} < C$ for some $0 < \delta < 1$ and $0 < c < \infty$.

Assumption 20 is a standard condition for ARMA models; see more discussions in Ing and Wei (2003). The mixing condition in Assumption 21 imposes a restriction on temporal dependence in $\{Y_t(\tau)\}$, which is commonly used in the literature (e.g., Chen and Hong (2012)).

Theorem 4. *Suppose Assumptions 3, 8, 9 and 15–21 hold. Then for any given time point t, the TVJMA estimator in this section satisfies the asymptotic optimality (OPT) property.*

As a main contribution, Theorem 4 extends Theorem 2 for the asymptotic optimality property of the TVJMA estimator from static regression models with constant parameters to the dynamic regression models with time-varying parameters and locally stationary regressors.

5. Monte Carlo Simulation

To examine the finite sample performance of the proposed TVJMA estimator, we consider the following DGPs:
DGP 1 (Smooth Structural Changes):

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

where $\tau = t/T$, $F(\tau) = \tau^3$, $X_{t1} = 1$, and observations on all other regressors $\{X_{tj}, j \geq 2\}$ are generated from $i.i.d.N(0, 1)$ sequences. Following Hansen and Racine (2012), $\theta_j = c\sqrt{2\alpha} j^{-\alpha-1/2}$, with $c > 0$ and $\alpha = 1.5$, and the coefficient c is selected to control the population coefficient of determination $R^2 = c^2/(1+c^2)$ to vary on a grid from 0.1 to 0.9.

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

We compare (1) the TVJMA estimator with a variety of popular model averaging estimators, namely (2) the nonparametric version of bias-corrected AIC in Cai and Tiwari (2000) (AICc); (3) a smoothed AICc (SAICc); (4) the JMA of Hansen and Racine (2012); (5) the MMA of Hansen (2007); (6) a smoothed Akaike information criterion (SAIC); and (7) a smoothed Bayesian information criterion (SBIC). The AICc for order selection is $\mathrm{AICc} = \ln \mathrm{RSS} + (T + \mathrm{tr}(S^*))/(T - (\mathrm{tr}(S^*) + 2))$, where $\mathrm{RSS} = \sum_{t=1}^{T}(Y_t - \widehat{Y}_t)^2$ is based on a local constant regression and $\mathrm{tr}(S^*)$ is the number of parameters in the model, which penalizes extra parameters for a larger value of $\mathrm{tr}(S^*)$. For the definition of S^*, see more discussions in Cai and Tiwari (2000). The SAICc method is the model averaging estimator with the weight $w^m = \exp(-\frac{1}{2}\mathrm{AICc}_m)/\sum_{m=1}^{M_T} \exp(-\frac{1}{2}\mathrm{AICc}_m)$, where AICc_m is obtained from Cai and Tiwari (2000) for the mth candidate model. The other four model averaging estimators are based on linear regressions with constant combination weights, including JMA, MMA, SAIC and SBIC. SAIC, proposed by Buckland et al. (1997), is the least squares model averaging estimator with the weight $w^m = \exp(-\frac{1}{2}\mathrm{AIC}_m)/\sum_{m=1}^{M_T} \exp(-\frac{1}{2}\mathrm{AIC}_m)$, where $\mathrm{AIC}_m = T\ln\widehat{\sigma}_m^2 + 2m$. SBIC is a simplified form of the Bayesian model averaging with the weight $w^m = \exp(-\frac{1}{2}\mathrm{BIC}_m)/\sum_{m=1}^{M_T} \exp(-\frac{1}{2}\mathrm{BIC}_m)$, where $\mathrm{BIC}_m = T\ln\widehat{\sigma}_m^2 + m\ln T$.

The number of candidate models is determined by the rule in Hansen and Racine (2012), i.e., $M_T = [3T^{1/3}]$, the nearest integer of $3T^{1/3}$. This yields $M_T = 11$, 14, 15 and 18 for $T = 50$, 75, 100 and 200, respectively. The candidate models are $Y_t = \sum_{j=1}^{m} \beta_j^m(\tau) X_{tj} + \varepsilon_t^m,\ t = 1, \ldots, T,\ m = 1, \ldots, M_T$. For our TVJMA estimator, parameters in these candidate models are estimated by the local constant method described in Section 3. For the JMA, MMA, SAIC and SBIC methods, the parameters $\{\beta_j^m(\tau)\}$ in candidate models are assumed to be constant (i.e., they do not depend on $\tau = t/T$), and as a result, the candidate models are simplified to $Y_t = \sum_{j=1}^{m} \beta_j^m X_{tj} + \varepsilon_t^m,\ t = 1, \ldots, T,\ m = 1, \ldots, M_T$.

For the TVJMA and AICc methods, we use the Epanechnikov kernel in smoothed nonparametric estimation; this kernel has been shown to be the optimal kernel for density estimation (Epanechnikov, 1969) and robust regression (Lehmann and Casella, 2006), although our experience suggests that the choice of $k(\cdot)$ has little impact on the performance of our TVJMA estimator. For space, we report results based on a rule-of-thumb bandwidth $h = 2.34T^{-1/5}$, which attains the optimal rate for MSE. We generate $N = 1000$ data sets from the random sample $\{Y_t, X_{tj}, j \geq 2\}$ of size T, and use the following MSE criterion to assess the accuracy of forecasts:

$$\frac{1}{N}\sum_{n=1}^{N} \|\widehat{\boldsymbol{\mu}}(\mathbf{w})^{(n)} - \boldsymbol{\mu}^{(n)}\|^2, \tag{22}$$

where $\widehat{\boldsymbol{\mu}}(\mathbf{w})^{(n)}$ and $\boldsymbol{\mu}^{(n)}$ denote the forecast value and the true value of the conditional expectation of $\mathbf{Y}$ in the nth replication, where $n = 1, \ldots, N$. To simplify comparisons, the risk (i.e., expected squared error loss) of all model averaging estimators are normalized by the MSE of the infeasible optimal least squares model averaging estimator, which is the same as in Hansen and Racine (2012). For space, we report the Monte Carlo results in graphical forms.

Figs. 1–3 report the results of simulations under DGP 1. Some MSE plots are not shown in these figures, because these methods perform so poorly that their results are beyond the range of the y-axis. In most cases, the TVJMA estimator delivers the most precise forecasts among all estimators considered, especially when R^2 is relatively large. Under both

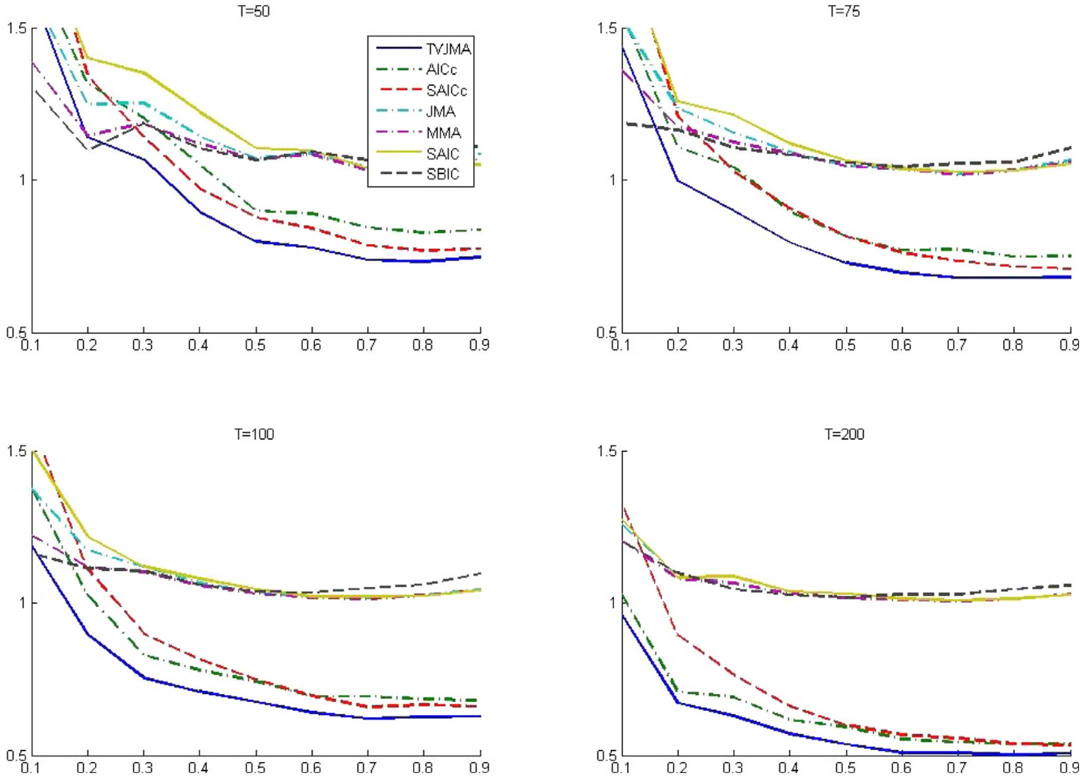

Fig. 1. Finite-sample performance under DGP 1 with Case (i).
Notes: (1) DGP 1 (Smooth Structural Changes):

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

where $\tau = t/T$, $F(\tau) = \tau^3$, $X_{t1} = 1$, and all other regressors $\{X_{tj}, j \geq 2\}$ are $i.i.d.N(0,1)$ sequences; $\theta_j = c\sqrt{2\alpha} j^{-\alpha-1/2}$, with $c > 0$ and $\alpha = 1.5$.
(2) In each figure, the sample sizes are shown in four panels. The sample size varies from $T = 50$, 75, 100 and 200.
(3) Three cases for $\{\varepsilon_t\}$: Case (i) $\varepsilon_t \sim i.i.d.N(0,1)$; Case (ii) $\varepsilon_t = e_{t,1} + e_{t,2}$, $e_{t,1} \sim N(0, X_{t2}^2)$, $e_{t,2} = \phi e_{t-1,2} + u_t$, $u_t \sim i.i.d.N(0,1)$ and $\phi = 0.5$; Case (iii) $\varepsilon_t = \sqrt{h_t} u_t$, $h_t = 0.2 + 0.5X_{t2}^2$, $u_t \sim i.i.d.N(0,1)$.
(4) In each panel, the y-axis and the x-axis display the MSE and the population R^2, respectively. Seven methods to estimate parameters are shown in these figures, including TVJMA, AICc in Cai and Tiwari (2000), SAICc, JMA, MMA, SAIC and SBIC estimators.

conditionally heteroscedastic errors and autocorrelated errors, our method displays the best performance in terms of the risk, as is expected. Also, when the sample size T is large enough, the AICc and SAICc estimators are sometimes marginally similar to the TVJMA estimator in the cases of large R^2. This happens because the parameters in DGP 1 are changing over time and the candidate models are time-varying parameter models as well. In most cases, the TVJMA estimator is preferred to any of the four estimators based on linear least squares, although occasionally small to moderate reductions in MSE can be achieved for the MMA and JMA estimators with small R^2 and small T; see $T = 50$ for example. We note that in some cases the TVJMA performances are a bit sensitive to bandwidth selection. The selection of an optimal bandwidth to estimate the time-varying combination weights is an important issue for future study. A possible solution is to consider model averaging bandwidths; see Henderson and Parmeter (2016) and Zhu et al. (2017).

Next, we consider a special case of time-varying parameter dynamic models that contain lagged dependent variables as regressors:
DGP 2 (Dynamic Regression with Smooth Structural Changes):

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

where $\theta_j = 1/\sqrt{2\alpha} j^{-\alpha-1/2}$, $F(\tau) = \tau$, $\epsilon_t = R\varepsilon_t/c$, $c = \sum_{j=1}^{\infty} \theta_j^2$, $\varepsilon_t \sim i.i.d.N(0,1)$ and $\alpha = 1.5$. We allow R^2 to vary on a grid from 0.1 to 0.9.

Furthermore, to investigate the finite sample performance of the TVJMA estimator under DGPs with various structural changes, we consider following three DGPs with Case (ii) for $\{\varepsilon_t\}$. For DGPs 3–5 below, $\theta_j = c\sqrt{2\alpha} j^{-\alpha-1/2}$, with various values of $c > 0$ and $\alpha = 1.5$. These parameter values are the same as those in DGP 1.

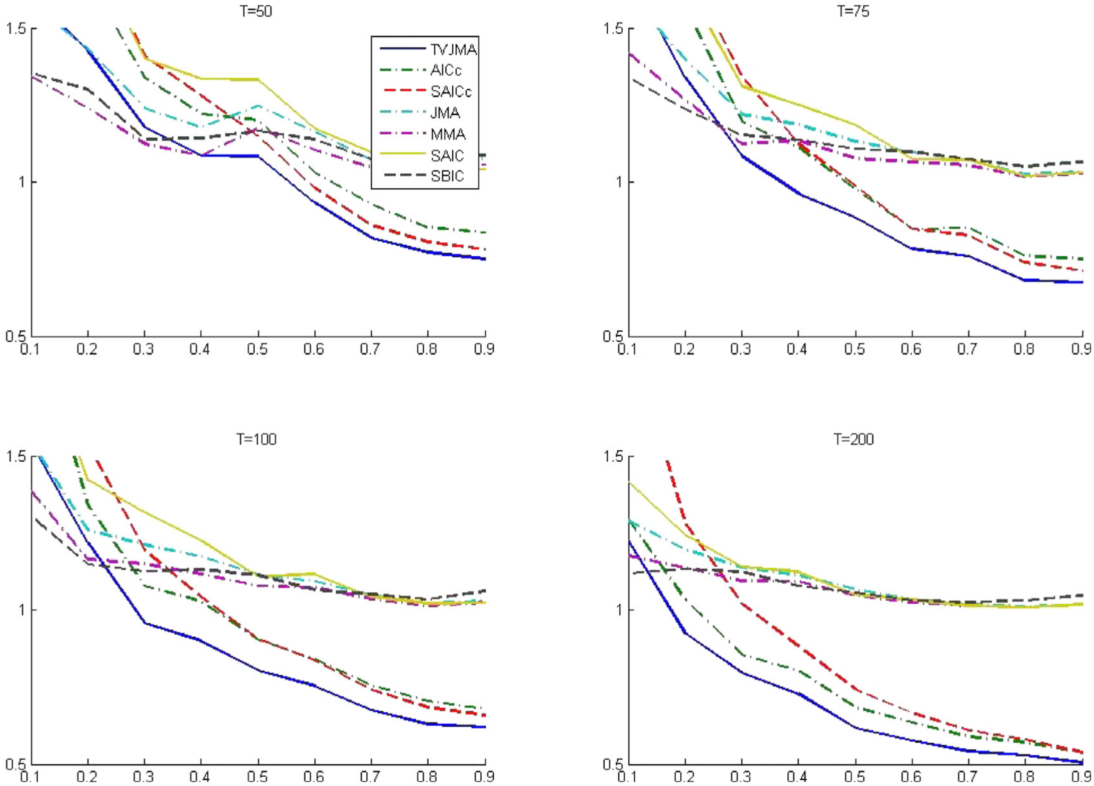

Fig. 2. Finite-sample performance under DGP 1 with Case (ii).
Notes: (1) DGP 1 (Smooth Structural Changes):

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

where $\tau = t/T$, $F(\tau) = \tau^3$, $X_{t1} = 1$, and all other regressors $\{X_{tj}, j \geq 2\}$ are $i.i.d.N(0,1)$ sequences; $\theta_j = c\sqrt{2\alpha}j^{-\alpha-1/2}$, with $c > 0$ and $\alpha = 1.5$.
(2) In each figure, the sample sizes are shown in four panels. The sample size varies from $T = 50$, 75, 100 and 200.
(3) Three cases for $\{\varepsilon_t\}$: Case (i) $\varepsilon_t \sim i.i.d.N(0,1)$; Case (ii) $\varepsilon_t = e_{t,1} + e_{t,2}$, $e_{t,1} \sim N(0, X_{t2}^2)$, $e_{t,2} = \phi e_{t-1,2} + u_t$, $u_t \sim i.i.d.N(0,1)$ and $\phi = 0.5$; Case (iii) $\varepsilon_t = \sqrt{h_t}u_t$, $h_t = 0.2 + 0.5X_{t2}^2$, $u_t \sim i.i.d.N(0,1)$.
(4) In each panel, the y axis and the x axis display the MSE and the population R^2, respectively. Seven methods to estimate parameters are shown in these figures, including TVJMA, AICc in Cai and Tiwari (2000), SAICc, JMA, MMA, SAIC and SBIC estimators.

DGP 3 (Single Structural Break):

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

where $F(\tau) = 0.5I(\tau \leq 0.3) + I(\tau > 0.3)$ and $\tau = t/T$.
DGP 4 (Smooth Transition Regression):

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

where $F(\tau) = 1.5 - 1.5\exp(-3(\tau - 0.3)^2)$ and $\tau = t/T$.
DGP 5 (Smooth Structural Changes with Periodicity):

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

where $F(\tau) = \sin(\pi\tau^2)$ and $\tau = t/T$.

For each of DGPs 2–5, we generate N data sets of the random sample $\{Y_t, X_{tj}, j \geq 2\}$ for each sample size $T = 50,\ 75,\ 100$ and 200, where $X_{t1} = 1$ and observations on all other regressors $\{X_{tj}, j \geq 2\}$ are generated from $i.i.d.N(0,1)$ sequences. The candidate models and their parameter estimation methods under DGPs 2–5 are the same as those under

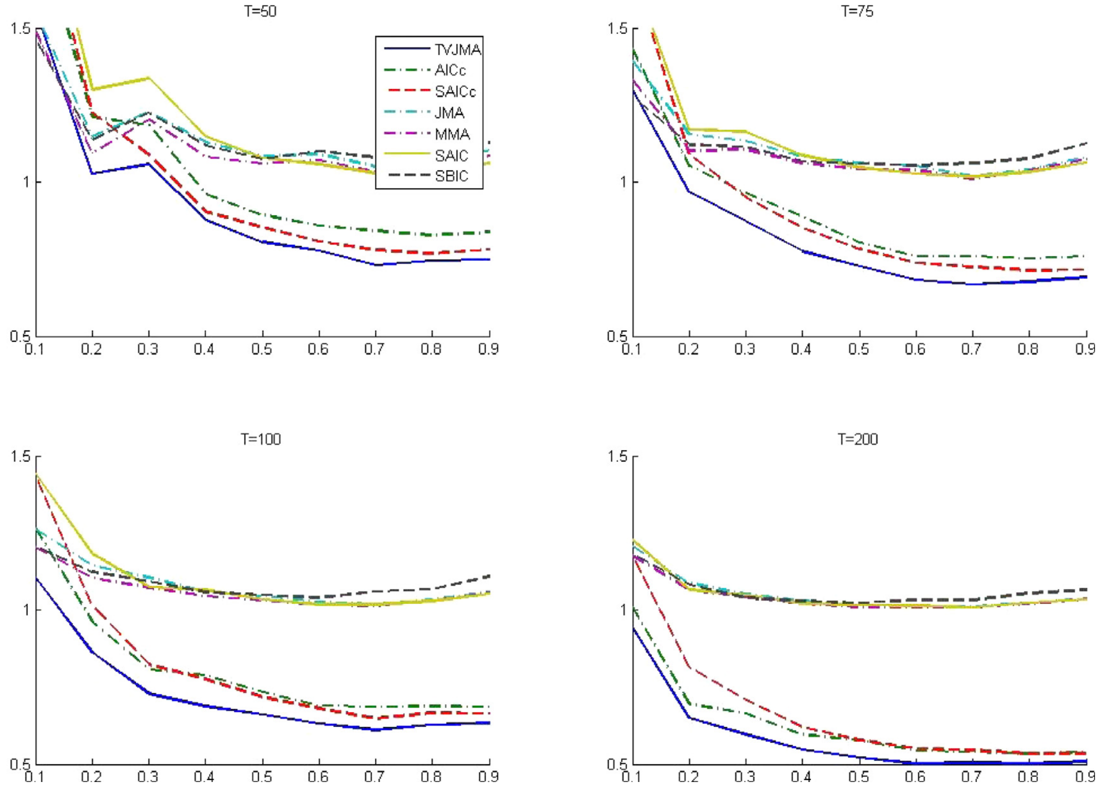

Fig. 3. Finite-sample performance under DGP 1 with Case (iii).
Notes: (1) DGP 1 (Smooth Structural Changes):

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

where $\tau = t/T$, $F(\tau) = \tau^3$, $X_{t1} = 1$, and all other regressors $\{X_{tj}, j \geq 2\}$ are $i.i.d.N(0, 1)$ sequences; $\theta_j = c\sqrt{2\alpha} j^{-\alpha-1/2}$, with $c > 0$ and $\alpha = 1.5$.
(2) In each figure, the sample sizes are shown in four panels. The sample size varies from $T = 50$, 75, 100 and 200.
(3) Three cases for $\{\varepsilon_t\}$: Case (i) $\varepsilon_t \sim i.i.d.N(0, 1)$; Case (ii) $\varepsilon_t = e_{t,1} + e_{t,2}$, $e_{t,1} \sim N(0, X_{t2}^2)$, $e_{t,2} = \phi e_{t-1,2} + u_t$, $u_t \sim i.i.d.N(0, 1)$ and $\phi = 0.5$; Case (iii) $\varepsilon_t = \sqrt{h_t} u_t$, $h_t = 0.2 + 0.5X_{t2}^2$, $u_t \sim i.i.d.N(0, 1)$.
(4) In each panel, the y-axis and the x-axis display the MSE and the population R^2, respectively. Seven methods to estimate parameters are shown in these figures, including TVJMA, AICc in Cai and Tiwari (2000), SAICc, JMA, MMA, SAIC and SBIC estimators.

DGP 1. Specifically, DGP 2 is a dynamic linear regression model with time-varying parameters, which is based on Section 4. DGPs 3–5 are based on the same set-up as that of DGP 1, except that DGPs 3–5 focus on various structural changes with Case (ii) for $\{\varepsilon_t\}$. The results are reported in Figs. 4–7.

In Fig. 4, we consider the dynamic regression model with smooth time-varying parameters under DGP 2. When the sample size T is large enough, the TVJMA estimator yields a smaller risk than all other four estimators. This is even more clear for small R^2.

In Fig. 5, we consider the deterministic single break under DGP 3, namely, a single break with a given breakpoint and size. The TVJMA estimator, not surprisingly, outperforms all other estimators when the sample size T is larger than 50 for all R^2, while AICc and SAICc yield smaller risks than SAIC and SBIC respectively; see, for example, the case with $T = 200$ and $R^2 > 0.4$.

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

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

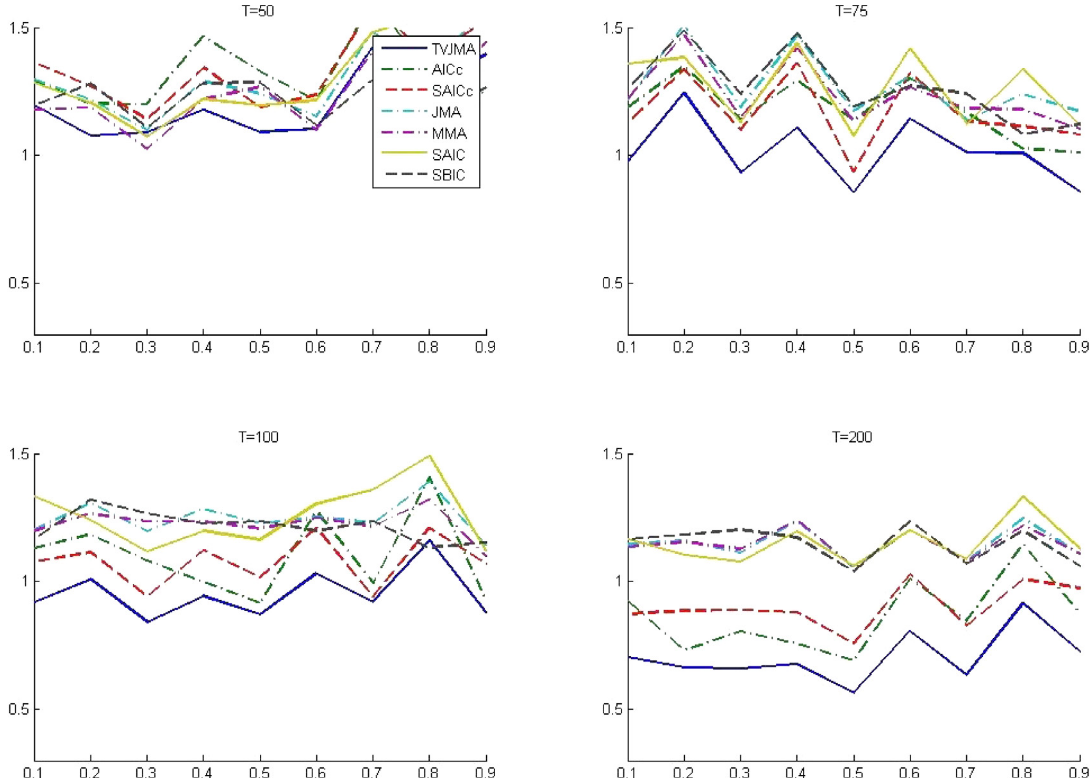

Fig. 4. Finite-sample performance under DGP 2.
Notes: (1) DGP 2 (Dynamic Regression with Smooth Structural Changes):

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

where $\theta_j = 1/\sqrt{2\alpha} j^{-\alpha-1/2}$, $c = \sum \theta_j^2$, $F(\tau) = \tau$, $\epsilon_t = R\varepsilon_t/c$ with R^2 varying on a grid from 0.1 to 0.9, $\varepsilon_t \sim i.i.d.N(0, 1)$ and $\alpha = 1.5$.
(2) In each figure, the sample sizes are shown in four panels. The sample size varies from $T = 50$, 75, 100 and 200.
(3) In each panel, the y-axis and the x-axis display the MSE and the population R^2, respectively. Seven methods to estimate parameters are shown in these figures, including TVJMA, AICc in Cai and Tiwari (2000), SAICc, JMA, MMA, SAIC and SBIC estimators.

To sum up, the TVJMA estimator achieves the lowest risk among all the model averaging estimators under various DGPs. When the sample size T increases, even for small R^2, the TVJMA appears to be the best estimator. When R^2 is large, the SAICc estimator achieves a lower risk than the AICc model selection, which is consistent with the findings in the earlier literature. However, both of them perform worse than the TVJMA estimator for large T and all R^2. We also consider a benchmark nonparametric local constant estimator without any model selection. It is shown that the local constant model without model selection performs quite poorly relative to other methods in most cases. Furthermore, following a referee's suggestion, we also compare the TVJMA estimator with a time-varying leave-k-out cross-validation model averaging (LkoMA) method (e.g., Gao et al. (2016)). We find that when R^2 is small, the TVJMA estimator outperforms the time-varying LkoMA estimator under different DGPs, especially DGP 2. Nevertheless, when R^2 is large, the time-varying LkoMA estimator achieves a slightly lower risk than the TVJMA estimator except for DGP 2. Developing optimal time-varying leave-k-out cross-validation weight selection methods and extending the proof technique for the asymptotic optimality property are important topics for future research.

6. Empirical application

It is widely accepted that stock return predictability is an important yet controversial issue in empirical finance. The conventional wisdom, studied by Campbell (1990) and Cochrane (1996), is that aggregate dividend yields strongly forecast excess stock return, even at longer horizons. Other commonly used predictive variables are financial ratios, such as dividend–price ratio, earnings–price ratio, and book-to-market ratio (Rozeff, 1984; Fama and French, 1988; Campbell and Shiller, 1988; Lewellen, 2004), as well as corporate payout and financing activity (Lamont, 1998; Baker and Wurgler, 2000). However, Wang (2003) and Welch and Goyal (2008) show that predictive regressions of excess stock returns perform poorly in out-of-sample forecasts of the U.S. equity premium while historical average returns generate superior forecasts, which causes vigorous debates in the literature (Campbell and Thompson, 2008). It is possible that the presence

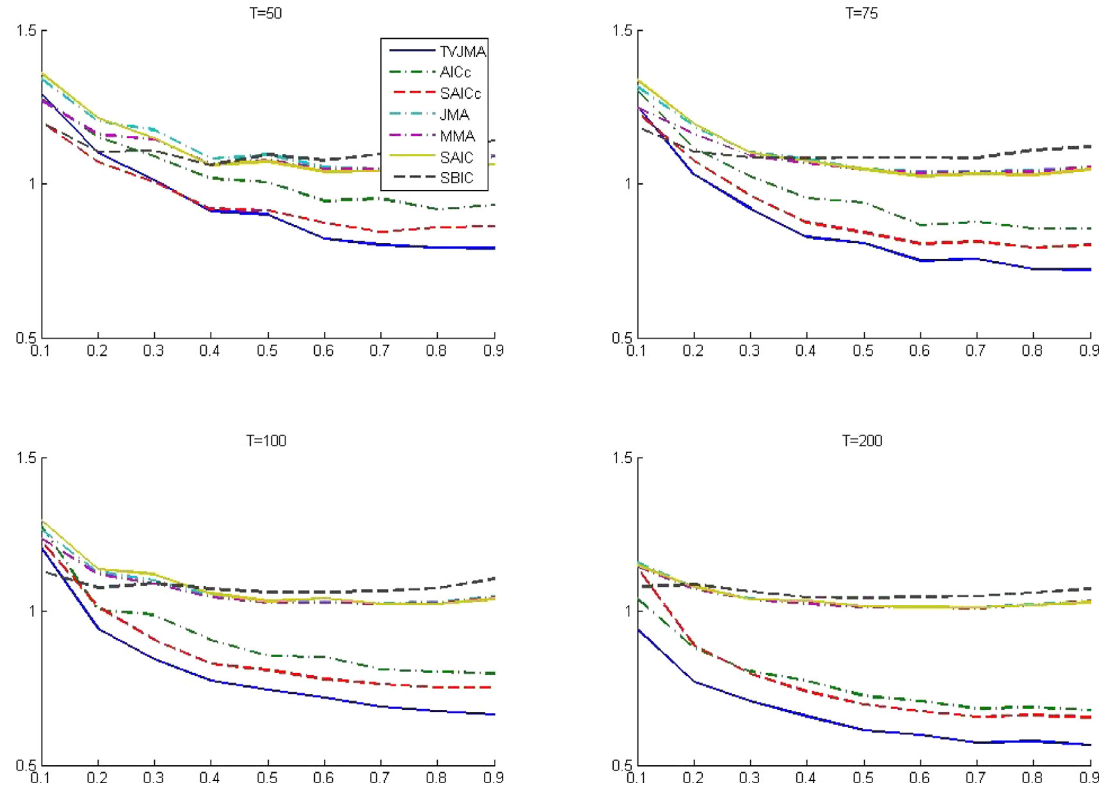

Fig. 5. Finite-sample performance under DGP 3 with Case (ii).
Notes: (1) DGP 3 (Single Structural Break):

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

where $F(\tau) = 0.5I(\tau \leq 0.3) + I(\tau > 0.3)$, $\tau = t/T$, $X_{t1} = 1$, and all other regressors $\{X_{tj}, j \geq 2\}$ are $i.i.d.N(0, 1)$ sequences; $\theta_j = c\sqrt{2\alpha} j^{-\alpha-1/2}$, with $c > 0$ and $\alpha = 1.5$.
(2) In DGP 3, $\{\varepsilon_t\}$ follows Case (ii) $\varepsilon_t = e_{t,1} + e_{t,2}$, $e_{t,1} \sim N(0, X_{t2}^2)$, $e_{t,2} = \phi e_{t-1,2} + u_t$, $u_t \sim i.i.d.N(0, 1)$ and $\phi = 0.5$.
(3) In each figure, the sample sizes are shown in four panels. The sample size varies from $T = 50$, 75, 100 and 200.
(4) In each panel, the y-axis and the x-axis display the MSE and the population R^2, respectively. Seven methods to estimate parameters are shown in these figures, including TVJMA, AICc in Cai and Tiwari (2000), SAICc, JMA, MMA, SAIC and SBIC estimators.

of structural changes leads to a changing predictive relationship. Indeed, Pesaran and Timmermann (2007) find that the size of parameter variations between the break points in models is considerably large, and the parameter estimates of dividend yields take even opposite signs before and after 1991. Chen and Hong (2012) find strong evidence against stability in univariate and multivariate predictor regressions for both the postwar and post-oil-shock sample periods. Furthermore, Rapach and Zhou (2013) point out that model instability and uncertainty seriously impair the forecasting ability of predictive regression models.

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

We employ Campbell and Thompson's (2008) popular data set, which is used in Chen and Hong (2012), Jin et al. (2014) and Lu and Su (2015), among many others. We consider the following predictive regression model:

$$Y_{t+1} = \alpha_t + \mathbf{X}_t \boldsymbol{\beta}_t + \varepsilon_{t+1},$$

where $Y_{t+1} = \ln[(P_{t+1} + D_{t+1})/P_t] - r_t$, P_t is the S&P 500 price index, D_t is the dividend paid on the S&P 500 price index, r_t is the 3-month treasury bill rates, $\mathbf{X}_t$ is a set of predictive variables, i.e., $\mathbf{X}_t = (X_{t1}, \ldots, X_{tp})$, and p is the number of predictive variables. Quarterly variables from Welch and Goyal (2008) are available for 1927Q1-2005Q4, since quarterly stock returns before 1927 are constructed by interpolation of lower-frequency data, which may be not reliable.

Following Welch and Goyal (2008) and Rapach et al. (2010), we consider 14 financial and economic variables, sorted by relevance to **Y**: default yield spread (X_1), treasury bill rate (X_2), net equity expansion (X_3), term spread (X_4), log dividend price ratio (X_5), log earnings price ratio (X_6), long-term yield (X_7), book-to-market ratio (X_8), inflation (X_9), log dividend

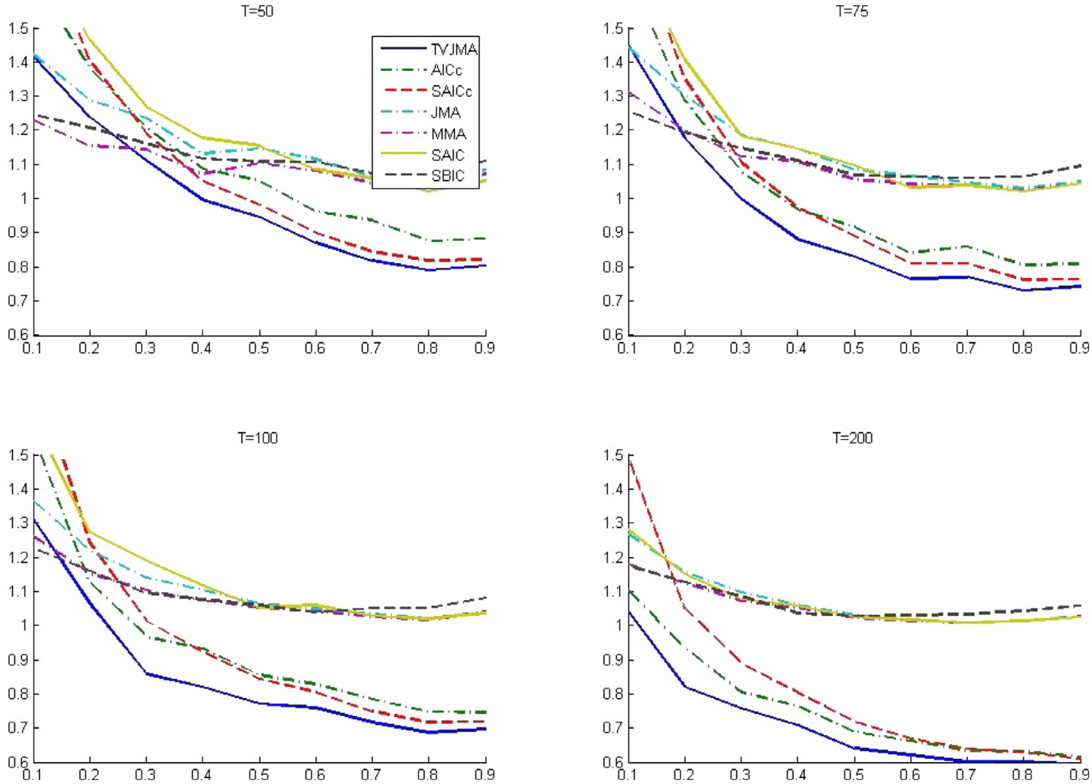

Fig. 6. Finite-sample performance under DGP 4 with Case (ii).
Notes: (1) DGP 4 (Smooth transition regression):

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

where $F(\tau) = 1.5 - 1.5\exp(-3(\tau - 0.3)^2)$, $\tau = t/T$, $X_{t1} = 1$ all other regressors $\{X_{tj}, j \geq 2\}$ are $i.i.d.N(0,1)$ sequences; $\theta_j = c\sqrt{2\alpha} j^{-\alpha-1/2}$, with $c > 0$ and $\alpha = 1.5$.
(2) In DGP 4, $\{\varepsilon_t\}$ follows Case (ii) $\varepsilon_t = e_{t,1} + e_{t,2}$, $e_{t,1} \sim N(0, X_{t2}^2)$, $e_{t,2} = \phi e_{t-1,2} + u_t$, $u_t \sim i.i.d.N(0,1)$ and $\phi = 0.5$.
(3) In each figure, the sample sizes are shown in four panels. The sample size varies from $T = 50$, 75, 100 and 200.
(4) In each panel, the y-axis and the x-axis display the MSE and the population R^2, respectively. Seven methods to estimate parameters are shown in these figures, including TVJMA, AICc in Cai and Tiwari (2000), SAICc, JMA, MMA, SAIC and SBIC estimators.

yield (X_{10}), log dividend payout ratio (X_{11}), stock variance (X_{12}), long-term return (X_{13}), default return spread (X_{14}). For simplicity, we consider the following 14 nested candidate models: $\{1, X_1\}$, $\{1, X_1, X_2\}, \ldots, \{1, X_1, \ldots, X_{14}\}$. All candidate models are time-varying parameter linear regression models, and parameters are estimated by the local constant method in (15) in Section 3.

The estimation sample starts from 1947Q1 and our estimation is based on subsamples with size $T_1 = 80$, 92, 104, 116, 128, 140, 152, 164, 176, 188, 200, 212 and 224, respectively. The remaining observations are used for out-of-sample recursive forecast accuracy assessment. For example, we use the model averaging weights for the time period T_1, $\widehat{\mathbf{w}}_{T_1}$, to construct a forecast of Y_{T_1+1}. After that we input a new observation and recalculate new model averaging weights for the time period $T_1 + 1$ and then obtain a forecast of Y_{T_1+2}. Thus, the out-of-sample forecast periods begins from 1967Q1, 1970Q1, 1973Q1, 1976Q1, 1979Q1, 1982Q1, 1985Q1, 1988Q1, 1991Q1, 1994Q1, 1997Q1, 2000Q1 and 2003Q1, respectively, and all end at 2005Q4. The postwar sample, covering 1947Q1–2005Q4, and the post-oil-shock subsample, covering 1976Q1–2005Q4, are commonly used in the literature, e.g., Welch and Goyal (2008), Chen and Hong (2012), etc. The bandwidth employed in TVJMA, AICc and smoothed AICc is set to be $2.34T_1^{-0.2}$. Following Ullah et al. (2017), we use the out-of-sample $\widetilde{R}^2$ measure:

$$\widetilde{R}^2 = 1 - \frac{\sum_{t=T_1}^{T-1}(Y_{t+1} - \widehat{Y}_{t+1})^2}{\sum_{t=T_1}^{T-1}(Y_{t+1} - \overline{Y})^2},$$

where $\widehat{Y}_{t+1}$ is the prediction of Y_{t+1} based on a given forecast method, and $\overline{Y}$ is the historical average of Y_t over the T_1 observations. This measure represents the relative difference in squared error predictive risks. The negative (positive)

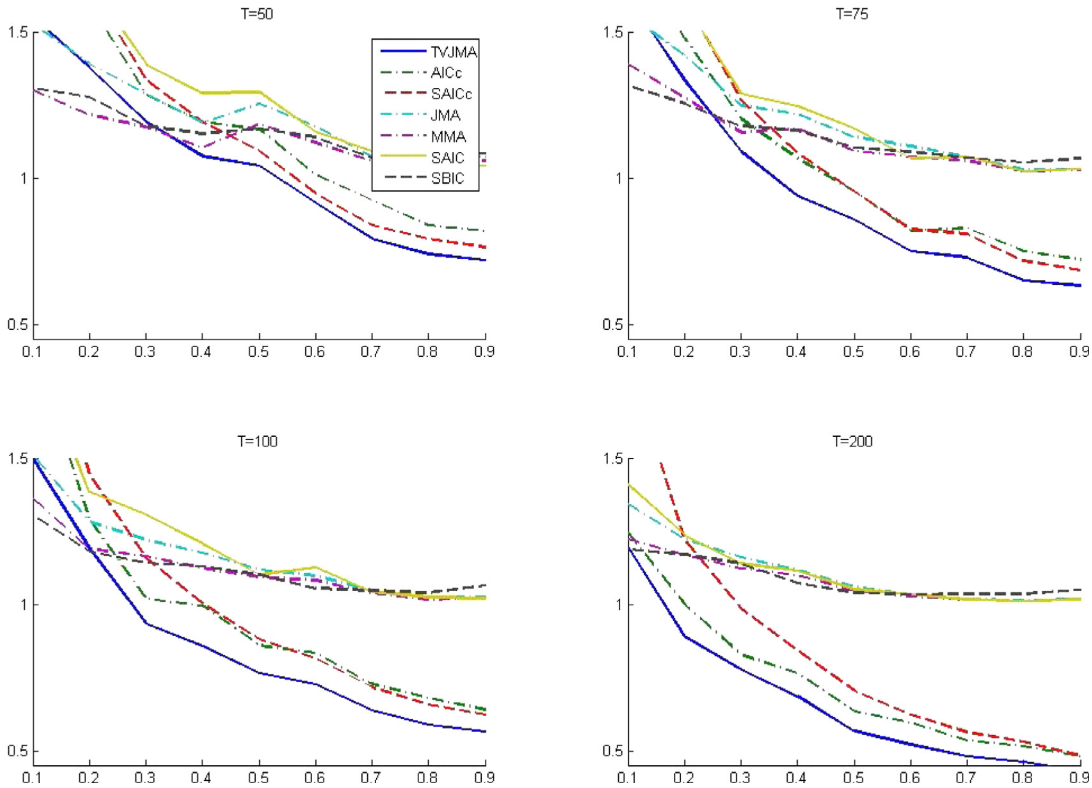

Fig. 7. Finite-sample performance under DGP 5 with Case (ii).
Notes: (1) DGP 5 (Smooth Structural Changes with Periodicity):

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

where $F(\tau) = \sin(\pi\tau^2)$, $\tau = t/T$, $X_{t1} = 1$, and all other regressors $\{X_{tj}, j \geq 2\}$ are $i.i.d.N(0, 1)$ sequences; $\theta_j = c\sqrt{2\alpha}j^{-\alpha-1/2}$, with $c > 0$ and $\alpha = 1.5$.
(2) In DGP 5, $\{\varepsilon_t\}$ follows Case (ii) $\varepsilon_t = e_{t,1} + e_{t,2}$, $e_{t,1} \sim N(0, X_{t2}^2)$, $e_{t,2} = \phi e_{t-1,2} + u_t$, $u_t \sim i.i.d.N(0, 1)$ and $\phi = 0.5$.
(3) In each figure, the sample sizes are shown in four panels. The sample size varies from $T = 50$, 75, 100 and 200.
(4) In each panel, the y-axis and the x-axis display the MSE and the population R^2, respectively. Seven methods to estimate parameters are shown in these figures, including TVJMA, AICc in Cai and Tiwari (2000), SAICc, JMA, MMA, SAIC and SBIC estimators.

value of $\widetilde{R}^2$ suggests that $\widehat{Y}$ yields a larger (smaller) sum of squared one-period forecast errors than the historical average method.

Another measure we use is the mean square predictive error (MSPE), which is widely used in the literature (e.g., Sun et al. (2018)).

$$\text{MSPE} = \frac{1}{T - T_1} \sum_{t=T_1}^{T-1} (Y_{t+1} - \widehat{Y}_{t+1})^2. \tag{23}$$

Tables 1 and 2 compare $\widetilde{R}^2$ and MSPE between the TVJMA estimator and other estimators. We find that in most cases, the TVJMA estimator is almost always the best estimator among all methods considered. Our finding supports the argument of Chen and Hong (2012) that instability exists in univariate predictor models for stock returns and smooth structural change is a possibility, which explains why the TVJMA estimator is more appropriate than JMA and MMA. We note that the JMA estimator yields the second smallest forecast errors in most cases, with the MMA estimator being a close fourth. In most cases, the AICc estimator yields the worst performance. It is possible that the evidence of instability is a bit weak in quarterly data, which is consistent with the findings in Chen and Hong (2012).

7. Conclusion

Although structural changes have received considerable attention in time series econometrics for a long time, no work has attempted to consider time-varying model averaging for both linear and nonlinear candidate models, including those with time-varying parameters. We propose a frequentist method for model averaging with time-varying jackknife

Table 1
Out-of-sample $\widetilde{R}^2$ of different methods.

Estimation	Prediction	TVJMA	AICc	SAICc	JMA	MMA	SAIC	SBIC
		$h = 2.34T_1^{-0.2}$						
1947Q1($T_1 = 80$)	1967Q1	0.1771°	0.0335	0.1018	0.1761°°	0.1657	0.1111	0.1024
1947Q1($T_1 = 92$)	1970Q1	0.1242°	−0.0312	0.0549	0.1228°°	0.1124	0.0512	0.0530
1947Q1($T_1 = 104$)	1973Q1	0.1212°	−0.0443	0.0770	0.1107°°	0.0977	0.0358	0.0367
1947Q1($T_1 = 116$)	1976Q1	0.0372°	−0.1574	−0.0397	0.0025°°	−0.0165	−0.1128	−0.0708
1947Q1($T_1 = 128$)	1979Q1	0.0357°	−0.1727	−0.0291	−0.0188°°	−0.0381	−0.1393	−0.1007
1947Q1($T_1 = 140$)	1982Q1	−0.1057°	−0.3579	−0.1375°°	−0.1830	−0.2064	−0.3210	−0.2220
1947Q1($T_1 = 152$)	1985Q1	−0.1833°	−0.4899	−0.2359°°	−0.2586	−0.2829	−0.4043	−0.2567
1947Q1($T_1 = 164$)	1988Q1	−0.2630°°	−0.6292	−0.2522 °	−0.3773	−0.4091	−0.5724	−0.3658
1947Q1($T_1 = 176$)	1991Q1	−0.2238°°	−0.4310	−0.2242	−0.2194°	−0.2347	−0.2945	−0.3095
1947Q1($T_1 = 188$)	1994Q1	−0.2181	−0.4080	−0.1579°	−0.2207	−0.2161°°	−0.2570	−0.3249
1947Q1($T_1 = 200$)	1997Q1	−0.0423°°	−0.1979	−0.1005	−0.0365°	−0.0538	−0.0735	−0.0990
1947Q1($T_1 = 212$)	2000Q1	0.0125°°	−0.1434	−0.1316	0.0694°	−0.0207	−0.0383	−0.0330
1947Q1($T_1 = 224$)	2003Q1	0.1859	−0.0714	−0.0560	0.1822	0.2909°°	0.3013°	−0.0543

Notes: (1) The estimation sample begins from 1947Q1, with T_1 observations. The prediction period begins from the quarter indicated in the second column.
(2) Seven methods are shown in Table 1: TVJMA, AICc in Cai and Tiwari (2000), SAICc, JMA, MMA, SAIC and SBIC estimators. The larger the criteria, the better the method.
(3) The bandwidth used here is $2.34T_1^{-0.2}$, the same as that in the simulation study.
(4) ° and °° denote the best and the second best forecast among these seven methods, respectively.

Table 2
Out-of-sample MSPE of different methods.

Estimation	Prediction	TVJMA	AICc	SAICc	JMA	MMA	SAIC	SBIC
		$h = 2.34T_1^{-0.2}$						
1947Q1($T_1 = 80$)	1967Q1	0.0758°	0.0891	0.0827	0.0759°°	0.0769	0.0819	0.0827
1947Q1($T_1 = 92$)	1970Q1	0.0803°	0.0945	0.0866	0.0804°°	0.0814	0.0870	0.0868
1947Q1($T_1 = 104$)	1973Q1	0.0797°	0.0947	0.0837	0.0806°°	0.0818	0.0874	0.0874
1947Q1($T_1 = 116$)	1976Q1	0.0693°	0.0833	0.0748	0.0717°°	0.0731	0.0800	0.0770
1947Q1($T_1 = 128$)	1979Q1	0.0708°	0.0861	0.0755	0.0748°°	0.0762	0.0836	0.0808
1947Q1($T_1 = 140$)	1982Q1	0.0740°	0.0908	0.0761°°	0.0791	0.0807	0.0884	0.0817
1947Q1($T_1 = 152$)	1985Q1	0.0779°	0.0981	0.0814°°	0.0829	0.0845	0.0925	0.0828
1947Q1($T_1 = 164$)	1988Q1	0.0695°°	0.0897	0.0690°	0.0758	0.0776	0.0866	0.0752
1947Q1($T_1 = 176$)	1991Q1	0.0699°°	0.0818	0.0700	0.0697°	0.0706	0.0740	0.0748
1947Q1($T_1 = 188$)	1994Q1	0.0816	0.0943	0.0776°	0.0818	0.0814°°	0.0842	0.0887
1947Q1($T_1 = 200$)	1997Q1	0.0875°°	0.1006	0.0924	0.0870°	0.0885	0.0901	0.0923
1947Q1($T_1 = 212$)	2000Q1	0.0821°°	0.0951	0.0941	0.0774°	0.0849	0.0863	0.0859
1947Q1($T_1 = 224$)	2003Q1	0.0395	0.0520	0.0513	0.0397	0.0344°°	0.0339°	0.0512

Notes: (1) For comparison, all results are multiplied by 10. The estimation sample begins from 1947Q1, with T_1 observations. The prediction period begins from the quarter indicated in the second column.
(2) Seven methods are shown in Table 2: TVJMA, AICc in Cai and Tiwari (2000), SAICc, JMA, MMA, SAIC and SBIC estimators. The smaller the criteria, the better the method.
(3) The bandwidth used here is $2.34T_1^{-0.2}$, the same as that in the simulation study.
(4) ° and °° denote the best and the second best forecast among these seven methods, respectively.

combination weights. This method is more appropriate than the conventional MMA and JMA methods under structural changes. It is shown that our TVJMA estimator is asymptotically optimal in the sense of achieving the lowest possible squared error loss in a class of time-varying model average estimators. In a simulation study, we document that the TVJMA method outperforms a variety of existing methods, including a nonparametric version of the bias-correct AIC method. An application to predicting stock returns also demonstrates that the TVJMA method outperforms many model averaging methods.

We conclude this paper by pointing out some important areas of future work. First, it would be interesting to propose a time-varying lasso-type method to select relevant regressors from a set of many potential predictive variables in the first step, and then consider time-varying model averaging in the second step. This would allow different sets of regressors (so different models) in different time periods. In these scenarios, time-varying model averaging weights are expected to yield robust and accurate forecasts. Second, this paper has only considered a global bandwidth for the TVJMA estimator, which may be severely affected by the existence of structural changes. It will be desirable to use a time-varying bandwidth for each time point. Finally, an extension of "leave-k-out" cross-validation model averaging (e.g., Gao et al. (2016)) to allow for time-varying combination weights would be highly interesting, which may be more appropriate for averaging time series models under structural changes.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.02.006.

References

Allen, D.M., 1974. The relationship between variable selection and data agumentation and a method for prediction. Technometrics 16 (1), 125–127.
Ando, T., Li, K.-C., 2014. A model-averaging approach for high-dimensional regression. J. Amer. Statist. Assoc. 109 (505), 254–265.
Andrews, D.W., 1991. Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. J. Econometrics 47 (2–3), 359–377.
Baker, M., Wurgler, J., 2000. The equity share in new issues and aggregate stock returns. J. Finance 55 (5), 2219–2257.
Blundell, R., Duncan, A., Pendakur, K., 1998. Semiparametric estimation and consumer demand. J. Appl. Econometrics 13 (5), 435–461.
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. J. 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., 1990. A variance decomposition for stock returns. Econom. J. 101 (405), 157–179.
Campbell, J.Y., Shiller, R.J., 1988. The dividend-price ratio and expectations of future dividends and discount factors. Rev. Financ. Stud. 1 (3), 195–228.
Campbell, J.Y., Thompson, S.B., 2008. Predicting excess stock returns out of sample: can anything beat the historical average? Rev. Financ. Stud. 21 (4), 1509–1531.
Chan, K.S., Tong, H., 1986. On estimating thresholds in autoregressive models. J. Time Series Anal. 7 (3), 179–190.
Chen, B., 2015. Modeling and testing smooth structural changes with endogenous regressors. J. Econometrics 185 (1), 196–215.
Chen, B., Hong, Y., 2012. Testing for smooth structural changes in time series models via nonparametric regression. Econometrica 80 (3), 1157–1183.
Chen, J., Li, D., Linton, O., Lu, Z., 2018. Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. J. Amer. Statist. Assoc. 113 (522), 919–932.
Cochrane, J.H., 1996. A cross-sectional test of an investment-based asset pricing model. J. Political Econ. 104 (3), 572–621.
Dahlhaus, R., 1996. On the Kullback-Leibler information divergence of locally stationary processes. Stochastic Process. Appl. 62 (1), 139–168.
Dahlhaus, R., 1997. Fitting time series models to nonstationary processes. Ann. Statist. 25 (1), 1–37.
Epanechnikov, V.A., 1969. Non-parametric estimation of a multivariate probability density. Theory Probab. Appl. 14 (1), 153–158.
Fama, E.F., French, K.R., 1988. Dividend yields and expected stock returns. J. Financ. Econ. 22 (1), 3–25.
Gao, Y., Zhang, X., Wang, S., 2016. Model averaging based on leave-subject-out cross-validation. J. Econometrics 192, 139–151.
Geisser, S., 1975. The predictive sample reuse method with applications. J. Amer. Statist. Assoc. 70 (350), 320–328.
Grant, A.P., 2002. Time-varying estimates of the natural rate of unemployment: a revisitation of Okun's law. Q. Rev. Econ. Finance 42 (1), 95–113.
Hansen, B.E., 2001. The new econometrics of structural change: dating breaks in US labor productivity. J. Econ. Perspect. 15 (4), 117–128.
Hansen, B.E., 2007. Least squares model averaging. Econometrica 75 (4), 1175–1189.
Hansen, B.E., 2008. Least-squares forecast averaging. J. Econometrics 146 (2), 342–350.
Hansen, B.E., Racine, J.S., 2012. Jackknife model averaging. J. Econometrics 167 (1), 38–46.
Henderson, D.J., Parmeter, C.F., 2016. Model averaging over nonparametric estimators. Adv. Econom. 36, 539–560.
Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. J. Amer. Statist. Assoc. 98 (464), 879–899.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statist. Sci. 14 (4), 382–401.
Ing, C.K., Wei, C.Z., 2003. On same-realization prediction in an infinite-order autoregressive process. J. Multivariate Anal. 85 (1), 130–155.
Jin, S., Su, L., Ullah, A., 2014. Robustify financial time series forecasting with bagging. Econometric Rev. 33 (5–6), 575–605.
Lamont, O., 1998. Earnings and expected returns. J. Finance 53 (5), 1563–1587.
Lehmann, E.L., Casella, G., 2006. Theory of Point Estimation. Springer Science & Business Media.
Lewellen, J., 2004. Predicting returns with financial ratios. J. Financ. Econ. 74 (2), 209–235.
Li, K.C., 1987. Asymptotic optimality for Cp, CL, cross-validation and generalized cross-validation: discrete index set. Ann. Statist. 15 (3), 958–975.
Lin, C.F.J., Teräsvirta, T., 1994. Testing the constancy of regression parameters against continuous structural change. J. Econometrics 62 (2), 211–228.
Liu, C.-A., 2015. Distribution theory of the least squares averaging estimator. J. Econometrics 186 (1), 142–159.
Liu, Q., Okui, R., 2013. Heteroskedasticity-robust *Cp* model averaging. Econom. J. 16, 462–473.
Lu, X., Su, L., 2015. Jackknife model averaging for quantile regressions. J. Econometrics 188 (1), 40–58.
Pesaran, M., Timmermann, A., 2007. Selection of estimation window in the presence of breaks. J. Econometrics 137 (1), 134–161.
Rapach, D.E., Strauss, J.K., Zhou, G., 2010. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. Rev. Financ. Stud. 23 (2), 821–862.
Rapach, D.E., Zhou, G., 2013. Forecasting stock returns. Handb. Econ. Forecast. 2 (Part A), 328–383.
Robinson, P.M., 1989. Nonparametric estimation of time-varying parameters. Stat. Anal. Forecast. Econ. Struct. Change 19 (2), 253–264.
Robinson, P.M., 1991. Consistent nonparametric entropy-based testing. Rev. Econom. Stud. 58 (3), 437–453.
Rossi, B., 2006. Are exchange rates really random walks? Some evidence robust to parameter instability. Macroecon. Dyn. 10 (01), 20–38.
Rossi, B., Sekhposyan, T., 2011. Understanding models' forecasting performance. J. Econometrics 164 (1), 158–172.
Rothman, P., 1998. Forecasting asymmetric unemployment rates. Rev. Econ. Stat. 80 (1), 164–168.
Rozeff, M.S., 1984. Dividend yields are equity risk premiums. J. Portf. Manag. 11 (1), 68–75.
Shao, J., 1997. An asymptotic theory for linear model selection. Statist. Sinica 7 (2), 221–242.
Stock, J.H., Watson, M.W., 1996. Evidence on structural instability in macroeconomic time series relations. J. Bus. Econom. Statist. 14 (1), 11–30.
Stock, J.H., Watson, M.W., 2002. Has the business cycle changed and why? NBER Macroecon. Annu. 17, 159–218.
Stock, J.H., Watson, M.W., 2003. Forecasting output and inflation: the role of asset prices. J. Econ. Lit. 41 (3), 788–829.
Stock, J.H., Watson, M.W., 2005. Understanding changes in international business cycle dynamics. J. Eur. Econ. Assoc. 3 (5), 968–1006.
Stock, J.H., Watson, M.W., 2007. Why has US inflation become harder to forecast? J. Money Credit Bank. 39 (s1), 3–33.
Stock, J.H., Watson, M.W., 2012. Generalized shrinkage methods for forecasting using many predictors. J. Bus. Econom. Statist. 30 (4), 481–493.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. B 36 (2), 111–147.
Sun, Y., Hong, Y., Wang, S., 2018. Selection of an Optimal Rolling Window in Time-varying Predictive Regression. Cornell University, Manuscript.
Twrdy, E., Batista, M., 2016. Modeling of container throughput in Northern Adriatic ports over the period 1990–2013. J. Transp. Geogr. 52, 131–142.
Ullah, A., Wan, A.T., Wang, H., Zhang, X., Zou, G., 2017. A semiparametric generalized ridge estimator and link with model averaging. Econometric Rev. 36 (1–3), 370–384.
Vogt, M., 2012. Nonparametric regression for locally stationary time series. Ann. Statist. 40 (5), 2601–2633.
Wan, A.T., Zhang, X., Zou, G., 2010. Least squares model averaging by Mallows criterion. J. Econometrics 156 (2), 277–283.
Wang, K.Q., 2003. Asset pricing with conditioning information: a new test. J. Finance 58 (1), 161–196.
Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financ. Stud. 21 (4), 1455–1508.
Yang, Y., 2001. Adaptive regression by mixing. J. Amer. Statist. Assoc. 96 (454), 574–588.
Yuan, Z., Yang, Y., 2005. Combining linear regression models: when and how? J. Amer. Statist. Assoc. 100 (472), 1202–1214.
Zhang, X., 2015. Consistency of model averaging estimators. Econom. Lett. 130, 120–123.
Zhang, X., Liu, C.-A., 2019. Inference after model averaging in linear regression models. Econometric Theory 35, 816–841.
Zhang, X., Wan, A.T., Zou, G., 2013. Model averaging by jackknife criterion in models with dependent data. J. Econometrics 174 (2), 82–94.
Zhu, R., Zhang, X., Wan, A.T., Zou, G., 2017. Kernel Averaging Estimators. Working Paper.