ESSAYS IN HIGH-DIMENSIONAL ECONOMETRICS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Haiqing Zhao

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2020

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Mohitosh Kejriwal, Chair
    Department of Economics
Dr. Yong Bao
    Department of Economics
Dr. Joshua Chan
    Department of Economics
Dr. Justin Tobias
    Department of Economics

Approved by:

Dr. Mohitosh Kejriwal
    Department of Economics

ACKNOWLEDGMENTS

To my thesis committee Mohitosh Kejriwal, Yong Bao, Joshua Chan, and Justin Tobias, thank you for your help and guidance during the past six years. I appreciate your comments which helped me on my way toward becoming a researcher, your patience with me when I made mistakes, and your generosity which made the department more like home to me.

To my mom, dad, and grandparents, thank you for all of the support. Even though I have never discussed my research area with you, talking with you has helped me in some of the hardest times.

To my officemates Pritha Chaudhuri, Tingmingke Lu, and Wumian Zhao, thank you for being so kind to me. You have brought so much fun to the office, and given me the chance to know such great friends. I would also like to thank Ben Raymond, Stanton Hudja, Clint Harris, Mary Kate Batistich, Xuewen Yu, and Daniel Kebede for always being open to talk and offering me help.

Finally, to Logan and his family, my gratitude is endless. Your company and help with my mental health are invaluable.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

ABSTRACT

Zhao, Haiqing Ph.D., Purdue University, August 2020. Essays in High-Dimensional Econometrics. Major Professor: Mohitosh Kejriwal.

My thesis consists of three chapters. The first chapter uses the Factor-augmented Error Correction Model in model averaging for predictive regressions, which provides significant improvements with large datasets in areas where the individual methods have not. I allow the candidate models to vary by the number of dependent variable lags, the number of factors, and the number of cointegration ranks. I show that the leave-h-out cross-validation criterion is an asymptotically unbiased estimator of the optimal mean squared forecast error, using either the estimated cointegration vectors or the nonstationary regressors. Empirical results demonstrate that including cointegration relationships significantly improves long-run forecasts of a standard set of macroeconomic variables. I also estimate simulation-based prediction intervals for six real and nominal macroeconomics variables. The results are consistent with the point estimates, which further support the usefulness of cointegration in long-run forecasts.

The second chapter is a Monte Carlo study comparing the finite sample performance of six recently proposed estimation methods designed for large-dimensional regressions with endogeneity. The methods are based on combining shrinkage estimation with two-stage least squares (2SLS) or generalized method of moments (GMM), where both the number of regressors and instruments can be large. The methods are evaluated in terms of bias and mean squared error of the estimators. I consider a variety of designs with practically relevant features such as weak instruments and heteroskedasticity as well as cases where the number of observations is smaller/larger than the number of regressors/instruments. The consistency results show that the methods using GMM with shrinkage provide smaller estimation errors than the methods using 2SLS with shrinkage. Moreover, the results support the use of cross-validation to select tuning parameters if theoretically derived parameters are unavailable. Lastly, the results indicate that all instruments should correlate with at least one endogenous regressor to ensure estimation consistency.

The third chapter is coauthored with Mohitosh Kejriwal. We present new evidence on the nexus between democracy and growth employing the dynamic common correlated effects (DCCE) approach advanced by Chudik and Pesaran (2015), which is robust to both parameter heterogeneity and cross-section dependence. The DCCE results indicate a positive and statistically significant effect of democracy on economic growth, with a point estimate between approximately 1.5-2% depending on the specification. We complement our estimates with a battery of diagnostic tests for heterogeneity and cross-section dependence that corroborate the use of the DCCE approach.

1. FACTOR-AUGMENTED ERROR CORRECTION MODEL AVERAGING IN PREDICTIVE REGRESSIONS

1.1 Introduction

The use of factor models in regressions has substantially increased in recent years, especially as big data has become widely available. Using a reduced number of common components to represent the comovement among a panel of predictors increases the degrees of freedom, thereby improving the parsimony of the regression model and mitigating the curse of dimensionality. Applications such as forecasting benefit from the use of factors extracted from a large predictor panel with cross-sectional dependence. Examples include Stock and Watson (2002a), Stock and Watson (2002b), Eickmeier and Ziegler (2008), Kim and Swanson (2014), and Leroux et al. (2017). These papers explore forecasting U.S. or Canadian macroeconomic variables and find that factor-augmented regressions (FARs) outperform autoregressive models and model selection methods.

Banerjee et al. (2014) introduce the Factor-augmented Error Correction Model (FECM) for forecasting, which builds on FARs using nonstationary information. The error correction terms represent the long-run relations between the nonstationary target variable and the predictors, and help guide the short-run changes of the target variable moving toward the long-run equilibrium. Empirically, Banerjee et al. (2014) demonstrate that including the error correction terms increases the forecasting accuracy for some real variables under long forecasting horizons.

Model averaging methods have also been widely proven to be useful in predictive regressions. These methods assign weights to each candidate model in order to optimize the trade-off between bias and variance and achieve the lowest weighted mean-squared forecast error (MSFE). Hansen (2008), Hansen (2010), and Hansen and

Racine (2012) derive the asymptotic properties of different model averaging methods. These papers show that under appropriate penalties, the weighted MSFE after normalization converges to the optimal MSFE. Tu and Yi (2017) are able to include nonstationary information by combining a constrained vector autoregressive model with an unconstrained error correction model. They employ Mallows model averaging and use the nonstationary variables directly as regressors, which is in contrast to the pre-estimations of cointegration vectors in Banerjee et al. (2014). Tu and Yi show that the OLS coefficients of the I(1) regressors are consistent assuming cointegrations. The population values for those OLS coefficients are the products of the adjustment parameters with the cointegration vectors. For their empirical application, Tu and Yi (2017) forecast inflation using the one-month treasury bill rate, and show that model averaging provides smaller MSFEs than models using pre-estimated unit root and cointegration test results.

Cheng and Hansen (2015) apply model averaging methods to FARs, thereby taking advantage of both regressor combinations and model combinations. Their simulation and empirical results demonstrate some forecasting improvements of FAR model averaging; however, they do not incorporate any nonstationary information. Leroux et al. (2017) also show that the combination of FARs with model averaging methods performs well in forecasting real activities, but not price variables. The one-year-ahead forecasts of price variables in particular, have larger efficiency losses compared to the one-month-ahead forecasts. Moreover, Swanson and Xiong (2018) and Elliott et al. (2015) show that forecasting nominal variables, such as the changes of interest rates or inflation, does not benefit from combining model averaging techniques with FAR as compared to using simple (vector) autoregressive models. The Tu and Yi (2017) framework does include potential nonstationary information in predictions, but is not suitable for a large panel of predictors. Additionally, their asymptotic theory is limited to two models with one-step-ahead forecasts.

In this paper, I bring the two ideas of using nonstationary information and using large datasets together, and show that adding error correction terms to FAR model averaging improves forecasting performance particularly for multi-step predictive regressions. I allow for the averaging of multiple FAR models and FECMs with different numbers of factors and cointegration vectors, with the largest model nesting both FAR models and FECMs. To incorporate the nonstationary information, I use the pre-estimated cointegration vectors to generate the regressors. As comparisons, I also use the estimated I(1) variables directly as regressors in the empirical work.

The value of using cointegration information in multi-step forecasting exercises is especially evident in forecasting nominal macroeconomic variables. Nominal variables are more likely to be influenced by mean shifts in the long-run trends, due to long-run economic restrictions such as price stickiness. The Beveridge-Nelson trend and cycle decomposition also shows that the short-run stationary dynamics can be influenced by the shocks to the long-run nonstationary trend components. This relation can be captured in the predictive regression by cointegrations. One example of nominal variable forecasting in the long run is trend inflation. Faust and Wright (2013) analyze the forecasting performance of inflation through a set of models, where they show that higher forecasting accuracy is generally accompanied by the inclusion of the slow-varying inflation trend. They argue that the low-frequency level shifts of the inflation trend usually involve fundamental changes in the overall economy, which the short-run dynamics are incapable of capturing. Additionally, the added long-run equilibrium relationships estimated through cointegrations can serve as shrinkage principles, which can enhance the out-of-sample predictions. For these reasons, I include the error correction terms in the predictive regressions in this paper.

I investigate four types of model averaging methods: Mallows model averaging, leave-h-out cross-validation model averaging, Bayesian model averaging, and simple averaging. My empirical results demonstrate that by applying FECM averaging, the forecasts of both real variables, such as industrial production, and nominal variables, such as consumer price index, achieve lower MSFEs over longer horizons compared to FAR model averaging and model selection methods.[1] These results contrast with Banerjee et al. (2014), where they find that FECM is only preferable for real variable predictions. Generally, the 5th, 25th, 50th, and 75th percentiles of the MSFE distributions using FECM averaging are smaller than using FAR model averaging for the macroeconomic variable forecasts, with improvements being particularly pronounced for nominal variables. Results are also consistent across different model averaging methods, which further supports the contribution of FECM averaging to the long-run forecasting.

I use four different panels of macroeconomic variables for the empirical analysis, at either the monthly or quarterly level. Broadly speaking, all four datasets contain output, labor market, and price variables. For brevity, the point forecasts presented in the main text are calculated using the monthly macroeconomic dataset taken from the Federal Reserve Economic Data website for the U.S. forecasting exercise. This dataset is also referred to as the FRED-MD dataset in McCracken and Ng (2016). They suggest that the factors extracted from this dataset have predictive advantages over factors extracted from other regressor panels. I also present the results using the monthly dataset provided by Fortin-Gagnon et al. (2018) to conduct Canadian macroeconomic variable forecasting. All of these four datasets are treated as stationary, and their recumulated counterparts are used as the nonstationary panels from which the nonstationary factors are extracted.[2] To check robustness, I use the panel taken from Stock and Watson (2012) and Cheng and Hansen (2015), where I manually transformed part of their monthly data into quarterly data. Additionally, I use the FRED-QD (quarterly) dataset in McCracken and Ng (2016). Furthermore, to evaluate the source of the FECM averaging forecasting improvement, I provide model averaging results using two subsets of the model candidates, where either the same optimal number of factors is used across models or the cointegration rank is

[1] The industrial production variable is transformed as the growth rate; the consumer price index variable is transformed as the growth rate of inflation.

[2] These choices are consistent with Banerjee et al. (2014), where they retained the I(1) variables from the Stock and Watson (2005) dataset. Essentially, the I(0) series are recumulated to construct the I(1) panel.

pre-determined. By comparing the two sets of results, I show that allowing the cointegration rank to vary across models contributes substantially more to the forecasting performance improvements. This again indicates that shifts in the long-run equilibrium relationships matter more than the short-run dynamics in macroeconomic predictive models.

The first step of the forecasting exercise is to estimate the stationary and nonstationary factors, as well as the cointegration vectors of the nonstationary factors and the predicted variable. I follow Bai and Ng (2004) to take first differences of the I(1) predictor panel, then estimate the stationary factors, and calculate the cumulative sums of the stationary factors to obtain their nonstationary counterparts.[3] Next, I calculate the covariance matrix of the estimated nonstationary factors and predicted variable as in Stock and Watson (1988). The eigenvectors of this covariance matrix can be treated as the cointegration vectors. This matrix is also used to estimate the number of common trends, following the Bai and Ng (2004) PANIC procedure.[4] Alternatively, in the second step forecasts, the recumulated I(1) factors can be used as regressors in place of the estimated cointegrations, as in Tu and Yi (2017).

The second step is to estimate the forecasting regression, wherein model averaging methods are applied. A rolling forecasting scheme is used to mitigate the effects of structural breaks. For each dataset, I report the distributions of the MSFE ratios of FECM averaging, FAR model averaging, and model selection methods over an autoregressive model with twelve lags (AR(12)).[5] Overall, empirical results uniformly indicate that by employing FECM averaging, MSFEs at longer horizons are significantly smaller than using the other methods. The improvement can be as large as 70

[3] The complete discussion of the consistency of these recumulated I(1) factors is presented in Bai and Ng (2004). Alternatively, nonstationary factors can be estimated directly through the procedure in Bai (2004) if the errors are stationary. The empirical results using Bai (2004) to estimate I(1) factors are available upon request.

[4] PANIC stands for Panel Analysis of Nonstationarity in Idiosyncratic and Common Components.

[5] The number of lags is four when using quarterly datasets. Correspondingly, the benchmark model is AR(4).

percent, and the superior performance holds for samples restricted to recession and expansion periods.[6]

Asymptotic properties are presented for the unbiasedness of the Mallows and leave-h-out cross-validation criteria. I show that after normalization, both these criteria converge in distribution to random variables whose means are the infeasible optimal MSFEs. These theoretical derivations are not for the purpose of demonstrating optimality since the predictive models are allowed to contain lags of the dependent variable.[7]

To further demonstrate the advantages of FECM averaging, prediction intervals are estimated using simulations as Zhang and Liu (2018). I use the nonstationary regressors and regressand directly in the forecasting model to apply the theoretical analysis developed by Tu and Yi (2017). The nonstationary factors are estimated according to Bai (2004), assuming the idiosyncratic components are stationary. Theoretical results show that under the fixed parameters setup, the leave-h-out cross-validation criterion asymptotically assigns zero weights to under-fitted models. The procedures are applied to estimate the prediction intervals for six real and nominal variables from the Stock and Watson (2012) panel, and the results show that FECM averaging provides narrower bands than FAR model averaging.

The paper is organized as follows: Section 1.2 discusses the related literature. Section 1.3 presents the estimation procedures, including the first step factor and cointegration estimation, and the second step model averaging methods. Section 1.4 discusses the asymptotic properties. Section 1.5 reports empirical results. Section 1.6 discusses the simulation-based confidence interval and related results. Section 1.7 concludes the chapter. Data preparations, extra tables for robustness, and proofs of the asymptotic properties are included in the appendix. For simplicity, all variables appearing in lowercase are stationary, and all variables appearing in the corresponding

[6] The U.S. recession periods are taken from the following URL: https://www.nber.org/cycles.html. The Canadian recession periods are taken from the following URL: https://www.cdhowe.org/council/business-cycle-council. Both are retrieved on September 3, 2019.

[7] Simulation results are available upon request.

uppercase are nonstationary. The notation ⇒ means converge in distribution, and → means converge in probability.

1.2 Literature Review

This paper contributes to four closely related literatures. First, and most broadly, is the growing literature on factor-augmented predictions. Stock and Watson (2002a) and Stock and Watson (2002b) show that FARs can help increase the forecasting accuracy of industrial production and CPI inflation over autoregressive (AR) predictive models. Eickmeier and Ziegler (2008) survey the FAR predictive performance with European datasets. They show that FARs deliver better inflation forecasting performance for the Euro-area. Using the FAR framework, researchers have proposed a variety of techniques to further improve the predictive power, such as determining the model specification by information criteria (Bai and Ng, 2008), and pre-selecting the number of factors (Bai and Ng, 2002). A recent survey paper by Swanson and Xiong (2018) summarizes these dimension reduction methods through penalized regression methods and FARs. Stock and Watson (2012) argue that several selection methods can be written as shrinkage formulae, and they compare FARs with pretest, bootstrap aggregation, and logit-type shrinkages. Empirical analyses presented in Boivin and Ng (2006), Bai and Ng (2008), Kim and Swanson (2014), and Leroux et al. (2017) also demonstrate that forecasting results from the direct application of FARs are inferior to those from applying shrinkage methods first, then applying FARs.

The second literature involves the use of model averaging methods in forecasting. One shortcoming of model selection methods is that the best model choice is sensitive to the order of factors. To overcome this disadvantage, Cheng and Hansen (2015) (henceforth CH) employ FAR structure in conjunction with model combination methods. Using the Stock and Watson (2012) dataset, CH compare five model averaging methods with the three model selection methods used in Stock and Watson (2012). Results show that about half the variables in the panel achieve higher predictive accuracy using the leave-h-out cross-validation model averaging on FAR models.

The third literature involves FECMs. Banerjee et al. (2014) suggest that adjustments to the long run disequilibrium can be utilized to improve real variable predictions. Specifically, FECM achieves lower MSFEs compared to an AR model more than half the time for longer forecasting horizons. To empirically calculate the forecasts, Banerjee et al. (2014) first estimate the cointegration rank and vectors to form the error correction terms. They then use the estimated error correction terms, as well as the factors and lags, as predictors. In a subsequent paper, Banerjee et al. (2016) continue to explore applying FECM to structural analysis.

While the FECM introduces nonstationary information to the prediction, it also requires additional estimation of cointegration ranks and vectors. In more recent work, Tu and Yi (2017) use nonstationary variables directly as regressors in an error correction model. This direct usage of I(1) variables has benefits and drawbacks. On the one hand, incorporating nonstationary variables avoids determining the cointegration rank and vectors, which also complements the idea of model averaging; on the other hand, estimations with I(1) variables assume that cointegration relationships exist. To both take advantage of this parsimonious estimation method with I(1) variables and accounting for its restriction, I present two sets of empirical results, using the I(1) regressors and estimated cointegration vectors, respectively.

Finally, to estimate the multi-step forecasting results, I consider only direct forecasting models rather than iterated forecasting models. It has been discussed in past literature that the trade-off between these two groups of models lies in the source of estimation errors (See McCracken and McGillicuddy, 2019, and Marcellino et al., 2006). The former model setup provides a larger efficiency gain when the one-step forecasting model is misspecified, and the latter model setup works better for the simple error structures. Given that model averaging methods alleviate the uncertainty involved in model misspecification, and the leave-h-out cross-validation criterion is robust to serial correlation and conditional heteroskedasticity, I focus on the direct forecasting models in this paper.

1.3 Estimation Procedures

1.3.1 Factors and Cointegration Vectors

Stock and Watson (1988)

Let Y_t and $\{X_{it}\}$ be the nonstationary observations for $t = 1, ..., T$ and $i = 1, ..., N$. Furthermore, assume that the unobserved factors satisfy a factor structure $X_{it} = \boldsymbol{\lambda}_i \boldsymbol{F}_t + e_{it}$, and the cointegration relationships between Y_t and $\boldsymbol{F}_t$ are $Y_t + \boldsymbol{\delta}_1' \boldsymbol{F}_t = \eta_{1t}$, ..., $Y_t + \boldsymbol{\delta}_K' \boldsymbol{F}_t = \eta_{Kt}$. Note that if the starting observations of $\boldsymbol{f}_t$ are zeros, then $\boldsymbol{F}_t = \sum_{s=2}^{t} \boldsymbol{f}_s$. The cointegration coefficient of Y_t is normalized to one for notation purposes. This normalization helps avoid additional rotation matrices when estimating the cointegration vectors $\{\boldsymbol{\delta_1, ...\ \delta_K}\}$. If the true cointegration coefficient of Y_t is zero, then the normalization can be written in different ways without loss of generality. The candidate predictors considered in this paper are the historical mean, the lags of the target variable, the estimated factors, and the estimated cointegrations between Y and $\tilde{\boldsymbol{F}}$.[8] Thus, the conditional forecast using the full model is

$$\hat{y}_{t+h|t} = \hat{c} + \hat{\alpha}_1 y_t + ... + \hat{\alpha}_{p+1} y_{t-p} + \sum_{r=1}^{R} \hat{\beta}_r \tilde{f}_{rt} + \sum_{i=1}^{K} \hat{\gamma}_i (Y_t + \tilde{\boldsymbol{\delta}}_i' \tilde{\boldsymbol{F}}_t),$$

where the number of dependent variable lags is p, the number of factors is R, and the cointegration rank is K.[9] From Bai and Ng (2004), $\tilde{\boldsymbol{f}}$ is estimated using Principal Component Analysis from the first-differenced panel $x_{it} = X_{it} - X_{it-1}$, $t = 2, ... T$. The nonstationary factors are then recumulated as $\tilde{\boldsymbol{F}}_t = \sum_{s=2}^{t} \tilde{\boldsymbol{f}}_t$.[10] Next, to estimate the

[8] Equations and formulae in this paper do not include factor lags for brevity.

[9] If a variable occurs with a tilde, it is the first step factor estimation's product. By comparison, a variable with a hat is a product from the second step forecasting regression.

[10] The estimated stationary factors have the time dimension of $T - 1$, and I assume the first period observations of the I(0) factors to be zeros. This setting is consistent with the FECM structure, wherein the I(1) factors' differences are the I(0) factors. These estimated I(0) factors are not exactly the same as the factors extracted directly from the stationary panel, but the results do not change

cointegration vector $\tilde{\boldsymbol{\delta}}_i$, I follow Stock and Watson (1988) and Bai and Ng (2004) to use the matrix $\tilde{\boldsymbol{\theta}}$ from the matrix $\boldsymbol{B} = [\tilde{\boldsymbol{\theta}}_{\perp}, \tilde{\boldsymbol{\theta}}]$, where $\boldsymbol{B}$ is comprised of the (normalized and orthogonal) eigenvectors of $1/T^2 \sum_{t=2}^{T} [Y_t^c,\ \tilde{\boldsymbol{F}}_t^{c'}]'[Y_t^c,\ \tilde{\boldsymbol{F}}_t^{c'}]$. The vector $[Y_t^c,\ \tilde{\boldsymbol{F}}_t^{c'}]$ is constructed using the demeaned Y_t and the demeaned estimated I(1) factors. Given that the maximum factor number is R, the cointegration rank ranges from 0 to R. Thus, there are two extreme cases. The FECM degenerates to a FAR model when the cointegration rank is zero; if the cointegration rank is R, then $\tilde{\boldsymbol{\theta}}$ contains the last R columns of the $\boldsymbol{B}$ matrix.

Tu and Yi (2017)

Tu and Yi (2017) use the nonstationary predicted and predictive variables directly as regressors, instead of estimating cointegration vectors. From the fitted equation in the last subsection, the final part regarding the error correction terms can be rewritten as $\hat{\gamma}_Y Y_t + \sum_{r=1}^{R} \hat{\gamma}_{Fr} \tilde{F}_{rt}$. This will not result in any imbalances of the predictive regressions given that Y_t and $\tilde{\boldsymbol{F}}_t$ are assumed to be cointegrated. Under this setting, the forecast becomes:

$$\hat{y}_{t+h|t} = \hat{c} + \hat{\alpha}_1 y_t + ... + \hat{\alpha}_{p+1} y_{t-p} + \textstyle\sum_{r=1}^{R} \hat{\beta}_r \tilde{f}_{rt} + \hat{\gamma}_Y Y_t + \sum_{r=1}^{R} \hat{\gamma}_{F_r} \tilde{F}_{rt}.$$

Assuming r factors are included, the number of forecasting models with the corresponding nonstationary information is one. This number can be significantly smaller than r, which is the number of extra models introduced to the FAR model averaging set using the estimated cointegration vectors.

1.3.2 Model Averaging Methods

I apply model averaging methods to combine predictions in the second step using the estimated regressors. Let the maximum lag of the target variable be p_{max}, the maximum factor number be R_{max}, and the corresponding cointegration rank be K_{max}.

much. Following Banerjee et al. (2014), in the empirical analysis I assume the recumulated factors to be I(1) and the idiosyncratic components to be I(0).

Then, using $\boldsymbol{z}_t$ to represent the true regressor vector at time t, and $\tilde{\boldsymbol{z}}_t$ to represent the estimated counterpart, the observations of the largest model approximation are:

$$\boldsymbol{z}_t = (1,\ y_t,\ ...,\ y_{t-pmax},\ f_{1t},\ ...,\ f_{R_{max}t},\ (Y_t + \boldsymbol{\delta}_1'\boldsymbol{F_t}), ...,\ (Y_t + \boldsymbol{\delta}_{K_{max}}'\boldsymbol{F_t}))\text{ and}$$

$$\tilde{\boldsymbol{z}}_t = (1,\ y_t,\ ...,\ y_{t-pmax},\ \tilde{f}_{1t},\ ...,\ \tilde{f}_{R_{max}t},\ (Y_t + \tilde{\boldsymbol{\delta}}_1'\tilde{\boldsymbol{F}}_t), ...,\ (Y_t + \tilde{\boldsymbol{\delta}}_{K_{max}}'\tilde{\boldsymbol{F}}_t)).$$

$\boldsymbol{z}_t$ and $\tilde{\boldsymbol{z}}_t$ can also be written as

$$\boldsymbol{z}_t = (1,\ y_t,\ ...,\ y_{t-pmax},\ f_{1t},\ ...,\ f_{R_{max}t},\ Y_t, ...,\ F_{1t},\ ...,\ F_{R_{max}t})\text{ and}$$

$$\tilde{\boldsymbol{z}}_t = (1,\ y_t,\ ...,\ y_{t-pmax},\ \tilde{f}_{1t},\ ...,\ \tilde{f}_{R_{max}t},\ Y_t, ...,\ \tilde{F}_{1t},\ ...,\ \tilde{F}_{R_{max}t})$$

if I(1) regressors are used. Furthermore, suppose there are M_0 model candidates, and $z_t(m)$ represents a subset of z_t for model m, where $m = 1, ...,\ M_0$. The m^{th} model approximation of FECM averaging is $y_{t+h}(m) = \boldsymbol{z}_t(m)'b(m) + \text{error}$. Plugging in the generated factors and cointegration vectors, the conditional forecast of period $T + h$ becomes $\hat{y}_{T+h|T}(m) = \tilde{\boldsymbol{z}}_T(m)'\hat{b}(m)$, where the estimated factors and their levels are consistent with the rotation matrices H_1 and H_2.

The aim of model averaging is to estimate the weights assigned to each model m as approximations to the infeasible weights which minimize the optimal MSFE. In this paper, I focus on showing the asymptotic unbiasedness of the Mallows the leave-h-out cross-validation model averaging criteria as the dynamics of the dependent variable are allowed in the models.

The objective function of the Mallows model averaging is the Mallows criterion: $w = argmin\ \boldsymbol{w'\hat{e}'\hat{e}w} + 2\sigma^2\boldsymbol{w'k}$, where $\hat{\boldsymbol{e}}$ is the OLS residual matrix stacked by all of the residuals obtained from estimating each model candidate. The variance of the error term σ^2 can be consistently estimated as long as the true model is nested in the largest model. After selecting the weights, the final forecast is $\hat{y}_{T+h|T}(\hat{w}) = \sum_{m=1}^{M_0} \hat{w}_m \hat{y}_{T+h|T}^m$.

As discussed in Hansen (2008), the Mallows model averaging can only asymptotically estimate the optimal weights when the forecasting regression error terms are conditional homoskedastic and serially uncorrelated. To adapt to the situation where there exists conditional heteroskedasticity or the forecasting horizons are longer than one, Hansen and Racine (2012) and Hansen (2010) propose the Jackknife model averaging and the leave-h-out cross-validation model averaging, respectively. Hansen and Racine (2012) show that the Jackknife model averaging is a special case of the leave-h-out cross-validation model averaging when the forecasting is one-step-ahead. Under the leave-h-out cross-validation criterion, weights are selected to minimize the objective function $w = argmin\ \boldsymbol{w'\breve{e}'\breve{e}w}$. Similar to the Mallows model averaging, $\boldsymbol{\breve{e}}$ is the residual matrix. For each model candidate, $\breve{e}_t = y_t - \boldsymbol{z}_{t-h}\boldsymbol{\breve{b}}_t$, where $\boldsymbol{\breve{b}}_t = (\sum_{|j-t|\geq h} \boldsymbol{\tilde{z}}_{j-h}\boldsymbol{\tilde{z}}^{'}_{j-h})^{-1}(\sum_{|j-t|\geq h} \boldsymbol{\tilde{z}}_{j-h}y^{'}_{j})$ is the leave-h-out estimator. Essentially, to estimate $\breve{e}_t$ given the moving average component of the model, observations at least h-step away from time t are kept in the estimation.

I also present the forecast performances using non-frequentist model averaging methods. For brevity, I only discuss the procedure in the text, and the empirical results are available upon request. As opposed to calculating the weights to minimize the weighted MSFE, Bayesian model averaging assigns weights based on the Bayesian information criterion (BIC) with $w_i = \frac{exp(-BIC_i/2)}{\sum_{i=1}^{M} exp(-BIC_i/2)}$, and simple model averaging assigns equal weights to all of the model candidates. One application of simple model averaging is the complete subset regression suggested by Elliott et al. (2013) and Elliott et al. (2015). Simulation results in these two papers show that the complete subset regression is superior to shrinkage methods such as bootstrap aggregation.

Lastly, I consider some model selection methods in this paper for comparison. Forecasting with these methods shares the shrinkage representation as stated in Stock and Watson (2012): $\hat{y}_{T+h|T} = \sum_{i=1}^{n} \psi(\kappa t_i)\hat{b}_i x_{iT} + o_p(1)$, where i is the index of regressors, $\hat{b}_i$ represents the OLS estimates from the full model regression, and $\phi(\kappa t_i)$ is a function specific to each model selection method. For the hard threshold pretest, the ψ function is $\psi^{PT}(\kappa t_i)\hat{b}_i = 1(|t_i| > c)$, where t_i is the t-statistic for each coefficient and c is the threshold. For the bootstrap aggregation, the ψ function is $\psi^{BG}(\kappa t_i)\hat{b}_i = \hat{b}_i(1 - \Phi(t_i + c) + \Phi(t_i - c) - t^{-1}[\phi(t_i - c) - \phi(t_i + c)])$, with Φ and ϕ being the cumulative and probability density functions of the standard normal distribution, respectively. For the logit-type shrinkage function, $\psi^{LG}(\kappa t_i)\hat{b}_i = \hat{b}_i \frac{exp(\theta_0+\theta_1|t_i|)}{1+exp(\theta_0+\theta_1|t_i|)}$, and θ_0 and θ_1 are pre-selected to minimize the MSFE of the prediction model.[11]

1.3.3 FECM Averaging Procedures

To conduct the FECM averaging, I first estimate the stationary and nonstationary factors from their corresponding panels. Then I apply Stock and Watson (1988) to estimate the cointegration vectors. The final regressor set includes the lags of the dependent variable, the estimated stationary factors, and the estimated cointegration relationships. Suppose the maximum number of dependent variable lags is p_{max}, and the maximum number of factors in the predictor panel is R_{max}. Tables 1.1 and 1.2 present the model sets to take average from. Each row presents the regressors contained in each model candidate. Table 1.1 contains the model set when the estimated cointegration vectors is used to generate the predictors. The smallest model contains the intercept, then the stationary factors are added as FARs, and finally the nonstationary information is included through the error correction matrix $\boldsymbol{c}$.

[11] In this paper's empirical analysis, I choose the truncation value c as 1.645; θ_0 is chosen from a sequence [-140:5:29], and θ_1 is chosen from a sequence [0:1:21]. These choices are the same as in Stock and Watson (2012).

Table 1.1.: The set of models with estimated cointegration vectors

1	# of models, culmulative
...	
$1\ y_{t-1}\ ...\ y_{t-p_{max}}$	$p_{max}+1$
$1\ y_{t-1}\ ...\ y_{t-p_{max}}\ f_{1t}$	
...	
$1\ y_{t-1}\ ...\ y_{t-p_{max}}\ f_{1t}\ \ ...\ \ f_{R_{max}t}$	$p_{max}+1+R_{max}$
$1\ y_{t-1}\ ...\ y_{t-p_{max}}\ f_{1t}\ ...\ c_{1t}$	
...	
$1\ y_{t-1}\ ...\ y_{t-p_{max}}\ f_{1t}\ \ ...\ \ f_{R_{max}t}\ ...\ \boldsymbol{c}_{R_{max}t}$	$p_{max}+1+R_{max}+(R_{max}+1)R_{max}/2$

From the last panel in Table 1.1, c_{1t} is the error correction term calculated using the eigenvector corresponding to the smallest eigenvalue of $\frac{1}{T^2}\sum_{t=2}^{T}[Y_t^c,\ \tilde{F}_{1t}^{c'}]'[Y_t^c,\ \tilde{F}_{1t}^{c'}]$; $\boldsymbol{c_{qt}}$ is vector containing the error correction terms calculated using the eigenvector matrix corresponding to the last R_{max} smallest eigenvalues of $\frac{1}{T^2}\sum_{t=2}^{T}[Y_t^c,\ \tilde{\boldsymbol{F}}_t^{c'}]'$ $[Y_t^c,\ \tilde{\boldsymbol{F}}_t^{c'}]$. The vector $\tilde{\boldsymbol{F}}_t^{c}$ is a $R_{max}\times 1$ demeaned vector of the estimated nonstationary factors. Given that the number of the included factors is R_{max}, the maximum cointegration rank of the system is also R_{max}, which leads to $(R_{max}+1)R_{max}/2$ number of FECMs in total. For example, if $R_{max}=3$ and $p_{max}=3$, the FAR model averaging set contains seven models, and the FECM averaging set contains six extra models.

As an alternative, the set of regressors considered in the FECM averaging using I(1) variables directly is presented in Table 1.2, with the total number of models being $p_{max}+2R_{max}+1$. This number is significantly smaller than $(R_{max}+1)R_{max}/2+R_{max}+p_{max}+1$. On the one hand, there is less efficiency loss in cointegration vector estimations using I(1) variables directly in the forecasting models. On the other hand, if some of the factor are not cointegrated with the predicted variable, then using the pre-estimated cointegration vectors offers more flexibility as it is conceptually equivalent to regressions with constraints.

Table 1.2.: The set of models with the I(1) regressors

1	# of models, culmulative
...	
$1\ \ y_{t-1}\ \ ...\ \ y_{t-p_{max}}$	$p_{max}+1$
$1\ \ y_{t-1}\ \ ...\ \ y_{t-p_{max}}\ \ f_{1t}$	
...	
$1\ \ y_{t-1}\ \ ...\ \ y_{t-p_{max}}\ \ f_{1t}\ \ ...\ \ f_{R_{max}t}$	$p_{max}+1+R_{max}$
$1\ \ y_{t-1}\ \ ...\ \ y_{t-p_{max}}\ \ f_{1t}\ \ Y_{1t}\ \ F_{1t}$	
...	
$1\ \ y_{t-1}\ \ ...\ \ y_{t-p_{max}}\ \ f_{1t}\ \ ...\ \ f_{R_{max}t}\ \ Y_{1t}\ \ F_{1t}\ \ ...\ \ F_{R_{max}t}$	$p_{max}+1+R_{max}+R_{max}$

1.4 Asymptotic Properties

This section consists of two parts. In the first subsection, I show that the Mallows model averaging criterion is an asymptotically unbiased estimator of the optimal MSFE for one-step-ahead forecasts with conditional homoskedasticity, when cointegration vectors are estimated. The leave-h-out cross-validation model averaging criterion has the similar property for multi-step forecasts and conditional heteroskedasticity. In the second subsection, I show that the asymptotic unbiasedness still holds using the generated I(1) regressors. The theorems are developed under the presumption that the true model is nested in the largest model candidate. I present the assumptions before stating the theorems in each subsection. Finally, I discuss the procedure of generating the critical values of the Bai and Ng (2004) MQ_c^c test through simulations to conduct robustness checks using the estimated cointegration ranks. All of the functions mentioned in this section are defined in the appendix.

1.4.1 Asymptotic Properties with Estimated Cointegration Vectors

Assumptions

The forecasting model considered in this subsection is $y_{t+h} = c+\alpha y_t+\boldsymbol{\beta}'\boldsymbol{f}_t+\gamma_1(Y_t+\boldsymbol{\delta}_1'\boldsymbol{F_t}) + ..\gamma_K(Y_t + \boldsymbol{\delta}_K'\boldsymbol{F_t}) + \epsilon_{t+h}$. Let $\boldsymbol{z}_t = (1,\ y_t,\ ...,\ y_{t-pmax},\ f_{1t},\ ...,\ f_{R_{max}t},\ (Y_t + \boldsymbol{\delta}_1'\boldsymbol{F_t}), ...,\ (Y_t + \boldsymbol{\delta}_{K_{max}}'\boldsymbol{F_t}))$. The assumptions contain two parts: the first part includes the assumptions for the predictive regression, and the second part includes the assumptions for factor estimations. To estimate the nonstationary factors, depending on whether the idiosyncratic components are stationary or not, either Bai and Ng (2004) or Bai (2004) procedure can be applied. For brevity, I only discuss the assumptions using the Bai and Ng (2004) PANIC procedure, given that PANIC allows more generalized error structures. Let M be a generic large constant, N be the cross-sectional dimension, and T be the time dimension. The assumptions for the predictive regression are as follows.

Assumption R. (i) $E(\epsilon_{t+h}|\mathscr{F}_t) = 0$, where $\mathscr{F}_t = \sigma(y_t,\ \boldsymbol{f}_t,\ x_{1t},\ x_{2t},\ ...)$ denote the information set at time t. (ii) $(\boldsymbol{z}_t,\ \epsilon_{t+h})$ is strictly stationary and ergodic. (iii) $E||\boldsymbol{z}_t||^4 \leq M$, $E||\epsilon_t||^4 \leq M$, and $E(\boldsymbol{z}_t\boldsymbol{z}_t') > 0$. (iv) $T^{-1/2}\sum_{t=1-h}^{T-h}\epsilon_{t+h} \xrightarrow{d} N(0,\ \Omega_\epsilon)$, where $\Omega_\epsilon = \sum_{|j|<h} E(\epsilon_{t+h}\epsilon_{t+h-j})$, and $T^{-1/2}\sum_{t=1-h}^{T-h}\boldsymbol{z}_t\epsilon_{t+h} \xrightarrow{d} N(0,\ \Omega)$, where $\Omega = \sum_{|j|<h} E(\boldsymbol{z}_t\boldsymbol{z}_{t-j}'\ \epsilon_{t+h}\epsilon_{t+h-j})$.

These assumptions are the same as CH. Assumptions R(i) and R(ii) assume the unpredictability of y_{t+h}; the stationarity assumption also connects the in-sample and out-of-sample MSFEs. Assumptions R(iii) and R(iv) are the standard moment bounds and the central limit theorems. Note that for this paper $\boldsymbol{z}_t$ contains the error correction terms $\boldsymbol{\eta}_t$. Assumptions R(i) to R(iv) also imply the assumption in Stock (1987) Theorem 1.

Assumption F1. For nonrandom $\boldsymbol{\lambda}_i$, $||\boldsymbol{\lambda}_i|| \leq M$; for random $\boldsymbol{\lambda}_i$, $E||\boldsymbol{\lambda}_i|| \leq M$; $N^{-1}\sum_{i=1}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i' \xrightarrow{p} \Sigma_\Lambda > 0$.

Assumption F2. $(1-L)\boldsymbol{F}_t = \boldsymbol{C}(L)\boldsymbol{\zeta}_t$, where L is the lag operator, and $C(L) = \sum \boldsymbol{C}_j L^j$. (i) $\boldsymbol{\zeta}_t \sim iid(0,\ \Sigma_u), E||\boldsymbol{\zeta}_t||^4 \leq M$. (ii) $var(\Delta\boldsymbol{F}_t) = \sum \boldsymbol{C}_j\Sigma_\zeta\boldsymbol{C}_j > 0$. (iii) $\sum j||\boldsymbol{C}_j|| < M$. (iv) $\boldsymbol{C}(1)$ has rank r_1, $0 \leq r_1 \leq r$.

Assumption F3. $(1-\rho_i L)e_{it} = D_i(L)\psi_{it}$, where $D_i(L) = \sum D_{ij}L^j$. (i) For each i, $\psi_{it} \sim iid(0,\ \sigma^2_{\psi i})$, $E|\psi_{it}|^8 \leq M$, $\sum j|D_{ij}| < M$. (ii) $E(\psi_{it}\psi_{jt}) = \tau_{ij}$ with $\sum_i|\tau_{ij}| \leq M$ for all j. (iii) $E|N^{-1/2}[\psi_{is}\psi_{it} - E(\psi_{is}\psi_{it})]|^4 \leq M$, for every (t,s).

Assumption F4. The errors $\{\psi_{it}\}$, $\{\boldsymbol{\zeta}_t\}$, and $\{\boldsymbol{\lambda}_i\}$ are three mutually independent groups.

Assumption F5. $E||\boldsymbol{F}_1|| \leq M$, and $E|e_{i1}| \leq M$ for every i.

Assumption F6. For all $(i,\ t)$, $E||(NT)^{-1/2}\sum_{t=1-h}^{T-h}\sum_{i=1}^{N}\boldsymbol{\lambda}_i\psi_{it}\epsilon_{t+h}||^2 \leq M$, where $E(\boldsymbol{\lambda}_i\psi_{it}\ \epsilon_{t+h}) = 0$.

Assumptions F1 to F5 are similar to the assumptions made by Bai and Ng (2004). Based on these assumptions, Bai and Ng (2004) Lemma 1 and Lemma 2 show that when $N,\ T \to \infty$, $min\{\sqrt{N},\ T\}(\tilde{f}_t - H_1 f_t) = O_p(1)$ for each t, and $max_{1\leq t\leq T}||\tilde{F}_t - H_1F_t + H_1F_1|| = O_p(T^{1/2}N^{-1/2}) + O_p(T^{-1/4})$. These properties suggest that $T/N \to 0$ is required for showing the consistency of the estimated I(1) factors, and that the cross-sectional dimension needs to be larger than the time dimension if the idiosyncratic components contain nonstationary elements. Given that the estimated levels of factors are recumulated from the estimated first-differenced factors, the rotation matrix for showing the consistency of the estimated I(1) factors is the same as the one for the estimated I(0) factors. If it is assumed that all of the idiosyncratic components are stationary, Bai (2004) suggests to directly apply Principal Component Analysis (PCA) to the nonstationary panel $\{X_{it}\}$. The requirement of the N and T relation is then relaxed under Assumptions A to E in Bai (2004).

Assumptions F1 to F4 and F6 imply Assumption F in CH except F(vi). The purpose of Assumption F(vi) in CH is to relax the requirement $\sqrt{T}/N \to 0$, which is automatically satisfied given $T/N \to 0$. In order to apply the asymptotic properties of factors estimated through PCA to the predictive regressions, I assume that the maximum number of factors used in the predictive regressions is no greater than the true number of factors.

The MSFE of the conditional forecast $\hat{y}_{T+h|T}(w)$ can be written using the relation $MSFE{=}\ E(y_{T+h|T} + \epsilon_{T+h} - \hat{y}_{T+h|T}(w))^2 \approx E(\epsilon^2_{T+h}) + E(y_{T+h|T} - \hat{y}_{T+h|T}(w))^2 = \sigma^2 + E\frac{1}{T}\sum_{t=1}^{t=T-h}(y_{t+h|t} - \hat{y}_{t+h|t}(w))^2$.[12] This shows that the expectation of the in-sample loss plus the unconditional variance of the error term is asymptotically equivalent to the out-of-sample MSFE. Thus, the weights calculated from the in-sample model averaging criteria can be applied to calculate the out-of-sample MSFEs.

Asymptotic Properties using PANIC to Estimate I(1) Factors

Theorem 1. Suppose $h = 1$, $E(\epsilon^2_{t+1}|\mathscr{F}_t) = \sigma^2$. Under Assumptions R and F1 to F6, for fixed M and weights (w), and N, $T \to \infty$ with $T/N \to 0$, $C_T(w) = L_T(w) + T^{-1}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} + 2T^{-1/2}r_{1T}(w) - T^{-1}r_{2T}(w)$, where $r_{1T}(w) \xrightarrow{d} \kappa_1(w)$, $r_{2T}(w) \xrightarrow{d} \kappa_2(w)$, $E(\kappa_1(w)){=}\ 0$, $E(\kappa_2(w)){=}\ 0$, and $L_T(w) = \frac{1}{T}\sum(y_{t+h|t} - \hat{y}_{t+h|t}(w))^2$.

Theorem 2. Under Assumptions R and F1 to F6, for fixed M and weights (w), and N, $T \to \infty$ with $T/N \to 0$, $CV_{h,T}(w) = \check{L}_T(w) + T^{-1}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} + 2T^{-1/2}\check{r}_{3T}(w)$, where $\check{r}_{3T}(w) \xrightarrow{d} \kappa_3(w)$, $E(\kappa_3(w)) = 0$, and $\check{L}_T(w) = \frac{1}{T}\sum(y_{t+h|t} - \check{y}_{t+h|t}(w))^2$. $\check{y}_{t+h|t}(m) = \tilde{z}_t(m)\check{b}_{t,h}(m)$ and $\check{b}_{t,h}(m)$ is the leave-h-out estimator of model m for time t.

[12] The last equality is true given the stationarity of $(y_t,\ \tilde{f}_t)$, and the full discussion can be found at CH. For the estimated cointegration vectors $\tilde{\boldsymbol{\eta}}_t$, Stock (1987) and Stock and Watson (1988) show that they are consistent.

Critical Values of MQ_c^c Tests

Table I in Bai and Ng (2004) provides the MQ_c^c test critical values with up to six common trends. For the purpose of the empirical analysis, I simulate the critical values of the MQ_c^c test with the maximum of eleven common trends. According to Bai and Ng (2004), the limit distribution of the MQ_c^c statistic under the null hypothesis of m common trends is the distribution of the smallest eigenvalue of the matrix $\frac{1}{2}[W_m^c(1)W_m^c(1)' - I_m][\int_0^1 W_m^c(s)W_m^c(s)'ds]^{-1}$, where $W_m^c = W_m - \int_0^1 W_m$ and W_m is an m-vector standard Brownian motion. The same limit distribution holds in this paper, as the system contains the observed nonstationary variable Y_t along with the generated factors. The simulated critical values are presented in Table 1.3.

Table 1.3.: MQ_c^c test critical values

Number of common trend	.01	.05	.10
1	-20.151	-13.73	-11.022
2	-31.621	-23.535	-19.923
3	-41.064	-32.296	-28.399
4	-48.501	-40.442	-36.592
5	-58.383	-48.617	-44.111
6	-66.978	-57.04	-52.312
7	-76.698	-65.749	-60.439
8	-87.382	-75.127	-69.642
9	-94.129	-82.956	-77.755
10	-105.298	-93.118	-86.605
11	-110.979	-100.290	-93.889

1.4.2 Asymptotic Properties with Estimated I(1) Variables

Assumptions

Let $\boldsymbol{z}_{1t} = (1,\ y_t,\ ...,\ y_{t-pmax},\ f_{1t},\ ...,\ f_{R_{max}t})'$, and $\boldsymbol{z}_{2t} = (Y_t,\ F_{1t},\ ...,\ F_{Rt})'$, where $R \in [1,\ R_{max}]$. The forecasting model considered in this subsection is $y_{t+h|t} = c + \alpha_0\beta_0'\boldsymbol{z}_{2t} + \alpha_1 y_t + ... + \alpha_{p+1}y_{t-p} + \sum_{r=1}^{R}\beta_r f_{rt} + \epsilon_{t+h}$. The cointegration structure is assumed to exist, with α_0 being the adjustment matrix, and β_0 being the cointegration matrix.[13]

Assumption R1'. (i) $E(\epsilon_{t+h}|\mathscr{F}_t) = 0$, where $\mathscr{F}_t = \sigma(y_t,\ \boldsymbol{f}_t,\ x_{1t},\ x_{2t},\ ...)$ denote the information set at time t. (ii) $(\boldsymbol{z}_{1t},\ \epsilon_{t+h})$ is strictly stationary and ergodic. (iii) $E||\boldsymbol{z}_{1t}||^4 \leq M$, $E||\epsilon_t||^4 \leq M$, and $E(\boldsymbol{z}_{1t}\boldsymbol{z}_{1t}') > 0$. (iv) $T^{-1/2}\sum_{t=1-h}^{T-h}\epsilon_{t+h} \xrightarrow{d} N(0,\ \Omega_\epsilon)$, where $\Omega_\epsilon = \sum_{|j|<h}E(\epsilon_{t+h}\epsilon_{t+h-j})$, and $T^{-1/2}\sum_{t=1-h}^{T-h}\boldsymbol{z}_{1t}\epsilon_{t+h} \xrightarrow{d} N(0,\ \Omega_1)$, where $\Omega_1 = \sum_{|j|<h}E(\boldsymbol{z}_{1t}\boldsymbol{z}_{1t-j}'\ \epsilon_{t+h}\epsilon_{t+h-j})$.

Assumption R2'. For each FECM candidate, let $\boldsymbol{z}_{2t}^m = (Y_t,\ F_{1t},\ ...,\ F_{mt})'$ with $m \leq R$, and β_0^m be the corresponding cointegration matrix. (i) $\boldsymbol{z}_{2t}^m\beta_0^m$ is strictly stationary and ergodic. (iii) $E||\boldsymbol{z}_{2t}^m\beta_0^m||^4 \leq M$, and $E(\boldsymbol{z}_{2t}^m\boldsymbol{z}_{2t}^{m'}) > 0$. (iv) $T^{-1/2}\sum_{t=1-h}^{T-h}\boldsymbol{z}_{2t}^m\beta_0^m\epsilon_{t+h} \xrightarrow{d} N(0,\ \Omega_2)$, where $\Omega_2 = \sum_{|j|<h}E(\boldsymbol{z}_{2t}^m\beta_0^m\beta_0^{m'}\boldsymbol{z}_{2t}^{m'}\ \epsilon_{t+h}\epsilon_{t+h-j})$.

Assumption R3'. For each FECM candidate, let $\boldsymbol{z}_{2t}^m = (Y_t,\ F_{1t},\ ...,\ F_{mt})'$ with $m \leq R$; let β_0^m and α_0^m be the corresponding cointegration matrix and the adjustment coefficient matrix, respectively. (i) Let the characteristic polynomial for the FECM system be $A(\lambda)$, and $det(A(\lambda)) = 0$ has roots on or outside the unit root circle. (ii) $\alpha_0^m\beta_0^{m'}$ has rank g, with $0 < g \leq m$. (iii) Let $\alpha_{0,\perp}^m$and $\beta_{0,\perp}^m$be two matrices satisfying

[13] This is also the first equation in the FECM system: $\Delta Z_{t+h} = \alpha_0\beta_0'Z_t + \sum_{j=1}^{p-1}\Pi_j\Delta Z_{t-j+1} + \epsilon_{t+h} +$ constant, with $Z_t = (Y_t,\ \boldsymbol{F}_t')'$.

$\beta_0^{m'}\beta_{0,\perp}^m = \mathbf{0}_{g\times(m+1-g)}$, $\beta_{0,\perp}^{m'}\beta_{0,\perp}^m = I_{m+1-g}$, $\alpha_0^{m'}\alpha_{0,\perp}^m = \mathbf{0}_{g\times(m+1-g)}$, and $\alpha_{0,\perp}^{m'}\alpha_{0,\perp}^m = I_{m+1-g}$. The matrix $\alpha_{0,\perp}^{m'}[I_m - \sum_{j=1}^{p-1}\Pi_j^m]\beta_{0,\perp}^m$ is nonsingular.[14]

Assumptions R1' and R2' are similar to Assumption R in the above subsection. They also imply Assumption 4 in Tu and Yi (2017). Assumption R3' is added since the I(1) variables are used directly as regressors. The following asymptotic theorems mimic Theorems 1 and 2.

Asymptotic Properties using PANIC to Estimate I(1) Factors

Theorem 3. Suppose $h = 1$, $E(\epsilon_{t+1}^2|\mathscr{F}_t) = \sigma^2$. Under Assumptions R1', R2', R3' and F1 to F6, for M and weights (w), and N, $T \to \infty$ with $T/N \to 0$, $C_T(w) = L_T(w) + T^{-1}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} + 2T^{-1/2}r_{4T}(w) - T^{-1}r_{5T}(w)$, where $r_{4T}(w) \xrightarrow{d} \kappa_4(w)$, $E(\kappa_4(w)) = 0$, $r_{5T}(w) \xrightarrow{d} \kappa_5(w)$, $E(\kappa_5(w)) = 0$, and $L_T(w) = \frac{1}{T}\sum(y_{t+h|t} - \hat{y}_{t+h|t}(w))^2$.

Theorem 4. Under Assumptions R1', R2', R3' and F1 to F6, for fixed M and weights (w), and N, $T \to \infty$ with $T/N \to 0$, $CV_{h,T}(w) = \check{L}_T(w) + T^{-1}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} + 2T^{-1/2}\check{r}_{6T}(w)$, where $\check{r}_{6T}(w) \xrightarrow{d} \kappa_6(w)$, $E(\kappa_6(w)) = 0$, and $\check{L}_T(w) = \frac{1}{T}\sum(y_{t+h|t} - \check{y}_{t+h|t}(w))^2$. $\check{y}_{t+h|t}(m) = \tilde{z}_t(m)\check{b}_{t,h}(m)$ and $\check{b}_{t,h}(m)$ is the leave-h-out estimator of model m at time t.

1.5 Empirical Results

In this section, I apply FECM averaging methods to forecast a number of U.S. and Canadian macroeconomic variables. The results indicate that FECM averaging improves the forecasting performance especially over longer horizons. The Canadian dataset is taken from Fortin-Gagnon et al. (2018), which is a monthly dataset of 139 macro series. For the U.S. forecasts, I use the Stock and Watson (2012) dataset and

[14] Assumptions R2' and R3' are specific to the nested model setup in this paper. Essentially, the cointegration matrix of any FECM candidates used in the model averaging set should be of full column rank. If the model candidates are not nested, then the notation of Assumptions R2' and R3' changes accordingly.

the datasets discussed in McCracken and Ng (2016). To incorporate the nonstationary information in the predictive regressions, I use either the estimated cointegration vectors to generate the regressors or the I(1) variables as predictive regressors directly. For brevity, only the MSFE results for the variables from the McCracken and Ng (2016) monthly dataset (FRED-MD) are reported in the main text.

There are 118 monthly series in the FRED-MD dataset, with the time span from 1959M1 to 2017M12. The monthly factors are extracted from the whole panel. To check for robustness, the same forecasting experiments are conducted using the quarterly dataset (FRED-QD) suggested by McCracken and Ng (2016), where the number of series which contain zero missing observations from 1959M1 to 2017M12 is 210. The other quarterly dataset I use is the CH dataset with the ending period being 2009Q4.[15] The quarterly factors are extracted from subsets of the quarterly panels, which contain only the lower-level disaggregate series. This choice helps avoid information overlapping as well as provide more efficient factor estimations (see Boivin and Ng, 2006).[16]

I include ten factors for each model, resulting in 66 models considered for FECM averaging if the cointegration vectors are estimated. Alternatively, there are 21 models if the nonstationary variables are used directly for predictions.[17] In separate robustness tests, all forecasting experiments are conducted with the inclusion of five factors, and the MSFE patterns remain the same. No lags of the factors are included in the predictions, as Kim and Swanson (2014) show that the lags of factors contribute

[15] The CH dataset is the same as the dataset used in Stock and Watson (2012). The raw dataset contains both monthly and quarterly series. To construct a larger quarterly panel, I take the last month observations /averages of the monthly observations and combine the transformed dataset with the existing quarterly subset. Results are not sensitive to how I construct the quarterly dataset.

[16] The Stock and Watson (2012) dataset can be found at the following URL: https://www.princeton.edu/˜mwatson/publi.html. Details of the FRED-MD and FRED-QD datasets can be found at the URL https://research.stlouisfed.org/econ/mccracken/fred-databases/.

[17] All empirical results are calculated based on Bai and Ng (2004) for factor estimations. I also present the MSFE results using Bai (2004) as robustness checks. The results are available upon request. MSFE distributions and ratios do not vary much. When using the estimated cointegrations, the overall number of model candidates is 66, which contains the smallest model with only the constant and the lags, 10 more models with the factors, and 1+2+. . . +10=55 more models containing the potential cointegrations between the nonstationary predicted variable and the factors. Similarly, when using the I(1) regressors directly, the overall number of model candidates is 21=1+10+10.

little to the FAR forecasting performance. All series in each of the panels are used as dependent variables to be predicted, and I follow McCracken and Ng (2016), Fortin-Gagnon et al. (2018), and Stock and Watson (2012) in transforming the datasets. The transformations used for each category of variables are listed in the appendix. I also provide model selection results to compare with the model averaging results, and the shrinkage or selection criteria for the model selection methods are chosen based on Stock and Watson (2012). Additionally, I calculate the MSFE ratios using an ARMA(1,1) model and the three-pass regression filter proposed by Kelly and Pruitt (2015), whose superiority has been shown in consumption, industrial production, and the market return forecasting.[18]

1.5.1 U.S. Variables Forecasting Performance

Tables 1.4 to 1.7 report the MSFE results for U.S. macroeconomic variable forecasting using the FRED-MD dataset, where the full sample period ranges from 1959M1 to 2017M12. For all estimation methods, I keep twelve lags of the dependent variable and a constant as fixed regressors. All values displayed in the tables are the MSFE values scaled by the MSFE obtained from estimating an AR(12) model, unless otherwise stated. These MSFEs are computed by a pseudo rolling forecasting scheme with a window size of 120 months; the forecasting horizons are one month, three months, six months, and one year.[19] In each table, Panels A, B, and H report results obtained using model selection methods. Panels C and D report results using model averaging methods. Panels E and F contain results of FECM averaging when factor numbers or the cointegration ranks are pre-estimated. In contrast to Panels B and D to F, Panel G reports results obtained using the generated I(1) factors.

Several patterns emerge from the MSFE results. First, model averaging outperforms model selection for both FAR models and FECMs in general. For example, the

[18] Here I set the factor number to be ten for comparison reasons. In Kelly and Pruitt (2015), the empirical work is reported with one factor.

[19] The rolling window size for quarterly datasets is 100 quarters. The benchmark model is AR(4) with four lags as fixed regressors.

ratios in Table 1.6 Panel C are smaller than those in Panel A, and similar conclusions can be drawn by comparing Panels D and B.[20] Second, among the model averaging results, the major contribution of FECM are shown at longer forecasting horizons. To illustrate, Table 1.6 Panels C and D report the model averaging results for FAR models and FECMs. The fifth percentile of the FECM averaging results indicates that about six variables have the relative MSFEs smaller than 0.45. The relative performance for FAR model averaging is worse by comparison, although FECM averaging provides larger median MSFE ratios in Tables 1.4 and 1.5. When the forecasting horizons are longer than three months, FECM averaging dominates FAR model averaging in a more uniform way. For example in Table 1.7, the MSFEs for FECM averaging are about 20 percent smaller compared to those for FAR model averaging. Third, using I(1) variables as predictors provides the best forecasting results in general. In Table 1.5, the MSFEs of applying the leave-h-out cross-validation model averaging in Panel G are on average five percent smaller than those in Panel D. Moving to longer forecasting horizons, these improvements are persistent and more significant. Finally, the forecasting performance of simple averaging and the leave-h-out cross-validation averaging are comparable. Smith and Wallis (2009) provide some explanations of the forecasting combination puzzle of real GDP, where using equal weights can outperform using estimated weights if the imposed restriction is approximately true. On the contrary, Panel G of Tables 1.4 to 1.7 presents different results. Simple averaging performs slightly better in the one-month forecasts, while using the leave-h-out cross-validation model averaging achieves smaller MSFEs in longer forecasting horizons. Thus, simple averaging losses in efficiency by trading off a large bias against a smaller estimation variance considering the long-run forecasting exercise.

Forecasting results using the quarterly datasets are also conducted, and are available upon request. The results display similar patterns. In the following subsection, I discuss the MSFE results of some robustness checks, including the forecasting perfor-

[20] Some of the MSFE ratios of "bagging" and "pretest" are outstandingly large, because these ratios depend on the threshold values which are chosen arbitrarily by researchers. For all forecasting experiments, I set the threshold value to be 1.645.

mances for recession periods, some individual macroeconomic and financial variables. MSFEs of these selected periods and variables are presented in Tables 1.12 to 1.18, and the results are formed using the generated I(1) variables.

1.5.2 Robustness Checks

Pre-select the Optimal Factor Number and Cointegration Rank

To check robustness, first, I report the MSFE results using all of the non-nested models with five factors. To reduce computation burden, I report the MSFE distributions using the Stock and Watson (2012) quarterly dataset with a constant and four lags. This adds to CH, where they focus on averaging across nested models. Figure 1.1 plots the two MSFE distributions using estimated cointegration vectors. The two distributions resemble each other, indicating that the improvement of FECM averaging is not sensitive to factor numbers.

Second, Panel E of Tables 1.4 to 1.11 presents the MSFE ratios where the number of factors is set to be the same across FECMs. These optimal numbers are selected by the PC2 criterion in Bai and Ng (2002), with ten factors to be the maximum set. Thus, the results reported in Panel E focus on averaging forecasts obtained from different cointegration vectors. Additionally, the MSFE ratios where the cointegration rank is estimated by the MQ_c^c tests from Bai and Ng (2004) are reported in Panel F of each table. In contrast to Panel E, Panel F fixes the cointegration rank, but allow the number of factors to vary.

In general, the MSFE patterns are similar among Panels E, F, and D; however, averaging across both the factors and cointegrations provides the smallest MSFE ratios for longer horizons. Moreover, by comparing the results in Panels E and F, forecasts for shorter horizons benefit from the flexibility of factor numbers, while forecasts for longer horizons have smaller MSFEs when the cointegration ranks are not estimated. For illustration, in Table 1.7, more than 75 percent of the MSFEs in Panel E are smaller than their counterparts in Panel F, indicating the long-run forecasting improvements stem from the inclusion of cointegration relationships and the flexibility of the cointegration ranks. Intuitively, cointegration tests can be sensitive to the number of lags included in the model, and different tests may suffer from different size or power distortions. As stated in Haug (1996), the Stock and Watson (1988) cointegration test tends to overestimate the cointegration rank, which again suggest the benefit of allowing cointegration ranks to be flexible.

To further demonstrate the contribution of cointegrations, I present some distributions of the weights assigned to each model candidate in the FECM averaging framework. The dataset where these weights are obtained is the McCracken and Ng (2016) monthly dataset, and the factors are estimated using PANIC with the maximum number being ten. In the first half of Table 1.19, I collect the model number for each of the out-of-sample periods, where the model is assigned the largest weight. I then calculate the median of these numbers, and present the distributions of the medians for the 118 series in the panel. Both the one-step and multi-step forecasting models have been evaluated. The second half of the table reports similar model selection distributions, with the interquartile range as the statistics. Given that the models are nested, the larger number indicates that the model contains more factors as well as more informative cointegration relationships.

From Table 1.19, the models associated with the largest weight contain error correction parts in the majority of the cases. The model number is larger when the forecasting horizon is longer, indicating that cointegration relationships benefit the long-run forecasts more. Moreover, the interquartile distributions show that the CVA model averaging criterion selects a variety of models. With the total number of model being 66, a range of 40 means that the smallest model can contain just one factor, while the largest model can contain ten factors.

Table 1.20 reports the variances of weights estimated from FECM averaging and FAR model averaging. To form the table, I first calculate the weight variance for each series and each out-of-sample period. Then, I calculate the average of these variances across the out-of-sample periods, and present the distributions of the averages of the 118 series in Table 1.20. The weight variances obtained from FECM averaging are smaller than the ones of FAR model averaging, meaning that the weights are assigned more evenly for the former method. Both Tables 1.19 and 1.20 demonstrate that modeling uncertainty exists in this forecasting experiment.

Recessions and Expansions

The U.S. recession periods are identified by NBER. There are six recessions occurred during the out-of-sample period: November 1973–March 1975, January 1980–July 1980, July 1981–November 1982, July 1990–March 1991, March 2001–November 2001, and December 2007–June 2009. Given the number of observations, I use the McCracken and Ng (2016) monthly dataset to calculate MSFEs for the U.S. recession periods. The Canadian recession periods are June 1981–October 1982, March 1990–April 1992, and October 2008–May 2009. These dates are identified by the C.D. Howe Institute. The expansion periods for U.S. and Canada are the complement sets of the corresponding recession periods.

The forecasting results specific to the recession periods are presented in Tables 1.12 and 1.13, and the results for the expansion periods are in Table2 1.14 and 1.15. In these tables, the upper and lower panels report the MSFEs of FAR model averaging and FECM averaging, respectively. As a baseline comparison, the MSFEs using FARs with the optimal factor numbers are presented as well. In these restricted sample periods, FECM averaging continues to outperform FAR model averaging for longer horizons. This finding is complementary to the existing literature, where FAR models are not performing well in forecasting expansions, and model averaging methods contribute less to forecast recessions (See Smith and Wallis, 2009, Pauwels and Vasnev, 2014, and Leroux et al., 2017).

Real Activities and Prices

To further illustrate the advantage of FECM averaging, I select four real variables and four nominal variables. These variables are: industrial production (IP), unemployment rate (UNRATE), real personal income (RPI), real manufacturing and trade industries sales (CMR), consumer price index (CPI), consumer price index less food and energy (CPI core), personal consumption expenditure (PCE), and producer price index (PPI). Depending on the different data transformations, the target variables are the changes for real variables and the changes of inflation for nominal variables. The MSFE ratios of FECM averaging and FAR model averaging are reported in Tables 1.16 to 1.18; a number smaller than one means that FECM averaging is preferable.

From the results, first, FECM averaging improves the forecasting performance for nominal variables at a broader range of horizons than real variables, except for CPI core. Moreover, the overall decreases in MSFE ratios are larger for nominal variables. For instance, in Table 1.18, the lowest MSFE ratios for CPI are about 40 percent, meaning that FECM averaging provides a smaller MSFE than FAR model averaging, where the improvement is about 60 percent. This is in contrast to the empirical results in Banerjee et al. (2014), wherein they suggest that FECMs outperforms FARs for real variable forecasts, but not for nominal variables in the more recent period. The results are also contrary to the inflation forecasting performance in Leroux et al. (2017), as they show that combing simple model averaging with FARs does not improve on using FARs themselves.

Exchange Rates, Interest Rates, and the S&P 500 Index

Forecasting results for the S&P 500 index are presented in Table 18, where FECM averaging outperforms ARMA(1,1) models and random walk models in the long term. Base on the arguments made in Leroux et al. (2017) about forecasting stock returns, these results further indicate the failure of the efficient market hypothesis.

Furthermore, I consider the forecasts of four exchange rates: Switzerland/U.S., Japan/U.S, U.S./U.K. and Canada/U.S. It is widely argued that exchange rates follow random walks and that factor models or model combinations do not contribute much to their forecasts. The middle panel in Table 18 reports the MSFE ratios for FECM averaging and its performance relative to FAR model averaging. Similar to the macroeconomic variables, the results favor FECM averaging for horizons longer than six months, suggesting exchange rates are predictable to some extend.

Finally, I select five interest rate series from the dataset, and show that FECM averaging continues to be preferable. Specifically, I compare the Dynamic Nelson-Siegel (DNS) model with the FECM averaging methods. The former method is suggested by Diebold et al. (2006), and some empirical evidence is provided by Swanson and Xiong (2018). The model is essentially a dynamic factor model with theoretical restrictions on the loading terms, and the factors represent the level, slope, and curvature of a yield curve. Swanson and Xiong (2018) show that the DNS models have the strongest forecasting performance than AR models, FARs, and hybrid models combining the DNS models with factors. This finding is supported by Table 1.18, where the DNS models produce smaller MSFEs than AR(12). On the contrary, Table 1.18 also shows the outperformance of FECM averaging comparing to the DNS models, indicating that model flexibility can contribute to forecasting term structures.

1.5.3 Canadian Variables Forecasting Performance

The Canadian forecasting results are reported in Tables 1.8 through 1.11. In general, FECM averaging improves on three-month-ahead forecasting and beyond. For instance, in Table 1.11 Panel G, about 90 percent of the series have better forecasting performances using FECM averaging than AR(12), and the decreases in the median MSFE ratios are about 20 percent larger than using FAR model averaging.

The superior performance of FECM averaging to FAR model averaging is less significant for the Canadian housing variables. The numbers in the last two columns of Table 1.17 are relatively larger than the ones in other columns, indicating that the Canadian housing market reacts more slowly to the disequilibrium in the long run. Thus, adding extra regressors to the prediction can reduce the signal-to-noise ratio. On the one hand, the prediction performance of the housing market is consistent with the results in Fortin-Gagnon et al. (2018), where they show that the improvements are smaller compared to the ones of forecasting real activity and inflation. On the other hand, FECM averaging provides the smallest MSFEs than FARs and AR models for the majority of the forecasts, which demonstrates the usefulness of cointegrations in the predictive regressions.

1.6 Prediction Interval Estimation

I construct prediction intervals to further demonstrate the usefulness of FECM averaging. Under the fixed parameter model settings, Zhang and Liu (2018) propose a simulation-based method to estimate the confidence intervals of coefficients obtained using Jackknife model averaging criterion. Bai and Ng (2006) derive the close form of the prediction intervals using FARs. These two papers' results are extended in this section to cover multi-step forecasting models with cointegrations.

The leave-h-out cross-validation model averaging method proposed by Hansen (2010) is applied to accommodate multi-step forecasting models. The nonstationary regressors and regressand are used directly in the forecasting model as Tu and Yi (2017), and the nonstationary factors are estimated using Bai (2004). The theoretical results of this section are thus build on the two papers. Results show that the weights assigned to under-fitted models converge to zeros. Moreover, the weights assigned to over-fitted models converge to some random variables, and converge to zeros by adding a penalizing term to the leave-h-out cross-validation criterion. Empirically, I apply the simulation method to six real and nominal macroeconomic variables in the Stock and Watson (2012) panel. The prediction intervals obtained from FECM averaging are narrower than the ones obtained from FAR model averaging for the out-of-sample period from 1985 to 2015.

1.6.1 Assumptions and Conditions

This subsection contains the assumptions and the regularity conditions. These assumptions and conditions are required by all of the model candidates, and are the prerequisite for all of the following propositions. The first M_0 model candidates are assumed to be under-fitted. Let $S = M - M_0$ and the small letter m be the m^{th} model candidate. The conditions are the modified regularity conditions of Zhang and Liu (2018) to accommodate multi-step forecasting models. The variance matrix Ω takes into consideration of the serial correlation among errors. The estimated I(0) and I(1) factors follow the distribution theories presented in Bai (2003) and Bai (2004), respectively.

Assumptions for factor estimation: Bai (2003) Assumption A-G, and Bai (2004) Assumption A-G.

Assumptions for FAR and FECM forecasting. Cheng and Hansen (2015) Assumption R and F. Tu and Yi (2017) Assumption 1, 4, and 5.

Remark: Given that $\frac{N}{T^3} \to 0$ is required in Bai (2004), and $\frac{T}{N^2} \to 0$ is required in Cheng and Hansen (2015), the requirement $\frac{N}{T^2} \to 0$ in Bai and Ng (2006) is implied.

The first set of conditions are with respect to multi-step forecasting models with observed regressors. The matrix $\boldsymbol{z}$ contains the all of the observed regressors, and e is the error term.

Condition 1. $Q_T = \frac{1}{T}\boldsymbol{z}'\boldsymbol{z} \to Q$, where $Q = E(\boldsymbol{z}_t\boldsymbol{z}_t')$ is a positive definite matrix.

Condition 2. $\boldsymbol{\xi}_T = \frac{1}{\sqrt{T}}\boldsymbol{z}'\boldsymbol{e} \to \boldsymbol{\xi} \sim N(0,\, \Omega)$, where $\Omega = \sum_{s=-(h-1)}^{h-1} E(\boldsymbol{z}_{t-h+s}\boldsymbol{z}_{t-h}' e_{t+s} e_t)$ is a positive definite matrix.

Condition 3. $E||\boldsymbol{z_t}||^4 < \infty$, and $\bar{h}_T = max_{1\leq m\leq M,\ 1\leq t\leq T} h_{tt}^m = o_p(T^{-1/2})$, where h_{tt}^m is the t^{th} diagonal element of the projection matrix $\boldsymbol{z}_m(\boldsymbol{z}'_m\boldsymbol{z}_m)^{-1}\boldsymbol{z}'_m$ and z_m is the regressors in model candidate m.

Condition 4. $\Omega_T = \sum_{s=-h+1}^{s=h-1} \frac{1}{T}\sum_t \boldsymbol{z}_{t-h}e_te_{t+s}\boldsymbol{z}'_{t-h+s} \to \Omega$.

The next set of conditions are the modified regularity conditions to accommodate multi-step forecasting models with estimated stationary factors. H_1 is the rotation matrix for factor estimations. These conditions mirror Conditions 1-4. The matrix $\tilde{\boldsymbol{f}}$ contains the all of the estimated regressors. e is the error term when the estimated factors are the regressors, and contains the factor estimation errors.

Condition f1. $Q_T = \frac{1}{T}\tilde{\boldsymbol{f}}'\tilde{\boldsymbol{f}} \to H_1QH_1'$, where $Q = E(\boldsymbol{f}_t\boldsymbol{f}'_t)$ is a positive definite matrix. Specifically, $\tilde{Q}_T = \frac{1}{T}\tilde{\boldsymbol{f}}'\tilde{\boldsymbol{f}} \boldsymbol{= I_r}$.

Condition f2. $\boldsymbol{\xi}_T = \frac{1}{\sqrt{T}}\tilde{\boldsymbol{f}}'\boldsymbol{e} \to H_1\boldsymbol{\xi} = \frac{1}{\sqrt{T}}H_1\boldsymbol{f}'\boldsymbol{e} \sim N(0,\ H_1\Omega H_1')$, where $\Omega = \sum_{s=-(h-1)}^{h-1} E(\boldsymbol{f}_{t-h+s}\boldsymbol{f}'_{t-h}e_{t+s}e_t)$ is a positive definite matrix.

Condition f3. $E||\tilde{\boldsymbol{f}}_{\boldsymbol{t}}||^4 < \infty$, and $\bar{h}_T = max_{1\leq m\leq M,\ 1\leq t\leq T} h_{tt}^m = o_p(T^{-1/2})$, where h_{tt}^m is the t^{th} diagonal element of the projection matrix $\tilde{\boldsymbol{f}}_m(\tilde{\boldsymbol{f}}'_m\tilde{\boldsymbol{f}}_m)^{-1}\tilde{\boldsymbol{f}}'_m$ and $\tilde{\boldsymbol{f}}_m$ is the regressors in model candidate m.

Condition f4. $\tilde{\Omega}_T = \sum_{s=-h+1}^{s=h-1} \frac{1}{T}\sum_t \tilde{\boldsymbol{f}}_{t-h}e_te_{t+s}\tilde{\boldsymbol{f}}'_{t-h+s} \to \sum_{s=-h+1}^{s=h-1} \frac{1}{T}\sum_t H_1\boldsymbol{f}_{t-h}e_te_{t+s}$ $\boldsymbol{f}'_{t-h+s}\ H_1' \to H_1\Omega H_1'$.

The last set of conditions are the modified regularity conditions to accommodate multi-step forecasting models with estimated factors and cointegrations. H_1 and H_2 are the rotation matrices for the stationary and nonstationary factor estimations, respectively. Denote $\mu^+ = [\boldsymbol{f}H'_{10},\ (Y,\ \boldsymbol{F}H'_{20})C]$[21], and $\tilde{\boldsymbol{\mu}}^+ = [\tilde{\boldsymbol{f}},\ (Y,\ \tilde{\boldsymbol{F}})C]$ where $\underset{(R+1)\times R}{C}$ contains the cointegration vectors between Y and $\tilde{F}$. Without loss of gener-

[21] H_{10} and H_{20} are the modified rotation matrices of Bai (2003) and Bai (2004). Specifically, $H_{10} = (\frac{\Lambda'\Lambda}{N})(\frac{f'f}{T})V_{NT,1}^{-1} = (\frac{\lambda'\lambda}{N})(\frac{f'\tilde{f}}{T})V_{NT,1}^{-1} + o_p(1)$, and $H_{20} = (\frac{\lambda'\lambda}{N})(\frac{F'F}{T})V_{NT,2}^{-1} = (\frac{\Lambda'\Lambda}{N})(\frac{F'\tilde{F}}{T^2})V_{NT,2}^{-1} + o_p(1)$. $V_{NT,1}^{-1}$ is an $r\times r$ diagonal matrix containing the first r largest eigenvalues of $(1/NT)xx'$ in decreasing order, and $V_{NT,2}^{-1}$ is an $r\times r$ diagonal matrix containing the first r largest eigenvalues of $(1/NT^2)XX'$ in decreasing order.

ality, assume that the matrix C is in the form $\begin{bmatrix} C_{11} & C_{21} & \dots & C_{R1} \\ C_{12} & C_{22} & \dots & C_{R2} \\ & C_{23} & \dots & C_{R3} \\ & & & \vdots \\ & & & C_{RR+1} \end{bmatrix}$ given the assumption that all models satisfy the FECM structure. e is the error term when the estimated stationary and nonstationary factors are the regressors, and contains the factor estimation errors.

Condition F1. $\tilde{Q}_T = \frac{1}{T}\tilde{\boldsymbol{\mu}}^{+'}\tilde{\boldsymbol{\mu}}^{+} \rightarrow Q_T = \frac{1}{T}\boldsymbol{\mu}^{+'}\boldsymbol{\mu}^{+} \rightarrow Q$, where $Q = E(\mu^{+}\mu^{+'})$ is a positive definite matrix.

Condition F2. $\tilde{\boldsymbol{\xi}}_T = \frac{1}{\sqrt{T}}\tilde{\boldsymbol{\mu}}^{+'}\boldsymbol{e} \rightarrow \boldsymbol{\xi}_T = \frac{1}{\sqrt{T}}\boldsymbol{\mu}^{+'}\boldsymbol{e} \rightarrow \xi \sim N(0,\ \Omega)$, where $\Omega = \sum_{s=-(h-1)}^{h-1} E(\boldsymbol{\mu}^{+}{}_{t-h+s}\boldsymbol{\mu}^{+'}{}_{t-h}e_{t+s}e_t)$ is a positive definite matrix.

Condition F3. $E||\tilde{\boldsymbol{\mu}}_t^{+}||^4 < \infty$, and $\bar{h}_T = max_{1\leq m\leq M,\ 1\leq t\leq T} h_{tt}^{m} = o_p(T^{-1/2})$, where h_{tt}^{m} is the t^{th} diagonal element of the projection matrix $\tilde{\boldsymbol{\mu}}_m^{+}(\tilde{\boldsymbol{\mu}}_m^{+'}\tilde{\boldsymbol{\mu}}_m^{+})^{-1}\tilde{\boldsymbol{\mu}}_m^{+'}$.

Condition F4. $\tilde{\Omega}_T = \sum_{s=-h+1}^{s=h-1} \frac{1}{T}\sum_t \tilde{\boldsymbol{\mu}}_{t-h}^{+}e_t e_{t+s}\tilde{\boldsymbol{\mu}}_{t-h+s}^{+} \rightarrow \Omega$.

Proposition 1 Consider a model averaging environment with M number of model candidates, where the first M_0 models are under-fitted. Let $m \in \{1, ..., M_0\}$, N, $T \rightarrow \infty$, and $\hat{w}_{CVA,m}$ be the weights estimated from the leave-h-out cross-validation model averaging criterion. Under the conditions and assumptions:

(1) For multi-step forecasting models with observed regressors, $\hat{w}_{CVA,m} = o_p(T^{-1/2})$.

(2) For multi-step forecasting models with estimated I(0) factors, and the true model being FAR, $\hat{w}_{CVA,m} = o_p(T^{-1/2})$.

(3) For multi-step forecasting models with estimated I(0) and I(1) factors, and the true model being FECM, $\hat{w}_{CVA,m} = o_p(T^{-1/2})$.

Proposition 2 This proposition adopts the notations in Conditions F1-F4. Assume the true model contains error corrections and the conditions and assumptions hold for all model candidates. Then the estimated coefficient vector $\hat{\alpha} = [\hat{\beta}',\ \hat{\gamma}_Y,\ \hat{\gamma}'_F]'$ has the property where

$$\begin{aligned}
\sqrt{T}(\hat{\alpha}_{CVA} - \alpha) &= \sum_{m=1}^{M_0} \hat{w}_{CVA,m}\sqrt{T}(\hat{\alpha}_m - \alpha) + \sum_{m=M_0+1}^{M} \hat{w}_{CVA,m}\sqrt{T}(\hat{\alpha}_m - \alpha) \\
&\to o_p(1) + \sum_{m=M_0+1}^{M} \hat{w}_{CVA,m} \underset{(r_m+1)\times r_m}{C_m} (\frac{\tilde{\boldsymbol{\mu}}^{+'}\tilde{\boldsymbol{\mu}}^{+}}{T})^{-1}\frac{1}{\sqrt{T}}\tilde{\boldsymbol{\mu}}^{+'} \\
&\quad \times \{(\boldsymbol{F} - \tilde{\boldsymbol{F}}\boldsymbol{H}_2^{-1'})\boldsymbol{\gamma_F} + (\boldsymbol{f} - \tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta} + e\} \\
&= o_p(1) + \sum_{m=M_0+1}^{M} \hat{w}_{CVA,m} \underset{(r_m+1)\times r_m}{C_m} (\frac{\tilde{\boldsymbol{\mu}}^{+'}\tilde{\boldsymbol{\mu}}^{+}}{T})^{-1}\frac{1}{\sqrt{T}}\tilde{\boldsymbol{\mu}}^{+'} e \\
&= o_p(1) + \sum_{m=M_0+1}^{M} \hat{w}_{CVA,m} \underset{(R+1)\times R}{C} \underset{R\times r}{\boldsymbol{\Pi}'_m}(\boldsymbol{\Pi}_m\tilde{\boldsymbol{Q}}_{\boldsymbol{T}}\boldsymbol{\Pi}'_m)^{-1}\boldsymbol{\Pi}_m\tilde{\boldsymbol{\xi}}_{\boldsymbol{T}} \\
&\to \sum_{s=1}^{S} \tilde{\zeta}_{CVA,s} \underset{(R+1)\times R}{C} \boldsymbol{V_s}\boldsymbol{\xi}
\end{aligned}$$

in distribution. Thus, the conditional point forecast $\hat{y}_{T+1|T,\ CVA}$ converges to a random variable

$$(\hat{y}_{T+1|T,\ CVA} - y_{T+1}) \to \sum_{s=1}^{S} \tilde{\zeta}_{CVA,s}\boldsymbol{\xi'}_{T+1}\boldsymbol{V_s}\boldsymbol{\xi}.$$

$\boldsymbol{\Pi}_m = (I_{r_m},\ 0_{r_m\times(R-r_m)})$ is the selection matrix. $Q_s = \boldsymbol{\Pi}_{M_0+s}Q\boldsymbol{\Pi}'_{M_0+s}$, $\Omega_s = \boldsymbol{\Pi}_{M_0+s}\Omega$ $\boldsymbol{\Pi}'_{M_0+s}$, and $\boldsymbol{V}_s = \boldsymbol{\Pi}'_{M_0+s}Q_s^{-1}\boldsymbol{\Pi}_{M_0+s}$. $\hat{\zeta}_{CVA,s} = argmin\ \zeta'\Sigma\zeta$ where Σ is an $S \times S$ matrix with the $(s,\ j)th$ element $\Sigma_{sj} = trace((Q_s)^{-1}\Omega_s) + trace((Q_j)^{-1}\Omega_j) - \boldsymbol{\xi}'V_{max\{s,j\}}\boldsymbol{\xi}$.

Proposition 3 Define $\tilde{CVA} = CVA + \phi_T\boldsymbol{w}'\boldsymbol{k}$, where the model averaging criterion is defined in section 2 and the vector $\boldsymbol{k}$ contains the numbers of regressors. Under the above assumptions and conditions, let $\phi_T \to \infty$ and $N,\ T \to \infty$, then $\tilde{w}_{CVA,m} = O_p(\phi_T^{-1})$ for any $m \in \{M_0 + 2, ...,\ M\}$.

1.6.2 Inference for the Leave-h-out Cross-Validation Model Averaging Estimator

Given the coefficient distributions derived by Tu and Yi (2017) Lemma 3, the estimated slopes of the cointegration system asymptotically converge to some multinormal distributed random variables. The estimated intercept, however, converges to a random variable which does not follow a normal distribution. Thus, the simulation procedures focus on the de-meaned sample to avoid estimating the intercept. To conduct the empirical analysis in the following subsection, I assume that the factors are cointegrated with the predicted variables for each model candidate. The choice of the estimated cointegration vectors is flexible as long as the consistency of Ω and Q estimates is ensured. The steps of estimating the simulated prediction intervals of $\hat{y}_T$ are following:

1. Demean the dataset by the averages using observations from $\{1, ..., \ T-1\}$.

2. Calculate the asymptotic variance of $\tilde{f}$ and $\tilde{F}$, with the formulae provided by Bai (2003) Theorem 1, and Bai (2004) Corollary 1, respectively.

3. Let $\hat{e}_t$ be the residuals from the full model. Then $\hat{\sigma}^2 = \sum_{t=1}^{T-1} \hat{e}_t^2/(T-k)$ is a consistent estimator of σ^2, where k is the number of regressors.

4. For each model m collect the coefficients for $(Y, \ \tilde{\boldsymbol{F}}_m)$. Use the coefficients as the cointegration vector to transform the nonstationary regressors into the stationary error correction terms. Combine the estimated error correction terms with the I(0) factors and create the matrix $\tilde{\boldsymbol{\mu}}^+$.

5. Calculate $\hat{Q} = \frac{1}{T}\sum_{t=1}^{T-1} \tilde{\boldsymbol{\mu}}_t^+ \tilde{\boldsymbol{\mu}}_t^{+'}$. Then, calculate $\hat{\Omega} = \frac{1}{T}\sum_{t=1}^{T-1} \tilde{\boldsymbol{\mu}}_t^+ \tilde{\boldsymbol{\mu}}_t^{+'} \hat{e}_t^2$ if the forecasting horizon is one, and $\hat{\Omega}_T = \sum_{s=-h+1}^{s=h-1} \frac{1}{T}\sum_{t=1}^{T-1} \tilde{\boldsymbol{\mu}}_{t-h}^+ \hat{e}_t \hat{e}_{t+s} \tilde{\boldsymbol{\mu}}_{t-h+s}^+$ if the forecasting horizon is beyond one. Note, the dimension of the regressors decreases by one unit with the transformation in step 4. The overall number of model candidates is still the same.

6. For steps 6-8, fix $M_0 = \{0,\ \ldots,\ M-1\}$. Calculate $\hat{\boldsymbol{V}}_s = \boldsymbol{\Pi}'_{M_0+s}\hat{Q}_s^{-1}\boldsymbol{\Pi}_{M_0+s}$, and simulate $\boldsymbol{\xi}^{(iter)} \sim N(0,\ \hat{\Omega}_T)$.

7. Demean the last observations of the dataset using the averages from $\{1, \ldots,\ T-1\}$. Similar to step 4, transform the nonstationary part using the estimated cointegration vectors. Denote the transformed last right-hand-side observations as $\tilde{\boldsymbol{\mu}}^{+}_{end}$.

8. Calculate the $\hat{y}^{(M_0,\ iter)}_{T+h|T} = \sum_{s=1}^{S}\hat{\zeta}^{(iter)}_{CVA,s}\tilde{\boldsymbol{\mu}}^{+'}_{end}\hat{\boldsymbol{V}}_s\boldsymbol{\xi}^{(iter)}$, where the superscript $(iter)$ represents each iteration. The weights $\hat{\zeta}^{(iter)}_{CVA,s}$ are estimated by applying Proposition 2.

9. Estimate $\boldsymbol{w}_{CVA}$ using Proposition 3, and calculate $\hat{y}^{(iter)}_{T+h|T}(\boldsymbol{w}_{CVA}) = \sum_{M_0=0}^{M-1} w_{M_0+1,\ CVA}\hat{y}^{(M_0,\ iter)}_{T+h|T}$.

10. Calculate the variance of $\hat{y}^{(iter)}_{T+h|T}(\boldsymbol{w}_{CVA})$, which is denoted as var_y. The simulation based prediction interval is

$$\begin{aligned}&[\hat{y}_{T+h|T} - 1.96 \times sqrt(\hat{\sigma}^2 + var_y/T + AsyVar(\tilde{f})/N + AsyVar(\tilde{F})/N),\\ &\hat{y}_{T+h|T} + 1.96 \times sqrt(\hat{\sigma}^2 + var_y/T + AsyVar(\tilde{f})/N + AsyVar(\tilde{F})/N)]\end{aligned}$$

1.6.3 Empirical Results with Six U.S. Macroeconomic Variables

This subsection estimates the prediction intervals of six macroeconomic variables from the Stock and Watson (2012) dataset. The factor number is set to be five. The prediction models are the same as Section 1.5. Figures 1.2-1.7 contain the predictive bands for industrial production, unemployment rate, CPI, core CPI, personal consumption expenditure, and S&P 500 index, where these variables are transformed as the section above. The green and the red shaded areas represent the prediction intervals for FECM averaging and FAR model averaging, respectively. The brown area is where the green and red bands overlap. The solid line in each of the graphs connects the true observations. h stands for the forecasting horizon, and the x-axis represents the out-of-sample forecasting period. The comparisons demonstrate that cointegrations contribute to forecasting macroeconomic variables.

First, the FECM averaging prediction bands are in general narrower than the corresponding FAR model averaging prediction bands, especially when the forecasting horizons are eight and twelve. The advantage is the least substantial in forecasting the growth rate of industrial production, and is the most significant in forecasting the growth rates of unemployment rate and core CPI changes.

Second, the advantages of using FECM averaging over FAR model averaging are more obvious during 1995 to 2000, an expansion period. Specifically, the growth rates of industrial production and personal consumption expenditure went through several troughs (and the growth rate of unemployment rate experienced a peak) around 2003. Even though the forecasts present lag effects, FECM averaging reacts faster and the changes are promptly picked up. As shown in Figure 1.3, the green band shifts up around 2004 in predicting the unemployment rate. In contrast, FECM averaging reacts with a less amount than the FAR model averaging if the adaption takes place too late. In Figure 1.6, neither the green or the red bands successfully cover the true values in the 2003 period. In later periods, both prediction bands shift downward, while the FECM averaging predictions are more conserved. In summary, FECM averaging tends to provide larger improvements to the predictions right away, or brings less damage to the predictions if the adaption takes a long time.

1.7 Conclusion

This paper combines model averaging methods with Factor-augmented Error Correction Models. The procedure estimates factors and error correction terms in the first step, and predictive regressions using the generated regressors in the second step. I show that both the Mallows and leave-h-out cross-validation criteria are asymptotically unbiased for the mean-squared forecast error. Forecasts of macroeconomic variables demonstrate that model averaging with the inclusion of error correction terms outperforms model averaging with only the factors in the long run, indicating adjustments to long-run equilibria is useful in prediction. In particular, FECM averaging provides superior results for exchange rate and interest rate forecasting when the forecasting horizon is greater than or equal to three months. The patterns of the forecasting performance also hold when analyzing recession and expansion periods separately. Additionally, I show that the improvements primarily come from the inclusion of the cointegration relationships. Lastly, this paper estimates the prediction intervals around the point forecasts using simulations. Results show that for the out-of-sample period from 1985 to 2015, FECM averaging not only produces smaller mean squared forecast errors but also narrower prediction intervals.

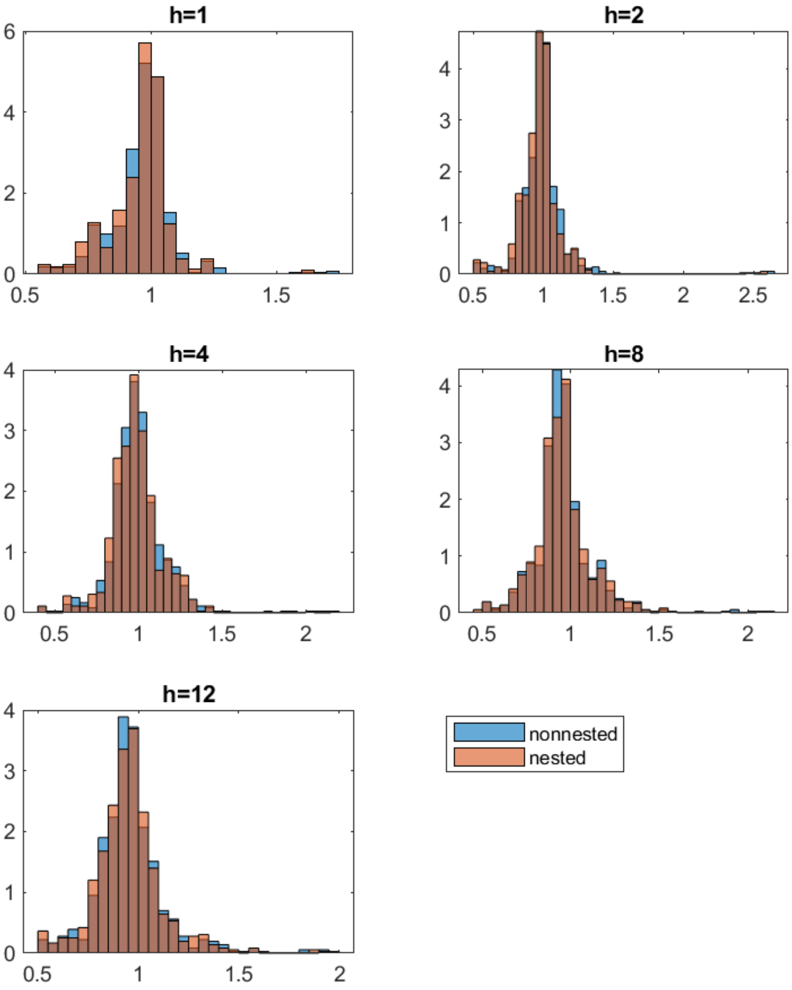

Fig. 1.1.: Nested and non-nested distribution

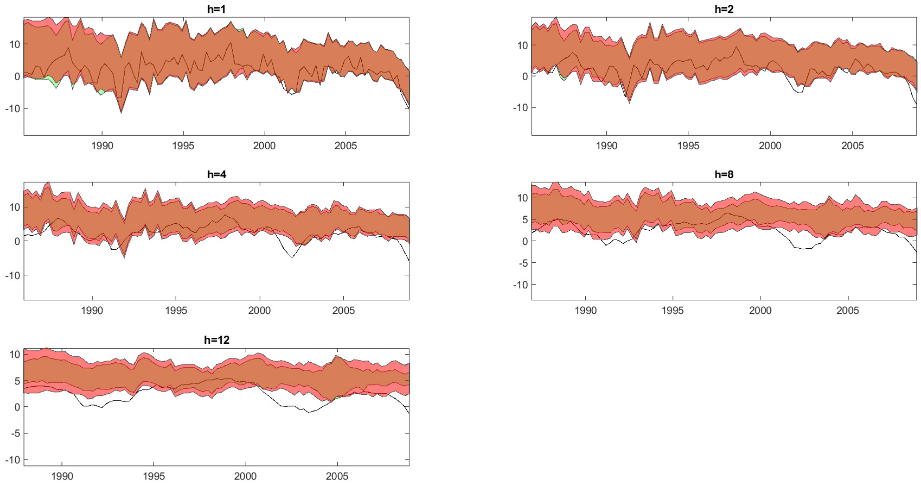

Fig. 1.2.: Prediction Interval, Industrial Production

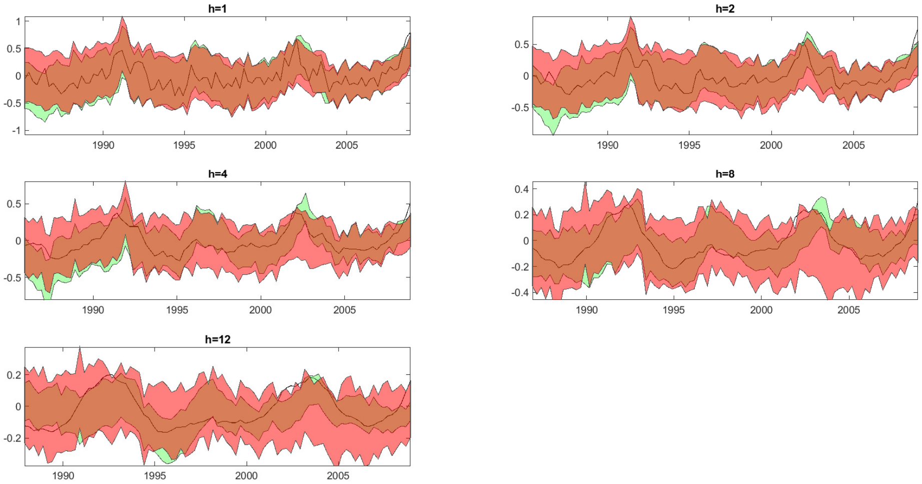

Fig. 1.3.: Prediction Interval, Unemployment Rate

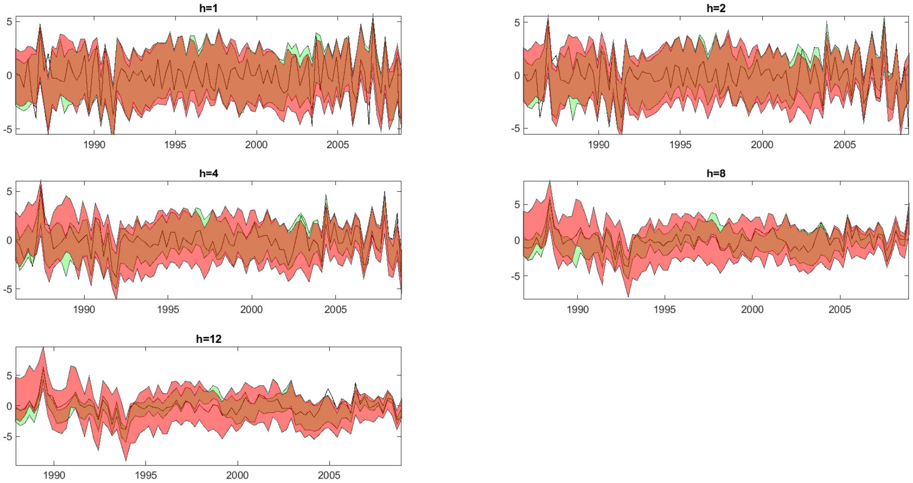

Fig. 1.4.: Prediction Interval, CPI

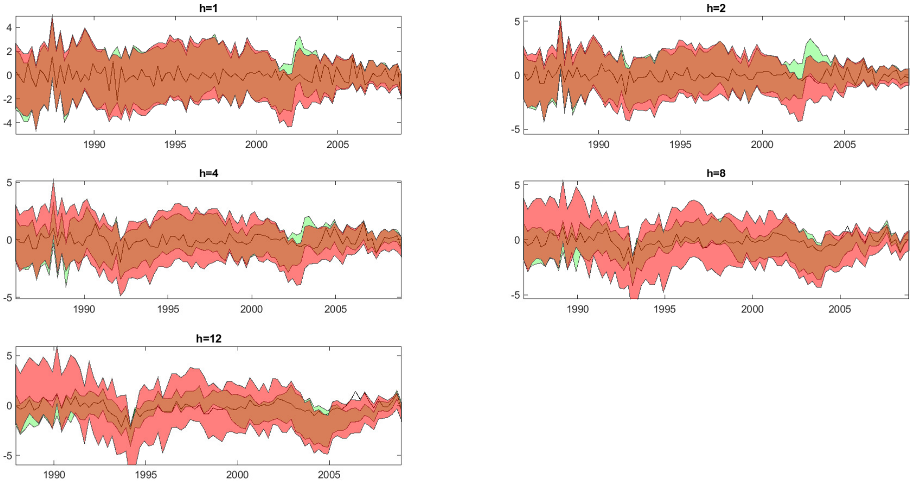

Fig. 1.5.: Prediction Interval, CPI core

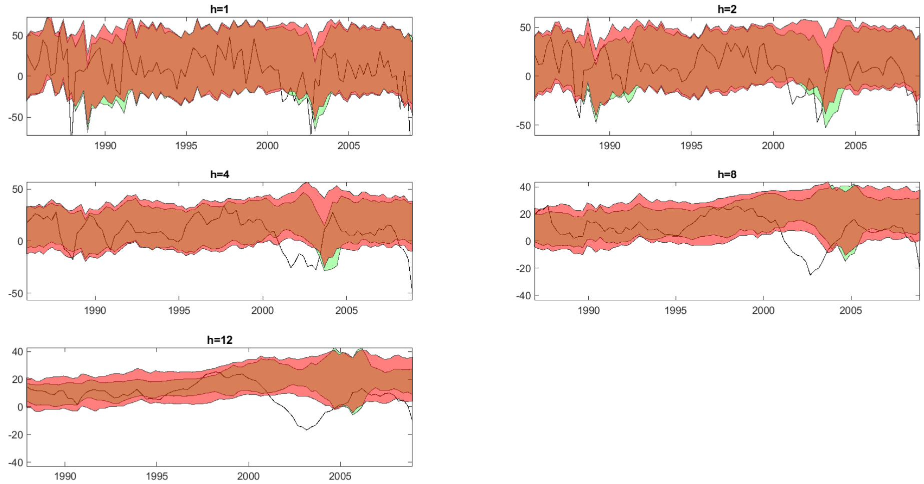

Fig. 1.6.: Prediction Interval, Personal Consumption Expenditure

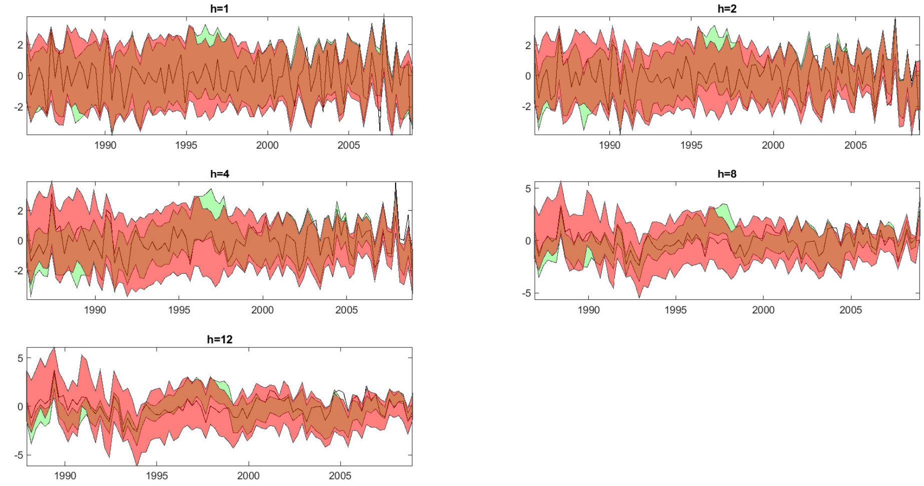

Fig. 1.7.: Prediction Interval, S & P 500

Table 1.4.: U.S. Forecasting MSFE Distribution, forecasting horizon = 1, monthly

	percentile	5	25	50	75	95		percentile	5	25	50	75	95
Panel A	FAR optimal	0.808	0.946	1.002	1.059	1.128	Panel B	FECM optimal	0.819	0.974	1.04	1.097	1.249
	logit	0.808	0.946	1.002	1.059	1.128		logit	0.819	0.974	1.04	1.097	1.249
FAR selection	bagging	0.817	0.902	0.981	1.042	34.929	FECM selection	bagging	0.819	1.045	1.177	1.388	103.198
	pretest	0.872	0.966	1.042	1.15	137.53		pretest	0.819	1.022	1.091	1.193	101.822
Panel C	MMA	0.791	0.913	0.963	1.011	1.043	Panel D	MMA	1.011	1.186	1.28	1.387	1.577
FAR MA	CVh	0.804	0.913	0.963	1.009	1.031	FECM MA	CVh	0.859	0.958	0.995	1.055	1.131
	EQ	0.799	0.906	0.949	1.015	1.063		EQ	0.824	0.924	0.968	1.049	1.125
Panel E	MMA	0.883	0.988	1.029	1.1	1.194	Panel F	MMA	0.832	0.909	0.986	1.044	1.348
FECM MA	CVh	0.893	0.985	1.03	1.104	1.187	FECM MA	CVh	0.815	0.925	0.962	1.013	1.092
pre-select # of fac	EQ	0.858	0.963	1.011	1.088	1.151	pre-select # of coin	EQ	0.782	0.913	0.95	1.015	1.079
							Panel G	CVh	0.853	0.936	0.984	1.042	1.113
							Tu Yi (2017)	EQ	0.809	0.884	0.93	1.012	1.076
Panel H	ARMA(1, 1)	0.804	1.004	1.1	1.494	2.32		3PRF	1.447	1.929	2.205	2.435	2.732

Note: Panel A and Panel B report MSFE distributions using model selection methods. “FAR optimal” selects the optimal number of factors for each forecasting exercise, using the PC2 criterion from Bai and Ng (2002). The “FECM optimal” model augments on the “FAR optimal” model, and uses the MQ_c^c test to select cointegration ranks. The shrinkage functions of “Pretest”, “bagging” and “logit” follow Stock and Watson (2012). These selection methods are based on the optimal FAR model or the optimal FECM. Panels C to G are MSFE distributions using model averaging methods. “MA” stands for “model averaging.” Panels C and D report the MSFE distributions using the estimated cointegration vectors.
Both of these panels allow for the maximum number of factors and cointegration relationships. Panels E and F report the robustness check results for the FECM averaging using the estimated cointegration vectors. Specially, Panel E assumes that the factor number is the optimal factor number determined by the PC2 criterion from Bai and Ng (2002). Panel F uses the pre-selected cointegration rank K_{max} with different factor numbers.
Panel G presents results using the I(1) variables suggested by Tu and Yi (2017). Panel H reports the MSFE distributions of ARMA(1,1) and 3PRF from Kelly and Pruitt (2015). “MMA”, “EQ”, and “CVh” stand for Mallows model averaging, simple averaging, and leave-h-out cross-validation model averaging, respectively.

Table 1.5.: U.S. Forecasting MSFE Distribution, forecasting horizon = 3, monthly

	percentile	5	25	50	75	95		percentile	5	25	50	75	95
Panel A	FAR optimal	0.706	0.792	0.909	1.016	1.121	Panel B	FECM optimal	0.707	0.811	0.892	1	1.237
	logit	0.706	0.792	0.909	1.016	1.121		logit	0.707	0.811	0.892	1	1.237
FAR selection	bagging	0.673	0.762	0.903	1.009	29.904	FECM selection	bagging	0.711	0.874	1.059	1.308	108.477
	pretest	0.749	0.817	0.973	1.078	139.92		pretest	0.707	0.866	0.985	1.11	104.02
Panel C	MMA	0.696	0.775	0.876	0.986	1.061	Panel D	MMA	0.596	0.767	0.88	1.075	1.407
FAR MA	CVh	0.693	0.788	0.89	0.988	1.035	FECM MA	CVh	0.609	0.718	0.81	0.944	1.085
	EQ	0.703	0.759	0.859	0.959	1.061		EQ	0.608	0.692	0.783	0.908	1.027
Panel E	MMA	0.636	0.729	0.828	0.978	1.167	Panel F	MMA	0.714	0.783	0.854	0.949	1.248
FECM MA	CVh	0.664	0.757	0.85	0.973	1.136	FECM MA	CVh	0.678	0.769	0.826	0.942	1.064
pre-select # of fac	EQ	0.639	0.735	0.834	0.954	1.049	pre-select # of coin	EQ	0.673	0.756	0.811	0.907	1.051
							Panel G	CVh	0.561	0.671	0.742	0.854	1.003
							Tu Yi (2017)	EQ	0.574	0.647	0.72	0.809	0.946
Panel H	ARMA(1, 1)	0.736	0.921	1.109	1.346	2.195		3PRF	1.34	1.701	1.894	2.173	2.972

Note: see Table 1.4.

Table 1.6.: U.S. Forecasting MSFE Distribution, forecasting horizon = 6, monthly

	percentile	5	25	50	75	95		percentile	5	25	50	75	95
Panel A	FAR optimal	0.606	0.743	0.857	1.003	1.111	Panel B	FECM optimal	0.58	0.723	0.8	0.941	1.259
	logit	0.606	0.743	0.857	1.003	1.111		logit	0.58	0.723	0.8	0.941	1.259
FAR selection	bagging	0.586	0.723	0.866	1.006	19.616	FECM selection	bagging	0.642	0.768	0.928	1.227	132.241
	pretest	0.627	0.767	0.925	1.053	85.343		pretest	0.626	0.761	0.884	1.034	130.203
Panel C	MMA	0.593	0.725	0.815	0.981	1.096	Panel D	MMA	0.414	0.528	0.643	0.841	1.196
FAR MA	CVh	0.64	0.747	0.855	0.979	1.035	FECM MA	CVh	0.454	0.582	0.685	0.795	0.998
	EQ	0.594	0.71	0.818	0.949	1.068		EQ	0.49	0.577	0.68	0.781	0.94
Panel E	MMA	0.451	0.576	0.665	0.792	1.082	Panel F	MMA	0.569	0.674	0.772	0.887	1.087
FECM MA	CVh	0.496	0.629	0.72	0.849	1.037	FECM MA	CVh	0.574	0.668	0.76	0.88	1.063
pre-select # of fac	EQ	0.524	0.615	0.714	0.816	0.955	pre-select # of coin	EQ	0.558	0.671	0.738	0.835	1.038
							Panel G	CVh	0.383	0.503	0.594	0.682	0.937
							Tu Yi (2017)	EQ	0.408	0.506	0.599	0.651	0.831
Panel H	ARMA(1, 1)	0.767	0.934	1.216	1.505	2.782		3PRF	1.105	1.362	1.659	1.957	2.44

Note: see Table 1.4.

Table 1.7.: U.S. Forecasting MSFE Distribution, forecasting horizon = 12, monthly

	percentile	5	25	50	75	95		percentile	5	25	50	75	95
Panel A	FAR optimal	0.518	0.642	0.781	0.894	1.065	Panel B	FECM optimal	0.454	0.598	0.712	0.898	1.056
FAR selection	logit	0.518	0.642	0.781	0.894	1.065	FECM selection	logit	0.454	0.598	0.712	0.898	1.056
	bagging	0.491	0.642	0.787	0.892	11.686		bagging	0.441	0.63	0.787	1.031	200.53
	pretest	0.521	0.697	0.837	0.931	51.484		pretest	0.445	0.624	0.761	0.959	178.397
Panel C	MMA	0.512	0.626	0.767	0.857	1.069	Panel D	MMA	0.256	0.353	0.435	0.548	0.89
FAR MA	CVh	0.553	0.675	0.788	0.908	1.006	FECM MA	CVh	0.339	0.432	0.511	0.639	0.852
	EQ	0.515	0.618	0.738	0.844	1.028		EQ	0.363	0.456	0.552	0.627	0.804
Panel E	MMA	0.297	0.406	0.486	0.622	0.913	Panel F	MMA	0.369	0.533	0.645	0.834	0.987
FECM MA	CVh	0.368	0.463	0.53	0.667	0.919	FECM MA	CVh	0.384	0.527	0.668	0.803	0.988
pre-select # of fac	EQ	0.386	0.489	0.575	0.67	0.832	pre-select # of coin	EQ	0.446	0.553	0.631	0.765	0.93
							Panel G	CVh	0.267	0.341	0.415	0.493	0.713
							Tu Yi (2017)	EQ	0.283	0.348	0.419	0.502	0.667
Panel H	ARMA(1, 1)	0.878	1.016	1.298	1.679	3.365		3PRF	0.745	0.968	1.141	1.426	2.096

Note: see Table 1.4.

Table 1.8.: Canada Forecasting MSFE Distribution, forecasting horizon =1, monthly

	percentile	5	25	50	75	95		percentile	5	25	50	75	95
Panel A	FAR optimal	0.931	1.006	1.036	1.057	1.088	Panel B	FECM optimal	0.93	1.009	1.041	1.059	1.114
	logit	0.931	1.006	1.036	1.057	1.088		logit	0.93	1.009	1.041	1.06	1.114
FAR selection	bagging	0.901	0.97	0.997	1.029	47.005	FECM selection	bagging	0.958	1.047	1.105	1.236	419.206
	pretest	0.916	0.994	1.032	1.112	214.915		pretest	0.955	1.019	1.054	1.245	424.09
Panel C	MMA	0.941	0.994	1.013	1.027	1.048	Panel D	MMA	1.233	1.375	1.469	1.57	1.706
FAR MA	CVh	0.939	0.993	1.012	1.025	1.042	FECM MA	CVh	0.989	1.054	1.09	1.123	1.18
	EQ	0.919	0.992	1.02	1.044	1.072		EQ	0.956	1.043	1.076	1.11	1.167
Panel E	MMA	0.982	1.051	1.088	1.122	1.184	Panel F	MMA	0.946	0.997	1.02	1.038	1.071
FECM MA	CVh	0.978	1.048	1.084	1.12	1.174	FECM MA	CVh	0.935	0.998	1.023	1.042	1.072
pre-select # of fac	EQ	0.968	1.027	1.069	1.106	1.153	pre-select # of coin	EQ	0.927	0.999	1.031	1.056	1.09
							Panel G	CVh	0.971	1.031	1.067	1.098	1.141
							Tu Yi (2017)	EQ	0.916	0.999	1.038	1.069	1.101
Panel H	ARMA(1, 1)	0.874	0.974	1.035	1.15	1.597		3PRF	1.53	1.808	2.004	2.259	3.001

Note: see Table 1.4.

Table 1.9.: Canada Forecasting MSFE Distribution, forecasting horizon =3, monthly

	percentile	5	25	50	75	95		percentile	5	25	50	75	95
Panel A	FAR optimal	0.873	0.965	1.014	1.049	1.114	Panel B	FECM optimal	0.874	0.965	1.015	1.051	1.117
FAR selection	logit	0.873	0.965	1.014	1.049	1.114	FECM selection	logit	0.874	0.965	1.015	1.051	1.117
	bagging	0.853	0.954	1	1.045	48.2		bagging	0.863	0.964	1.037	1.164	455.728
	pretest	0.9	0.987	1.054	1.12	239.29		pretest	0.88	0.966	1.023	1.119	456.415
Panel C	MMA	0.873	0.95	0.995	1.03	1.062	Panel D	MMA	0.935	1.107	1.22	1.358	1.593
FAR MA	CVh	0.883	0.961	1	1.022	1.045	FECM MA	CVh	0.876	0.967	1.018	1.094	1.242
	EQ	0.854	0.939	0.985	1.03	1.068		EQ	0.822	0.894	0.947	1.012	1.105
Panel E	MMA	0.876	0.982	1.038	1.127	1.277	Panel F	MMA	0.884	0.952	0.999	1.034	1.091
FECM MA	CVh	0.901	0.981	1.047	1.104	1.202	FECM MA	CVh	0.893	0.959	1.004	1.034	1.078
pre-select # of fac	EQ	0.868	0.933	0.997	1.053	1.13	pre-select # of coin	EQ	0.862	0.945	0.995	1.038	1.083
							Panel G	CVh	0.793	0.887	0.948	0.998	1.048
							Tu Yi (2017)	EQ	0.746	0.831	0.894	0.942	1.004
Panel H	ARMA(1, 1)	0.903	0.983	1.03	1.149	1.963		3PRF	1.579	1.891	2.084	2.326	2.646

Note: see Table 1.4.

Table 1.10.: Canada Forecasting MSFE Distribution, forecasting horizon =6, monthly

	percentile	5	25	50	75	95		percentile	5	25	50	75	95
Panel A	FAR optimal	0.832	0.931	0.988	1.036	1.118	Panel B	FECM optimal	0.826	0.932	0.987	1.041	1.121
FAR selection	logit	0.832	0.931	0.988	1.036	1.118	FECM selection	logit	0.826	0.932	0.987	1.041	1.121
	bagging	0.819	0.929	0.991	1.049	22.647		bagging	0.84	0.952	1.018	1.146	565.384
	pretest	0.834	0.965	1.042	1.122	106.278		pretest	0.843	0.939	1.009	1.085	566.041
Panel C	MMA	0.805	0.92	0.964	1.018	1.078	Panel D	MMA	0.622	0.776	0.89	1.078	1.387
FAR MA	CVh	0.855	0.945	0.984	1.017	1.047	FECM MA	CVh	0.691	0.813	0.894	0.992	1.181
	EQ	0.796	0.903	0.944	1.007	1.067		EQ	0.67	0.764	0.83	0.907	1.005
Panel E	MMA	0.694	0.825	0.92	1.041	1.213	Panel F	MMA	0.809	0.913	0.962	1.023	1.1
FECM MA	CVh	0.722	0.868	0.953	1.044	1.185	FECM MA	CVh	0.83	0.931	0.983	1.023	1.081
pre-select # of fac	EQ	0.711	0.829	0.892	0.966	1.068	pre-select # of coin	EQ	0.796	0.904	0.95	1.011	1.077
							Panel G	CVh	0.583	0.703	0.781	0.855	0.982
							Tu Yi (2017)	EQ	0.582	0.659	0.739	0.795	0.908
Panel H	ARMA(1, 1)	0.894	0.967	1.024	1.122	2.082		3PRF	1.328	1.663	1.879	2.111	2.49

Note: see Table 1.4.

Table 1.11.: Canada Forecasting MSFE Distribution, forecasting horizon =12, monthly

	percentile	5	25	50	75	95		percentile	5	25	50	75	95
Panel A	FAR optimal	0.756	0.86	0.924	0.995	1.123	Panel B	FECM optimal	0.74	0.855	0.921	0.99	1.113
FAR selection	logit	0.756	0.86	0.924	0.995	1.123	FECM selection	logit	0.74	0.855	0.921	0.99	1.113
	bagging	0.738	0.858	0.928	1.024	28.944		bagging	0.771	0.868	0.942	1.097	725.921
	pretest	0.77	0.897	0.979	1.101	136.326		pretest	0.759	0.864	0.929	1.056	727.714
Panel C	MMA	0.722	0.807	0.88	0.941	1.071	Panel D	MMA	0.32	0.429	0.588	0.804	1.135
FAR MA	CVh	0.805	0.884	0.938	0.989	1.066	FECM MA	CVh	0.456	0.57	0.681	0.831	1.017
	EQ	0.726	0.811	0.865	0.933	1.055		EQ	0.541	0.602	0.667	0.766	0.953
Panel E	MMA	0.449	0.597	0.739	0.907	1.168	Panel F	MMA	0.708	0.805	0.878	0.944	1.088
FECM MA	CVh	0.566	0.685	0.784	0.925	1.115	FECM MA	CVh	0.751	0.863	0.919	0.977	1.085
pre-select # of fac	EQ	0.569	0.653	0.731	0.868	1.024	pre-select # of coin	EQ	0.706	0.802	0.864	0.933	1.056
							Panel G	CVh	0.403	0.515	0.606	0.713	0.882
							Tu Yi (2017)	EQ	0.414	0.48	0.544	0.622	0.787
Panel H	ARMA(1, 1)	0.923	0.999	1.055	1.186	2.367		3PRF	1.035	1.216	1.412	1.692	2.291

Note: see Table 1.4.

Table 1.12.: U.S. recession, Distribution, I(1) regressors

percentiles	5	25	50	75	95	5	25	50	75	95	5	25	50	75	95	5	25	50	75	95
FAR optimal	0.803	1.045	1.127	1.161	1.249	0.617	0.951	1.074	1.124	1.187	0.518	0.904	0.992	1.099	1.194	0.45	0.705	0.861	0.943	1.124
FAR MA	h=1					h=3					h=6					h=12				
MMA	0.658	0.806	0.913	1	1.104	0.561	0.627	0.755	0.885	1.069	0.43	0.574	0.686	0.841	1.11	0.393	0.505	0.618	0.754	1.195
CVh	0.669	0.8	0.916	0.997	1.069	0.587	0.664	0.787	0.908	1.05	0.491	0.616	0.738	0.899	1.067	0.441	0.538	0.65	0.853	1.033
EQ	0.684	0.807	0.88	0.988	1.09	0.548	0.627	0.725	0.852	1.07	0.429	0.561	0.672	0.852	1.092	0.404	0.496	0.586	0.745	1.076
FECM MA	h=1					h=3					h=6					h=12				
CVh	0.637	0.762	0.913	1.018	1.187	0.436	0.539	0.636	0.728	1.038	0.298	0.38	0.46	0.595	1.022	0.194	0.245	0.302	0.405	0.806
EQ	0.642	0.763	0.866	0.978	1.137	0.431	0.516	0.614	0.699	0.965	0.315	0.402	0.465	0.577	0.887	0.216	0.262	0.316	0.411	0.678

Note: the notation of the estimation methods follows Table 1.4.

Table 1.13.: U.S. expansion, Distribution, I(1) regressors

percentiles	5	25	50	75	95	5	25	50	75	95	5	25	50	75	95	5	25	50	75	95
FAR optimal	0.944	1.014	1.045	1.082	1.141	0.839	0.949	1.014	1.084	1.181	0.732	0.825	0.965	1.066	1.168	0.586	0.719	0.87	1.026	1.103
FAR MA	h=1					h=3					h=6					h=12				
MMA	0.77	0.955	0.993	1.025	1.044	0.72	0.868	0.937	1.014	1.081	0.712	0.799	0.874	1.006	1.088	0.545	0.691	0.801	0.941	1.021
CVh	0.762	0.95	0.995	1.023	1.046	0.683	0.873	0.938	1.01	1.054	0.732	0.816	0.901	0.99	1.094	0.578	0.724	0.856	0.95	1.028
EQ	0.776	0.943	0.979	1.023	1.057	0.742	0.848	0.91	1.003	1.071	0.694	0.768	0.858	0.995	1.075	0.528	0.681	0.784	0.907	0.986
FECM MA	h=1					h=3					h=6					h=12				
CVh	0.845	0.999	1.047	1.083	1.153	0.554	0.763	0.864	0.948	1.046	0.439	0.575	0.69	0.761	0.914	0.268	0.37	0.495	0.57	0.756
EQ	0.822	0.939	0.984	1.026	1.08	0.611	0.73	0.822	0.888	0.973	0.457	0.567	0.666	0.741	0.838	0.294	0.388	0.469	0.575	0.675

Note: the notation of the estimation methods follows Table 1.4.

Table 1.14.: CAN recession, Distribution, I(1) regressors

percentiles	5	25	50	75	95		5	25	50	75	95		5	25	50	75	95		5	25	50	75	95
FAR optimal	0.725	0.893	1.032	1.146	1.327	0	0.571	0.812	0.952	1.098	1.441	0	0.521	0.708	0.892	1.074	1.499	0	0.42	0.621	0.842	1.011	1.516
FAR MA	h=1						h=3						h=6						h=12				
MMA	0.749	0.927	1	1.055	1.186		0.594	0.843	0.932	1.045	1.267		0.497	0.744	0.878	1.045	1.334		0.399	0.621	0.787	0.988	1.399
CVh	0.759	0.93	0.997	1.053	1.191		0.646	0.885	0.965	1.035	1.189		0.586	0.854	0.966	1.03	1.196		0.493	0.775	0.923	1.013	1.204
EQ	0.737	0.896	1.003	1.101	1.244		0.601	0.8	0.925	1.059	1.288		0.532	0.739	0.856	1.042	1.314		0.432	0.624	0.812	0.987	1.361
FECM MA	h=1						h=3						h=6						h=12				
CVh	0.668	0.904	1.042	1.155	1.429		0.479	0.696	0.838	1.054	1.404		0.319	0.516	0.691	0.961	1.48		0.152	0.352	0.54	0.818	1.402
EQ	0.655	0.888	1.025	1.15	1.381		0.441	0.656	0.786	0.957	1.446		0.331	0.456	0.631	0.813	1.292		0.2	0.33	0.474	0.642	1.091

Note: the notation of the estimation methods follows Table 1.4.

Table 1.15.: CAN expansion, Distribution, I(1) regressors

percentiles	5	25	50	75	95		5	25	50	75	95		5	25	50	75	95		5	25	50	75	95
FAR optimal	0.956	1.01	1.036	1.054	1.087	0	0.917	0.976	1.017	1.054	1.104	0	0.866	0.954	1.003	1.048	1.122	0	0.763	0.857	0.934	1.019	1.148
FAR MA	h=1						h=3						h=6						h=12				
MMA	0.95	0.999	1.014	1.027	1.049	0	0.902	0.967	1	1.032	1.067	0	0.829	0.933	0.978	1.029	1.078	0	0.713	0.81	0.888	0.986	1.081
CVh	0.95	0.997	1.014	1.026	1.046	0	0.907	0.97	1.005	1.026	1.053	0	0.867	0.951	0.993	1.025	1.075	0	0.784	0.864	0.943	0.999	1.071
EQ	0.942	0.993	1.02	1.042	1.068	0	0.888	0.952	0.991	1.028	1.067	0	0.821	0.92	0.957	1.006	1.073	0	0.733	0.805	0.867	0.956	1.072
FECM MA	h=1						h=3						h=6						h=12				
CVh	0.957	0.999	1.009	1.021	1.053		0.85	0.96	1.027	1.081	1.175		0.586	0.731	0.825	0.934	1.141		0.296	0.395	0.511	0.657	0.897
EQ	0.932	1.003	1.043	1.068	1.103		0.772	0.852	0.901	0.95	1.01		0.605	0.683	0.751	0.81	0.92		0.404	0.492	0.566	0.646	0.806

Note: the notation of the estimation methods follows Table 1.4.

Table 1.16.: U.S., MSFE ratio, FECM:FAR, I(1) regressors

Horizon	MN monthly	IP	UNRATE	CMR	RPI	CPI	CPI core	PPI	PCE
h=1	CVh	**1.056**	**1.043**	**1.071**	1.031	**0.956**	**0.985**	0.98	1.009
	EQ	**0.978**	**1**	**1.026**	1.044	**0.953**	**0.954**	**0.968**	0.959
h=3	CVh	**0.845**	**0.892**	**0.964**	**0.954**	**0.783**	**0.809**	**0.796**	**0.81**
	EQ	**0.83**	**0.86**	**0.877**	**0.95**	**0.79**	**0.799**	**0.806**	**0.813**
h=6	CVh	**0.647**	**0.657**	**0.733**	**0.861**	**0.616**	**0.688**	**0.627**	**0.692**
	EQ	**0.706**	**0.714**	**0.732**	**0.865**	**0.643**	**0.675**	**0.655**	**0.681**
h=12	CVh	**0.506**	**0.459**	**0.541**	**0.666**	**0.499**	**0.564**	**0.467**	**0.574**
	EQ	**0.56**	**0.565**	**0.571**	**0.633**	**0.543**	**0.569**	**0.489**	**0.566**

Note: The variables from left to right: industrial production, unemployment rate, real manufacturing and trade industries sales, real personal income, consumer price index, consumer price index less food and energy, producer price index, and personal consumption expenditure. The benchmark is the FAR model averaging. The bold numbers mean that FECM averaging improves on both FAR model averaging and AR(12).

Table 1.17.: CAN, MSFE ratio, FECM:FAR, I(1) regressors

Horizon	CAN monthly	IP	EMP	UNRATE	CPI	CPI core	Houst	5Y MORT
h=1	CVh	1.085	1.044	1.042	**1.001**	**1.006**	**1.054**	1.026
	EQ	1.045	1.023	**0.98**	**0.999**	**0.995**	**1.018**	**0.985**
h=3	CVh	**0.915**	**0.949**	**0.848**	**0.931**	**0.867**	1.016	**0.907**
	EQ	**0.878**	**0.874**	**0.811**	**0.912**	**0.865**	**0.937**	**0.806**
h=6	CVh	**0.677**	**0.758**	**0.718**	**0.816**	**0.712**	**0.884**	**0.626**
	EQ	**0.704**	**0.703**	**0.658**	**0.769**	**0.743**	**0.802**	**0.676**
h=12	CVh	**0.482**	**0.515**	**0.537**	**0.727**	**0.591**	**0.793**	**0.566**
	EQ	**0.523**	**0.497**	**0.53**	**0.668**	**0.631**	**0.701**	**0.576**

Note: The variables from left to right: industrial production, employment, unemployment rate, consumer price index, consumer price index less food and energy, housing starts, and five year mortgage rate. The benchmark is the FAR model averaging. The bold numbers mean that FECM averaging improves on both FAR model averaging and AR(12).

Table 1.18.: U.S., MSFE, I(1) regressors

		S&P 500	Exchange rates					Interest rates				
Horizon	MN monthly	S&P 500	Swi	Japan	UK	CAN		3 months	6 months	1 year	5 years	10 years
h=1	ARMA(1,1)	1.002	0.988	1.07	1.129	1.019		1.056	1.129	1.431	1.243	1.178
	RW	**0.922**	**0.882**	**0.961**	**0.976**	**0.955**		0.875	0.928	1.014	0.997	**0.945**
	CVh, FAR MA	0.973	1.005	1.029	1.028	1.04		0.819	0.804	0.88	0.967	0.972
	EQ, FAR MA	0.999	1.028	1.058	1.036	1.043		0.802	0.792	**0.851**	**0.938**	**0.945**
	CVh, FECM MA	1.057	1.063	1.104	1.114	1.111		0.852	0.878	0.973	1.007	1.016
	EQ, FECM MA	1.018	1.027	1.077	1.044	1.074		**0.781**	**0.79**	0.865	0.941	0.956
							FECM optimal	0.807	0.806	0.889	0.995	1.031
							DNS VAR(1)	1.129	1.221	1.324	1.277	1.143
							DNS AR(1)	0.944	0.994	1.075	1.057	0.992
h=3	ARMA(1,1)	0.921	0.853	0.949	0.929	0.951		0.774	0.773	0.817	0.828	0.822
	RW	0.894	**0.848**	**0.923**	0.922	**0.936**		0.746	0.773	0.776	0.821	0.811
	CVh, FAR MA	0.983	1.002	1.021	1.03	1.037		0.859	0.853	0.872	0.941	0.947
	EQ, FAR MA	0.963	1.015	1.025	1.037	1.035		0.807	0.803	0.833	0.926	0.948
	CVh, FECM MA	0.91	0.907	1.016	0.928	0.96		**0.631**	**0.654**	0.754	**0.776**	**0.803**
	EQ, FECM MA	**0.871**	0.882	0.952	**0.907**	0.941		0.659	0.663	**0.713**	0.793	0.817
							FECM optimal	0.838	0.842	0.85	0.995	1.036
							DNS VAR(1)	0.752	0.783	0.789	0.837	0.821
							DNS AR(1)	0.748	0.777	0.778	0.825	0.812
h=6	ARMA(1,1)	0.907	0.85	0.929	0.93	0.939		0.816	0.863	0.859	0.862	0.866
	RW	0.9	0.845	0.891	0.926	0.942		0.82	0.87	0.845	0.855	0.863
	CVh, FAR MA	1.002	0.988	0.989	1.016	1.07		0.917	0.847	0.859	0.929	0.93
	EQ, FAR MA	0.95	1.004	0.974	1.037	1.069		0.827	0.803	0.826	0.931	0.943
	CVh, FECM MA	0.723	0.743	0.937	**0.704**	**0.764**		**0.633**	**0.615**	0.644	**0.65**	**0.674**
	EQ, FECM MA	**0.704**	**0.717**	**0.818**	0.749	0.807		0.649	0.623	**0.635**	0.703	0.708
							FECM optimal	0.914	0.899	0.908	1.056	1.07
							DNS VAR(1)	0.82	0.877	0.85	0.863	0.869
							DNS AR(1)	0.82	0.875	0.846	0.858	0.867
h=12	ARMA(1,1)	1.038	0.901	0.976	0.977	1.026		0.836	0.886	0.877	0.901	0.911
	RW	1.026	0.899	0.904	0.975	1.024		0.835	0.891	0.872	0.893	0.905
	CVh, FAR MA	0.991	0.995	0.961	0.962	1.01		0.86	0.868	0.874	0.948	0.932
	EQ, FAR MA	0.858	0.948	0.896	0.924	0.869		0.816	0.808	0.816	0.875	0.891
	CVh, FECM MA	**0.474**	0.625	0.726	**0.495**	0.683		**0.467**	**0.479**	**0.491**	**0.543**	**0.561**
	EQ, FECM MA	0.498	**0.54**	**0.618**	0.548	**0.548**		0.531	0.533	0.534	0.559	0.565
							FECM optimal	0.9	0.926	0.935	1.01	1.033
							DNS VAR(1)	0.837	0.902	0.875	0.899	0.912
							DNS AR(1)	0.836	0.9	0.873	0.898	0.911

Note: The variables from left to right: the S&P 500 index, the exchange rates between Switzerland/U.S., Japan/U.S, U.S./U.K., and Canada/U.S, the 3 months and 6 months T-bill rates, and the 1 year, 5 years, and 10 years T-bond rates. I also present the MSFE ratios of interest rates using the Dynamic Nelson-Siegel models. I use the five interest rates to form the variant, and use AR(1) / VAR(1) to model the dynamics of the factors. The decay parameter is set to be 0.0609. The benchmark estimation method is AR(12). Bold numbers are the smallest numbers of each column under the same forecasting horizon.

Table 1.19.: Distributions, the model with the largest weight, FECM MA

Median	Percentile	5	25	50	75	95
	hori=1 month	3.4	13	19	25	33.3
	hori=3 months	12	18	25.25	31	38
	hori=6 months	14.4	20	25	31	38
	hori=12 months	15.4	20	24	27	35.2
Interquartile range	Percentile	5	25	50	75	95
	hori=1 month	11.4	22	30	36	43
	hori=3 months	19	26	32.5	37	42
	hori=6 months	22.4	29.75	33	36	41.75
	hori=12 months	20	29	33.75	38	41

Table 1.20.: Distributions, the variances of weights

FECM weight variance	Percentile	5	25	50	75	95
	hori=1 month	0.031	0.036	0.042	0.047	0.059
	hori=3 months	0.035	0.041	0.045	0.051	0.062
	hori=6 months	0.041	0.051	0.056	0.062	0.071
	hori=12 months	0.045	0.055	0.062	0.069	0.08
FAR weight variance	Percentile	5	25	50	75	95
	hori=1 month	0.059	0.071	0.083	0.094	0.119
	hori=3 months	0.062	0.072	0.081	0.092	0.113
	hori=6 months	0.058	0.071	0.084	0.093	0.112
	hori=12 months	0.065	0.081	0.092	0.101	0.116

2. APPROACHES TO ESTIMATING LARGE-DIMENSIONAL REGRESSIONS WITH ENDOGENEITY: A SIMULATION COMPARISON

2.1 Introduction

The objective of this paper is to compare several recently proposed estimation methods designed for large-dimensional regressions with endogeneity. When using instrumental variables, two-stage least squares (2SLS) and generalized method of moments (GMM) are the two most common ways to cope with endogeneity. The former method includes first-stage regressions of the endogenous regressors on the instruments, and a second-stage regression between the dependent variable and the estimated regressors from the first-stage. The latter method is the generalized form of the first method, where the error distribution assumptions are relaxed. Both approaches assume the number of observations to be large to obtain estimation consistency.

Empirically, endogeneity is a common issue in reduced-form analyses. Applications include, but are not limited to, growth empirics, demand/ supply curve estimation, and epidemiological studies. Specifically, the GMM framework is often applied to dynamic panel settings. The two widely-applied GMM methods, proposed by Arellano and Bond (1991) and Blundell and Bond (1998), use the lags of dependent and independent variables as instruments. These strategies have proven to be useful in multiple literatures, but they suffer from weak instruments and over-identification problems. Roodman (2009) discusses these issues, and reproduces the analysis of growth and national income inequality from Forbes (2000). He argues that by expanding or collapsing the original Forbes (2000) instrument set, the significance of national income inequality is unstable, and it is unclear if reversibility or third-variable causation exist. In a more recent simulation study, Hauk Jr (2017) generates simulated datasets which meet the moment conditions of the real-world growth data. He shows that the within-group estimators, such as system-GMM (see Blundell and Bond, 1998), are dominated by between-group estimators when estimating all coefficients. Whether or not this result comes from the over-identification problem associated with the GMM estimators remains to be explored.

In a similar fashion, the instrument selection process of the 2SLS method can be ad-hoc. For example, to find out the determinants of GDP and income growth, researchers have investigated various factors in the presence of reverse causality. Common determinants include foreign aid, trade volume, access to financial intermediaries, and governance quality. The Solow-model-induced productivity factors have also been closely examined from a growth accounting perspective. Other covariates may contribute to growth as well, as macroeconomic variables comove with each other. Which variables and instruments are the most informative? Four of the most well-cited papers using instruments to evaluate the causality of foreign aids to GDP growth, namely, Burnside and Dollar (2000) , Collier and Dollar (2002), Hansen and Tarp (2001), and Dalgaard et al. (2004), have four different sets of covariates and instruments. Twenty-six different covariates are used to predict growth while only seven of them overlap, and the total number of regressors used in the first-stage regressions is even greater. This is a large number of variables considering that the sample size of low-income countries is only 189, according to the dataset used by Burnside and Dollar (2000), and can be even smaller in the presence of structural breaks. As a consequence, the choices of regressors can be highly influential to the final results (see related arguments in Christensen and Miguel, 2018). Additionally, it may be beneficial to include other functional forms of these regressors, which requires high-dimensional econometric methods.

Some micro-related topics also rely on the proper analysis of large dimensional datasets. Chernozhukov et al. (2015) re-evaluate the analysis of automobile market share studied in Berry et al. (1995). After applying shrinkage methods to both the control variables and instruments, their results show that the estimates of own-price elasticities are larger in absolute value, which fit the automobile industry market structure better than the Berry et al. (1995) results. Additionally, recent papers, such as Hansen and Kozbur (2014) and Carrasco (2012), re-analyze the regression model in Angrist and Keueger (1991) using shrinkage estimation methods. These papers have not yet investigated the use of shrinkage methods in the second-stage regression.

Last but not least, instrumental variables have been intensively used for making causal inferences in epidemiology. Lawlor et al. (2008) discuss the limitations of using randomized controlled trials to make causal inference in medical studies, and propose the use of instrumental variables. Moreover, high-dimensional models have allowed researchers to determine the origins of certain diseases. For example, Lin et al. (2015) use 2SLS with shrinkage estimation methods and identify genes that are related to obesity. It is also desirable for estimation methods to be robust to weak instruments, as it is a concern that individual genes can only capture a small amount of the variance contained in genetically complex traits.

Theoretically, a strand of literature has developed shrinkage estimation methods which allow for many regressors and many instruments simultaneously. Lin et al. (2015) and Zhu (2018), among others, combine the least absolute shrinkage and selection operator (Lasso) with 2SLS in this context. Shi (2016) and Caner et al. (2018) also use shrinkage techniques, combining Lasso and elastic net with GMM, respectively. The simulations conducted in these papers provide supportive evidence of coefficient estimation consistency. Another collection of papers, e.g., Caner and Kock (2018), move one step further and construct confidence intervals around the consistently estimated coefficients.

With these newly developed econometric tools, how do applied researchers determine the most appropriate approach to use? One way is to examine the assumptions and determine which one best fit the real-world dataset. The first assumption to check is regarding the dimensions of the regressors and instruments. The sparsity requirements, which represent the level of useful information contained in the regressors and the instruments, also varies across approaches. For example, Lin et al. (2015) require the sparsity of both the first step and the second step regressions to be $o(\sqrt{n})$ in a 2SLS framework, where n is the number of observations. In contrast, Zhu (2018) does not impose such requirements. Instead, the paper imposes stricter assumptions on the number of instruments. The sparsity and dimension assumptions of Shi (2016) are somewhere in between. Other assumptions include the allowance of conditional heteroskedasticity. For instance, Gold et al. (2020) focus on homoskedasticity to develop statistical inference, while GMM-based shrinkage methods allow for second-stage conditional heteroskedasticity.

In this paper, I conduct Monte Carlo simulations comparing the finite sample performance of six shrinkage methods that are designed for regressions with large dimensional regressors and instrumental variables. The simulations are aimed at answering the following questions: between the two types of shrinkage methods based on 2SLS and GMM, which one has the better finite sample properties? How sensitive are the results to weak instruments, heteroskedasticity, different dimensions for the number of observations and regressors/instruments, and the choice of tuning parameters?

The rest of this paper is organized as follows. Section 2.2 describes the six estimation procedures and assumptions. Section 2.3 presents the simulation designs, results, an empirical example, and discussions. Section 2.4 concludes the chapter. All result tables are located subsequently.

2.2 Estimation Procedures

This section presents the assumptions and estimation procedures of each method. The regressions considered are $\underset{n\times 1}{Y} = \underset{n\times p}{X}\underset{p\times 1}{\beta} + u$ and $\underset{n\times 1}{X_j} = \underset{n\times q}{Z}\underset{q\times 1}{\alpha_j} + e_j,\ j = 1,\ \ldots, p$, where n is the number of observations, p is the dimension of regressors X, and q is the dimension of the instruments Z. The first-stage regressions can be stacked as $X = [X_1, \ldots,\ X_p] = Z[\alpha_1, \ldots,\ \alpha_p] + e = ZA + e$.

The six estimation methods are categorized into two groups: the two-stage least squares method combined with shrinkage (2SLS-shrinkage) and the generalized method of moments method combined with shrinkage (GMM-shrinkage). Zhu (2018), Lin et al. (2015), and Gold et al. (2020) use the 2SLS-shrinkage in the large-dimensional data environment. In the first stage, $\hat{\alpha}_j = \underset{\alpha_j \in R^q}{argmin} \frac{1}{2n}(X_j - Z \underset{q\times 1}{\alpha_j})'(X_j - Z \underset{q\times 1}{\alpha_j}) + \lambda_j p(\alpha_j)$, $j \in 1,\ ...,\ p$. The second-stage objective is to find $\hat{\beta} = \underset{\beta \in R^p}{argmin} \frac{1}{2n}(Y - \hat{X}\beta)'(Y - \hat{X}\beta) + \lambda_0 p(\beta)$, where $p(\bullet)$ is the penalty function, $\hat{X}$ is the first-stage estimated regressors, and $\{\lambda_1,\ ...,\ \lambda_p,\ \lambda_0\}$ are the tuning parameters.

Caner et al. (2018), Shi (2016), and Caner and Kock (2018) adopt the GMM-shrinkage methods. These methods also involve two estimation stages, where the second-stage is designed for conditional heteroskedasticity in u. In the first stage, the weighting matrix W is set to be the identity matrix with dimension q, and the objective is to find $\hat{\beta}^{1st} = \underset{\beta \in R^p}{argmin} \frac{1}{n^2}(Y - X\beta)'ZZ'(Y - X\beta) + \lambda^{1st} p(\beta)$, where $p(\bullet)$ is the penalty function and λ^{1st} is the first-stage tuning parameter. In the second stage, the weighting matrix is constructed as $diag(\ \frac{1}{\hat{\sigma}_1^2},\ \cdots,\ \frac{1}{\hat{\sigma}_l^2},\ \cdots,\ \frac{1}{\hat{\sigma}_q^2}\ )$, with $\hat{\sigma}_l^2 = \frac{1}{n}\sum_{i=1}^n Z_{il}^2 \hat{u}_i^2$ and $\hat{u} = Y - X\hat{\beta}^{1st}$. The objective is then to find $\hat{\beta}^{2nd} = \underset{\beta \in R^p}{argmin} \frac{1}{n^2}(Y - X\beta)'ZWZ'(Y - X\beta) + \lambda^{2nd} p(\beta)$, where $p(\bullet)$ is the same penalty function as the first stage and λ^{2nd} is the second-stage tuning parameter.

The following six subsections contain the estimation procedures in detail. I then discuss the procedures of two bias-corrected estimators. The last subsection discusses the assumptions of each method.

2.2.1 Lin et al. (2015)

Lin et al. (2015) analyze three shrinkage methods in the 2SLS environment, including the Lasso-type shrinkage, smoothly clipped absolute deviation penalty, and minimax concave penalty (MCP). The largest rate that the numbers of instruments and regressors can grow of is $e^{o(n)}$. The authors derive the l_1 loss upper bound of the coefficient estimation error, and the weak oracle property regarding the nonzero subset of $\{\beta_i\}_{i=1}^p$. Cross-validation is suggested to select the tuning parameters. I focus on the MCP penalty in this paper, as Lin et al. (2015) show that MCP provides the smallest estimation losses among the three penalties.

The MCP penalty takes the form $\lambda p(t) = \int_0^t \frac{(a\lambda-\theta)_+}{a} d\theta$, $j = 1,\ ...,\ p$, where $a = 3.7$ and λ is the tuning parameter. Thus, the first stage objective is to find $\hat{\alpha}_j = \underset{\alpha_j \in R^q}{argmin} \frac{1}{2n}(X_j - Z\alpha_j)'(X_j - Z\alpha_j) + \sum_{i=1}^q \int_0^{|\alpha_{ij}|} \frac{(3.7\lambda_j-\theta)_+}{3.7} d\theta$ for each $j \in \{1,\ ...,\ p\}$. The second-stage objective is to find $\hat{\beta} = \underset{\beta \in R^p}{argmin} \frac{1}{2n}(Y - \hat{X}\beta)'(Y - \hat{X}\beta) + \sum_{j=1}^p \int_0^{|\beta_j|} \frac{(3.7\lambda_0-\theta)_+}{3.7}$ $d\theta$.

2.2.2 Zhu (2018)

Zhu (2018) adopts the Lasso-type penalty, and uses theoretically derived tuning parameters. The detailed sparsity assumptions can be found in Assumption 2.5 of the paper, and the theoretical results include the upper bounds for both l_1 and l_2 losses of $\{\hat{\beta}_i\}_{i=1}^p$.

The algorithm includes an iterative estimation of the tuning parameters. The details of calculating the tuning parameters can be found in the paper. In the first stage, for each $j = 1,\ ...,\ p$, $\hat{\alpha}_j = \underset{\alpha_j \in R^q}{argmin} \frac{1}{2n}(X_j - Z\alpha_j)'(X_j - Z\alpha_j) + \sum_{i=1}^q \lambda_j^{(k)} |\alpha_{ij}|$, where λ_j is the tuning parameter and the superscript (k) represents the k^{th} iteration of estimating the tuning parameters. The second-stage objective is to find $\hat{\beta} = \underset{\beta \in R^p}{argmin} \frac{1}{2n}(Y - \hat{X}\beta)'(Y - \hat{X}\beta) + \sum_{j=1}^p \lambda_0^{(k)} |\beta_j|$.

2.2.3 Gold et al. (2020)

Gold et al. (2020) analyze the asymptotic behavior of two-stage Lasso estimators, with the tuning parameters selected based on cross-validation. The two stage estimation objectives are the same as Zhu (2018).

2.2.4 Shi (2016)

Shi (2016) uses the Lasso-type penalty with GMM-shrinkage, and suggests the GMM-AIC criterion to select the tuning parameter ρ. The paper focuses on the standard instrument regressions where $n > q$, allowing q and p to be large. The theoretical analysis in Shi (2016) is built on approximate sparsity, and shows the consistency of $\hat{\beta}$ under the l_1 norm.

The two stage estimations can be written as $\hat{\beta} = \underset{\beta \in R^p}{argmin} \frac{1}{n^2}(Y - X\beta)'ZWZ'(Y - X\beta) + \hat{\rho}\sum_1^p |\beta_i|$, where $\hat{\rho} = argmin \frac{1}{n^2}(Y - X\check{\beta}(\rho))'ZWZ'(Y - X\check{\beta}(\rho)) + \frac{2}{n}ln[ln(n)]$ $||\check{\beta}(\rho)||_0$, $\check{\beta}(\rho)$ is the estimated β with the chosen tuning parameter ρ, and $|| \bullet ||_0$ is the number of nonzero components in a vector.

2.2.5 Caner and Kock (2018)

Caner and Kock (2018) use the Lasso-type penalty with cross-validation to select the tuning parameters and derive the l_1 estimation loss upper bound of $\hat{\beta}$ for both $n > q$ and $q > n$. There are no sparsity requirements nor full rank assumptions of A. This allows Σ_{xz} and A to be rank deficient under higher order moments conditions. As a result, their procedure is robust to a fixed number of weak instruments.

2.2.6 Caner et al. (2018)

Caner et al. (2018) use the adaptive elastic net as the penalty function, where $\hat{\beta} = (1 + \frac{\lambda_2}{n^2})\underset{\beta \in R^p}{argmin}(Y - X\beta)'ZWZ'(Y - X\beta) + \lambda_1 \sum_{j=1}^p \pi_j|\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2$. The model does not allow q or p to be greater than n. The theoretical results demonstrate the model selection consistency, and the simulation results demonstrate the superior performance of the elastic net penalty as compared to the Lasso-type penalty.

There are two estimation steps at each stage. First, $\hat{\beta}$ is obtained by setting $\pi_i = 1$ without the scaling factor $(1 + \frac{\lambda_2}{n^2})$. Denote $\mathring{\beta}$ as the estimated coefficient at this step.

In the next step, $\pi_i = |\mathring{\beta}_i|^{-2}$ is calculated and shrinkage estimation is conducted one more time to obtain the final $\hat{\beta}$.

The tuning parameters are selected based on a BIC-type information criterion, $(\hat{\lambda}_1,\ \hat{\lambda}_2) = argmin\ IC_\lambda = \frac{1}{n}(Y - X\check{\beta}(\lambda_1,\ \lambda_2))'ZWZ'(Y - X\check{\beta}(\lambda_1,\ \lambda_2)) + ||\check{\beta}||_0 ln(n)$ $max\{1,\ ln[ln(p)]\}$, where $\check{\beta}(\lambda_1,\ \lambda_2)$ is the estimated β with the chosen tuning parameters $(\lambda_1,\ \lambda_2)$, and $||\bullet||_0$ is the number of nonzero components in a vector.

2.2.7 Bias Corrections

Gold et al. (2020) and Caner and Kock (2018) also propose the bias-corrected estimators. Specifically, Gold et al. (2020) construct the bias-corrected estimator as $\tilde{\beta} = \hat{\beta} + (\frac{\hat{X}'\hat{X}}{n})^{-1}\hat{X}'(y - \hat{X}'\hat{\beta})/n$, where $\hat{\beta}$ is the second-stage estimator. To estimate $(\frac{\hat{X}'\hat{X}}{n})^{-1}$ with the possibility that $\hat{X}'\hat{X}$ is rank deficient, Gold et al. (2020) propose a modification of the CLIME estimator of Cai et al. (2011).

Similar to Gold et al. (2020), Caner and Kock (2018) suggest the bias-corrected estimator $\tilde{\beta} = \hat{\beta} + (\frac{X'ZWZ'X}{n^2 q})^{-1}(\frac{X'ZWZ'(Y - X\hat{\beta})}{n^2 q})$, where W is the second-stage weighting matrix. The estimation procedure of $(\frac{X'ZWZ'X}{n^2 q})^{-1}$ is the same as the estimation procedure of $(\frac{\hat{X}'\hat{X}}{n})^{-1}$ in Gold et al. (2020).

To make the estimation procedures comparable across the six papers, I focus on the consistency results obtained from the non-bias-corrected estimators. The results of the two bias-corrected estimators are available upon request.

2.2.8 Model Assumptions

Table 2.1 lists the main assumptions and objectives of the six papers to highlight the differences among them. These assumptions and objectives serve as guidelines to help researchers select the most suitable method based on the data characteristics and research goals.

Table 2.1.: Model Requirements

Papers	$q > n$	Hetero	Sparsity	A rank	Gaussian	Theorem Objectives
Caner and Kock (2018)	x	Y	flexible		N	l_1 loss upper bound
Caner et al. (2018)		Y	exact	Full	N	model selection consistency
Shi (2016)		Y	approximate		N	l_1 loss upper bound
Lin et al. (2015)	x	Y	exact		Y	l_1 loss upper bound, weak oracle
Zhu (2018)	x	Y	exact	Full	Y	l_1, l_2 losses upper bound
Gold et al. (2020)	x	N	exact	Full	Y	l_1 loss upper bound

Notes: This table contains the modeling requirements. n is the number of observations and q is the number of instruments. The first column lists the papers. The second column is marked if the method is valid under $q > n$. "Hetero" column is "Y" when the method allows for second-stage conditional heteroskedasticity. "A rank" column is "Full" if the method requires the first-stage coefficient matrix to be full rank. "Guassian" is "Y" if the method requires sub-gaussian distribution assumptions. "Sparsity" and "Specific Theorem Objectives" columns list the sparsity assumptions and the main theoretical results for each paper, respectively.

Data Dimensions

The first assumption is with respect to the relationship between n and q, where Caner et al. (2018) and Shi (2016) require $n > q$ and p. The other four papers allow the numbers of regressors (p) and instruments (q) to be larger than the number of observations (n).

Conditional Heteroskedasticity

The three GMM-based methods allow for conditional heteroskedasticity given the weighting matrix. Zhu (2018) and Caner et al. (2018) focus on the consistency of $\hat{\beta}$, and their methods are also robust to conditional heteroskedasticity. Gold et al. (2020) derive the inference analysis for 2SLS-shirinkage estimators, and the asymptotic theory is constructed under conditional homoskedasticity. Note that this paper focuses on consistency, thus the simulation results present in the next section should not vary much between homoskedasticity and heteroskedasticity.

Sparsity of A and β

A model is considered of "exact sparsity" if many components of these vectors are exactly zero. Alternatively, the model is considered of "approximate sparsity" if some of the components are small but not exact zero.[1] Four out of the six papers assume exact sparsity. The two exceptions are Shi (2016), which allows approximate sparsity, and Caner and Kock (2018), which do not have sparsity requirements. In Remark 4.1 of Gold et al. (2020), the authors note that the exact sparsity requirements in their paper may be relaxed and more complicated theoretical estimation error bounds can be derived.

Rank Conditions

Caner et al. (2018), Zhu (2018), and Gold et al. (2020) use 2SLS-shrinkage, and assume A and the covariance matrix of X and Z to be full rank. The other three papers instead make assumptions regarding the population restricted eigenvalue.

Data Distributions

The last assumption is about the distributions of the regressors and errors, where sub-gaussian distributions are often chosen to facilitate the derivations of the estimation error's upper bounds. Caner and Kock (2018), Caner et al. (2018), and Shi (2016) do not require specific distributions on X, Z, u, or e. Instead, these papers impose stricter assumptions on moment conditions and tuning parameters than the other three papers.

[1] The definition of sparsity is taken from Shi (2016), Assumption 2.

Main Theoretical Results

Finally, Table 2.1 reports the main theoretical results of each paper. Two out of the six papers discuss the model selection consistency. Specifically, Lin et al. (2015) prove the weak oracle property, and Caner et al. (2018) show the model selection consistency derived using the tuning parameters selected by the BIC-type information criterion. The other four papers emphasize the l_1 estimation losses of $\hat{\beta}$, and Zhu (2018) reports the upper bound of $||\hat{\beta} - \beta||_2$ as well as the upper bound of $||\hat{\beta} - \beta||_1$.

2.3 Monte Carlo Simulations

2.3.1 Simulation Designs

The data generation processes (DGPs) considered in this paper follow the modeling structures $Y = X\beta + u$ and $X = ZA + e$. The parameter choices are selected to reflect different modeling assumptions in Table 2.1. Based on these assumptions, Table 2.2 contains the DGP designs, which mainly follow Zhu (2018) and Caner and Kock (2018).

There are two main aspects of these DGP designs. The first is the relative magnitudes between n and q, and two situations are considered in this paper: large instruments ($q > n$), and standard asymptotics ($n > q$). Shi (2016) and Caner et al. (2018) require q and p to be smaller than n, although p and q can both be large.

The second aspect of the DGP designs relates to the first-stage coefficient matrix A. Given that the sparsity requirements are explicitly discussed in each of the six papers, the simulations in this paper focus on exact sparsity to save computation time. Moreover, the matrix A can be full column rank or rank deficient. The latter case is considered for the situation where researchers do not screen the instrument set in advance and include useless instruments. Lastly, the element values of A determine the relative strength of the instruments. To distinguish between these settings, I consider four cases whose details are listed in the second half of Table 2.2.

The instruments Z are generated following the standard normal distribution. The level of endogeneity, which is characterized by the correlation coefficient between u and $\{e_j\}_{j=1}^{p}$, is set to be 0.5. Assuming homeskedasticity,

$$(u,\ e_j) \sim_{iid}\ N\left(\begin{pmatrix} 0 \\ 0 \\ \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \sigma_u^2 & \rho\sigma_u\sigma_e & ... & ... & \rho\sigma_u\sigma_e \\ \rho\sigma_u\sigma_e & \sigma_e^2 & 0 & ... & 0 \\ & 0 & \sigma_e^2 & ... & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \rho\sigma_u\sigma_e & 0 & ... & 0 & \sigma_e^2 \end{matrix}\right)$$

with $\sigma_u = \sigma_e = 1$ and $\rho = 0.5$. The error terms under conditional heteroskedasticity are generated as $u_i^{hetero} = \eta_i u_i ||Z_i||_2/\sqrt{q}$, where $\eta_i \sim U[0.5,\ 2]$, $i = 1,\ ...,\ n$. I conduct 500 Monte Carlo simulations for each DGP design. The parameter choices are also made due to computational limitations.

I mainly use two R packages for the shrinkage estimations: "glmnet" and "ncpen". The results are trimmed to exclude outliers in order to avoid non-convergent estimations. Specifically, I calculate the distances between the estimates with their medians, and compare the values with six times the interquartile ranges. For the methods proposed by Caner and Kock (2018) and Gold et al. (2020), I report the consistency results without the bias-corrections, which are essentially the Lasso estimators with tuning parameters selected by cross-validation.[2]

2.3.2 Simulation Results

Tables 2.3A, 2.3B, 2.4A, and 2.4B report the bias, l_1-loss, and l_2-loss for consistency evaluations. Each column contains the estimation results for one estimation method. The combination of the first and the third row represents whether 2SLS or GMM is used, and how the tuning parameters are selected. A new notation, "CV", is introduced, which stands for cross-validation to select tuning parame-

[2] I focus on the results without bias correction because the consistency theories in these papers are derived for the non-bias-corrected estimators. Five-fold cross-validation is applied when estimating the Caner and Kock (2018), Gold et al. (2020), and Lin et al. (2015) methods.

ters. With these abbreviations, "GMM-Lasso+AIC" means that Shi (2016) uses the GMM-shrinkage method with the Lasso-type penalty, and uses AIC to select tuning parameters; "2SLS-Lasso+CV" means that the Gold et al. (2020) estimator uses cross-validation to select tuning parameters in a 2SLS framework.

Several patterns emerge from the results. First, GMM-shrinkage provides smaller estimation losses than 2SLS-shrinkage when $n > q$, especially for the methods proposed by Shi (2016) and Caner et al. (2018). This observation is most distinctive when A is full rank and the instruments are strong. When $q > n$, Zhu (2018) and "2SLS-Lasso+CV" provide the smallest l_1 losses as well as MSEs across the GMM and 2SLS methods, while Caner et al. (2018) and Shi (2016) are not applicable. The two procedures also preform the best among all 2SLS-shirinkage methods in most cases when $n > q$.

Second, the advantages of using GMM-shrinkage remain under heteroskedasticity. In Table 2.3B Case D, the estimation losses of "GMM-Lasso+AIC" are about half of those produced by "2SLS-Lasso+Zhu". In contrast, differences between the estimation losses are less distinctive under homoskedasticity, with the MSE of "GMM-Lasso+AIC" being slightly smaller in Table 2.3A Case D. Moreover, Gold et al. (2020) produce larger estimation errors than Caner and Kock (2018) in Tables 2.3B and 2.4B Cases C when $\alpha_j = 0.5$, which is consistent with the simulation results presented in Caner and Kock (2018). This relation is reversed when $\alpha_j = 5$, indicating that GMM-shrinkage is sensitive to weak instruments.

Third, the consistency accuracy is generally worse when the instruments are weaker. This is shown by comparing the losses in Case A (C) with their counterparts in Case B (D). Relatedly, the losses in Case B are larger than the ones in Case D, indicating that $\hat{X}$ needs to be full rank and the instruments should be correlated with at least one endogenous regressor.

Fourth, tuning parameters can influence the results significantly. Generally speaking, cross-validation is the preferred method, although the GMM-AIC and BIC criteria proposed by Shi (2016) and Caner et al. (2018), respectively, perform better when $n > q$.

Lastly, one would expect that the consistency performance deteriorates for smaller sample sizes. Even though this paper does not investigate the simulation performance for larger sample sizes given computational constraints,[3] some evidence can be found in Table 2.1 of Caner and Kock (2018), where the size of "D2GMM" is 17% when $n = 75$, $q = 100$ and the size drops to 0% when $n = 150$, $q = 200$.

2.3.3 An Empirical Illustration

In this section, I apply the methods proposed by Shi (2016), Caner and Kock (2018), and Caner et al. (2018) to analyze how foreign aid affects GDP growth. León-González and Montolio (2015) show that the magnitude of the foreign aid coefficient is unstable across different groups of regressors and instruments using GMM. Instead, they propose a Bayesian model averaging method and demonstrate that the foreign aid coefficient is small. Moreover, the authors find that the interaction term between the "good policy index" proposed by Burnside and Dollar (2000) and foreign aid also has no major impact on GDP growth either. This result differs from the policy implications in previous literature, which suggest that foreign aid simulates GDP growth for countries which have "good" macroeconomic polices.

I use the same unbalanced panel dataset as León-González and Montolio (2015) and apply the three GMM-shrinkage methods studied in this paper. There are 291 observations, and the dataset covers 63 countries with from 1974-1994. The dependent variable is the GDP growth rate, and the independent variables of interest are foreign aid and foreign aid interacted with the good policy index. The instruments are lags of initial GDP, lags of foreign aid amount, lags of the natural log of population, lags of armed imports, and lags of various interactions involving the policy index. The details of the instruments and regressors are listed in León-González and Montolio

[3] In this paper, I set $n = 75$ when $q > n$ to save computational cost. As reference, the numbers of observations are set to be at least 100 in five of the six papers.

(2015), Appendix A. I also conduct wild bootstrapping to test the significance of the estimates.[4]

Table 2.5 reports the estimation results. The number of lags used as instruments is one, two, three, or four, which results in 54, 99, 135, or 162 instruments. There are 16 regressors for all regressions reported in the table. The first half of Table 2.5 reports that the effect of foreign aid on growth using Shi (2016) and Caner and Kock (2018) is much smaller than the GMM estimates presented in León-González and Montolio (2015), and is closer to their Bayesian model averaging results. The method proposed by Caner et al. (2018) produces a larger effect, although the magnitude is unstable across different instrument sets. All of the coefficient estimates are insignificant at the 1%, 5%, or 10% level.

Additionally, the coefficient of the policy interaction is nearly zero and insignificant for all model settings, as reported in the second half of Table 2.5. This result supports León-González and Montolio (2015) where they show that the policy interaction has a nearly zero posterior probability of entering in the regression.

In summary, these findings highlight the over-identification problem of the GMM estimators, which is mentioned in León-González and Montolio (2015) and confirmed by their diagnostic tests. The results also show the usefulness of the GMM-shrinkage methods, which introduce parsimony to the model.

2.3.4 Discussion

How will researchers select the most suitable methods? One intuitive direction is to refer to the assumptions of each paper. First, the sample dimension should satisfy the $(n,\ p,\ q)$ requirements. Next, GMM-shrinkage methods are the natural choices with dynamic panel regressions. If heteroskedasticity is a component for the

[4] To conduct the bootstrapping to test a coefficient's significance, I first collect the residuals from the original shrinkage estimation results. Then, I resample these residuals and multiply each element with a randomly generated standard normal variable. Next, the bootstrapped dependent variable is generated under the null hypothesis that the coefficient is zero, and shrinkage methods are applied to the bootstrapped sample. Finally, the original estimates are compared with the $\alpha/2$ and $(1-\alpha/2)$ percentiles of their corresponding bootstrapped coefficients.

second-stage regression, GMM-shrinkage is also advantageous given the choice of the weighting matrix. On the contrary, 2SLS-shrinkage is preferable when researchers are interested in knowing which instruments contribute to each endogenous regressor.[5] Finally, the sparsity requirements of A and β need to be met.

There are two other less obvious but equally important data characteristics. The first is regarding the full rank assumption of $\Sigma_{xz} = E(X'Z)$, indicating that all instruments should correlate with at least one endogenous regressor. It is well documented that GMM and 2SLS estimators are volatile with weak instruments, which is also shown in the above simulation results. In addition, 2SLS-shrinkage procedures maybe even more affected by the possibility that A is rank deficient, since the second-stage estimation requires the the first-stage estimated $\hat{X}$. This negative effect may be severe when weak instruments are used.

The other data assumption made by some of the papers is sub-gaussianality. Heavy tail distributions, such as the log-normal distribution commonly used in the finance literature, are usually not sub-gaussian. Last but not least, the elastic net procedure proposed by Caner et al. (2018) is theoretically preferable when instruments are highly correlated, considering that the elastic net shrinkage encourages a grouping effect.

The next set of decisions researchers face is computational, when converting the selected model to a computer program. The most crucial parameters in any shrinkage method are the tuning parameters, which control how parsimonious the model will be. The 2SLS-shrinkage procedures have more tuning parameters than the GMM-shrinkage procedures, and so GMM-shrinkage estimators are more efficient from this perspective. Researchers can refer to the theoretical results from Shi (2016) and Zhu (2018) and analytically pick out the suitable tuning parameters. However, such choices may still require researchers to assign (other) parameter values, such as the ζ and C in Zhu (2018), to calculate the value of the tuning parameters. These values can play a critical role in finite sample estimations, although theoretically they should

[5] All of the methods are robust to serial correlation for the second-stage regression.

produce the same asymptotic properties with enough iterations. Non-parametrically, cross-validation is the most common choice even though there is no thorough proof of its validity to shrinkage methods. Despite its wide-spread applicability, cross-validation tends to select too many regressors (see a related discussion in Spindler (2016)).

Convergence thresholds can also affect finite sample results. For some non-convex penalties, such as the MCP used in Lin et al. (2015), the convergence to a global minimum is not guaranteed. Different software packages allow researchers to set different convergence criteria. For example, the R package "glmnet" does not allow researchers to set the maximum number of iterations, but they can set the convergence thresholds. This naturally invokes the trade-off between computation time and estimation accuracy.

Moreover, the cutoff value where coefficients can be treated as zeros also requires attention. The common choice is $1e^{-4}$ for most of the existing software programs, and this concern is specific to the 2SLS-shrinkage methods. On the one hand, it is desirable to keep only a few selected instrument for the first-stage shrinkage estimations. On the other hand, if too few of the original instruments enter the second stage, $\hat{X} = Z\hat{A}$ may not be full column rank. This conclusion is in accordance with the consistency results reported in Tables 2.3A and 2.3B, where Lin et al. (2015) and Zhu (2018) provide larger estimation losses when the first-stage coefficient matrix is rank deficient. These all encourage empirical researchers to select a strong and relevant instrument set before using any shrinkage estimation methods.

2.4 Conclusion

In this paper, I compare six shrinkage estimation methods with large-dimensional regressors and instruments through Monte Carlo simulations. The results show that combining shrinkage methods with generalized method of moments offers smaller estimation errors, as compared to combining shrinkage methods with two-stage least squares. The difference is especially large when the number of observation is greater than the number of instruments, when the instruments are strong, and when at least some of the instruments are plausibly correlated with the regressors. The results also support using cross-validation to select tuning parameters. Lastly, I apply a GMM-shrinkage method to reestimate an empirical macroeconomic model, and show that foreign aid has no significant impact on economic growth.

Table 2.2.: DGP Designs

I	II		
$(n,\ p,\ q)$	#of "nonzero"$(\beta,\ \alpha_j)$	$(\beta,\ \alpha_j)$ value	exact sparsity
$(100,\ 20,\ 40)$	Case A: $(2,\ 2)$	$(0.5,\ 0.5)$, "weak instrument",, $j=1,\ 2$, A rank deficient	$A=\begin{bmatrix} 0.5 & 0.5 & \dots & 0.5 & 0.5 \\ 0.5 & 0.5 & \dots & 0.5 & 0.5 \\ 0 & \dots & \dots & 0 & 0 \\ & & \vdots & & \\ 0 & \dots & \dots & 0 & 0 \end{bmatrix}$
$(75,\ 50,\ 100)$	Case B: $(2,\ 2)$	$(0.5,\ 5)$, $j=1,\ 2$, A rank deficient	$A=\begin{bmatrix} 5 & 5 & \dots & 5 & 5 \\ 5 & 5 & \dots & 5 & 5 \\ 0 & \dots & \dots & 0 & 0 \\ & & \vdots & & \\ 0 & \dots & \dots & 0 & 0 \end{bmatrix}$
	Case C: $(2,\ 2)$	$(0.5,\ 0.5)$, "weak instrument", A full column rank	$A=\begin{bmatrix} 0.5 & 0.5 & \dots & 0 & 0 \\ 0 & 0.5 & 0.5 & \dots & 0 \\ 0 & \dots & \dots & 0.5 & 0.5 \\ 0.5 & 0.5 & \dots & 0 & 0 \\ 0 & 0.5 & 0.5 & \dots & 0 \\ 0 & \dots & \dots & 0.5 & 0.5 \\ & & \vdots & & \end{bmatrix}$
	Case D: $(2,\ 2)$	$(0.5,\ 5)$, $j=1,\ 2$ A full column rank	$A=\begin{bmatrix} 5 & 5 & \dots & 0 & 0 \\ 0 & 5 & 5 & \dots & 0 \\ 0 & \dots & \dots & 5 & 5 \\ 5 & 5 & \dots & 0 & 0 \\ 0 & 5 & 5 & \dots & 0 \\ 0 & \dots & \dots & 5 & 5 \\ & & \vdots & & \end{bmatrix}$

This table contains the main DGP designs. Panel I reports the $(n,\ p,\ q)$ choices, where n is the number of observations, p is the number of regressors, and q is the number of instruments. Panel II reports the choices of coefficients. A is the first-stage coefficient matrix, and β is the second-stage coefficient vector. The first column contains the number of nonzero elements in β and A. The second column contains the values assigned to the nonzero elements. The last column lists the detailed composition of matrix A.

Table 2.3A.: Consistency, $(n,\ p,\ q)$=(100, 20, 40)

	GMM			2SLS		
	Shi (2016)	Caner et al. (2018)	Caner and Kock (2018)	Lin et al. (2015)	Zhu (2018)	Gold et al. (2020)
Case A	Lasso+AIC	Elastic Net+BIC	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	-0.0049	-0.02853	-0.00107	-0.0106	0.00412	0.01621
l1 loss	0.34373	0.66045	0.32096	1.56033	1.76233	0.9748
MSE	0.0042	0.01405	0.00315	0.0542	0.0554	0.03116
Case B	Lasso+AIC	Elastic Net+BIC	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	-0.00148	0.00017	-0.00953	0.00097	-6e-05	0.0015
l1 loss	0.176	0.32257	0.4906	1.97521	1.65951	0.82172
MSE	0.00095	0.00351	0.00828	0.07708	0.04786	0.01976
Case C	Lasso+AIC	Elastic Net+BIC	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	-0.00251	-0.02766	0.01837	-0.02398	-0.03212	-0.02856
l1 loss	0.32855	0.59185	0.85971	0.68932	0.77235	0.61976
MSE	0.00223	0.01249	0.00593	0.0136	0.01662	0.01308
Case D	Lasso+AIC	Elastic Net+BIC	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	-0.00211	-0.00319	-0.01216	0.00104	-0.00218	0.00235
l1 loss	0.04381	0.06427	0.24313	0.18833	0.05676	0.35665
MSE	7e-05	0.00013	0.00174	0.00032	0.0001	0.00053

Notes: This table presents the consistency results when $n > q$ under second-stage conditional homoskedasticity. The first three rows of this table describe the estimation method, how the tuning parameters are selected, and the corresponding paper. "CV" stands for cross-validation. Zhu (2018) uses theoretically derived tuning parameters.

Table 2.3B.: Consistency, $(n,\ p,\ q)$=(100, 20, 40)

	GMM			2SLS		
	Shi (2016)	Caner et al. (2018)	Caner and Kock (2018)	Lin et al. (2015)	Zhu (2018)	Gold et al. (2020)
Case A	Lasso+AIC	Elastic Net+BIC	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	0.0043	-0.0224	-0.00383	-0.01261	-0.00104	0.00268
l1 loss	0.78354	0.53555	0.47325	1.56831	1.72325	0.78023
MSE	0.00916	0.011	0.00632	0.05603	0.05578	0.02095
Case B	Lasso+AIC	Elastic Net+BIC	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	-0.0002	-0.00058	-0.00837	0.00107	-0.00053	0.00161
l1 loss	0.80991	0.4057	0.516	2.22809	1.73665	0.9279
MSE	0.01048	0.0071	0.00865	0.10959	0.05147	0.02185
Case C	Lasso+AIC	Elastic Net+BIC	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	0.02618	-0.03796	0.01981	-0.03397	-0.03737	-0.03259
l1 loss	1.18173	0.76344	1.1527	0.85868	0.8712	0.77154
MSE	0.01071	0.01777	0.01062	0.01899	0.02003	0.01768
Case D	Lasso+AIC	Elastic Net+BIC	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	-0.00141	-0.00336	-0.01222	0.00083	-0.00333	-0.00162
l1 loss	0.07227	0.06765	0.24443	0.22826	0.0788	0.05864
MSE	9e-05	0.00016	0.00174	0.00045	0.0002	0.00015

Notes: This table presents the consistency results when $n > q$ under second-stage conditional heteroskedasticity. The first three rows of this table describe the estimation method, how the tuning parameters are selected, and the corresponding paper. “CV” stands for cross-validation. Zhu (2018) uses theoretically derived tuning parameters.

Table 2.4A.: Consistency, $(n,\ p,\ q)$=(75, 50, 100)

	GMM	2SLS		
	Caner and Kock (2018)	Lin et al. (2015)	Zhu (2018)	Gold et al. (2020)
Case A	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	0.00594	0.00686	0.00123	0.00287
l1 loss	6.72049	8.71486	1.88005	0.56494
MSE	0.04148	0.0503	0.02499	0.00398
Case B	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	-0.00154	0.00012	-0.00015	0.00138
l1 loss	0.58501	3.76987	1.82068	1.02852
MSE	0.00339	0.02583	0.02038	0.01039
Case C	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	0.02161	0.01847	-0.01728	-0.01518
l1 loss	5.38486	6.29962	0.96133	0.80384
MSE	0.02533	0.02607	0.00954	0.00762
Case D	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	-0.00314	0.00338	-0.0023	-0.00043
l1 loss	0.15717	1.0036	0.11907	0.07015
MSE	0.00032	0.0007	0.00019	8e-05

Notes: This table presents the consistency results when $q > n$ under second-stage conditional homoskedasticity. The first three rows of this table describe the estimation method, how the tuning parameters are selected, and the corresponding paper. “CV” stands for cross-validation. Zhu (2018) uses theoretically derived tuning parameters.

Table 2.4B.: Consistency, heteroskedasticity, $(n,\ p,\ q)$=(75, 50, 100)

	GMM	2SLS		
	Caner and Kock (2018)	Lin et al. (2015)	Zhu (2018)	Gold et al. (2020)
Case A	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	0.00599	-0.0055	-0.00392	0.00576
l1 loss	6.85032	1.58081	1.68267	1.00515
MSE	0.04274	0.02117	0.02236	0.01522
Case B	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	-0.0016	0.00045	-0.00041	0.00177
l1 loss	0.59517	3.29084	1.82248	1.04233
MSE	0.00354	0.09285	0.02077	0.0104
Case C	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	0.02163	-0.01303	-0.01812	-0.01659
l1 loss	5.45186	0.99136	0.98575	0.86888
MSE	0.02581	0.00907	0.00995	0.0084
Case D	Lasso+CV	MCP+CV	Lasso+Zhu	Lasso+CV
Bias	-0.00314	0.00051	-0.00292	-0.0006
l1 loss	0.15741	0.87535	0.14838	0.07533
MSE	0.00031	0.00112	0.00028	8e-05

Notes: This table presents the consistency results when $q > n$ under second-stage conditional heteroskedasticity. The first three rows of this table describe the estimation method, how the tuning parameters are selected, and the corresponding paper. "CV" stands for cross-validation. Zhu (2018) uses theoretically derived tuning parameters.

Table 2.5.: Foreign Aid to GDP Growth

	Methods	# of lags	1	2	3	4
$\hat{\beta}_{aid}$	Shi (2016)		0.106	-0.083	-0.113	-0.113
	Caner and Kock (2019)		0.296	-	-	0.005
	Caner et al. (2018)		-3.754	-0.13	-3.85	-1.39
$\hat{\beta}_{polaid}$	Shi (2016)		-	0.052	-0.011	-0.021
	Caner and Kock (2019)		0.019	-	0.042	0.002
	Caner et al. (2018)		-0.081	-0.007	-0.019	0.068
	# of instrument		54	99	135	162

Notes: $\hat{\beta}_{aid}$ and $\hat{\beta}_{polaid}$ are the estimated coefficients for foreign aid and foreign aid interacts with the good policy index. "-" means the shrinkage procedure does not select the variable. The coefficients are marked with "*", "**", or "***" if they are significant with 10%, 5%, or 1% significance level. Otherwise the coefficient is not significant. The bootstrapping iteration number is 399.

3. REVISITING THE DEMOCRACY-GROWTH NEXUS: NEW EVIDENCE FROM A DYNAMIC COMMON CORRELATED EFFECTS APPROACH

with Mohitosh Kejriwal

3.1 Introduction

The question of whether democracy is beneficial for economic growth has spurred a large theoretical and empirical literature over the past five decades (see, e.g., Doucouliagos and Ulubaşoğlu, 2008, for a review). While proponents of democracy argue that political rights and civil liberties are necessary to preserve the motivation of citizens to work and invest while maintaining an effective allocation of resources in the marketplace, opponents promote the view that democracies are vulnerable to popular demands at the expense of profitable investments and are unable to suppress ethnic, religious and class conflicts that are detrimental to growth. There is also a third so-called "skeptic view" that points to the importance of the institutional structure in facilitating growth rather than the regimes per se. The literature does not appear to have reached a consensus yet among these different views.

In an influential recent article, (Acemoglu et al., 2019, henceforth ANRR) take a major step forward by empirically examining the effect of democracy on economic growth based on a new comprehensive panel dataset covering 175 countries over the period 1960-2010. Their analysis employs standard dynamic panel data estimation methods such as within groups and Arellano-Bond GMM as well as the more recent bias-correction approach proposed by Hahn et al. (2001). These methods assume that the model parameters are homogeneous across countries and rule out strong cross section dependence among the countries.[1] The baseline estimates reported in ANRR suggest that democracy has a positive and statistically significant effect on economic growth, with GDP per capita being approximately 20% higher in the 25 years following a permanent democratization.

This paper reconsiders the nexus between democracy and growth using a recently proposed econometric approach that allows for both parameter heterogeneity and strong cross section dependence. Parameter heterogeneity can arise from economic, cultural and political institutional differences across countries. As shown in Pesaran and Smith (1995), pooled estimators are biased in a dynamic model with random coefficients. On the other hand, strong cross section dependence can emanate from common global shocks that affect different countries to different degrees. This notion of dependence is distinct from spatially correlated shocks that essentially capture weak dependence (see Chudik et al., 2011). Chudik and Pesaran (2015) demonstrate through Monte Carlo experiments the serious biases associated with the within groups estimator in the presence of strong cross section dependence. Our analysis employs the dynamic common correlated effects (DCCE) approach proposed by Chudik and Pesaran (2015) that models the cross section dependence in terms of a small (unknown) number of unobserved common factors with heterogeneous loadings. The DCCE estimator is computed by augmenting the dynamic panel regression with cross-sectional averages of the current and lagged values of the dependent variable and regressors. Our findings confirm the statistically significant positive effect of democracy on growth documented in ANRR, with a DCCE point estimate between approximately 1.5-2%, depending on the specification. We complement our estimates with a battery of diagnostic tests for heterogeneity and cross-section dependence that corroborate the use of the DCCE approach in evaluating the effect of democracy on growth.

[1] An exception is when the dependence is not country-specific in which case a specification that includes time fixed effects (as in ANRR) is sufficient to address the issue.

The rest of the paper is organized as follows. Section 3.2 lays out the econometric framework and the DCCE estimation procedure. Section 3.3 presents the empirical results. Section 3.4 concludes. Appendix A reports results from a set of diagnostic tests for parameter heterogeneity and cross-section dependence as well as estimates of the degree of cross section dependence using the approach proposed by Bailey et al. (2016).

3.2 The Dynamic Common Correlated Effects (DCCE) Approach

Consider the dynamic panel data model

$$\begin{aligned} y_{it} &= \alpha_i + \sum_{j=1}^{p} \gamma_{ij} y_{t-j} + \beta_i D_{it} + u_{it} && (3.1) \\ u_{it} &= \lambda_i' f_t + \varepsilon_{it} && (3.2) \end{aligned}$$

for $i = 1, ..., N$ and $t = p+1, ..., T$, where y_{it} is the log of GDP per capita (or the growth rate) in country i at time t and D_{it} is a dummy variable which equals unity if country i is democratic at period t and zero otherwise. The α_i denote the country fixed effects representing the time-invariant country characteristics. The error u_{it} is composed of a common component $\lambda_i' f_t$ and an idiosyncratic component ε_{it}. Here, f_t represents a $(m \times 1)$ vector of unobserved common factors and λ_i denotes a $(m \times 1)$ vector of associated factor loadings. The number of factors m is assumed unknown. The factors are allowed to be correlated with the dichotomous democracy measure. The traditional dynamic panel framework adopted by ANRR can be obtained as a special case of (3.1) and (3.2) by setting $\lambda_i = \lambda$, $\beta_i = \beta$ for all i and $\gamma_{ij} = \gamma_j$ for all i and $j = 1, ..., p$.

Chudik and Pesaran (2015) consider consistent estimation of the means of the parameters in (3.1). They propose proxying for the common factors by augmenting the regression (3.1) with cross-sectional averages of y_{it} and D_{it}:

$$y_{it} = \alpha_i + \sum_{j=1}^{p} \gamma_{ij} y_{t-j} + \beta_i D_{it} + \sum_{l=0}^{q_T} \delta_i'(L) \bar{z}_{t-l} + e_{it} \qquad (3.3)$$

where $\bar{z}_t = N^{-1}\sum_{i=1}^{N} z_{it},\ z_{it} = (y_{it}, D_{it})'$ and q_T, the number of lags of cross-sectional averages included, is assumed to grow with the sample size at a particular rate: $q_T \to \infty$ and $q_T^3/T \to \kappa$, with $0 < \kappa < \infty$.[2]

Denote $\pi_i = (\gamma_{i1}, ..., \gamma_{ip}, \beta_i)'$. The common correlated effects mean group (CCEMG) estimator of $\pi = E(\pi_i)$ is given by

$$\hat{\pi} = N^{-1}\sum_{i=1}^{N}\hat{\pi}_i$$

where $\hat{\pi}_i$ is the ordinary least squares estimate of π_i from (3). Chudik and Pesaran (2015) establish the consistency of $\hat{\pi}$ under two alternative sets of assumptions. The first set consists of a rank condition on the matrix of factor loadings which, in the current context, requires that the number of factors $m \leq 2$. The second set does not require the rank condition but assumes that the factors are serially uncorrelated. In both cases, $\hat{\pi}$ is shown to be $\sqrt{N}$-consistent and its asymptotic variance can be estimated by

$$\hat{\sum} = (N-1)^{-1}\sum_{i=1}^{N}(\hat{\pi}_i - \hat{\pi})(\hat{\pi}_i - \hat{\pi})'$$

In order to correct the small sample bias of $\hat{\pi}$, a "half-panel jackknife" procedure is adopted in which the bias-corrected estimator is obtained as

$$\tilde{\pi} = 2\hat{\pi} - 0.5(\hat{\pi}_a + \hat{\pi}_b)$$

where $\hat{\pi}_a$ denotes the CCEMG estimator computed over the period $t = 1, ..., [T/2]$, and $\hat{\pi}_b$ is the CCEMG estimator computed over the period $t = [T/2]+1, ..., T$. Based on Monte Carlo experiments, Chudik and Pesaran (2015) propose using the jackknife bias corrected estimates for the coefficients of the lagged dependent variable while the uncorrected estimate is preferred for the coefficient on democracy.[3] As per their recommendation, we set $q_T = [T^{1/3}]$.

[2] While the theoretical analysis in Chudik and Pesaran (2015) allows weighted cross-sectional averages, their Monte Carlo experiments are based on simple averages.

[3] The authors also consider bias correction based on recursive mean adjustment which is, however, dominated by the jackknife.

3.3 Empirical Results

Our empirical analysis is based on a balanced sample of countries appearing in the dataset compiled by ANRR.[4] Each country in our sample has experienced a change in democratic status at least once. The reason for concentrating on this subsample is that the CCEMG estimator is based on country-wise time series regressions so that if a country's democratic status remains unchanged over the sample period, it cannot be separately identified from the country-specific effect α_i. This constraint combined with the focus on a balanced sample led us to a set of 41 countries over the period 1975-2010.[5] ANRR report results based on three estimators: the fixed effects or within groups (WG) estimator, the Arellano-Bond GMM (AB) estimator and the Hahn, Hausman and Kuersteiner (HHK) bias-corrected instrumental variables estimator. They also present estimates of the long run effect of democracy and the effect after 25 years (say the medium run effect) for each of the estimators. ANRR consider four choices of the lag order p: 1,2,4,8. Since the DCCE approach is based on country-specific time series regressions, we only consider $p = 1, 2, 4$ out of a degrees of freedom consideration.

Table 3.1 presents our findings based on the three estimators considered by ANRR where Panel A reports the results for GDP measured in levels while Panel B refers to GDP growth. Considering first the estimates in Panel A, the effect of democracy is smaller for a given estimator and lag order, relative to the original ANRR estimates. For instance, with $p = 4$, the WG estimate is about .48% while the corresponding estimate in ANRR is .78%. The medium and long run effects are also smaller. These differences reflect the fact that our analysis is based on a smaller balanced sample.

[4] The focus on a balanced sample is due to the fact that the DCCE estimator is derived assuming a balanced sample and its statistical properties are known in this case. To the best of our knowledge, the corresponding properties in the unbalanced case are yet unknown.

[5] The countries are: Argentina, Burundi, Benin, Bangladesh, Bolivia, Brazil, Central African Rep., Chile, Dominican Republic, Ecuador, Spain, Ghana, Gambia, Guatemala, Honduras, Hungary, Indonesia, Kenya, Lesotho, Madagascar, Mexico, Mali, Mauritania, Malawi, Niger, Nigeria, Nicaragua, Nepal, Pakistan, Peru, Philippines, Portugal, Sudan, Senegal, Sierra Leone, Thailand, Turkey, Uruguay, Venezuela, South Africa, Zambia.

Among the three estimators, for a given lag specification, the immediate effect on democracy as well as the medium and long run effects are largest for WG and smallest for HHK based on the current dataset. Given the possibility of a unit root for the data in levels as indicated by the high persistence estimates, Panel B presents the results based on the growth rate of GDP. Again, the effect of democracy is smaller than the original ANRR estimates. This is true for the current period effect as well as the medium and long run effects. Relative to the results in levels, the parameter estimates are more similar between the three estimation methods. The differences are particularly small between the WG and AB estimates, with the instantaneous effects ranging between 63% and .69% regardless of the number of lags used. Finally, as expected, the estimates of the persistence parameter are much lower (¡.15 in all cases) than the corresponding estimates from the specification in levels.

Table 3.2 presents results obtained from two estimation methods: (1) the mean group (MG) estimate that is obtained by taking the average of the country-specific effects from least squares time series regressions estimated separately for each country; (2) the DCCE estimate that accounts for both parameter heterogeneity and cross section dependence. The role of the MG estimate is to isolate the impact of parameter heterogeneity from that of cross section dependence. The MG estimates are all bias-uncorrected. The standard errors for the MG and DCCE estimates are computed nonparametrically based on the standard deviation of the country-specific estimates. The medium and long run effects are based on bias-uncorrected estimates.

The results for GDP in levels are reported in Panel A of Table 3.2. The MG estimates of the current period effect of democracy are considerably larger than those reported in Table 3.1. The lowest MG estimate across the three specifications is about 1.71% while the highest among the homogeneous estimators is about .58%. Further, the MG estimates are all statistically significant at the 1% level. The medium and long run effects are also markedly larger across the three lag orders relative to those reported in Table 3.1.[6] For example, the estimated medium run effects range

[6] To compute the medium and long run effects using the MG estimator, we eliminated three countries Pakistan, Sierra Leone and Sudan as the country-specific effects in these cases were implausibly large

between 0.9%-8.6% for the homogeneous estimators while the MG estimates are all between 22-25%. These results indicate that parameter heterogeneity can have a substantial impact on the estimated effect of democracy. Turning to the DCCE estimates, the current period effects are between 1.5%-2% and broadly comparable to the corresponding MG estimates although for $p = 2$, the MG effect is somewhat more pronounced (∼1.87%) than the DCCE effect (∼1.51%). The magnitude of the DCCE medium and long run effects are strongly dependent on the lag order employed with the estimated effects being much larger (∼38%) when $p = 1$. When $p = 2$ or 4, the WG estimates from Table 3.1 are larger than the corresponding DCCE estimates while the opposite is true when $p = 1$.

Moving to results for the GDP growth rate presented in Panel B of Table 3.2, we find that the MG estimates are larger than the homogeneous estimates in Table 3.1, echoing the findings obtained from the data in levels. While the DCCE estimates are comparable in magnitude to the MG estimates when $p < 4$, the DCCE estimates are more notable with $p = 4$. The current period DCCE effects lie between 1.46%-1.65% while the corresponding range for the MG estimate is 1.42%-1.59%. In contrast to the results in levels, the DCCE estimates based on the growth rate are much less sensitive to the number of lags of the dependent variable used in the estimation.

In summary, the foregoing results suggest that parameter heterogeneity and cross section dependence can have important implications for the impact of democracy on economic growth and the DCCE framework appears to provide a useful extension to traditional dynamic panel approaches that can be used to quantify the influence of these features when evaluating the economic consequences of democratization.

3.4 Conclusion

This paper investigates the robustness of the democracy-growth relationship using an econometric approach that accounts for the twin features of parameter heterogene-

and negative which dominated the average based on all countries. The DCCE estimates are, however, computed using all countries.

ity and cross section dependence. The estimates show that the finding of a positive and statistically significant effect of democracy on growth is robust to the presence of these features. It is important to stress, however, that our findings are specific to the dataset under consideration and do not necessarily generalize to countries outside our sample.

Table 3.1.: The effect of democracy on GDP [WG, AB & HHK estimates]

Panel A				[GDP in levels]					
	WG, lag1	WG, lag2	WG, lag4	AB, lag1	AB, lag2	AB, lag4	HHK, lag1	HHK, lag2	HHK, lag4
Democracy	0.584	0.488	0.482	0.259	0.200	0.185	0.259	0.135	0.083
	(0.391)	(0.360)	(0.365)	(0.426)	(0.391)	(0.400)	(0.409)	(0.382)	(0.383)
Longrun effect	11.935	9.589	9.509	4.095	3.305	3.059	3.442	1.735	1.009
	(8.569)	(7.484)	(7.606)	(6.881)	(6.535)	(6.710)	(5.663)	(5.007)	(4.688)
Effect of democracy after 25 years	8.528	7.496	7.383	3.296	2.768	2.525	2.956	1.541	0.917
	(5.877)	(5.670)	(5.761)	(5.482)	(5.438)	(5.507)	(4.790)	(4.428)	(4.256)
Persistence of GDP process	0.951***	0.949***	0.949***	0.937***	0.939***	0.940***	0.925***	0.922***	0.918***
	(0.008)	(0.008)	(0.007)	(0.010)	(0.009)	(0.009)	(0.013)	(0.013)	(0.014)
Panel B				[GDP in growth]					
	WG, lag1	WG, lag2	WG, lag4	AB, lag1	AB, lag2	AB, lag4	HHK, lag1	HHK, lag2	HHK, lag4
Democracy	0.687*	0.674*	0.638*	0.691*	0.673*	0.639	0.552	0.488	0.495
	(0.354)	(0.367)	(0.370)	(0.379)	(0.392)	(0.392)	(0.481)	(0.469)	(0.535)
Longrun effect	0.785*	0.734*	0.665*	0.804*	0.758*	0.685	0.630	0.570	0.580
	(0.406)	(0.408)	(0.393)	(0.442)	(0.452)	(0.432)	(0.562)	(0.558)	(0.660)
Effect of democracy after 25 years	19.508*	18.320*	16.716*	19.963*	18.872*	17.202*	15.650	14.126	14.399
	(10.094)	(10.157)	(9.858)	(10.968)	(11.235)	(10.817)	(13.945)	(13.812)	(16.271)
Persistence of GDP process	0.124***	0.082*	0.040	0.140***	0.111**	0.067	0.122	0.143	0.146
	(0.038)	(0.049)	(0.061)	(0.038)	(0.046)	(0.064)	(0.078)	(0.097)	(0.148)
Observations	1476	1476	1476	1476	1476	1476	1476	1476	1476
Countries in sample	41	41	41	41	41	41	41	41	41

Notes: This table presents estimates of the effect of democracy on GDP per capita in levels and the growth rate of GDP using the balanced sample of 41 countries over 1975-2010. The reported coefficient on democracy is multiplied by 100. All specifications include a full set of country and year fixed effects. The estimators are denoted as: WG = Within groups; AB = Arellan-Bond GMM; HHK = Hahn et al. (2001) bias corrected estimator. We use *, ** and *** to denote significance at the 10%, 5% and 1% level, respectively.

Table 3.2.: The effect of democracy on GDP [Mean Group & DCCE estimates]

Panel A	[GDP in levels]					
	MG, lag1	MG, lag2	MG, lag4	CCE, lag1	CCE, lag2	CCE, lag4
Democracy	1.861***	1.868***	1.715***	1.926**	1.514*	1.693*
	(0.425)	(0.409)	(0.459)	(0.808)	(0.824)	(0.872)
Longrun effect on GDP	23.798	44.965***	110.728*	38.214	5.375	5.992
	(19.819)	(11.715)	(58.092)	(24.761)	(3.500)	(3.990)
Effect of democracy after 25 years	24.816***	23.313***	22.480**	10.185	2.680	2.204
	(8.066)	(8.229)	(9.292)	(7.186)	(3.962)	4.029
Persistence	0.948***	0.935***	0.932***	0.948***	0.898***	1.059***
	(0.012)	(0.012)	(0.014)	(0.062)	(0.075)	(0.122)
Panel B	[GDP in growth]					
	MG, lag1	MG, lag2	MG, lag4	CCE, lag1	CCE, lag2	CCE, lag4
Democracy	1.585***	1.504***	1.428***	1.545**	1.463**	1.642**
	(0.368)	(0.363)	(0.370)	(0.661)	(0.703)	(0.827)
Longrun effect on GDP growth rate	1.897***	1.675***	1.533***	1.805**	1.734**	2.087**
	(0.468)	(0.415)	(0.378)	(0.702)	(0.743)	(0.916)
Effect of democracy after 25 years	46.769***	41.309***	37.198***	44.448**	42.314**	48.109**
	(11.518)	(10.260)	(9.360)	(17.342)	(18.178)	(21.115)
Persistence of growth rate process	0.161***	0.105***	0.076	0.194***	0.181**	0.446***
	(0.041)	(0.049)	(0.064)	(0.065)	(0.091)	(0.163)
Observations	1476	1476	1476	1476	1476	1476
Countries in sample	41	41	41	41	41	41

Notes: This table presents the mean group (MG) and the dynamic common correlated effects (DCCE) estimates of the effect of democracy on log GDP per capita and effect of democracy on log GDP per capita and the growth rate of GDP. The reported coefficient on democracy is multiplied by 100. To compute the long run effects using the MG estimator, we eliminated three countries, Pakistan, Sierra Leone, and Sudan, as the country-specific effects in these cases were implausible. The DCCE estimates are computed using all 41 countries. We use *, ** and *** to denote significance at the 10%, 5% and 1% level, respectively.

BIBLIOGRAPHY

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. (2019). Democracy does cause growth. *Journal of Political Economy*, 127(1):47–100.

Ando, T. and Bai, J. (2015). A simple new test for slope homogeneity in panel data models with interactive effects. *Economics Letters*, 136:112–117.

Angrist, J. D. and Keueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2):277–297.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics*, 122(1):137–183.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2004). A panic attack on unit roots and cointegration. *Econometrica*, 72(4):1127–1177.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317.

Bailey, N., Kapetanios, G., and Pesaran, M. (2016). Exponent of cross-sectional dependence. *Journal of Applied Econometrics*, 31(6):929–960.

Banerjee, A., Marcellino, M., and Masten, I. (2014). Forecasting with factor-augmented error correction models. *International Journal of Forecasting*, 30(3):589–612.

Banerjee, A., Marcellino, M., Masten, I., et al. (2016). An overview of the factor-augmented error-correction model. *Dynamic Factor Models*, 35:3–41.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1):115–143.

Boivin, J. and Ng, S. (2006). Are more data always better for factor analysis? *Journal of Econometrics*, 132(1):169–194.

Burnside, C. and Dollar, D. (2000). Aid, policies, and growth. *American Economic Review*, 90(4):847–868.

Cai, T., Liu, W., and Luo, X. (2011). A constrained $\ell 1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Caner, M. and Kock, A. B. (2018). High dimensional linear gmm. *arXiv preprint arXiv:1811.08779*.

Caner, M., Xu, H., and Lee, Y. (2018). Adaptive elastic net GMM estimation with many invalid moment conditions: Simultaneous model and moment selection. *Journal of Business & Economic Statistics*, 36(1):24–46.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398.

Castagnetti, C., Rossi, E., and Trapani, L. (2015). Inference on factor structures in heterogenenous panels. *Journal of Econometrics*, 184(1):145–157.

Cheng, X. and Hansen, B. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics*, 186(2):280–293.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*.

Christensen, G. and Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–80.

Chudik, A. and Pesaran, M. H. (2015). Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. *Journal of Econometrics*, 188(2):393–420.

Chudik, A., Pesaran, M. H., and Tosetti, E. (2011). Weak and strong cross-section dependence and estimation of large panels.

Collier, P. and Dollar, D. (2002). Aid allocation and poverty reduction. *European Economic Review*, 46(8):1475–1500.

Dalgaard, C.-J., Hansen, H., and Tarp, F. (2004). On the empirics of foreign aid and growth. *The Economic Journal*, 114(496):F191–F216.

Diebold, F. X., Rudebusch, G. D., and Aruoba, S. B. (2006). The macroeconomy and the yield curve: A dynamic latent factor approach. *Journal of Econometrics*, 131(1-2):309–338.

Doucouliagos, H. and Ulubaşoğlu, M. A. (2008). Democracy and economic growth: a meta-analysis. *American Journal of Political Science*, 52(1):61–83.

Eickmeier, S. and Ziegler, C. (2008). How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach. *Journal of Forecasting*, 27(3):237–265.

Elliott, G., Gargano, A., and Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, 177(2):357–373.

Elliott, G., Gargano, A., and Timmermann, A. (2015). Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics & Control*, 54:86–110.

Faust, J. and Wright, J. H. (2013). Forecasting inflation. In *Handbook of economic forecasting*, volume 2, pages 2–56. Elsevier.

Forbes, K. J. (2000). A reassessment of the relationship between inequality and growth. *American Economic Review*, 90(4):869–887.

Fortin-Gagnon, O., Leroux, M., Stevanovic, D., and Surprenant, S. (2018). *A large Canadian database for macroeconomic analysis.* Centre interuniversitaire de recherche en analyse des organisations.

Gold, D., Lederer, J., and Tao, J. (2020). Inference for high-dimensional instrumental variables regression. *Journal of Econometrics*, 217(1):79–111.

Gonçalves, S. and Perron, B. (2014). Bootstrapping factor-augmented regression models. *Journal of Econometrics*, 182(1):156–173.

Hahn, J., Hausman, J. A., and Kuersteiner, G. M. (2001). Bias corrected instrumental variables estimation for dynamic panel models with fixed effects.

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

Hansen, B. E. (2010). Multi-step forecast model selection. In *20th Annual Meetings of the Midwest Econometrics Group.*

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

Hansen, C. and Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics*, 182(2):290–308.

Hansen, H. and Tarp, F. (2001). Aid and growth regression. *Journal of Development Economics*, 64(2):547–570.

Haug, A. (1996). Tests for cointegration: A Monte Carlo comparison. *Journal of Econometrics*, 71(1-2):89–115.

Hauk Jr, W. R. (2017). Endogeneity bias and growth regressions. *Journal of Macroeconomics*, 51:143–161.

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

Kim, H. H. and Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: new empirical evidence. *Journal of Econometrics*, 178:352–367.

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163.

Leroux, M., Kotchoni, R., and Stevanović, D. (2017). *Forecasting economic activity in data-rich environment.* Center for Interuniversity Research and Analysis on Organizations.

León-González, R. and Montolio, D. (2015). Endogeneity and panel data in growth regressions: A bayesian model averaging approach. *Journal of Marcoeconomics*, 46:23–39.

Lin, W., Feng, R., and Li, H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509):270–288.

Marcellino, M., Stock, J. H., and Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1-2):499–526.

McCracken, M. W. and McGillicuddy, J. T. (2019). An empirical investigation of direct and iterated multistep conditional forecasts. *Journal of Applied Econometrics*, 34(2):181–204.

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

Pauwels, L. and Vasnev, A. (2014). Forecast combination for U.S. recessions with real-time data. *North American Journal of Economics and Finance*, 28:138–148.

Pesaran, M. H. (2015). Testing weak cross-sectional dependence in large panels. *Econometric Reviews*, 34(6-10):1089–1117.

Pesaran, M. H. and Smith, R. (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 68(1):79–113.

Roodman, D. (2009). A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics*, 71(1):135–158.

Shi, Z. (2016). Estimation of sparse structural parameters with many endogenous variables. *Econometric Reviews*, 35(8-10):1582–1608.

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

Spindler, M. (2016). Lasso for instrumental variable selection: A replication study. *Journal of Applied Econometrics*, 31(2):450–454.

Stock, J. H. (1987). Asymptotic properties of least squares estimators of cointegrating vectors. *Econometrica: Journal of the Econometric Society*, pages 1035–1056.

Stock, J. H. and Watson, M. W. (1988). Testing for common trends. *Journal of the American Statistical Association*, 83(404):1097–1107.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Stock, J. H. and Watson, M. W. (2005). Implications of dynamic factor models for var analysis. Technical report, National Bureau of Economic Research.

Stock, J. H. and Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493.

Su, L. and Chen, Q. (2013). Testing homogeneity in panel data models with interactive fixed effects. *Econometric Theory*, pages 1079–1135.

Swanson, N. and Xiong, W. (2018). Big data analytics in economics: What have we learned so far, and where should we go from here? *Canadian Journal of Economics*, 51(3):695–746.

Tu, Y. and Yi, Y. (2017). Forecasting cointegrated nonstationary time series with time-varying variance. *Journal of Econometrics*, 196(1):83–98.

Zhang, X. and Liu, C.-A. (2018). Inference after model averaging in linear regression models. *Econometric Theory*, 35(4):816–841.

Zhu, Y. (2018). Sparse linear models and l1-regularized 2SLS with high-dimensional endogenous regressors and instruments. *Journal of Econometrics*, 202(2):196–213.

A. APPENDIX FOR: FACTOR-AUGMENTED ERROR CORRECTION MODEL AVERAGING IN PREDICTIVE REGRESSIONS

A.1 Proof of Theorem 1 and 2

Fist, I introduce some notations before the proofs: if a variable is with a "tilde", it is the first step factor estimation's product. As comparisons, a variable with a hat is a product from the second step forecasting regression. All variables are assumed to have zero means. Eigenvectors are ordered such that the corresponding eigenvalues are decreasing.

In this appendix, I assume there is no predictors other than the factors and cointegration relationships for simplicity. Additionally, I assume there is one cointegration relationship between Y_t and $\boldsymbol{F}_t$: $Y_t = \boldsymbol{\delta}'\boldsymbol{F}_t + e_t$, where I fix the coefficient for Y_t to be one without loss of generality. With the estimated factors, the equation is rewritten as: $Y_t = \boldsymbol{\delta}'\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t + \boldsymbol{\delta}'(\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) + e_t$. Under different assumptions of the dataset and different I(1) factor estimations, $(\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t)$ has different orders. If we assume there is no nonstationarity among the idiosyncratic components when estimating the factors, Bai (2004) proves that $(\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) = O_p(\frac{1}{min(\sqrt{N},\ T^{3/2})})$, where $\boldsymbol{H}_2 = (\boldsymbol{\lambda}'b/N)(\boldsymbol{F}'\tilde{\boldsymbol{F}}/T^2)V_{NT}^{-1}$, and V_{NT} is an $r \times r$ diagonal matrix consisting of the first r eigenvalues of $(1/NT^2)\boldsymbol{X}\boldsymbol{X}'$. If we assume there is potential nonstationarity in the idiosyncratic components, then $(\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) = O_p(\frac{\sqrt{T}}{min(\sqrt{N},\ T^{3/4})})$, where $\boldsymbol{H}_2 = (\boldsymbol{\lambda}'\boldsymbol{\lambda}/N)(\boldsymbol{f}'\tilde{\boldsymbol{f}}/T)V_{NT}^{-1}$, and V_{NT} is an $r \times r$ diagonal matrix consisting of the first r eigenvalues of $(1/NT)\boldsymbol{x}\boldsymbol{x}'$. In the rest of the proof where the second estimation method is used, I implicitly assume $T < N$.

Start from the forecasting equation :

$$\begin{aligned}
y_{t+h} &= \boldsymbol{\alpha}'\boldsymbol{f}_t + \gamma'(Y_t + \boldsymbol{\delta}'\boldsymbol{F}_t) + \epsilon_t \\
&= \boldsymbol{\alpha}'\boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t + \gamma'(Y_t + \boldsymbol{\delta}'\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) + \{\gamma'\boldsymbol{\delta}'(\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) + \\
&\quad +\boldsymbol{\alpha}'(\boldsymbol{f}_t - \boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t) + \epsilon_t\} \\
&= \boldsymbol{\alpha}'\boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t + \gamma'(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t) \\
&\quad +\{\gamma'(\boldsymbol{\delta}'\boldsymbol{H}_2^{-1} - \tilde{\boldsymbol{\delta}}')\tilde{\boldsymbol{F}}_t + \gamma'\boldsymbol{\delta}'(\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) + \boldsymbol{\alpha}'(\boldsymbol{f}_t - \boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t) + \epsilon_t\} \\
&= \hat{\boldsymbol{\alpha}}'\tilde{\boldsymbol{f}}_t + \hat{\gamma}'(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t) \\
&\quad +\{(\boldsymbol{H}_1^{-1'}\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})'\tilde{\boldsymbol{f}} + (\gamma - \hat{\gamma})'(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t) \\
&\quad +\gamma'(\boldsymbol{\delta}'\boldsymbol{H}_2^{-1} - \tilde{\boldsymbol{\delta}}')\tilde{\boldsymbol{F}}_t + \gamma'\boldsymbol{\delta}'(\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) + \boldsymbol{\alpha}'(\boldsymbol{f}_t - \boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t) + \epsilon_t\} \\
&= \hat{\boldsymbol{\alpha}}'\tilde{\boldsymbol{f}}_t + \hat{\gamma}'(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t) + \hat{\epsilon}_t
\end{aligned}$$

, where $\boldsymbol{H}_1 = (\boldsymbol{\lambda}'\boldsymbol{\lambda}/N)(\boldsymbol{f}'\tilde{\boldsymbol{f}}/T)\boldsymbol{V}_{NT}^{-1}$, and $\boldsymbol{V}_{NT}$ is an $r \times r$ diagonal matrix consisting of the first r eigenvalues of $(1/NT)\boldsymbol{x}\boldsymbol{x}'$. $\boldsymbol{H}_2$ follows different formulae depending on which method is used for the estimation of nonstationary factors. In the error structure, the first two terms come from the estimation of the forecasting regression, the third to the fifth terms are errors from the cointegration and factor estimations, and the last term is the original error term in the true forecasting regression. We know that $\tilde{\boldsymbol{\delta}} \xrightarrow{p} \boldsymbol{\delta}\boldsymbol{H}_2^{-1}$ from Stock and Watson (1988). It is shown in Stock and Watson (1988) and Stock (1987) that the estimated cointegration vectors have the follow asymptotic property: $\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\boldsymbol{H}_2^{-1} = O_p(\frac{1}{T^{1-\Delta}})$, for all $\Delta < 1$.

The rest of the proof follows Cheng and Hansen (2015). From $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} + \boldsymbol{u} - \hat{\boldsymbol{u}}$, we have $\frac{1}{T}\hat{\boldsymbol{\epsilon}}(w)'\hat{\boldsymbol{\epsilon}}(w) = \frac{1}{T}(\boldsymbol{u} - \hat{\boldsymbol{u}}(w))'(\boldsymbol{u} - \hat{\boldsymbol{u}}(w)) + \frac{1}{T}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} + \frac{2}{T}(\boldsymbol{u} - \hat{\boldsymbol{u}}(w))'\boldsymbol{\epsilon}$. Using the notation of CH,

$$C_{1T}(w) \quad = \quad L_{1T}(w) + \tfrac{1}{T}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} + \tfrac{2}{\sqrt{T}}r_{1T}(w) - \tfrac{2}{T}r_{2T}(w) \ ,$$

where $r_{1T}(w) = \frac{1}{\sqrt{T}}\boldsymbol{u}'(\boldsymbol{I} - \tilde{\boldsymbol{P}}(w))\boldsymbol{\epsilon}$, $r_{2T}(w) = \boldsymbol{\epsilon}'\tilde{\boldsymbol{P}}(w)\boldsymbol{\epsilon} - \hat{\delta}^2\sum_{m=1}^{M} w(m)k(m)$, $\tilde{\boldsymbol{P}}(w)$ is the projection matrix of $[\tilde{\boldsymbol{f}}_t', \ (Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)']'$, and $\boldsymbol{P}$ is the projection matrix of the true data $[\boldsymbol{f}_t', \ (Y_t + \boldsymbol{\delta}'\boldsymbol{F}_t)']'$. As in CH , the goal is to show $r_{1T}(w) \xrightarrow{p} \kappa_1(w)$ and $r_{2T}(w) \xrightarrow{p} \kappa_2(w)$, where $\kappa_1(w)$ and $\kappa_s(w)$ are zero mean variables. The proofs have two parts. The first part contains the proof of the above two convergences when the cointegration rank is r, which means all the cointegration relationships used as regressors are "valid" stationary variables. The second part is to show that even there are some "invalid" nonstationary terms in the predictive regression, those terms do not affect the asymptotic theory.

There are two lemmas that are useful:

Lemma 1. Under assumptions R and Fs, for $N, T \to \infty$, and $\frac{T}{N} \to 0$ if Bai and Ng (2004) is used for estimating the levels of the factors,

(a) $\frac{1}{\sqrt{T}}\boldsymbol{\epsilon}'(\tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'} - \boldsymbol{f}) = o_p(1)$

(b) $\frac{1}{\sqrt{T}}\boldsymbol{\epsilon}'(\tilde{\boldsymbol{F}}\tilde{\boldsymbol{\delta}} - \boldsymbol{F}\boldsymbol{\delta}) = o_p(1)$

Proof of Lemma 1 (a). $\frac{1}{\sqrt{T}}\boldsymbol{\epsilon}'(\tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'} - \boldsymbol{f}) = \boldsymbol{H}_1^{-1}\frac{1}{\sqrt{T}}\sum\tilde{\boldsymbol{f}}_t\epsilon_t - \frac{1}{\sqrt{T}}\sum\boldsymbol{f}_t\epsilon_t = \boldsymbol{H}_1^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t)\epsilon_t$. Bai (2003) proves that $\boldsymbol{H}_1 = O(1)$. The asymptotic of $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t)\epsilon_t$ is discussed in both Gonçalves and Perron (2014) and Bai and Ng (2006).

Proof of Lemma 1 (b). $\frac{1}{\sqrt{T}}\boldsymbol{\epsilon}'(\tilde{\boldsymbol{F}}\tilde{\boldsymbol{\delta}} - \boldsymbol{F}\boldsymbol{\delta}) = \boldsymbol{\epsilon}'\frac{1}{\sqrt{T}}(\tilde{\boldsymbol{F}}\tilde{\boldsymbol{\delta}} - \boldsymbol{F}\boldsymbol{H}_2'\tilde{\boldsymbol{\delta}} + \boldsymbol{F}\boldsymbol{H}_2'\tilde{\boldsymbol{\delta}} - \boldsymbol{F}\boldsymbol{\delta}) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(Y_t + \tilde{\boldsymbol{\delta}}\tilde{\boldsymbol{F}}_t - Y_t - \boldsymbol{\delta}\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t)\epsilon_t + \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(Y_t + \boldsymbol{\delta}\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t - Y_t - \delta\boldsymbol{F}_t)\epsilon_t$ =I+II,

For I, $(Y_t + \tilde{\boldsymbol{\delta}}\tilde{\boldsymbol{F}}_t - Y_t - \boldsymbol{\delta}\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) = o_p(1)$, for all t. Thus, with the Assumption R, I $= o_p(1)$. For II, $\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t - \boldsymbol{F}_t = \boldsymbol{H}_2^{-1}(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t) = O_p(\frac{\sqrt{T}}{\sqrt{N}})$ or $O_p(\frac{1}{min(\sqrt{N}, T^{3/2})})$, where $\boldsymbol{H}_2 = O(1)$ is shown in Bai (2003) and Bai (2004), depending on how the I(1) factors are estimated.

Lemma 2. Under assumptions R and Fs, for $N, T \to \infty$, and $\frac{T}{N} \to 0$ if Bai and Ng (2004) is used for estimating the levels of the factors,

(a) $\boldsymbol{H}_1^{-1}\sum_{t=1}^{T}\tilde{\boldsymbol{f}}_t\tilde{\boldsymbol{f}}_t'\boldsymbol{H}_1^{-1'}/T \xrightarrow{p} \sum_{t=1}^{T}\boldsymbol{f}_t\boldsymbol{f}_t'/T$

(b) $\boldsymbol{H}_1^{-1}\sum_{t=1}^{T}\tilde{\boldsymbol{f}}_t(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)/T \xrightarrow{p} \sum_{t=1}^{T}\boldsymbol{f}_t(Y_t + \boldsymbol{\delta}'\boldsymbol{F}_t)/T$

(c) $\sum_{t=1}^{T}(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)/T \xrightarrow{p} \sum_{t=1}^{T}(Y_t + \boldsymbol{\delta}'\boldsymbol{F}_t)(Y_t + \boldsymbol{\delta}'\boldsymbol{F}_t)/T$

Proof of Lemma 2 (a). $\boldsymbol{H}_1^{-1}\sum_{t=1}^T \tilde{\boldsymbol{f}}_t\tilde{\boldsymbol{f}}_t'\boldsymbol{H}_1^{-1'} = \boldsymbol{H}_1^{-1}\frac{1}{T}\sum_{t=1}^T(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t + \boldsymbol{H}_1\boldsymbol{f}_t)(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t + \boldsymbol{H}_1 f_t)'\boldsymbol{H}_1^{-1'} \xrightarrow{p} \sum_{t=1}^T \boldsymbol{f}_t\boldsymbol{f}_t'/T$, these proofs are straightforward from Gonçalves and Perron (2014), Lemma A.2.

Proof of Lemma 2 (b). We can decompose the left-hand side into:

$$\begin{aligned}\boldsymbol{H}_1^{-1}\sum_{t=1}^T \tilde{\boldsymbol{f}}_t(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)/T = H_1^{-1}\frac{1}{T}\sum_{t=1}^T\{\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t + \boldsymbol{H}_1\boldsymbol{f}_t\}\{&Y_t + \boldsymbol{\delta}'\boldsymbol{F}_t \\ &+ (\tilde{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1})\boldsymbol{F}_t + (\tilde{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1})(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t) \\ &+ \boldsymbol{\delta}'\boldsymbol{H}_2^{-1}(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t)\}'\end{aligned}$$

Thus it is sufficient to show that $\frac{1}{T}\sum_{t=1}^T\{\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t\}(Y_t + \boldsymbol{\delta}'\boldsymbol{F}_t)'$, $\frac{1}{T}\sum_{t=1}^T(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t)\boldsymbol{F}_t'(\tilde{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1})$, $\frac{1}{T}\sum_{t=1}^T(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t)(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t)'(\tilde{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1})'$, $\frac{1}{T}\sum_{t=1}^T(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t)(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t)'\boldsymbol{\delta}'$, $\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}_t\boldsymbol{F}_t'(\tilde{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1})$, $\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}_t(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t)'(\tilde{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1})$, and $\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}_t(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t)'\boldsymbol{\delta}'$ are all $o_p(1)$.

The first four terms are $o_p(1)$, and there are intermediate results from Bai (2003), Bai (2004), and Bai and Ng (2004). We know that $(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t) = o_p(1)$ and $(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t) = O_p(\frac{\sqrt{T}}{\sqrt{N}})$ or $O_p(\frac{1}{min(\sqrt{N}, T^{3/2})})$, thus

$$\frac{1}{T}\sum_{t=1}^T(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t)\boldsymbol{F}_t'(\tilde{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1}) = \frac{1}{T}\sum_{t=1}^T(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t)\frac{\boldsymbol{F}_t'}{\sqrt{T}}\sqrt{T}(\tilde{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1}) = o_p(1).$$

Similarly,

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^T(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t)(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t)'(\tilde{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1})' \\ =&\frac{1}{T}\sum_{t=1}^T(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t)(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t)'(\tilde{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1}) = o_p(1)\end{aligned}$$

and $\frac{1}{T}\sum_{t=1}^T(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t)(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t)'\boldsymbol{\delta}' = o_p(1)$.

For the last three terms, first we have $||\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}_t\boldsymbol{F}_t'(\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\boldsymbol{H}_2^{-1})'|| =||\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\boldsymbol{H}_2^{-1}|| \, ||\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}_t\boldsymbol{F}_t'|| = o_p(1)$ and $||(\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\boldsymbol{H}_2^{-1})'||||\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}_t(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t)'|| = O_p(\frac{1}{T^{1-\Delta}})||\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t|| = o_p(1)$. Then, $||\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}_t(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t)'\boldsymbol{\delta}'|| = O_p(1)||\frac{1}{T}\sum_{t=1}^T(\tilde{\boldsymbol{F}}_t -$

$\boldsymbol{H}_2\boldsymbol{F}_t)'||$, which is $o_p(1)$ depending on if we assume there is nonstationary idiosyncratic components.

Proof of Lemma 2 (c). Similar to the proof of Lemma 2 (b), $(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)$ can be decomposed to $\{Y_t + \boldsymbol{\delta}'\boldsymbol{F}_t + (\tilde{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1})\boldsymbol{F}_t + (\tilde{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1})(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t) + \boldsymbol{\delta}'\boldsymbol{H}_2^{-1}(\tilde{\boldsymbol{F}}_t - \boldsymbol{H}_2\boldsymbol{F}_t)\}$. Thus, we need to prove that each of the last three terms is $o_p(1)$,which is shown in Stock and Watson (1988), Bai (2003), and Bai (2004).

Proof of Theorem 1. Consider the first part where all predictors are stationary. First, I show that $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$, where $\hat{\sigma}^2 = \frac{\sum_{t=1}^T \hat{\epsilon}_t^2}{T}$. By examining the terms in the residuals, we have $||\boldsymbol{f}_t - \boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t|| = O_p(\frac{1}{min(\sqrt{N},\ T)})$, $||\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t|| = O_p(\frac{1}{min(\sqrt{N},\ T^{3/2})})$ or $O_p(\frac{\sqrt{T}}{\sqrt{N}})$, $(\hat{\boldsymbol{\delta}}' - \boldsymbol{\delta}'\boldsymbol{H}_2^{-1})\tilde{\boldsymbol{F}}_t = O_p(\frac{1}{T^{1-\Delta}})\sqrt{T}$, $(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) = O_p(\frac{1}{\sqrt{T}})$ and $(\hat{\gamma} - \gamma) = O_p(\frac{1}{\sqrt{T}})$. Thus, with the largest model nests the true model, $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$.

Second, I show that $r_{2T}(w) \xrightarrow{p} \kappa_2(w)$, which means $\boldsymbol{\epsilon}'\tilde{\boldsymbol{P}}(w)\boldsymbol{\epsilon} \xrightarrow{p} \boldsymbol{\epsilon}'\boldsymbol{P}(w)b$. From the definition of the projection matrix,

$$\begin{aligned}
\tilde{\boldsymbol{P}}(w) &= [\tilde{\boldsymbol{f}}\ (\boldsymbol{Y} + \tilde{\boldsymbol{F}}\tilde{\boldsymbol{\delta}})] \begin{bmatrix} \sum_{t=1}^T \tilde{\boldsymbol{f}}_t\tilde{\boldsymbol{f}}_t' & \sum_{t=1}^T \tilde{\boldsymbol{f}}_t(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t) \\ \sum_{t=1}^T (Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)\tilde{\boldsymbol{f}}_t' & \sum_{t=1}^T (Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t) \end{bmatrix}^{-1} \\
&\quad \times [\tilde{\boldsymbol{f}}\ (\boldsymbol{Y} + \tilde{\boldsymbol{F}}\tilde{\boldsymbol{\delta}})]' \\
&= [\tilde{\boldsymbol{f}}\ (\boldsymbol{Y} + \tilde{\boldsymbol{F}}\tilde{\boldsymbol{\delta}})] \begin{bmatrix} H_1^{-1}/\sqrt{T} & 0 \\ 0 & 1/\sqrt{T} \end{bmatrix} \\
&\quad \times \begin{bmatrix} \boldsymbol{H}_1^{-1}\sum_{t=1}^T \tilde{\boldsymbol{f}}_t\tilde{\boldsymbol{f}}_t'\boldsymbol{H}_1^{-1'}/T & \boldsymbol{H}_1^{-1}\sum_{t=1}^T \tilde{\boldsymbol{f}}_t(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)'/T \\ \sum_{t=1}^T (Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)\tilde{\boldsymbol{f}}_t'\boldsymbol{H}_1^{-1'}/T & \sum_{t=1}^T (Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)'/T \end{bmatrix}^{-1} \\
&\quad \times \begin{bmatrix} H_1^{-1'}/\sqrt{T} & 0 \\ 0 & 1/\sqrt{T} \end{bmatrix} [\tilde{\boldsymbol{f}}\ (\boldsymbol{Y} + \tilde{\boldsymbol{F}}\tilde{\boldsymbol{\delta}})]' \qquad .
\end{aligned}$$

By applying Lemma 1 and Lemma 2, we can separately show that $\boldsymbol{\epsilon}'(\tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1} - \boldsymbol{f}) = o_p(1)$, $\boldsymbol{\epsilon}'(\tilde{\boldsymbol{F}}\tilde{\boldsymbol{\delta}} - \boldsymbol{F}\boldsymbol{\delta}) = o_p(1)$. Thus,

$$\begin{bmatrix} \boldsymbol{H}_1^{-1}\sum_{t=1}^T \tilde{\boldsymbol{f}}_t\tilde{\boldsymbol{f}}_t'\boldsymbol{H}_1^{-1'}/T & \boldsymbol{H}_1^{-1}\sum_{t=1}^T \tilde{\boldsymbol{f}}_t(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)'/T \\ \sum_{t=1}^T (Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)\tilde{\boldsymbol{f}}_t'\boldsymbol{H}_1^{-1'}/T & \sum_{t=1}^T (Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)'/T \end{bmatrix}$$
$$- \begin{bmatrix} \sum_{t=1}^T \boldsymbol{f}_t\boldsymbol{f}_t'/T & \sum_{t=1}^T \boldsymbol{f}_t(Y_t + \boldsymbol{\delta}'\boldsymbol{F}_t)/T \\ \sum_{t=1}^T (Y_t + \boldsymbol{\delta}'\boldsymbol{F}_t)f_t'/T & \sum_{t=1}^T (Y_t + \boldsymbol{\delta}\boldsymbol{F}_t)(Y_t + \boldsymbol{\delta}\boldsymbol{F}_t)/T \end{bmatrix} = o_p(1)$$

Next, by applying Lemma C.1 in Bai and Ng (2004),

$$\begin{bmatrix} \boldsymbol{H}_1^{-1}\sum_{t=1}^T \tilde{\boldsymbol{f}}_t\tilde{\boldsymbol{f}}_t'\boldsymbol{H}_1^{-1'}/T & \boldsymbol{H}_1^{-1}\sum_{t=1}^T \tilde{\boldsymbol{f}}_t(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{F}_t)/T \\ \sum_{t=1}^T (Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)\tilde{\boldsymbol{f}}_t'\boldsymbol{H}_1^{-1'}/T & \sum_{t=1}^T (Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)(Y_t + \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t)/T \end{bmatrix}^{-1}$$
$$- \begin{bmatrix} \sum_{t=1}^T f_tf_t'/T & \sum_{t=1}^T f_t(Y_t + \delta F_t)'/T \\ \sum_{t=1}^T (Y_t + \delta F_t)f_t'/T & \sum_{t=1}^T (Y_t + \delta F_t)(Y_t + \delta F_t)'/T \end{bmatrix}^{-1} = o_p(1)$$

.

This proves $\boldsymbol{\epsilon}'\tilde{\boldsymbol{P}}(w)\boldsymbol{\epsilon} \xrightarrow{p} \boldsymbol{\epsilon}'\boldsymbol{P}(w)\boldsymbol{\epsilon}$. The proof of $r_{1T}(w) \xrightarrow{p} \kappa_1(w)$ is similar.

The last part of the proof of Theorem 1 is to show that when we incorrectly include some nonstationary regressors in the forecasting regression, those "useless" regressors do not affect the asymptotic properties of the forecasts. Now assume the true model is

$$y_{t+h} \quad = \quad \alpha'\boldsymbol{f}_t + \gamma_1'(Y_t + \boldsymbol{\delta}_1'\boldsymbol{F}_t) + \epsilon_t \ ,$$

but we are estimating

$$y_{t+h} \quad = \quad \alpha'\boldsymbol{f}_t + \gamma_1'(Y_t + \boldsymbol{\delta}_1'\boldsymbol{F}_t) + \gamma_2'(Y_t + \boldsymbol{\delta}_2'\boldsymbol{F}_t) + \epsilon_t \ .$$

In the prediction, the regressors are $\tilde{\boldsymbol{f}}_t$, $a_t = (Y_t + \tilde{\boldsymbol{\delta}}_1'\tilde{\boldsymbol{F}}_t)$, and $b_t = (Y_t + \tilde{\boldsymbol{\delta}}_2'\tilde{\boldsymbol{F}}_t)$.

Let $\ddot{u}_{1t} = \ddot{\gamma}_1 a_t$ and $\ddot{u}_{2t} = \ddot{\gamma}_2 b_t$,

$$\begin{aligned} \tfrac{1}{T}\ddot{\boldsymbol{\epsilon}}(w)'\ddot{\boldsymbol{\epsilon}}(w) &= \tfrac{1}{T}(\boldsymbol{u} - \ddot{\boldsymbol{u}}_1(w) - \ddot{\boldsymbol{u}}_2(w))'(\boldsymbol{u} - \ddot{\boldsymbol{u}}_1(w) - \ddot{\boldsymbol{u}}_2(w)) + \tfrac{1}{T}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} \\ &\quad + \tfrac{2}{T}(\boldsymbol{u} - \ddot{\boldsymbol{u}}_1(w) - \ddot{\boldsymbol{u}}_2(w))'\boldsymbol{\epsilon} \\ &= \tfrac{1}{T}(\boldsymbol{u} - \ddot{\boldsymbol{u}}_1(w))'(\boldsymbol{u} - \ddot{\boldsymbol{u}}_1(w)) + \tfrac{1}{T}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} \\ &\quad + \tfrac{2}{T}(\boldsymbol{u} - \ddot{\boldsymbol{u}}_1(w))'\boldsymbol{\epsilon} + \tfrac{2}{T}(\ddot{\boldsymbol{u}}_2(w) + \boldsymbol{\epsilon})'\ddot{\boldsymbol{u}}_2(w) \end{aligned} \quad ,$$

where the last term is of interest. Since we are using OLS for the estimations, the formula of the estimated coefficients are:

$$\begin{matrix} \sqrt{T}(\ddot{\gamma}_1-\gamma_1) \\ T(\ddot{\gamma}_2-0) \end{matrix} = \begin{bmatrix} \sum_{t=1}^T a_t a_t'/T & \sum_{t=1}^T a_t b_t'/T^{3/2} \\ \sum_{t=1}^T b_t a_t'/T^{3/2} & \sum_{t=1}^T b_t b_t'/T^2 \end{bmatrix}^{-1} \begin{pmatrix} \sum_{t=1}^T a_t \epsilon_t/\sqrt{T} \\ \sum_{t=1}^T b_t \epsilon_t/T \end{pmatrix}$$

, thus $\ddot{u}_{2t} = \ddot{\gamma}_2(Y_t + \tilde{\delta}_1 \tilde{F}_t) = O_p(\frac{1}{\sqrt{T}})$ and $\frac{2}{T}(\ddot{\boldsymbol{u}}_2(w)+\boldsymbol{\epsilon})'\ddot{\boldsymbol{u}}_2(w) = O_p(\frac{1}{\sqrt{T}})$.

The final step is to show that with the "invalid" I(1) regressors, $\ddot{\gamma}_1$ are asymptotically equivalent to $\hat{\gamma}_1$ when there are just the $\tilde{\boldsymbol{f}}_t$ and $a_t = (Y_t + \tilde{\boldsymbol{\delta}}_1' \tilde{\boldsymbol{F}}_t)$ as regressors. Back to the above distribution formula, applying the formula for inverse of partitioned matrix,

$$\begin{aligned} \sqrt{T}(\ddot{\gamma}_1-\gamma_1) &= \{\tfrac{\sum_{t=1}^T a_t a_t'}{T} - \tfrac{\sum_{t=1}^T a_t b_t'}{T^{3/2}}(\tfrac{\sum_{t=1}^T b_t b_t'}{T^2})^{-1}\tfrac{\sum_{t=1}^T b_t a_t'}{T^{3/2}}\}^{-1}\tfrac{\sum_{t=1}^T a_t\epsilon_t}{\sqrt{T}} \\ &\quad -(\tfrac{\sum_{t=1}^T a_t a_t'}{T})^{-1}\tfrac{\sum_{t=1}^T a_t b_t'}{T^{3/2}}\{\tfrac{\sum_{t=1}^T b_t b_t'}{T^2} - \tfrac{\sum_{t=1}^T a_t b_t'}{T^{3/2}}(\tfrac{\sum_{t=1}^T a_t a_t'}{T})^{-1} \\ &\quad \tfrac{\sum_{t=1}^T b_t a_t'}{T^{3/2}}\}^{-1}\tfrac{\sum_{t=1}^T b_t\epsilon_t}{T} \\ &= (\tfrac{\sum_{t=1}^T a_t a_t'}{T})^{-1}\tfrac{\sum_{t=1}^T a_t\epsilon_t}{\sqrt{T}} - (\tfrac{\sum_{t=1}^T a_t a_t'}{T})^{-1}\tfrac{\sum_{t=1}^T a_t b_t'}{T^{3/2}}(\tfrac{\sum_{t=1}^T b_t b_t'}{T^2})^{-1}\tfrac{\sum_{t=1}^T b_t\epsilon_t}{T} \\ &\quad +O_p(\tfrac{1}{\sqrt{T}}) \\ &= (\tfrac{\sum_{t=1}^T a_t a_t'}{T})^{-1}\tfrac{\sum_{t=1}^T a_t\epsilon_t}{\sqrt{T}} + O_p(\tfrac{1}{\sqrt{T}}) \\ &\to \sqrt{T}(\hat{\gamma}_1-\gamma_1) \end{aligned}$$

,with $\frac{\sum_{t=1}^T a_t b_t'}{T^{3/2}} = \frac{1}{\sqrt{T}}\frac{\sum_{t=1}^T a_t b_t'}{T} = O_p(\frac{1}{\sqrt{T}})$. Thus, $\ddot{u}_{1t} \to \hat{u}_{1t}$, and $\frac{1}{T}\ddot{\boldsymbol{\epsilon}}(w)'\ddot{\boldsymbol{\epsilon}}(w) \xrightarrow{p} \frac{1}{T}\hat{\boldsymbol{\epsilon}}(w)'\hat{\boldsymbol{\epsilon}}(w) \xrightarrow{p} \sigma^2$.

Proof of Theorem 2. The proof of Theorem 2 is similar to Theorem 1. First, if the regressors contain any nonstationarity, the above proof shows that this extra predicted part has a smaller order than $O(\frac{1}{\sqrt{T}})$. Thus, we can write the CVA criterion as: $CV_{h,T}(w) = \frac{1}{T}\check{\boldsymbol{\epsilon}}_h(w)'\check{\boldsymbol{\epsilon}}_h(w) = \check{L}(w) + \frac{1}{T}\epsilon'\epsilon + \frac{2}{\sqrt{T}}\check{r}_{3T}(w) + o_p(\frac{1}{\sqrt{T}})$, where $\check{L}(w) = \frac{1}{T}(\boldsymbol{u} - \check{\boldsymbol{u}}_h(w))'(\boldsymbol{u} - \check{\boldsymbol{u}}_h(w))$ and $\check{r}_{3T}(w) = \sum_{m=1}^M w(m)\frac{1}{\sqrt{T}}\sum_{t=1-h}^{T-h}(\boldsymbol{u}_t - \tilde{\boldsymbol{z}}_t(m)'\check{\boldsymbol{b}}_{t,h}(m))\epsilon_{t+h}$.

Next, we can further decompose $\check{r}_{3T}(w)$ into four terms: $\check{r}_{3T}(w) = \check{r}_{3T}^0(w) + \check{s}_{1T}(w) + \check{s}_{2T}(w) + \sum_{m=1}^M w(m)\frac{1}{\sqrt{T}}\sum_{t=1-h}^{T-h}\{\boldsymbol{z}_{Ht}(m) - \tilde{\boldsymbol{z}}(m)\}'\check{\boldsymbol{b}}_{t,h}(m)\epsilon_{t+h}$, where $\check{r}_{3T}^0(w)$, $\check{s}_{1T}(w)$, and $\check{s}_{2T}(w)$ have the same formula of CH equations (4.18), (4.19) and (4.20).

For the last term $\frac{1}{\sqrt{T}}\sum_{t=1-h}^{T-h}\{\boldsymbol{z}_{Ht}(m)-\tilde{\boldsymbol{z}}(m)\}'\check{\boldsymbol{b}}_{t,h}(m)\epsilon_{t+h} = \frac{1}{\sqrt{T}}\sum_{t=1-h}^{T-h}\{[\boldsymbol{z}_{Ht}(m)-\tilde{\boldsymbol{z}}_t(m)]'\boldsymbol{b}(m)+\tilde{\boldsymbol{z}}(m)'[\boldsymbol{b}(m)-\check{\boldsymbol{b}}_{t,h}(m)]\}\epsilon_{t+h}$, from the Lemma 1 and Theorem 2 in CH, $[\boldsymbol{b}(m)-\check{\boldsymbol{b}}_{t,h}(m)] = o_p(1)$ and $\boldsymbol{\delta}'\boldsymbol{F}_t - \tilde{\boldsymbol{\delta}}'\tilde{\boldsymbol{F}}_t = \boldsymbol{\delta}'(\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) + (\boldsymbol{\delta}'\boldsymbol{H}_2^{-1} - \tilde{\boldsymbol{\delta}}')\tilde{\boldsymbol{F}}_t = O_p(\frac{\sqrt{T}}{\sqrt{N}}) + O_p(\frac{1}{T^{1/2-\Delta}})$ or $O_p(\frac{1}{min(\sqrt{N},T^{3/2})}) + O_p(\frac{1}{T^{1/2-\Delta}})$ for all t. With$\check{\boldsymbol{b}}_{t,h}(m) = O(1)$ and Assumption F, we have $\check{r}_{3T}(w) \xrightarrow{p} \check{r}^0_{3T}(w)$, and $\check{r}^0_{3T}(w) \xrightarrow{d} \kappa_3(w)$, which proves $E(\kappa_1(w)) = 0$.

A.2 Proof of Theorem 3 and 4

In this appendix, I show the unbiasedness of MMA and CVA criteria using the I(1) regressors as Tu and Yi (2017). Assume the true model is $y_{t+h} = \alpha_y\beta_0'[Y_{t-1}\ \boldsymbol{F}_{t-1}']'+\epsilon_t$, and α_y is the first row of the adjustment matrix α_0.

Proof the Theorem 3. From Cheng and Hansen (2015), we know that $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}+\boldsymbol{u}-\hat{\boldsymbol{u}}$, and $\frac{1}{T}\hat{\boldsymbol{\epsilon}}(w)'\hat{\boldsymbol{\epsilon}}(w) = \frac{1}{T}(\boldsymbol{u}-\hat{\boldsymbol{u}}(w))'(\boldsymbol{u}-\hat{\boldsymbol{u}}(w)) + \frac{1}{T}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} + \frac{2}{T}(\boldsymbol{u}-\hat{\boldsymbol{u}}(w))'\boldsymbol{\epsilon}$. Following the notation of CH,

$$C_{1T}(w) \quad = \quad L_{1T}(w) + \tfrac{1}{T}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} + \tfrac{2}{\sqrt{T}}r_{4T}(w) - \tfrac{2}{T}r_{5T}(w) \ ,$$

where $r_{4T}(w) = \frac{1}{\sqrt{T}}\boldsymbol{u}'(\boldsymbol{I}-\tilde{\boldsymbol{P}}(w))\boldsymbol{\epsilon}$, $r_{5T}(w) = \boldsymbol{\epsilon}'\tilde{\boldsymbol{P}}(w)\boldsymbol{\epsilon}-\hat{\sigma^2}\sum_{m=1}^{M}w(m)k(m)$. The true regressors can be considered as the error correction term $\beta_0'\boldsymbol{Z}_t = W_t$. Correspondingly, the set of regressors are now formed by subsets of the full error-correction matrix, and the mapping should be one-to-one. Without loss of generosity, let $\tilde{\boldsymbol{P}}(w)$ be the projection matrix of $[\tilde{\boldsymbol{f}}_t{}',\ \tilde{\boldsymbol{W}}_t{}']'$, and $\boldsymbol{P}$ is the projection matrix of the true data $[\boldsymbol{f}_t',\ \boldsymbol{W}_t']'$. As CH , the goal is to show $r_{4T}(w) \xrightarrow{p} \kappa_4(w)$ and $r_{5T}(w) \xrightarrow{p} \kappa_5(w)$. The proof have two parts. The first part is to show the consistency of $\hat{\sigma}^2$. The second part is to prove $r_{4T}(w) \xrightarrow{d} \kappa_4(w)$, $r_{5T}(w) \xrightarrow{d} \kappa_5(w)$, with $E(\kappa_4(w)) = 0$, $E(\kappa_5(w)) = 0$.

For the first part, the proof is similar to Theorem 1. For the second part, define $r^0_{4T}(w) = \frac{1}{\sqrt{T}}\boldsymbol{u}'(\boldsymbol{I}-\boldsymbol{P}(w))\boldsymbol{\epsilon}$, $r^0_{5T}(w) = \boldsymbol{\epsilon}'\boldsymbol{P}(w)\boldsymbol{\epsilon} - \sigma^2\sum_{m=1}^{M}w(m)k(m)$, and rewrite $u = \boldsymbol{Z}b = \boldsymbol{W}\alpha_y'$. Now, with $\tilde{\boldsymbol{P}}(w) = \boldsymbol{W}(w)(\boldsymbol{W}'(w)\boldsymbol{W}(w))^{-1}\boldsymbol{W}'(w)$, $r_{4T}(w)$ turns into:

$$
\begin{aligned}
r_{4T}(w) &= \frac{1}{\sqrt{T}}\boldsymbol{u}'(\boldsymbol{I} - \boldsymbol{P}(w))\boldsymbol{\epsilon} \\
&= \frac{1}{\sqrt{T}}\alpha_y W'(\boldsymbol{\epsilon} - \boldsymbol{W}(w)(\boldsymbol{W}'(w)\boldsymbol{W}(w))^{-1}\boldsymbol{W}'(w)\boldsymbol{\epsilon}) \\
&= \frac{1}{\sqrt{T}}\alpha_y \boldsymbol{W}'\boldsymbol{\epsilon} \\
&\quad - \frac{1}{\sqrt{T}}\alpha_y \boldsymbol{W}'\frac{\boldsymbol{W}(w)}{\sqrt{T}}(\frac{\boldsymbol{W}'(w)\boldsymbol{W}(w)}{T})^{-1}\frac{\boldsymbol{W}'(w)\boldsymbol{\epsilon}}{\sqrt{T}} \\
&\to \kappa_4
\end{aligned}
$$

Similarly, $\boldsymbol{\epsilon}'\boldsymbol{P}(w)\boldsymbol{\epsilon}$ is

$$
\boldsymbol{\epsilon}'\boldsymbol{P}(w)\boldsymbol{\epsilon} = \frac{\boldsymbol{\epsilon}'\boldsymbol{W}(w)}{\sqrt{T}}(\frac{\boldsymbol{W}'(w)\boldsymbol{W}(w)}{T})^{-1}\frac{\boldsymbol{W}'(w)\boldsymbol{\epsilon}}{\sqrt{T}} \to \kappa_5.
$$

These together prove that $r^0_{4T}(w)$ and $r^0_{5T}(w)$ converge in distribution to some zero mean random variables.

Next, similar to CH, we show that the terms A_T , B_{1T} and B_{2T} in equation (4.12) of CH are $o_p(1)$. The rest follows the proof of Theorem 1.

Proof of Theorem 4. The proof of Theorem 4 is similar to Theorem 2. First, since the I(1) variables are assumed to be cointegrated, with the omission of the rotation matrices of factor estimations, we can write the CVA criterion as:

$CV_{h,T}(w) = \frac{1}{T}\check{\boldsymbol{\epsilon}}_h(w)'\check{\boldsymbol{\epsilon}}_h(w) = \check{L}(w)+\frac{1}{T}\epsilon'\epsilon+\frac{2}{\sqrt{T}}\check{r}_{6T}(w)+o(\frac{1}{\sqrt{T}})$, where $\check{L}(w) = \frac{1}{T}(\boldsymbol{u}-\check{\boldsymbol{u}}_h(w))'(\boldsymbol{u}-\check{\boldsymbol{u}}_h(w))$ and $\check{r}_{6T}(w) = \sum_{m=1}^{M} w(m)\frac{1}{\sqrt{T}}\sum_{t=1-h}^{T-h}(\boldsymbol{u}_t - \tilde{\boldsymbol{z}}_t(m)'\check{\boldsymbol{b}}_{t,h}(m))\epsilon_{t+h}$.

We can further decompose $\check{r}_{6T}(w)$ as

$$
\check{r}_{6T}(w) = \check{r}^0_{6T}(w) + \check{s}_{3T}(w) + \check{s}_{4T}(w) + \sum_{m=1}^{M} w(m)\frac{1}{\sqrt{T}}\sum_{t=1-h}^{T-h}\{z_t(m) - \tilde{z}(m)\}'\check{b}_{t,h}(m)\epsilon_{t+h},
$$

where $\check{r}^0_{6T}(w)$, $\check{s}_{3T}(w)$, and $\check{s}_{4T}(w)$ have the same formula of CH equations (4.18), (4.19) and (4.20). Note that the term $\boldsymbol{b}(m) - \hat{\boldsymbol{b}}(m)$ still has the same order when the regressors are I(1) and cointegrated, but the asymptotic formula follows equation (A.5) in Tu and Yi (2017).

For the last term $\frac{1}{\sqrt{T}}\sum_{t=1-h}^{T-h}\{\boldsymbol{z}_t(m) - \tilde{\boldsymbol{z}}(m)\}'\check{\boldsymbol{b}}_{t,h}(m)\epsilon_{t+h} = \frac{1}{\sqrt{T}}\sum_{t=1-h}^{T-h}\{[\boldsymbol{z}_t(m) - \tilde{\boldsymbol{z}}_t(m)]'\boldsymbol{b}(m) + \tilde{\boldsymbol{z}}(m)'[\boldsymbol{b}(m) - \check{\boldsymbol{b}}_{t,h}(m)]\}\epsilon_{t+h}$, from Lemma 1 and Theorem 2 in CH, $[\boldsymbol{b}(m) - \check{\boldsymbol{b}}_{t,h}(m)] = o_p(1)$ and $\boldsymbol{z}_t(m) - \tilde{\boldsymbol{z}}_t(m) = O_p(\frac{\sqrt{T}}{\sqrt{N}})$ or $O_p(\frac{1}{min(\sqrt{N}, T^{3/2})})$ for all t. With $\check{\boldsymbol{b}}_{t,h}(m) = O(1)$ and Assumption F, we have $\check{r}_{6T}(w) \xrightarrow{p} \check{r}^0_{6T}(w)$, and $\check{r}^0_{6T}(w) \xrightarrow{d} \kappa_6(w)$, with $E(\kappa_6(w)) = 0$.

A.3 Proof of Propositions 1-3

In this appendix, I discuss the proofs of Propositions 1-3 and their related lemmas. The theoretical work is built on the assumptions regarding factor estimations from Bai (2003) and Bai (2004), as well as the Conditions 1-4, f1-f4, and F1-F4. I assume the first M_0 candidate models are under-fitted, and $S = M - M_0$.

Lemma 3 Assume the largest model is $y_{t+h} = \sum_r^R \beta_r \tilde{f}_{rt} + e_t$ and R is the total number of factors. Denote $\tilde{\boldsymbol{P}}_m$ as the projection matrix of the estimated stationary factors $\tilde{\boldsymbol{P}}_m = \tilde{\boldsymbol{f}}_m(\tilde{\boldsymbol{f}}_m{}'\tilde{\boldsymbol{f}}_m)^{-1}\tilde{\boldsymbol{f}}_m{}' = \frac{1}{T}\tilde{\boldsymbol{f}}_m\tilde{\boldsymbol{f}}_m'$, and $a_m = y'(I - \tilde{\boldsymbol{P}}_m)y$. Under the assumptions and $N,\ T \to \infty$, $a_m - a_M = O_p(T)$ when $m \in \{1,\ ...,\ M_0\}$.

Proof of Lemma 3 Consider an FAR structure:

$$\begin{aligned}
y_{t+h} &= \beta'\boldsymbol{f}_t + e_t \\
&= \boldsymbol{\beta}'\boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t + \{\boldsymbol{\beta}'(\boldsymbol{f}_t - \boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t) + e_t\} \\
&= \tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{f}}_t + \tilde{e}_t
\end{aligned}$$

Let $\tilde{\boldsymbol{\beta}}_{m^c} = \Pi_{m^c}\tilde{\boldsymbol{\beta}}$, where $\tilde{\boldsymbol{\beta}}$ is the true coefficient vector when the estimated factors are regressors, and $\Pi_m = (I_{r_m},\ 0_{r_m \times (R - r_m)})$ is the selection matrix with r_m being the number of factors in model m.

$$\begin{aligned}
a_m - a_M &= (\tilde{\boldsymbol{e}} + \tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c})'(\boldsymbol{I} - \tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{e}} + \tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c}) - \tilde{\boldsymbol{e}}'(\boldsymbol{I} - \boldsymbol{P}_M)\tilde{\boldsymbol{e}} \\
&= (\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c})'(\boldsymbol{I} - \tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c}) + 2(\tilde{\boldsymbol{e}})'(\boldsymbol{I} - \tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c}) - \tilde{\boldsymbol{e}}'(\tilde{\boldsymbol{P}}_m - \tilde{\boldsymbol{P}}_M)\tilde{\boldsymbol{e}} \\
&= I + II + III
\end{aligned}$$

It is not hard to show that $I = (\tilde{\boldsymbol{f}}_{m^c}\boldsymbol{\beta}_{m^c})'(\boldsymbol{I} - \tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\boldsymbol{\beta}_{m^c}) = O_p(T)$.

$$\begin{aligned} II &= 2(\tilde{\boldsymbol{e}})'(\boldsymbol{I} - \tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c}) \\ &= 2(\boldsymbol{e} + (\boldsymbol{f} - \tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta})'(\boldsymbol{I} - \tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c}) \\ &= 2\boldsymbol{e}'(\boldsymbol{I} - \tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c}) + 2((\boldsymbol{f} - \tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta})'(\boldsymbol{I} - \tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c}) \\ &= O_p(\sqrt{T}) + o_p(\sqrt{T}) \end{aligned}$$

where the second part is discussed in Gonçalves and Perron (2014) and Bai and Ng (2006).

$$\begin{aligned} III &= \tilde{\boldsymbol{e}}'(\tilde{\boldsymbol{P}}_m - \tilde{\boldsymbol{P}}_M)\tilde{\boldsymbol{e}} \\ &= (\boldsymbol{e} + (\boldsymbol{f} - \tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta})'(\tilde{\boldsymbol{P}}_m - \tilde{\boldsymbol{P}}_M)(\boldsymbol{e} + (\boldsymbol{f} - \tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta}) \\ &= \boldsymbol{e}'(\tilde{\boldsymbol{P}}_m - \tilde{\boldsymbol{P}}_M)\boldsymbol{e} + 2((\boldsymbol{f} - \tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta})'(\tilde{\boldsymbol{P}}_m - \tilde{\boldsymbol{P}}_M)\boldsymbol{e} \\ &\quad + ((\boldsymbol{f} - \tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta})'(\tilde{\boldsymbol{P}}_m - \tilde{\boldsymbol{P}}_M)((\boldsymbol{f} - \tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta})) \\ &= O_p(1) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad . \end{aligned}$$

These prove Lemma 3.

Lemma 4 Assume the largest model is $y_{t+h} = \sum_r \beta_r \tilde{f}_{rt} + \gamma_Y Y_t + \boldsymbol{\gamma}_F \tilde{\boldsymbol{F}}_t + e_t$. Denote $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ as the rotation matrices for the estimated stationary and nonstationary factors, respectively. Let $C_{m^c} = C\Pi'_{m^c}$, and $\theta_{m^c} = \Pi_{m^c}\theta$, where $\Pi_m = (I_{r_m},\ 0_{r_m \times (R - r_m)})$ is the selection matrix. Moreover, denote $\tilde{\boldsymbol{P}}_m$ as the projection matrix of the estimated stationary factors and cointegrations $\tilde{\boldsymbol{P}}_m = [\tilde{\boldsymbol{f}}_m,\ (Y, \tilde{\boldsymbol{F}}_m)C_m]\{[\tilde{\boldsymbol{f}}_m,\ (Y, \tilde{\boldsymbol{F}}_m)C_m]'$ $[\tilde{\boldsymbol{f}}_m,\ (Y, \tilde{\boldsymbol{F}}_m)C_m]\}^{-1}\ [\tilde{\boldsymbol{f}}_m,\ (Y, \tilde{\boldsymbol{F}}_m)C_m]'$, where C is the cointegration vector matrix between Y and $\tilde{F}$. Let $a_m = y'(I - \tilde{P}_m)y$, then $a_m - a_M = O_p(T)$ when $m \in \{1,\ ...,\ M_0\}$ and $N,\ T \to \infty$.

Proof of Lemma 4 Consider an FECM structure:

$$\begin{aligned}
y_{t+h} &= \boldsymbol{\beta}'\boldsymbol{f}_t + \gamma_Y Y_t + \boldsymbol{\gamma_F}'\boldsymbol{F}_t + e_t \\
&= \boldsymbol{\beta}'\boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t + \gamma_Y Y_t + \boldsymbol{\gamma}'_{\boldsymbol{F}}\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t \\
&+ \{\boldsymbol{\gamma}'_{\boldsymbol{F}}(\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) + \boldsymbol{\beta}'(\boldsymbol{f}_t - \boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t) + e_t\} \\
&= \boldsymbol{\beta}'\boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t + \gamma_Y Y_t + \boldsymbol{\gamma_F}'\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t + \tilde{e}_t \\
&= \boldsymbol{\beta}'\boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t + \gamma_Y Y_t + \tilde{\boldsymbol{\gamma}}_{\boldsymbol{F}}'\tilde{\boldsymbol{F}}_t + \tilde{e}_t \\
&= \boldsymbol{\beta}'\boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t + \theta'[Y_t, \tilde{\boldsymbol{F}}_t']'\boldsymbol{C} + \tilde{e}_t \\
&= \tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{f}}_t + \theta'[Y_t, \tilde{\boldsymbol{F}}_t']'\boldsymbol{C} + \tilde{e}_t
\end{aligned}$$

Denote θ as the adjustment vector. The error correction structure indicates that the same subset of $\tilde{\boldsymbol{f}}$ and $\tilde{\boldsymbol{F}}$ exists in the model. Note that Y and $\tilde{F}$ are assumed to be cointegrated with rank R, and C has form $\begin{bmatrix} C_{11} & C_{21} & \dots & C_{R1} \\ C_{12} & C_{22} & \dots & C_{R2} \\ & C_{23} & \dots & C_{R3} \\ & & & \vdots \\ & & & C_{RR+1} \end{bmatrix}$ without loss of generality.

$$\begin{aligned}
a_m - a_M &= (\tilde{\boldsymbol{e}} + \tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c} + [Y, \tilde{\boldsymbol{F}}]C_{m^c}\theta_{m^c})'(\boldsymbol{I} - \tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{e}} + \tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c} + [Y, \tilde{\boldsymbol{F}}]C_{m^c}\theta_{m^c}) \\
&\quad - \tilde{\boldsymbol{e}}'(\boldsymbol{I} - \boldsymbol{P}_M)\tilde{\boldsymbol{e}} \\
&= (\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c} + [Y, \tilde{\boldsymbol{F}}]C_{m^c}\theta_{m^c})'(\boldsymbol{I} - \tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c} + [Y, \tilde{\boldsymbol{F}}]C_{m^c}\theta_{m^c}) \\
&\quad + 2\tilde{\boldsymbol{e}}'(\boldsymbol{I} - \tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c} + [Y, \tilde{\boldsymbol{F}}]C_{m^c}\theta_{m^c}) - \tilde{\boldsymbol{e}}'(\tilde{\boldsymbol{P}}_m - \tilde{\boldsymbol{P}}_M)\tilde{\boldsymbol{e}} \\
&= I + II + III
\end{aligned}$$

$$I = (\tilde{\boldsymbol{f}}_{m^c}\boldsymbol{\beta}_{m^c} + [Y, \tilde{\boldsymbol{F}}]C_{m^c}\theta_{m^c})'(\boldsymbol{I} - \tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\boldsymbol{\beta}_{m^c} + [Y, \tilde{\boldsymbol{F}}]C_{m^c}\theta_{m^c}) = O_p(T).$$

$$\begin{aligned}
II &= 2(\tilde{\boldsymbol{e}})'(\boldsymbol{I}-\tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c}+[Y,\tilde{\boldsymbol{F}}]C_{m^c}\theta_{m^c})\\
&= 2(\boldsymbol{e}+(\boldsymbol{f}-\tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta}+(\boldsymbol{F}-\tilde{\boldsymbol{F}}\boldsymbol{H}_2^{-1'})\boldsymbol{\gamma}_F)'(\boldsymbol{I}-\tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c}+[Y,\tilde{\boldsymbol{F}}]C_{m^c}\theta_{m^c})\\
&= 2\boldsymbol{e}'(\boldsymbol{I}-\tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c}+[Y,\tilde{\boldsymbol{F}}]C_{m^c}\theta_{m^c})\\
&\quad +2((\boldsymbol{f}-\tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta})'(\boldsymbol{I}-\tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c}+[Y,\tilde{\boldsymbol{F}}]C_{m^c}\theta_{m^c})\\
&\quad +2((\boldsymbol{F}-\tilde{\boldsymbol{F}}\boldsymbol{H}_2^{-1'})\boldsymbol{\gamma}_F)'(\boldsymbol{I}-\tilde{\boldsymbol{P}}_m)(\tilde{\boldsymbol{f}}_{m^c}\tilde{\boldsymbol{\beta}}_{m^c}+[Y,\tilde{\boldsymbol{F}}]C_{m^c}\theta_{m^c})\\
&= O_p(\sqrt{T})+o_p(\sqrt{T})+o_p(1)
\end{aligned}$$

The proof of $\frac{1}{T}\sum_t(\boldsymbol{f}_t'(\boldsymbol{F}_t-\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t)) = o_p(T^{-1})+o_p(N^{-1/2}T^{-1/2})$ is similar to Lemma B.4 (i) of Bai (2004) with the application of Bai (2003) Assumption F. Then $\frac{1}{T}\sum_t(\tilde{\boldsymbol{f}}_t'(\boldsymbol{F}_t-\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t)) = \frac{1}{T}\sum_t((\tilde{\boldsymbol{f}}_t-\boldsymbol{H}_1\boldsymbol{f}_t)'(\boldsymbol{F}_t-\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t)) + \frac{1}{T}\sum_t\boldsymbol{f}_t'\boldsymbol{H}_1'(\boldsymbol{F}_t-\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) = o_p(T^{-1})+o_p(N^{-1/2}T^{-1/2})$. Finally, from Theorem 2 of Bai (2004),

$$\begin{aligned}
III &= \tilde{\boldsymbol{e}}'(\tilde{\boldsymbol{P}}_m-\tilde{\boldsymbol{P}}_M)\tilde{\boldsymbol{e}}\\
&= (\boldsymbol{e}+(\boldsymbol{f}-\tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta})'(\tilde{\boldsymbol{P}}_m-\tilde{\boldsymbol{P}}_M)(\boldsymbol{e}+(\boldsymbol{f}-\tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta})\\
&\quad +((\boldsymbol{F}-\tilde{\boldsymbol{F}}\boldsymbol{H}_2^{-1'})\boldsymbol{\gamma}_F)'(\tilde{\boldsymbol{P}}_m-\tilde{\boldsymbol{P}}_M)((\boldsymbol{F}-\tilde{\boldsymbol{F}}\boldsymbol{H}_2^{-1'})\boldsymbol{\gamma}_F)\\
&\quad +2(\boldsymbol{e}+(\boldsymbol{f}-\tilde{\boldsymbol{f}}\boldsymbol{H}_1^{-1'})\boldsymbol{\beta})'(\tilde{\boldsymbol{P}}_m-\tilde{\boldsymbol{P}}_M)((\boldsymbol{F}-\tilde{\boldsymbol{F}}\boldsymbol{H}_2^{-1'})\boldsymbol{\gamma}_F)\\
&= O_p(1)+o_p(1) \qquad .
\end{aligned}$$

These prove Lemma 4.

Proof of Proposition 1

(1) From Hansen (2010) equation (13), $\hat{e}_{t,h} = \hat{e}_{ols,t}+\boldsymbol{x}_{t-h}(\boldsymbol{z}'\boldsymbol{z}-\boldsymbol{z}_{t,h}'\boldsymbol{z}_{t,h})^{-1}z_{t,h}'\hat{\boldsymbol{e}}_{ols,t:h}$. Let C_{mj} denote the (m,j) element of the CVA criteria, and the superscripts (j) and (m) denote the model candidates, then

$$\begin{aligned}
CV_{mj} &= \sum_t \hat{e}_{t,h}^{(m)} \hat{e}_{t,h}^{(j)} \\
&= \sum_t \{\hat{e}_{ols,t}^{(m)} + \boldsymbol{z}_{t-h}^{(m)} (\boldsymbol{z}^{(m)\prime} \boldsymbol{z}^{(m)} - \boldsymbol{z}_{t,h}^{(m)\prime} \boldsymbol{z}_{t,h}^{(m)})^{-1} \boldsymbol{z}_{t,h}^{(m)\prime} \hat{\boldsymbol{e}}_{ols,t:h}^{(m)}\} \\
&\quad \times \{\hat{e}_{ols,t}^{(j)} + \boldsymbol{z}_{t-h}^{(j)} (\boldsymbol{z}^{(j)\prime} \boldsymbol{z}^{(j)} - \boldsymbol{z}_{t,h}^{(j)\prime} \boldsymbol{z}_{t,h}^{(j)})^{-1} \boldsymbol{z}_{t,h}^{(j)\prime} \hat{\boldsymbol{e}}_{ols,t:h}^{(j)}\} \\
&= \sum_t \hat{e}_{ols,t}^{(m)} \hat{e}_{ols,t}^{(j)} + \sum_t \{\boldsymbol{z}_{t-h}^{(m)} (\boldsymbol{z}^{(m)\prime} \boldsymbol{z}^{(m)} - \boldsymbol{z}_{t,h}^{(m)\prime} \boldsymbol{z}_{t,h}^{(m)})^{-1} \boldsymbol{z}_{t,h}^{(m)\prime} \hat{\boldsymbol{e}}_{ols,t:h}^{(m)}\} \\
&\quad \times \{\boldsymbol{z}_{t-h}^{(j)} (\boldsymbol{z}^{(j)\prime} \boldsymbol{z}^{(j)} - \boldsymbol{z}_{t,h}^{(j)\prime} \boldsymbol{z}_{t,h}^{(j)})^{-1} \boldsymbol{z}_{t,h}^{(j)\prime} \hat{\boldsymbol{e}}_{ols,t:h}^{(j)}\} \\
&\quad + \sum_t \hat{e}_{ols,t}^{(m)} \{\boldsymbol{z}_{t-h}^{(j)} (\boldsymbol{z}^{(j)\prime} \boldsymbol{z}^{(j)} - \boldsymbol{z}_{t,h}^{(j)\prime} \boldsymbol{z}_{t,h}^{(j)})^{-1} \boldsymbol{z}_{t,h}^{(j)\prime} \hat{\boldsymbol{e}}_{ols,t:h}^{(j)}\} \\
&\quad + \sum_t \hat{e}_{ols,t}^{(j)} \{\boldsymbol{z}_{t-h}^{(m)} (\boldsymbol{z}^{(m)\prime} \boldsymbol{z}^{(m)} - \boldsymbol{z}_{t,h}^{(m)\prime} \boldsymbol{z}_{t,h}^{(m)})^{-1} \boldsymbol{z}_{t,h}^{(m)\prime} \hat{\boldsymbol{e}}_{ols,t:h}^{(m)}\} \\
&= \{\Phi_{mj} + \Psi_{mj} + \hat{\sigma}^2 (K_m - K_j)\}
\end{aligned}$$

Here Φ_{mj} follows the same definition of Zhang and Liu (2018). Ψ_{mj} is self-defined through the equation.

From matrix algebra,

$$\begin{aligned}
&\{\boldsymbol{z}_{t-h}^{(j)\prime} (\boldsymbol{z}^{(j)\prime} \boldsymbol{z}^{(j)} - \boldsymbol{z}_{t,h}^{(j)\prime} \boldsymbol{z}_{t,h}^{(j)})^{-1} \boldsymbol{z}_{t,h}^{(j)\prime} \hat{\boldsymbol{e}}_{ols,t:h}^{(j)}\} \\
&= trace\{\boldsymbol{z}_{t-h}^{(j)\prime} (\boldsymbol{z}^{(j)\prime} \boldsymbol{z}^{(j)} - \boldsymbol{z}_{t,h}^{(j)\prime} \boldsymbol{z}_{t,h}^{(j)})^{-1} \boldsymbol{z}_{t,h}^{(j)\prime} \hat{\boldsymbol{e}}_{ols,t:h}^{(j)}\} \\
&= trace\{(\boldsymbol{z}^{(j)\prime} \boldsymbol{z}^{(j)} - z_{t,h}^{(j)\prime} \boldsymbol{z}_{t,h}^{(j)})^{-1} \sum_{s=-h+1}^{s=h-1} \boldsymbol{z}_{t-h+s}^{(j)\prime} \hat{e}_{ols,t+s}^{(j)} \boldsymbol{z}_{t-h}^{(j)}\} \\
&= trace\{(\frac{\boldsymbol{z}^{(j)\prime} \boldsymbol{z}^{(j)} - \boldsymbol{z}_{t,h}^{(j)\prime} \boldsymbol{z}_{t,h}^{(j)}}{T})^{-1} \frac{1}{T} \sum_{s=-h+1}^{s=h-1} \boldsymbol{z}_{t-h+s}^{(j)\prime} \hat{e}_{ols,t+s}^{(j)} \boldsymbol{z}_{t-h}^{(j)}\} \\
&= O_p(T^{-1}).
\end{aligned}$$

Thus, $\sum_t \{\boldsymbol{z}_{t-h}^{(m)\prime} (\boldsymbol{z}^{(m)\prime} \boldsymbol{z}^{(m)} - \boldsymbol{z}_{t,h}^{(m)\prime} \boldsymbol{z}_{t,h}^{(m)})^{-1} \boldsymbol{z}_{t,h}^{(m)\prime} \hat{\boldsymbol{e}}_{ols,t:h}^{(m)}\} \{\boldsymbol{z}_{t-h}^{(j)\prime} (\boldsymbol{z}^{(j)\prime} \boldsymbol{z}^{(j)} - \boldsymbol{z}_{t,h}^{(j)\prime} \boldsymbol{z}_{t,h}^{(j)})^{-1} \boldsymbol{z}_{t,h}^{(j)\prime} \hat{\boldsymbol{e}}_{ols,t:h}^{(j)}\} = O_p(1)$, and $\sum_t \hat{e}_{ols,t}^{(j)} \{\boldsymbol{z}_{t-h}^{(m)\prime} (\boldsymbol{z}^{(m)\prime} \boldsymbol{z}^{(m)} - \boldsymbol{z}_{t,h}^{(m)\prime} \boldsymbol{z}_{t,h}^{(m)})^{-1} \boldsymbol{z}_{t,h}^{(m)\prime} \hat{\boldsymbol{e}}_{ols,t:h}^{(m)}\} = \sum_t \hat{e}_{ols,t}^{(j)} \{\boldsymbol{z}_{t-h}^{(m)\prime} (\frac{\boldsymbol{z}^{(m)\prime} \boldsymbol{z}^{(m)} - \boldsymbol{z}_{t,h}^{(m)\prime} \boldsymbol{z}_{t,h}^{(m)}}{T})^{-1} \frac{1}{T} \sum_{s=-h+1}^{s=h-1} \boldsymbol{z}_{t-h+s}^{(m)\prime} \hat{e}_{ols,t+s}^{(m)}\} = o_p(\sqrt{T})$. Thus, applying the similar proof as Zhang and Liu (2018) Theorem 3, with the same definition of a_m, equation (A.13) of Zhang and Liu (2018) remains the same:

$\hat{w}_{CVA,m} \leq (a_m - a_M)^{-1}\{2\hat{\sigma}^2(K_M - K_m) + \hat{w}_{CVA,m}(\Psi_{MM} + \Psi_{mm} - \Psi_{Mm} - \Psi_{mM}) + 2\sum_{j=1}^{M} \hat{w}_{CVA,j}(\Psi_{Mj} - \Psi_{mj})\}$. Part (1) of Proposition 1 is then implied .

(2) Now consider an FAR:

$$\begin{aligned}
y_{t+h} &= \beta'\boldsymbol{f}_t + e_t \\
&= \boldsymbol{\beta}'\boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t + \{\boldsymbol{\beta}'(\boldsymbol{f}_t - \boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t) + e_t\} \\
&= \tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{f}}_t + \tilde{e}_t \\
&= \hat{\boldsymbol{\beta}}'\tilde{\boldsymbol{f}}_t + \{\boldsymbol{\beta}'(\boldsymbol{f}_t - \boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t) + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})'(\tilde{\boldsymbol{f}}_t - \boldsymbol{H}_1\boldsymbol{f}_t) + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})'\boldsymbol{H}_1\boldsymbol{f}_t + e_t\} \\
&= \hat{\boldsymbol{\beta}}'\tilde{\boldsymbol{f}}_t + \hat{e}_t
\end{aligned}$$

where $(\boldsymbol{f}_t - \boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t)$ and $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})$ have orders smaller than e_t.
From **Lemma 3,** $(a_m - a_M)^{-1} = O_p(1/T)$. The formula of CV_{mj} is revised by replacing $\boldsymbol{x}$ with $\tilde{\boldsymbol{f}}$, and the formula for $\hat{e}_t$ is updated according to the above notation.

$$\begin{aligned}
&\{\tilde{\boldsymbol{f}}_{t-h}^{(j)'}(\tilde{\boldsymbol{f}}^{(j)'}\tilde{\boldsymbol{f}}^{(j)} - \tilde{\boldsymbol{f}}_{t,h}^{(j)'}\tilde{\boldsymbol{f}}_{t,h}^{(j)})^{-1}\tilde{\boldsymbol{f}}_{t,h}^{(j)'}\hat{\boldsymbol{e}}_{ols,t:h}^{(j)}\} \\
&= trace\{(I_r - \tilde{\boldsymbol{f}}_{t,h}^{(j)'}\tilde{\boldsymbol{f}}_{t,h}^{(j)})^{-1}(\sum_{s=-h+1}^{s=h-1}\tilde{f}_{t-h+s}^{(j)'}\hat{e}_{ols,t:h+s}^{(j)})\tilde{\boldsymbol{f}}_{t-h}^{(j)'}\} \\
&= O_p(T^{-1}),
\end{aligned}$$

where $max_{1\leq t\leq T}||\tilde{f}_t|| \leq max_{1\leq t\leq T}||H_1 f_t| + max_{1\leq t\leq T}||\tilde{f}_t - H_1 f_t|| = O_p(max_{1\leq t\leq T}||f_t||)$. Thus, $\sum_t\{\tilde{\boldsymbol{f}}_{t-h}^{(m)'}(\tilde{\boldsymbol{f}}^{(m)'}\tilde{\boldsymbol{f}}^{(m)} - \tilde{\boldsymbol{f}}_{t,h}^{(m)'}\tilde{\boldsymbol{f}}_{t,h}^{(m)})^{-1}\tilde{\boldsymbol{f}}_{t,h}^{(m)'}\hat{\boldsymbol{e}}_{ols,t:h}^{(m)}\}\{\tilde{\boldsymbol{f}}_{t-h}^{(j)'}(\tilde{\boldsymbol{f}}^{(j)'}\tilde{\boldsymbol{f}}^{(j)} - \tilde{\boldsymbol{f}}_{t,h}^{(j)'}\tilde{\boldsymbol{f}}_{t,h}^{(j)})^{-1}\tilde{\boldsymbol{f}}_{t,h}^{(j)'}\hat{\boldsymbol{e}}_{ols,t:h}^{(j)}\} = O_p(1)$, as well as $\sum_t \hat{e}_{ols,t}^{(j)}\{\tilde{\boldsymbol{f}}_{t-h}^{(m)'}(\tilde{\boldsymbol{f}}^{(m)'}\tilde{\boldsymbol{f}}^{(m)} - \tilde{\boldsymbol{f}}_{t,h}^{(m)'}\tilde{\boldsymbol{f}}_{t,h}^{(m)})^{-1}\tilde{\boldsymbol{f}}_{t,h}^{(m)'}\hat{\boldsymbol{e}}_{ols,t:h}^{(m)}\} = \sum_t \hat{e}_{ols,t}^{(j)}\{\tilde{\boldsymbol{f}}_{t-h}^{(m)'}(\frac{\tilde{\boldsymbol{f}}^{(m)'}\tilde{\boldsymbol{f}}^{(m)} - \tilde{\boldsymbol{f}}_{t,h}^{(m)'}\tilde{\boldsymbol{f}}_{t,h}^{(m)}}{T})^{-1}\frac{1}{T}\sum_{s=-h+1}^{s=h-1}\tilde{\boldsymbol{f}}_{t-h+s}^{(m)'}\hat{e}_{ols,t+s}^{(m)}\} = o_p(\sqrt{T})$,
the rest of proof is similar to the proof of part (1).

(3) Move to the FECM structure:

$$\begin{aligned}
y_{t+h} &= \boldsymbol{\beta}'\boldsymbol{f}_t + \gamma_Y Y_t + \boldsymbol{\gamma_F}'\boldsymbol{F}_t + e_t \\
&= \boldsymbol{\beta}'\boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t + \gamma_Y Y_t + \boldsymbol{\gamma_F}'\boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t \\
&+ \{\boldsymbol{\gamma}'_{\boldsymbol{F}}(\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) + \boldsymbol{\beta}'(\boldsymbol{f}_t - \boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t) + e_t\} \\
&= \hat{\boldsymbol{\beta}}'\tilde{\boldsymbol{f}}_t + \hat{\gamma}_Y Y_t + \hat{\boldsymbol{\gamma}}_{\boldsymbol{F}}'\tilde{\boldsymbol{F}}_t \\
&+ \{\boldsymbol{\gamma}'_{\boldsymbol{F}}(\boldsymbol{F}_t - \boldsymbol{H}_2^{-1}\tilde{\boldsymbol{F}}_t) + \boldsymbol{\beta}'(\boldsymbol{f}_t - \boldsymbol{H}_1^{-1}\tilde{\boldsymbol{f}}_t) \\
&+ (\boldsymbol{\beta}'\boldsymbol{H}_1^{-1} - \hat{\boldsymbol{\beta}})'\tilde{\boldsymbol{f}}_t + (\boldsymbol{\gamma_F}'\boldsymbol{H}_2^{-1'} - \hat{\gamma}_F)'\tilde{\boldsymbol{F}}_t + (\gamma_Y - \hat{\gamma}_Y)Y_t + e_t\}
\end{aligned}$$

$(\boldsymbol{\gamma_F}'\boldsymbol{H}_2^{-1'} - \hat{\gamma}_F)'\tilde{\boldsymbol{F}}_t + (\gamma_Y - \hat{\gamma}_Y)Y_t$ is $o_p(1)$ given the cointegration assumption and Tu and Yi (2017) Lemma 3. $\hat{\gamma}_Y$ and $\hat{\boldsymbol{\gamma}}_{\boldsymbol{F}}'$ have the same asymptotic as $\hat{\boldsymbol{\beta}}$, and the terms in the bracket of the above equation is $\{e_t + o_p(1)\}$. Thus, from **Lemma 4,** $(a_m - a_M)^{-1} = O_p(1/T)$, we can replace the $\tilde{\boldsymbol{f}}$ by $\tilde{\boldsymbol{F}}$. The proof is similar to the proof of part (2).

Proof of Proposition 2

The proof of the first half of the proposition is straightforward under the assumptions and conditions. For the second half of the proposition, we know that the weights are $o_p(1)$ for the under-fitted models, and the proof that $CV_{mj} = o_p(\frac{1}{\sqrt{T}})$ for $m,\ j \in \{1,\ ...,\ M_0\}$ is the same as Zhang and Liu (2018). The focus is on how to simulate weights for the just-fitted and over-fitted models.

From Hansen (2010), $CV_{mj} = \sum_t \hat{e}_{t,h}^{(m)}\hat{e}_{t,h}^{(j)} \approx trace((\check{Q}_s)^{-1}\check{\Omega}_s) + trace((\check{Q}_j)^{-1}\check{\Omega}_j)$ $-e'[\tilde{\boldsymbol{f}},\ (Y,\ \tilde{\boldsymbol{F}})]'\,\check{V}_{max\{s,j\}}[\tilde{\boldsymbol{f}},\ (Y,\ \tilde{\boldsymbol{F}})]e$, where $\check{Q}$, $\check{V}$ and $\check{\Omega}$ are functions of $[\tilde{\boldsymbol{f}},\ (Y,\ \tilde{\boldsymbol{F}})]$. Denote C contains the cointegration vectors and $\check{C}$ as its Moore–Penrose inverse. We can transform CV_{mj} into by playing with the term $\begin{pmatrix} I_R & \\ & C \end{pmatrix}_{(2R+1)\times 2R} \times$ $\begin{pmatrix} I_R & \\ & \check{C} \end{pmatrix}_{2R\times(2R+1)}$. Note $\begin{pmatrix} I_R & \\ & C \end{pmatrix}$ is full column rank, and $\begin{pmatrix} I_R & \\ & \check{C} \end{pmatrix}$ is full row rank. Thus, as an example,

$$\begin{aligned}
&(\begin{pmatrix} I_R & \\ & \check{C}' \end{pmatrix}\begin{pmatrix} I_R & \\ & C' \end{pmatrix}\check{Q}\begin{pmatrix} I_R & \\ & C \end{pmatrix}\begin{pmatrix} I_R & \\ & \check{C} \end{pmatrix})^+ \\
&= \begin{pmatrix} I_R & \\ & C \end{pmatrix}\left[\begin{pmatrix} I_R & \\ & \check{C}' \end{pmatrix}\begin{pmatrix} I_R & \\ & C' \end{pmatrix}\check{Q}\begin{pmatrix} I_R & \\ & C \end{pmatrix}\right]^+ \\
&= \begin{pmatrix} I_R & \\ & C \end{pmatrix}\left[\begin{pmatrix} I_R & \\ & \check{C}' \end{pmatrix}\begin{pmatrix} I_R & \\ & C' \end{pmatrix}\check{Q}\begin{pmatrix} I_R & \\ & C \end{pmatrix}\right]^+ \\
&= \begin{pmatrix} I_R & \\ & C \end{pmatrix}\left[\begin{pmatrix} I_R & \\ & C' \end{pmatrix}\check{Q}\begin{pmatrix} I_R & \\ & C \end{pmatrix}\right]^+\begin{pmatrix} I_R & \\ & C' \end{pmatrix} \\
&= \begin{pmatrix} I_R & \\ & C \end{pmatrix}\tilde{Q}^{-1}\begin{pmatrix} I_R & \\ & C' \end{pmatrix}
\end{aligned}$$

Then, $CV_{mj} \approx trace((\tilde{Q}_s)^{-1}\tilde{\Omega}_s) + trace((\tilde{Q}_j)^{-1}\tilde{\Omega}_j) - e'[\tilde{\boldsymbol{f}},\ (Y,\ \tilde{\boldsymbol{F}})C]'\tilde{V}_{max\{s,j\}}[\tilde{\boldsymbol{f}},\ (Y,\ \tilde{\boldsymbol{F}})C]e$, where $\tilde{Q}$, $\tilde{V}$ and $\tilde{\Omega}$ are functions of $[\tilde{\boldsymbol{f}},\ (Y,\ \tilde{\boldsymbol{F}})C]$. Given $\tilde{\boldsymbol{F}} \to FH_{20}$ and $\tilde{\boldsymbol{f}} \to fH_{10}$, the $(s,j)^{th}$ element of $\Sigma_{sj} = trace((Q_s)^{-1}\Omega_s) + trace((Q_j)^{-1}\Omega_j) - \boldsymbol{\xi}'V_{max\{s,j\}}\boldsymbol{\xi} + o_p(1)$.

Proof of Proposition 3 The proof are essentially the same as the proof of Theorem 5 of Zhang and Liu (2018).

A.4 Variable transformations

Table A.1.: The right-hand side

Categories	Stationarity	I(1)
Income, output	$\Delta log(x_t)$	$log(x_t)$
Labor market	$\Delta log(x_t)$	$log(x_t)$
Construction, inventory and orders	$\Delta log(x_t)$, Δx_t	$log(x_t)$, x_t
Interest rates and asset prices	Δx_t	x_t
Prices, wages and money	$\Delta^2 log(x_t)$	$\Delta log(x_t)$

Table A.2.: The left-hand side

stationary transformations	y^h_{t+h} transformation
$\Delta log(y_t)$	$400/k * \{log(y_{t+h}) - log(y_t)\}$
Δy_t	$1/k * (y_{t+h} - y_t)$
$\Delta^2 log(y_t)$	$400/k * \{log(y_{t+h}) - log(y_t)\} - 400\{log(y_t) - log(y_{t-1})\}$

B. APPENDIX FOR: REVISITING THE DEMOCRACY-GROWTH NEXUS: NEW EVIDENCE FROM A DYNAMIC COMMON CORRELATED EFFECTS APPROACH

This appendix contains supplementary results pertaining to formal diagnostic tests for parameter heterogeneity and cross section dependence and estimates of the degree of cross section dependence.

B.1 Diagnostic Tests

In order to motivate the use of Chudik and Pesaran (2015) dynamic common correlated effects (DCCE) approach, we conduct a set of diagnostic tests for parameter heterogeneity and cross section dependence, the two potential features of the data that the approach is designed to account for. When testing for the presence of one of these features, it is important to allow for the presence of the other so that the outcome of the test is not affected by model misspecification emanating from ignoring one of these features. We therefore test for parameter heterogeneity while allowing for cross section dependence and vice-versa. We only briefly describe the tests here and refer the reader to the original papers for details.

First, we conduct tests of the null hypothesis of slope homogeneity that allow for the potential presence of cross section dependence through an interactive fixed effects specification. Two procedures are employed in this regard: (1) the LM test of Su and Chen (2013) that is based on testing if the slope coefficients in the regression of the restricted (imposing homogeneity) residuals on the observable regressors are zero; (2) the Swamy-type test of Ando and Bai (2015) that is calculated from the dispersion of country-specific slope estimates from a pooled estimate, both of which are obtained from estimating an interactive effects model. The pooled estimate is taken to be the mean of the individual slope estimates. Both tests possess a standard normal limiting distribution under the null hypothesis. The results are presented in Panel A of Table B.1. When GDP is measured in levels, both tests are significant across the three lag order specifications at the 1% level, except the Su and Chen test with four lags that rejects only at the 5% level. For GDP growth, the evidence against slope homogeneity is weaker when based on the Su and Chen test although the Ando and Bai test still rejects the null at the 1% level in all cases.

Next, we consider procedures for testing cross section dependence. We use two tests to this end: (1) Pesaran (2015) CD test which is based on estimated pairwise error correlations that allows for weak cross section dependence under the null hypothesis; (2) Castagnetti et al. (2015) test for homogeneous factor loadings computed using the maximum deviation of the estimated loadings from its mean so that the factor structure reduces to a time effect under the null hypothesis. To construct the CD test, the residuals are obtained from country-wise estimation of the dynamic heterogeneous model (1) which entails estimating N country-specific time series least squares regressions. The test has a standard normal limiting distribution under the null hypothesis so that standard critical values can be used. To construct the test based on factor loadings, we employ the DCCE estimates to obtain the residuals in the first step which are then used to estimate the factor loadings by principal components in a second step. The critical values of the test are obtained from the Gumbel distribution. The results are presented in Panel B of Table B.1. Regardless of whether GDP is measured in levels or first differences, both tests comprehensively reject the null hypothesis for all three lag order specifications.

B.2 Estimates of the Degree of Cross Section Dependence

Here we present estimates of the exponent or degree of cross section dependence using the approach proposed by Bailey et al. (2016). In particular, these authors propose a bias-corrected estimate of α, where α denote the rate at which the largest eigenvalue of the covariance matrix of the data grows with the cross-section sample size (N) with $1/2 < \alpha \leq 1$. The closer α is to unity, the higher is the degree of cross-section dependence and hence the more plausible is the presence of a common factor structure relative to a spatial structure. Table B.2 reports the estimate for each lag order specification when GDP is measured in levels or first differences. The results are suggestive of strong cross section dependence, with the exponent estimates in the range [.79,.83]. The estimates appear to be quite robust to the lag order as well as to the way in which GDP is measured.

Table B.1.: Diagnostic tests

Panel A Slope heterogeneity tests: Dependent variable	GDP level		GDP growth	
	Su and Chen (2013)	Ando and Bai (2015)	Su and Chen (2013)	Ando and Bai (2015)
one GDP lag	2.758***	39.344***	1.685*	12.210***
two GDP lags	2.966***	44.125***	1.797*	19.145***
four GDP lags	2.447**	52.818***	1.482	36.407***

Panel B Cross section dependence tests: Dependent variable	GDP level		GDP growth	
	Pesaran (2015)	Castagnetti et al. (2015))	Pesaran (2015)	Castagnetti et al. (2015)
one GDP lag	17.425***	33.580***	16.403***	55.022***
two GDP lags	17.433***	31.282***	16.188***	52.850***
four GDP lags	17.456***	33.997***	16.203***	58.374***

Notes: This table reports results of diagnostic tests for parameter heterogeneity and cross-sectional dependence. Panel A presents the Su and Chen (2013) and Ando and Bai (2015) tests for slope heterogeneity. The critical values for both tests are obtained from the standard normal distribution. Panel B reports the Pesaran (2015) CD test and the Castagnetti et al. (2015) test for cross-sectional dependence. The critical values of the former are obtained from the standard normal distribution and those of the latter are obtained from the Gumbel distribution. We use *, ** and *** to denote significance at the 10%, 5% and 1% level, respectively.

Table B.2.: Estimates of the degree of cross-sectional dependence

Dependent variable	GDP level	GDP growth
one GDP lag	0.834	0.803
two GDP lags	0.797	0.794
four GDP lags	0.801	0.791

Notes: This table reports the degree of cross-sectional dependence suggested by Bailey et al. (2016). The estimates are calculated from equation (13) in that paper.

VITA

Education

Ph.D. Economics, **Purdue University, West Lafayette, IN** August 2020

Committee: Mohitosh Kejriwal (Chair), Yong Bao, Joshua Chan, Justin Tobias

M.A. Economics, **Boston University, Boston, MA** January 2014

B.A. Economics, **Wuhan University, China** July 2012

B.S. Mathematics, **Wuhan University, China** July 2012

Research Fields

Applied Econometrics, Econometrics

Working Papers

"Factor-augmented Error Correction Model Averaging in Predictive Regressions"

"Approaches to Estimating Large-dimensional Regressions with Endogeneity: A Simulation Comparison"

"Revisiting the Democracy-Growth Nexus: New Evidence from a Dynamic Common Correlated Effects Approach" with Mohitosh Kejriwal

Conference and Seminar Activities

2019 Krannert Ph.D. Research Symposium

2019 Midwest Econometrics Groups, presenter

2019 Midwest Econometrics Groups Mentoring Workshop

2017 Joint Statistical Meeting

2016 Krannert Ph.D. Research Symposium

Awards

Certificate for Outstanding Teaching (Econometrics)	Summer 2020
Certificate for Outstanding Teaching (Principles of Macroeconomics)	Summer 2017
Ross Fellowship	2014-2015

Teaching Experience

Instructor:

Econometrics	Fall 2019, Summer 2020
Macroeconomics	Summer 2017
Macroeconomics, Online	Summer 2016
Recitation instructor, Principles Of Economics	Spring 2015

Teaching Assistant:

Time series Econometrics (PhD)	Fall 2016, Spring 2019
Panel data Econometrics (PhD)	Spring 2017, 2018, 2019
Financial Econometrics (Master)	Spring 2019
Advanced Panel data Econometrics (PhD)	Spring 2017, 2018
Probability And Statistics (PhD)	Fall 2018
Microeconomics	Spring 2016
Macroeconomics	Spring 2016
Econometrics	Fall 2014

Skills

Programming: Matlab (Advanced), Stata (Advanced), R (Intermediate), SQL (Intermediate), Eviews (Intermediate), Python (Basic), Mathematica (Basic)
Languages: English (fluent), Chinese (native)