UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays in asset pricing and forecasting

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Management

by

Ritong Qu

Committee in charge:

Professor Allan Timmermann, Chair
Professor James Hamilton
Professor Jun Liu
Professor Alexis Toda
Professor Rossen Valkanov

2021

The dissertation of Ritong Qu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my grandfather, Jianqing Jiang.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I owe a lot to my advisor Allan Timmermann for his guidance through my graduate school. I first met Allan in 2014 and learned a lot from numerous discussions with him. Allan is a great mentor and makes a scholar out of me. I am grateful to my committe members James Hamilton, Jun Liu, Alexis Toda and Rossen Valkanov for spending their valuable time reading my thesis and sharing their insights. I want to thank Yinchu Zhu who is not only a talented coauthor but also a great friend. I would also like to thank faculty and visiting scholars at UCSD for suggestions and guidance, especially Snehal Banerjee, Kai Li and Simon Smith. My family and friends are always there for me along the way.

Chapter 1 is currently being prepared for submission of publication of the material. It is solely authored by the dissertation author.

Chapter 2 is currently being prepared for submission of publication of the material, and is coauthored with Allan Timmermann and Yinchu Zhu. The dissertation author is a primary investigator of this material.

Chapter 3 is currently being prepared for submission of publication of the material, and is coauthored with Allan Timmermann and Yinchu Zhu. The dissertation author is a primary investigator of this material.

All errors are my own.

VITA

2014	B. S. in Mathematics, Peking University, China
2015	B. S. in Finance, University of California San Diego
2021	Ph. D. in Management, University of California San Diego

ABSTRACT OF THE DISSERTATION

Essays in asset pricing and forecasting

by

Ritong Qu

Doctor of Philosophy in Management

University of California San Diego, 2021

Professor Allan Timmermann, Chair

My thesis has two themes: The first theme is about studying investors' expectations and the relation to asset prices; while the second theme is about evaluating forecasting performance. Both themes focus on what we can learn from a panel of data. The first chapter of my dissertation studies rational investors' expectation of consumption growth at the presence of structure breaks and asset pricing implications. While the first chapter studies how rational individuals should do, the second and third chapters focus on forecasters' behavior in real world, by developing tools to evaluate forecasters' performance about multiple variables, across many forecasters and at single time periods.

In Chapter 1, we use data on multiple consumption goods to identify infrequent, but

persistent breaks to consumption growth dynamics. Over a sixty-year sample, we find four breaks, all of which are associated with major macroeconomic and financial market events such as oil price shocks, the Great Moderation, the end of the tech stock market bubble, and the Covid pandemic. The impact of the breaks on consumption growth is highly uncertain and heterogeneous across consumption goods. We explore the asset pricing implications of our novel empirical evidence in the context of a Lucas tree model in which investors use information on multiple consumption goods to learn about model parameters. We find that break risk in consumption growth, combined with investor learning, helps resolve a number of asset pricing puzzles such as high risk premium and volatility of market returns, as well as cross-sectional anomalies such as momentum.

Chapter 2 is joint work with Allan Timmermann and Yinchu Zhu. Forecasting skills are often identified by comparing predictive accuracy across large numbers of forecasts. This generates a multiple hypothesis testing problem that can trigger many false positives. We develop a new bootstrap test approach for identifying superior predictive accuracy that applies to multi-dimensional panel settings with arbitrarily many forecasts, outcome variables, horizons, and time periods. Our approach controls the family-wise error rate while retaining the ability to identify truly skilled forecasters. An empirical analysis of the IMF's World Economic Outlook forecasts across 185 countries, five variables and several forecast horizons shows how our approach can be used to identify variables and countries for which the IMF's forecasts improve significantly at shorter horizons as well as cases where they fail to improve.

Chapter 3 is also joint work with Allan Timmermann and Yinchu Zhu. We develop new methods for pairwise comparisons of predictive accuracy with cross-sectional data. Using a common factor setup, we establish conditions on cross-sectional dependencies in forecast errors which allow us to test the null of equal predictive accuracy on a single cross-section of forecasts. We consider both unconditional tests of equal predictive accuracy as well as tests that condition on the realization of common factors and show how to decompose forecast errors into exposures to

common factors and idiosyncratic components. An empirical application compares the predictive accuracy of financial analysts' short-term earnings forecasts across six brokerage firms.

Chapter 1

Breaks in Consumption Growth and Asset Prices

1.1 Introduction

Consumption growth dynamics is a key determinant of discount rates in the canonical asset pricing model, and many studies have attempted to explain puzzles such as the high equity risk premium and excess market volatility through the time series dynamics of aggregate consumption growth. One strand of the literature explores shocks to consumption growth that are highly persistent (Bansal and Yaron, 2004 and Hansen, Heaton and Li, 2008). These types of shocks make it difficult for agents to smooth their consumption intertemporally and lead to high risk premia and volatility of asset returns. A second strand of the literature explores large but rare disasters, which have a sudden and sharp but short-lived effect on consumption growth (Barro, 2006 and Wachter, 2013).

This paper investigates a new type of consumption growth dynamics which takes the form of infrequent, but moderately-sized and persistent regime shifts in consumption growth. In common with the long-run risk model, our consumption growth process is quite persistent. However, we recognize that economic growth is not steady or continuous and identify infrequent shifts in consumption growth dynamics, which is distinctly different from the long-run risk model. These shocks are also much smaller than the disaster breaks identified by authors such as Barro and Ursúa (2012) and can affect consumption growth both positively and negatively.

Our analysis uses a multivariate specification that includes major types of consumption goods. The multivariate dimension has two advantages. First, it allows for a richer set of asset pricing implications, as shocks can stem from a subset of major type of goods, i.e., the recent Covid-19 pandemic hit transportation services and recreation services especially hard. Second, by using a panel of consumption growth series, our approach gains power in identifying regime changes that are common across different types of goods. By construction, breaks that affect multiple consumption series are more likely to be identified than breaks that only affect one or just a few series. Such common breaks are more likely to have strong asset pricing implications

because they cannot easily be hedged against. The statistical power gained from pervasive breaks is especially important when breaks occur infrequently and the historical data is relatively short.

We rely on recent development in identifying common breaks using a Bayesian framework as in Smith and Timmermann (2020) and Smith (2017). We model conditional consumption growth of each series as an MA process whose conditional mean and volatility are subject to breaks. Each break can affect all or a subset of consumption series. We identify 4 breaks in the sample period from 1959Q2 to 2020Q3. Coinciding with major economic events: our model captures the oil crisis, the monetary experiment in the early eighties, the burst of tech bubble and the recent Covid-19 pandemic. The average regime length is around 15 years. The annualized expected aggregate consumption growth is highest in the regime between 1959 and 1972, reaching 2.25%, and lowest in the regime between 2000 and 2019, reaching 1.3%. For the most recent regime starting in 2020Q1, expected aggregate consumption growth further decreases to -1.5% though the estimation error is large given limited data. In contrast to the sharp result based on 11 types of goods, the model only identify 2 breaks when using a single time series of aggregate consumption and the locations of breaks are uncertain.

We find that breaks affect different goods unevenly. On average, each break affects 7 out of the 11 goods. The breaks in the early 1970s and 1980s affect the nondurable goods more strongly, led by the oil and energy goods. The break in the early 2000s affect the services more strongly, especially for financial services. While the recent Covid-19 break is the most pervasive: with the exception of housing and utility services, and financial services, all the major types of goods are affected. Due to lockdowns, the Covid-19 break hits transportation, recreation, and food services and accomodation especially hard: personal consumption in the three catagories is only around 80% of their levels before the pandemic.

Having established that the consumption growth process is subject to infrequent but pervasive breaks, we next develop an asset pricing model with an infinite-dimensional set of non-recurring consumption growth states using a Lucas-tree model with breaks in consumption

dynamics. We assume investors observe the timing of the breaks, but need to learn the conditional mean in the new regime as new consumption data arrives. The ensuing parameter uncertainty combined with investors' dynamic learning leads to persistent changes in investors' beliefs about long-run consumption growth. Under recursive preferences, investors are averse to long-run risks and pay a premium for the risk embedded in equity returns. We show that our model can explains asset pricing puzzles like the high equity risk premium and low risk free rate discovered by Mehra and Prescott (1985), as well as the high equity market volatility found by Shiller (1981) and LeRoy and Porter (1981).

Parameter uncertainty is highest following a break. As a result, investors adjust their belief more actively by putting a higher weight on coming consumption signals, which leads to higher price of consumption risk. Under reasonable assumptions about investors' risk aversion and EIS, breaks and ensuing changes in investors' beliefs of parameters explain more than 80% of the variance in the pricing kernel. Consistent with data, the price-dividend ratio is procyclical, and decreases with the degree of parameter uncertainty, whereas a higher level of parameter uncertainty increases both the risk premium and volatility. Parameter uncertainty plays a key role in driving price-dividend ratios. The model can also explain predictable excess returns and unpredictable consumption growth by price-dividend ratios, as found in Campbell and Shiller (1988) and Beeler and Campbell (2012).

In the cross section, types of goods affected by breaks are more informative about long run aggregate consumption growth and attract more investor attention. The resulting pricing kernel is tilted away from aggregate consumption growth and overweights types of goods whose dynamics are more fragile to breaks. To adjust for the information heterogeneity across goods, we extend the CCAPM with an additional fragility factor defined by consumption growth in types of goods with higher parameter uncertainty relative to types of goods with lower uncertainty. Using the extended CCAPM model with the fragility factor, we explore sources of risk underlying widely used risk factors such as size, value, and momentum, as discovered in Fama and French

(1992) and Jegadeesh and Titman (1993). Empirical tests with Fama-French portfolios and momentum portfolios show aggregate consumption growth accounts for value premium, while relative consumption growth accounts for momentum premium.

This paper is related to a large literature that use various features in consumption dynamics to explain asset pricing puzzles. For example, Bansal and Yaron (2004) model expected growth of aggregate consumption as driven by an AR(1) process with small, frequent, and persistent shocks. Wachter (2013) models consumption growth as subject to rare disasters: large, infrequent shocks with short term effects. We estimate a model that nests long-run risk and breaks. The model estimates show that the persistent component is subject to infrequent and discontinuous shifts, while the long-run risk component is less persistent than perceived in Bansal and Yaron (2004). The break model complements the rare disaster model: we demonstrate that rare macroeconomic events can be break points when the long-term expected consumption growth shifts.

Both rare disasters or long run risk (small but persistent changes) are difficult to identify when the sample length is small[1]: Hansen, Heaton and Li (2008) and Beeler and Campbell (2012) show the persistence of the long-run component is not high enough to explain the risk premium puzzle under reasonable risk aversion, while Mehra and Prescott (1988) and Ju and Miao (2012) that show the magnitude of rare disasters estimated from century-long US data is not large enough to fit the data without additional assumptions of ambiguity aversion. To improve the fit of models, Bollerslev and Todorov (2011), Wachter (2013), Bansal et al. (2014), Campbell et al. (2018), and Gallant, Jahan-Parvar and Liu (2019) use consumption data reinforced by equity data and other financial variables to measure disaster probabilities and the persistence of the long-run component. The methodology depends on assumptions about investors' preference and may generate endogeneity issues given other risk factors are also embedded in asset prices. Exploiting the panel structure for more statistical power, we find strong evidence for persistent shocks in

[1]Barro (2006) and Barro and Ursúa (2012) use international data to measure frequency and magnitude of rare disasters. Recently, Schorfheide, Song and Yaron (2018) find long-run risk by modeling measure errors as an MA process in monthly aggregate consumption data.

consumption growth by solely examining consumption data.

Large and infrequent breaks of persistent effects naturally imply that investors are uncertain about regimes' parameters. Our paper is related to a large literature of model uncertainty and investor learning. Recently Collin-Dufresne, Johannes and Lochstoer (2016) and Johannes, Lochstoer and Mou (2016) show that subjective parameter uncertainty implies high price of risk under recursive preference. We emphasize that our model uncertainty is not Knightian, but objective, generated by recurring breaks in model parameters.

This paper is also related to solving asset pricing puzzles using multiple goods. Accounting for other types of goods in addition to the combination of nondurables and services can increase the volatility of the pricing kernel. For example, Ait-Sahalia, Parker and Yogo (2004) examine luxury goods; Yogo (2006) examines durable goods; Piazzesi, Schneider and Tuzel (2007) examine housing. Recently, Belo and Donangelo (2020) introduce unobserved consumption components. Instead of introducing new consumption components, we disaggregate the aggregate consumption of nondurable goods and services into 11 major types of goods. We explore the channel that various types of goods serves as economic signals for the persistent component in consumption growths.

The paper proceeds as the following. Section 1.2 summarizes consumption data of major types of goods. Section 1.3 examine breaks in consumption growth dynamics. Building on the empirical evidence, Section 1.4 constructs a Lucas tree model to derive and test asset pricing implications on market returns. Section 1.5 extends the Lucas tree model to account for multiple goods and test asset pricing implication on cross-sectional returns of characteristic-sorted portfolios. Section 1.6 concludes.

1.2 Consumption growth dynamics: Empirical evidence

In this section, we examine persistent shocks to consumption growth in major types of product by summarizing statistics and figures.

1.2.1 Data

The consumption data of various nonduarble goods and services are from the US national accounts. From NIPA Table 1.2.3.3 "Real Personal Consumption Expenditures by Major Type of Product, Quantity Indexes," we collect real personal consumption expenditures of 11 types of nondurable goods and services including:

1. Food and beverages purchased for off-premises consumption
2. Clothing and footwear
3. Gasoline and other energy goods
4. Other nondurable goods
5. Housing and utilities
6. Health care
7. Transportation services
8. Recreation services
9. Food services and accommodations
10. Financial services and insurance
11. Other services.

Quarterly data for major types of goods is from 1959Q2 to 2020Q3. Annual data starts earlier in 1929 and ends in 2019. Aggregate consumption is the sum over the 11 types of nondurable goods and services. Figure 1.1 presents shares of consumption expenditures overtime. Food, housing and utility services, and health care services have the highest expenditure shares. Their aggregate accounts for more than half of consumption expenditures. Over the sample period, the share of services increased relative to nondurables: the ratio of services over nondurables was one in 1960, while the ratio is three in 2020.

Summary statistics for the 11 series are provided in Table 1.1. The top panel presents statistics based on quarterly data. The mean and standard deviation reported at an annual rate. The mean of log aggregate consumption growth is 1.81%. Clothing, health care, recreation, and financial service have a higher mean growth rate, ranging from 2.43% to 2.64% while energy, food and, food services and accomodation have a lower growth rate, ranging from -0.05% to 1.14%. The large skewness and kurtosis are due to the severe effect of Covid-19 when aggregate consumption drops more than 10% and transportation, recreation, and food and accomodation services drop more than 40% in 2020Q2. To exclude the effect of Covid-19, the middle panel of Table 1.1 presents the same set of statistics using data from 1959Q2 to 2019Q4. The 2020 data greatly affect higher order moments, while the mean of consumption growth changes little. The volatility of aggregate consumption is 0.87, less than half of the one computed based on 1959Q2 to 2020Q3. The autocorrelations are positive based on data before 2020, while they are negative when including the Covid-19 period. The skewness and kurtosis are also much smaller after excluding the Covid-19 data. The bottom panel of Table 1.1 summarizes statistics that are estimated from annual data from 1929 to 2019. It is worth noting that volatility estimated from the longer annual sample is generally higher due to the war.

1.2.2 Consumption growth dynamics

To examine the long-run component in expected consumption growth, we present the 20-year moving average of consumption growth of selected series in Figure 1.2. To expand our sample period, we use annual data from BEA that starts in 1929. The numbers are presented at an annual rate, and shaded areas represent 95% confidence interval of the estimates. Panel (a) presents the moving average of aggregate consumption growth. The twenty-year moving average of aggregate consumption growth is highest in 1954, reaching 2.8%. Starting in 1980, it gradurally decreases to 1.2% in 2020. Similar long-term swings are also present for major types of goods as presented in panels (b)-(f) in Figure 1.2. The growth rate of gasoline and other energy goods gradually decreased to -1% in the twenty-year window that ends in 2019. Long-term growth rate of housing and utilities services were increasing from 1950 to 1960, and then gradually decreased to 0.8%. Long-term growth rate of health care peaks at 4.5% in 1975 and decreased to 1.5% in 2000. Long-term growth rate of financial services and insurance peaks at 2000 at 4.5% and then decreased to 0.5%.

Figure 1.3 examines the long-run component in volatility of consumption growth using a sample standard deviation within 20-year moving windows. The numbers are presented at an annual rate. For the aggregate consumption, as well as major types of goods, volatility is high in the interwar period and decreases sharply after the war. Volatility started to stablize in the twenty-year window that ends after 1960. The aggregate volatility is highest in the beginning of the sample, reaching 4%. After 1965, it is lower than 1.5%. The post 1965 peak is in 1980, reaching 1.5%. The volatility of consumption growth in major types of goods follow similar patterns.

In summary, our 20-year rolling window analysis of consumption growth suggests persistent changes in conditional means and volatility of aggregate consumption growth, as well as its major components.

1.3 Modeling breaks in consumption growth

Dynamics of consumption growth is subject to shocks of a spectrum of frequencies, among which investors are especially averse to low-frequency shocks with persistent effects and attach high price to the risk. According to Gordon (2017), "It has long been recognized that economic growth is not steady or continuous." To investigate the nature of discontinuous and persistent changes in consumption dynamics, we build a model for consumption growth of multiple goods featuring infrequent changes in conditional mean and volatility. Structural break models are especially suitable for capturing infrequent breaks in model parameters. By setting the priors of regime durations to long periods, we allow the model to focus on those infrequent shocks with persistent effects.

Consider a panel of growth rates for $i = 1, ..., N$ consumption goods observed over $t = 1, ..., T$ time periods. Let Δc_{it} denote log consumption growth of good i. Log consumption growth are driven by

$$\Delta c_{it} = g_{it} + \gamma_{i0} f_t + \gamma_{i1} f_{t-1} + \sigma_{it} \varepsilon_{it}, \ i = 1, ..., N. \tag{1.1}$$

The common component f_t are i.i.d. $N\left(0, \sigma_{ft}^2\right)$ and idiosyncratic innovations ε_{it}, $i = 1, ..., N$, are i.i.d. standard normal.

We model the mean and volatility of consumption growth, $\left(g_{it}, \sigma_{it}, \sigma_{ft}\right)$, as a regime switching process of infinite number of states states. A total of K breaks happen at times $(\tau_1, ..., \tau_K)$. When a break happens, the data generating processes of a subset of N goods and their common component shift to a new regime. Let the indicator function $\mathbb{1}_{ik}$ be equal to 1 if the kth break hits goods i, and 0 otherwise. The parameters of the dynamics of Δc_{it} and the common component are governed by:

$$(g_{it}, \sigma_{it}) = \begin{cases} (g_{ik}, \sigma_{ik}) & \text{if } t = \tau_k \text{ and } \mathbb{1}_{ik} = 1 \\ (g_{it-1}, \sigma_{it-1}) & \text{if } \mathbb{1}_{ik} = 0 \end{cases} \tag{1.2}$$

$$\sigma_{ft} = \begin{cases} \sigma_{fk} & \text{if } t = \tau_k \text{ and } \mathbb{1}_{fk} = 1 \\ \sigma_{ft-1} & \text{if } \mathbb{1}_{fk} = 0 \end{cases} \tag{1.3}$$

Regime durations

Regime durations characterize the persistence of breaks, which is especially important when investors have recursive utility as in many asset pricing models (Hansen, Heaton and Li, 2008, Bansal and Yaron, 2004, Wachter, 2013, Collin-Dufresne, Johannes and Lochstoer, 2016). Because investors prefer early resolution of uncertainty, persistent shocks bear higher risk premium.

Let l_k denotes the duration of kth regime, $\tau_k - \tau_{k-1}$. We follow Koop and Potter (2007) and Smith and Timmermann (2020) by assuming that the length of regime k follow a Poisson distribution of parameter λ_k

$$p(l_k|\lambda_k) = \frac{\lambda_k^{l_k}}{l_k!} \exp(-\lambda_k), \ k = 1, ..., K, \tag{1.4}$$

where λ_k is drawn from a conjugate Gamma distribution

$$p(\lambda_k) = \frac{d^c}{\Gamma(c)} \lambda_k^{c-1} \exp(-\lambda_k d),, \ k = 1, ..., K. \tag{1.5}$$

As Koop and Potter (2007) point out, when modeling regimes of long durations, the Poisson-Gamma distribution is more preferable to the Geometric distribution used in Chib (1998), which implies a declining probability on regime duration so that higher weight is placed on shorter durations.

Pervasiveness of breaks

Breaks of different nature affect different sets of goods. Oil crises hit gasoline and transportation services especially hard, and the recent lockdowns due to the Covid-19 pandemic affect transportation, recreation, and food and accommodation services more than other goods. We allow for each break to hit a subset of types of goods randomly drawn from N goods.

Once an uncommon break k happens at t, the joint distribution of $\left(\mathbb{1}_{1k},...,\mathbb{1}_{Nk},\mathbb{1}_{fk}\right)$ are truncated i.i.d. Bernoulli distribution with parameter π where the point of all zeros are excluded from the support. We assume π is generated from a uniform distribution on $[0,1]$.

Expected growth rate and volatility

We assume regression coefficients and volatility of a new regime are drawn from a Normal-Gamma distribution following Geweke and Jiang (2011) and Smith and Timmermann (2020) to preserve conjugacy and improve computation speed. Let $K_{i(f)}$ denotes the set of indices of breaks that hit series i (F) and $|K_{i(f)}|$ denote the number of breaks in set $K_{i(f)}$ for $i=1,...,N$. The idiosyncratic variances of N consumption series and the common component have an inverse gamma distribution

$$p\left(\sigma_{ik}^2\right)=\frac{v^u}{\Gamma(u)}\left(\sigma_{ik}^2\right)^{-(u+1)}\exp\left(-\frac{v}{\sigma_{ik}^2}\right),\ k\in K_i,\ i=1,...,N\text{or }f. \tag{1.6}$$

The intercept and regression coefficients have a Gaussian distribution conditional on σ_{ik}^2:

$$p\left(g_{ik}|\sigma_{ik}^2\right)=\left(2\pi\sigma_{ik}^2\right)^{-1/2}|V_\beta|^{-1/2}\exp\left(-\frac{\left(g_{ik}-\bar{g}_i\right)'V_\beta^{-1}\left(g_{ik}-\bar{g}_i\right)}{2\sigma_{ik}^2}\right),\ k\in K_i,\ i=1,...,N, \tag{1.7}$$

$$V_\beta=\sigma_\beta^2,$$

where $\sigma_\beta^2\cdot\sigma_{ik}^2$ chanracterizes the variance of g_{ik} conditional on σ_{ik}^2.

1.3.1 Model estimation

We estimate the model under Bayesian framework. The Appendix B presents details of priors of parameters and how to implement the MCMC algorithm.

Location of Breaks

Using quarterly data from 1959Q2 to 2020Q3, the model identified 4 breaks, which coincide with major economic events. Panel (a) of Figure 1.4 shows the posterior locations of breaks. The first break is around 1972 coinciding with the first oil crisis. The second break is around 1981 when the Fed tightened monetary policy and shifted its goal to maintain low inflation. The third break is around 2000 coinciding with the dotcom crisis. The most recent break is in 2020 due to the Covid-19 pandemic.

To illustrate the statistical power from using multiple goods, we apply the same approach to aggregate consumption only. Panel (b) of Figure 1.4 shows the posterior of locations of breaks. In contrast to posteriors based on multiple goods (Panel (a)), we only identify 2 breaks. The estimation errors in locations of breaks are much larger when using aggregate consumption alone. Except for year 2020, posterior probability that a break happens in each year is below 0.5.

We examine the pervasiveness of breaks in Figure 1.5 . Panel (a) of Figure 1.5 presents the expected number of goods (including common factor) affected by each break. Each break affects 6-10 series. Panel (b) of Figure 1.5 presents a heat map of the posterior probability of break across years and types of goods. Certain types of goods are more prone to breaks than others: energy, health care, transportation, and food and accommodation services are affected by all of the 4 breaks, while financial services is only affected by the break in 2000. Notably, the break in 2000 only affects services. The recent Covid-19 break is the most pervasive: with the exception of housing and utilities and financial services, all major types of goods are affected.

Regime parameters

We define that a series is hit by a break if the posterior probability exceeds 0.5. Conditional on the location of breaks, Table 1.2 presents estimates of g_{it} of each regime at an annual rate. If a series is not affected by a break, the corresponding values are left blank. The top panel presents the estimates, and corresponding standard errors are in parenthesis. The bottom panel presents summary statistics including the mean absolute changes in parameters and the three most-affected types of goods by each break.

For the first three breaks, the mean absolute changes in regimes' expected growth rates are around 1 to 2 percent. The 2020 break is much larger in magnitude, with mean absolute changes of 9.89%. Breaks have heterogeneous effects on the expected growth rate of different goods. Energy goods and transportation services are affected most by the break in early 1970s and early 1980s: the expected growth of energy goods is 2.75% before 1972Q4, drops to -2.06% in the regime from 1972Q4 to 1981Q1, and then bounces back to 0.36%. Expected growth in consumption of financial services is affected most by the break in 2000. In the regime from 1959Q2 to 1999Q4, expected consumption growth in financial services is at 4.21%, while, post 2000Q1, the expected growth rate is -0.65%. The Covid-19 break sees unprecedented shocks to transportation services, recreation services, and food services and accommodation due to lockdowns and fear of contracting the disease. Expected consumption growth rate of the three types of goods is -28.8%, -25.97%, and -8.36%, respectively. The estimates are based on three quarters of data in 2020 and the resulting estimation errors are large. But the sheer magnitude of the shock is striking and demonstrates that changes in parameters can be large and abrupt.

Table 1.3 presents estimates of σ_{ik} and σ_{fk} for each regime. The volatility estimates are at an annual rate by multiplying quarterly estimates by a factor of 2. For the first three breaks, the mean absolute changes in regimes' volatility are in the range of 0.5 to 1.5 percent. The 2020 break is much larger in magnitude with a mean absolute change of 15.53 percent. As expected, consumption growth in energy goods is affected the most in the breaks of 1972 and 1981. The

volatilty of energy consumption growth increased dramatically to 7.09% in the regime from 1972 to 1981 from 3.7 in the previous regime and then decreased to 3.72% after the break in 1981. The break in 1981 is a negative shock to volatility for all series affected. The result is consistent with the effect of Great Moderation as documented in Stock and Watson (2002) and the finding of Lettau, Ludvigson and Wachter (2008) that consumption growth volatility is especially low in the 1990s, leading to higher price dividend ratios. The break in 2020 is an unprecedented positive shock to volatility for all series affected. Due to lockdowns, the types of goods affected most are transportation and recreation services whose volatility increases to 48.8 and 39.64, respectively. The caveat is the volatility estimates of the last regime are based on only 3 quarters of data in 2020, and we don't know how long the regime will last.

We proceed to analyze the combined effect of major components' breaks on aggregate consumption. We calculate expected growth in aggregate consumption as

$$g_t = \sum_{i=1}^{N} w_{it} g_{it}, \tag{1.8}$$

where w_{it} is the expenditure share of goods i. The volatility of aggregate consumption growth is calculated as the combined effect of idiosyncratic components and the common component

$$\sigma_t = \sqrt{w_{it}^2 \sigma_{it}^2 + \left(\sum_{i=1}^{N} w_{it} \gamma_{i0}\right)^2 \sigma_{ft}^2 + \left(\sum_{i=1}^{N} w_{it} \gamma_{i1}\right)^2 \sigma_{ft-1}^2}. \tag{1.9}$$

Figure 1.5 plot time series of g_t and σ_t before 2020. The estimates are at annual rates by multiplying the quarterly estimates by a factor of 4 for mean and 2 for volatility. The parameters of the regime starting in 2020Q1 is not shown given the values are extreme ralative to paths of historical estimates. g_t is highest in the early 1970s, reaching around 2.2%, and decreased to around 1.3% in the regime between 2000 and 2019. g_t is lowest in the latest regime starting from 2020Q1, reaching -1.5%. The volatility of aggregate consumption growth is highest in the 1970s,

reaching 1.05%, and then decreased following breaks in 1980 and 2000. In the regime between 2000 and 2019, the volatility of aggregate consumption growth is 0.75%. Covid-19 is a huge positive volatility shock to aggregate consumption, σ_t is highest in the regime that starts from 2020Q1, reaching 4%.

In summary, the model identifies infrequent, pervasive and persistent breaks in consumption growth of major types of goods. The impact of the breaks on consumption growth is highly uncertain and heterogeneous across consumption goods, leading to unevenly distributed parameter uncertainty across different goods.

1.3.2 Comparison with long-run risk model in Bansal and Yaron (2004) and rare disaster model in Barro (2006)

Comparison with long-run risk

Similar to our break model, the long-run risk model proposed in Bansal and Yaron (2004) also characterizes a persistent component in expected consumption growth. Bansal and Yaron (2004) model aggregate consumption growth as

$$\Delta c_{t+1} = \mu_t + x_t + \sigma_t \eta_{t+1}, \tag{1.10}$$

$$x_{t+1} = \rho x_t + \varphi_e \sigma_t e_{t+1}, \tag{1.11}$$

$$\sigma^2_{t+1} = \bar{\sigma}^2 + \nu\left(\sigma^2_t - \bar{\sigma}^2\right) + \sigma_w w_{t+1}, \tag{1.12}$$

and calibrate monthly persistence ρ at 0.979, implying a half-life of around 3 years. Because the predictable component x_t is highly persistent and involves high observation error σ_t, it is difficult to measure ρ accurately. Using quarterly data, Schorfheide, Song and Yaron (2018) estimate that shocks to x_t have a half-life of around 5 months. Hansen, Heaton and Li (2008) show the persistence of long-run component is not high enough to explain the risk premium puzzle under

reasonable risk aversion. Beeler and Campbell (2012) show that the variance ratios of long-run consumption growth over short-run consumption growth is lower than the implication of Bansal and Yaron (2004).

To compare long-run risk model with the break model, we revise the equation (1.1) to introduce a the long-run risk, x_t, into the break model:

$$\Delta c_{it} = g_{it} + \gamma_{i0} f_t + \gamma_{i1} x_{t-1} + \sigma_{it} \varepsilon_{it}, \ i = 1, ..., N. \tag{1.13}$$

f_t is the short-run common component, and x_t is the long-run component driven by

$$x_t = \rho x_{t-1} + \varepsilon_{x,t},$$

where $\varepsilon_{x,t}$ is i.i.d. standard normal. The regime parameters $\left(g_{it}, \sigma_{it}, \sigma_{ft}\right)$ follow the same dynamic as in equations (1.2) and (1.3). The estimated break locations of model (1.13) is the same to the estimates of (1.1). With a uniform prior of ρ, the posterior mean of ρ is 0.81 at a quarterly rate, with 95% confidence interval of $(0.71, 0.90)$. The estimate implies a half-life of around 3 quarters, with 95% confidence interval of $(1.9, 6.6)$ quarters. Hence, the persistent component in expected consumption growth features infrequent and discontinuous breaks, rather than continuous long-run risk.

Comparison with rare disaster model

The rare disaster model characterizes sudden and large drops in consumption. The theory was first proposed by Rietz (1988). Barro (2006) and Barro and Ursúa (2008) use an NBER-style peak-to-trough measurement of the sizes of macroeconomic contractions: "Starting from the annual time series, proportionate decreases in C and GDP were computed peak to trough over one or more years, and declines by 10% or greater were considered." Based on the definition, US experienced one rare disaster after 1929: a trough in 1933 with consumption declining 21%. As

shown by a strand of literature (Mehra and Prescott, 1988; Cecchetti, Lam and Mark, 2000; Ju and Miao, 2012), the magnitude and probability of US disasters is not large enough to account for asset-pricing puzzles.

The break model complements the rare disaster model by incoporating persistent effects of macroeconomic events. Gordon (2017) states "Reseach conducted half a century ago concluded that American growth was steady but relatively slow until 1920, when it began to take off. Scholars struggled for decades to identify the factors that caused the productivity growth to decline significantly after 1970." The year 1973 is generally accepted as the starting date of a pronounced slowdown in productivity growth[2]. Using the Great Depression and the 1973 Oil Crisis, we separate the sample period into three parts: 1890-1929, 1934-1972, and 1980-2019[3]. The average consumption growth during the three periods are 2.04%, 2.66% and 1.48% respectively. Such large and persistent shifts in expected consumption growth have significant implications on asset prices, as will be shown in the next few sections.

1.4 Implications for market returnsn

The following two sections explore the asset pricing implications of infrequent, persistent shocks to consumption growth in a setting of investors with Epstein-Zin preferences and parameter uncertainty. We consider both implications on aggregate (market) prices and cross-sectional asset returns. For computation reasons, we simplify our model to a single good when analyzing aggregate returns and use a multiple goods model when explaining cross-sectional risk premium.

[2]See, e.g., Jorgenson, Ho and Samuels (2014)

[3]The annual consumption data from 1989 to 1929 is from Gordon (2007)

1.4.1 Preference, information and asset prices

Investors' utility is governed by Epstein-Zin preferences

$$V_t = \left\{ (1-\beta) C_t^{1-\rho} + \beta R_t \left(V_{t+1}\right)^{1-\rho} \right\}^{\frac{1}{1-\rho}}, \tag{1.14}$$

where the function R_t function characterize investors' risk aversion

$$R_t \left(V_{t+1}\right) = E_t \left\{ V_{t+1}^{1-\gamma} \right\}^{\frac{1}{1-\gamma}}. \tag{1.15}$$

Let π_t denote the price level of consumption bundle C_t, the pricing kernel satisfies

$$M_{t+1} = \beta \left(\frac{C_{t+1}}{C_t} \right)^{-\rho} \left(\frac{V_{t+1}}{R_t \left(V_{t+1}\right)} \right)^{\rho-\gamma}. \tag{1.16}$$

For computation reasons, we ignore short run serial correlations of aggregate consumption growth and focus on the break component, which is economically more significant when investors have a recursive preference. The log consumption consumption growth, denoted as $\Delta c_t = \log C_t - \log C_{t-1}$, is governed by

$$\Delta c_t = g_k + \sigma_k \varepsilon_t, \text{ if } \tau_k \leq t < \tau_{k+1}. \tag{1.17}$$

We model log dividend growth of aggregate market as leveraged aggregate consumption growth plus idiosyncratic component

$$\Delta \log D_t = \bar{g}_d + L(\Delta c_t - \bar{g}) + \sigma_{dk} \varepsilon_{dt}, \text{ if } \tau_k \leq t < \tau_{k+1}. \tag{1.18}$$

The error terms ε_t and ε_{dt} are i.i.d. standard normal.

At each period, a break happens with probability λ. Parameters, $(g_k, \sigma_k, \sigma_{dk})$, are randomly

generated when regimes shift. $(g_k - \bar{g})/\sigma_k$ is independently drawn from a truncated normal distribution $N(0,\Sigma_0)$ with support $\left[\underline{G},\bar{G}\right]$. The truncation is introduced to ensure price-dividend ratio converge. For computation reasons, we simplify the volatility regime to 2 states: the scaled conditional volatility $\sigma_k/\bar{\sigma}_c$ and $\sigma_{dk}/\bar{\sigma}_d$ in equations (1.17) and (1.18) are generated from a discrete distribution of two outcome $\left\{\sigma_{Low},\sigma_{High}\right\}$ with probability π_{Low} and $1-\pi_{Low}$.

We assume investors observe historical consumption growth, the timing of breaks, and volatility. Without the assumption of observing the timing of breaks, investors' posterior distribution has infinite dimensions, which makes calculation of price-dividend ratios computationally infeasible. Investors learn g_{Kt} by Bayes rule from the data. Investors' information set is defined as $F_t = \{c_\tau, \tau_k, \sigma_k, \sigma_{dk}, |\tau \leq t, \tau_k \leq t\}$. The posterior of g_k follows a truncated normal with $N\left(\mu_t, \Sigma_t \sigma_k^2\right)$ when $\tau_k \leq t < \tau_{k+1}$. The dynamics of parameter uncertainty Σ_t is driven by

$$\frac{1}{\Sigma_t} = \begin{cases} \frac{1}{\Sigma_0} & \text{if } t \in \{\tau_1, ..., \tau_K\}, \\ \frac{1}{\Sigma_{t-1}} + 1 & \text{otherwise.} \end{cases} \tag{1.19}$$

The μ_t is driven by

$$\mu_t = \begin{cases} \bar{g} & \text{if } t \in \{\tau_1, ..., \tau_K\}, \\ \mu_{t-1} + \Sigma_t (\Delta c_t - \mu_{t-1}) & \text{otherwise.} \end{cases}$$

Let v_t denote the log value-consumption ratio $v_t \equiv \ln(V_t/C_t)$. From the definition of Epstein-Zin preference (1.14) and (1.15), we have

$$v_t = \frac{1}{1-\rho} \ln\left\{(1-\beta) + \beta \exp\left[(1-\rho) Q_t\left(v_{t+1} + \Delta c_{t+1}\right)\right]\right\}, \tag{1.20}$$

where $Q(x_{t+1}) \equiv \ln(R_t(\exp(x_{t+1})))$. v_t is a function of $v(\mu_t, \Sigma_t, \sigma_{Kt})$. The pricing kernel can be

expressed as

$$M_{t+1} = \frac{\beta \exp\left[-\gamma \Delta c_{t+1} + (\rho - \gamma) v_{t+1}\right]}{\exp\left[(\rho - \gamma) Q_t \left(\Delta c_{t+1} + v_{t+1}\right)\right]}. \tag{1.21}$$

Market portfolio price P_t can be expressed as

$$P_t = E_t \left(M_{t+1} \left(D_{t+1} + P_{t+1}\right)\right). \tag{1.22}$$

Let φ_t denote price dividend ratio $\frac{P_t}{D_t}$. φ_t follows

$$\varphi_t = E_t \left(M_{t+1} \left(1 + \varphi_{t+1}\right) \exp\left(\Delta \ln D_{t+1}\right)\right). \tag{1.23}$$

We are able to solve the $\varphi(\mu_t, \Sigma_t, \sigma_{Kt})$ with numerical methods. Appendix A presents the details of numerical methods.

1.4.2 Model Calibration

We calibrate the model parameters and study its performance in reproducing moments of market returns as observed in the data.

The model is solved with time unit of one quarter. To match the mean and volatility of the annual log consumption growth from 1929 to 2019 (1.8% and 2.2% respectively), we set $\bar{g}$ to 0.45% and $\bar{\sigma}_c$ to 1. The empirical results in Section 1.3 shows that parameter σ_β^2 in equation (1.7) is around 0.1. Hence we set Σ_0 equal to 0.1. The probability of regime shifts, λ, is set to 1.67% to match the average duration of regimes which is around 15 years. Based on the time series of volatility of aggregate consumption growth presented in Figure 1.5, we find two high-volatility regimes covering 1959-1981, which constitutes a third of the sample. Hence, we set $\pi_{Low} = 2/3$ with $\sigma_H = 1.13$ while $\sigma_L = 0.91$ such that the volatility of high volatility regime is 25% higher than during the low-volatility regime, while the unconditional volatility is still $\bar{\sigma}_c$. Following Bansal and Yaron (2004), we set the leverage parameter to $L = 3$. The unconditional

quarterly volatility of idiosyncratic component in dividend growth is set to 4.5 percent to match the unconditional volatility in dividend growth.

The preference parameters are set consistent with recent work in explaining market risk premiums (Collin-Dufresne, Johannes and Lochstoer 2016 and Bansal and Yaron, 2004). We use EIS equals to 2 and set the risk aversion parameter γ to be 6 in our benchmark calibrations. As in Bansal and Yaron (2004), we set β to 0.994. Table 1.4 presents list of parameters for the benchmark break model.

1.4.3 Unconditional moments

We test our model by evaluating model-implied moments of asset returns. Table 1.5 analyzes the relation between the parameters governing consumption dynamics and key asset price moments by perturbing the frequency and magnitude of breaks around the ones used in the benchmark break model. The first row presents average log excess market returns, market return volatility, average real risk-free rates, and the Sharpe ratio estimated from the data. The second row shows asset price moments implied by our benchmark break model. Our model fits the data well, with model-implied risk premium of 5.41% and model-implied market return volatility of 17.82%. The model-implied price of risk is 0.51 higher than the Sharpe ratio of market portfolio, 0.30.

The third panel in Table 1.5 analyzes the effect of varying magnitudes of breaks while holding λ constant at 0.017 (average regime length of 15 years), as suggested by the empirical analysis. Larger break magnitude in expected consumption growth leads to a higher risk premium, price of risk, and volatility in market returns and lower level of risk free rates.

The bottom panel in Table 1.5 holds $\sqrt{\Sigma_0}$ constant to 0.3 and varies the magnitude of λ from 0.05 to 0.012, which corresponds to the expected length of regimes of 20 years and 5 years. We find that the risk premium and the price of risk increases with the persistence of shocks, consistent with investors' aversion to long-run risk.

Table 1.6 presents the effect of preference parameters γ and ρ on long-term moments while holding parameters of consumption dynamics to constant. The top panel holds EIS to 2 and varies risk aversion from 4 to 8. Rising risk aversion increases price of risk and risk premium but decreases volatility of returns. The bottom panel repeats the similar exercise with EIS fixed to 1.5, which leads to higher risk free rate and lower risk premium. Higher EIS reduces risk premium by lower risk exposure of market returns: for a large value of the EIS, higher growth prospects decrease price-dividend ratios through a higher discount rate.

To analyze the source of price of risk, we decompose the log pricing kernel M_t in (1.21) into the idiosyncratic component of consumption growth, sudden changes to belief when observing new breaks, gradual changes in investors' beliefs due to learning:

$$\begin{aligned}\Delta \log M_t - \mathrm{E}_{t-1}\left(\Delta \log M_t\right) = & \underbrace{-\gamma(\Delta c_t - \mathrm{E}_{t-1}\left(\Delta c_t\right))}_{\text{consumption growth}} + \underbrace{(\rho-\gamma)\left[\bar{v} - \mathrm{E}_{t-1}\left(v_t\right)\right]\mathbb{1}(t \in \{\tau_1, ..., \tau_K\})}_{\text{sudden structural break}} \\ & + \underbrace{(\rho-\gamma)\left[\Delta v_t - \mathrm{E}_{t-1}\left(\Delta v_t\right)\right]\mathbb{1}(t \notin \{\tau_1, ..., \tau_K\})}_{\text{gradual learning}}. \end{aligned} \tag{1.24}$$

Variance of each component can be estimated by simulation. The first term is same to the log pricing kernel under power utility without breaks and the resulting model uncertainty. Under parameters of the benchmark break model, the variance of first term is only 11% of the variance of log pricing kernel. The second term accounts for 2% of total variance in pricing kernel. The number seems small because breaks are infrequent events and investors can't observe the each regime's conditional mean of consumption growth. Breaks generate most of the variance in pricing kernel through subsequent learning: 87% of the variance of pricing kernel comes from the changes in investors' beliefs due to learning and its covariance with the consumption growth.

1.4.4 Predictability of excess returns and consumption growth

Our consumption model implies that the price-dividend ratio is driven by investors' beliefs about future consumption growth. It is natural to test the model by evaluating the ability of

price-dividend ratio in predicting excess returns and consumption growth.

As documented in Campbell and Shiller (1988), log dividend price ratios predict multi-horizon returns. Let $r^{ex}_{t+1:t+H}$ denotes $r^{ex}_{t+1:t+H} = \sum_{h=1}^{H} r^{ex}_{t+h}$. We estimate the predictive regression

$$r^{ex}_{t+1:t+H} = \alpha_H + \beta_H pd_t + \varepsilon_{t+1:t+H} \tag{1.25}$$

of different values of H. The top panel in Table 1.7 presents regression results of long-horizon excess return forecasts of horizons 1 quarter, 1 year, 2 years, 3 years, and 5 years. Our model captures positive relation between dividend yields and risk premium: The magnitude of slope coefficients and R^2 increases with forecast horizons as in the data.

Beeler and Campbell (2012) find little predictive power of price-dividend ratios in predicting consumption growth. Let $\Delta c_{t+1:t+H}$ denotes $\Delta c_{t+1:t+H} = \sum_{h=1}^{H} \Delta c_{t+h}$. The bottom panel in Table 1.7 presents estimates of the predictive regression

$$\Delta c_{t+1:t+H} = \alpha_H + \beta_H pd_t + \varepsilon_{t+1:t+H} \tag{1.26}$$

of different values of H. Consistent the data, the break model generates little predictability in consumption growth.

To explore why price the dividend ratio is able to predict excess returns but not consumption growth, we plot model implied price-dividend ratios, risk premia, and volatility of market returns against state variables, namely investors' beliefs of expected consumption growth rate, μ_t, and uncertainty around it, Σ_t. Figure 1.7 presents the relation between the price dividend ratio and various state variables. The left panel plots price-dividend ratios against μ_t while fixing Σ_t at 3.5%, which is roughly 5 years after a break. The red line is the low volatility regime, and the blue line is the high volatility regime. Both lines increase with expected conditional consumption growth, suggesting procyclical price dividend ratios. The right panel plots price dividend ratio against $\sqrt{\Sigma_t}$ holding μ_t to 1.8 percent. For both volatility regimes, price dividend

ratio decreases with parameter uncertainty.

Figure 1.8 presents the relation between risk premium, return volatility, and state variables. Panel (a) plot risk premia against μ_t (left) and $\sqrt{\Sigma_t}$ (right). The risk premium is non-monotonic in μ_t and increasing in $\sqrt{\Sigma_t}$. This is consistent with our hypothesis that both the price of risk and risk exposures are increasing in $\sqrt{\Sigma_t}$. Panel (b) plots conditional volatility against μ_t (left) and $\sqrt{\Sigma_t}$ (right). It is evident that volatility of equity returns is decreasing in expected consumption growth and increasing in uncertainty.

Figure 1.7 and 1.8 suggest that predictable excess returns and unpredictable consumption growth can be explained by the dominant effect of parameter uncertainty on price-dividend ratios and the price of risk, in line with theoretical implications of Jahan-Parvar and Liu (2014) and Collin-Dufresne, Johannes and Lochstoer (2016). The driver behind the negative relation between risk premium and price dividend ratio is that high parameter uncertainty decreases the price-dividend ratio while increasing the risk premium. The explanation for little predictability in consumption growth is that while the price dividend ratio is increasing in expected consumption growth, it also contains information of parameter uncertainty, as demonstrated in Figure 1.8. As a result, the price-dividend ratio is a noisy predictor of expected consumption growth.

1.4.5 The Covid shock

The Covid-19 pandemic is a sharp example of breaks and offers a good test of the model: In 2020Q2, log aggregate consumption growth is -11%, with 44%, 62% and 41% decline in log consumption of transportation, recreation services, and food services and accomodation. The resulting posterior of a break happening in 2020Q1 is 1 and all types of goods, with the exceptions of other nondurables and houting and utility services, are affected by the break.

Figure 1.9 plots forecasts of excess returns (top panel) and market volatility (bottom panel) in forecast horizons of 1 to 5 years. The forecasts are made at 2020Q3, assuming that a break happened on 2020Q1 and the mean of investors' beliefs of consumption growth is -1.5% in

annual units and the regime is of high volatility. The model predicts a risk premium of 20% in the immediate year after the break and gradually decreases to 12% in five years. This is consistent with the 30% plunge of S&P500 index in March 2020 and the subsequent recovery of loss. The model predicts market volatility of 65% in the immediate year after the break and gradually decreases to 50% in five years. The short term forecast seems in line with 90% and 40% realized volatility in March and April 2020. Over the long-run, the model-implied effects of breaks on market volatility seem too persistent relative to the data. In practice, investors may access more information such that uncertainty is resolved sooner than the model would imply.

1.5 Implications for the cross-section of returns

To explore asset pricing implications of learning from multiple goods, we assume investors' total consumption expenditure is the aggregation of two types of goods:

$$C_t = C_{1t} + C_{2t}, \tag{1.27}$$

where C_{it} is expenditure on goods i at time t. Combined with the utility function specified in (1.14) and (1.15), investors' utility is a function of aggregate consumption expenditure, and include the CES utility function used in Yogo (2006) and Piazzesi, Schneider and Tuzel (2007) as a special case.

The consumption dynamics follow:

$$\Delta c_{it} = g_{ik} + \sigma_{ik}\varepsilon_{it},\ i = 1,2 \text{ if } \tau_k \leq t < \tau_{k+1} \tag{1.28}$$

where ε_{it} are i.i.d standard normal.

To solve the model in close form, we assume EIS $= 1$ and breaks are non-recurring. Let

$N\left(\mu_{it}, \Sigma_{it}\sigma_{it}^2\right)$ denote investors posterior of g_i, $i = 1, 2$. Investors' beliefs evolve as

$$\begin{aligned} \Delta\mu_{it} &= \Sigma_{it}\left(\Delta c_{it} - \mu_{it-1}\right), \\ \frac{1}{\Sigma_{it}} &= \frac{1}{\Sigma_{it-1}} + 1. \end{aligned} \tag{1.29}$$

Following the same method in Collin-Dufresne, Johannes and Lochstoer (2016), it can be proven that the log value-consumption ratio v_t takes the form $v_t = \frac{\beta}{1-\beta}\left(\sum_{i=1}^{2}\alpha_{it}\Delta\mu_{it}\right) + f\left(\Sigma_{1t}, \Sigma_{2t}\right)$, where α_{it} is the expenditure share of each type of good and $f(\cdot)$ is a deterministic function of model uncertainty. We have the following propostion of the pricing kernel.

Proposition 1.5.1. *When the break is nonrecurring* $(\lambda = 0)$ *and investors' EIS=1, investors' log pricing kernel* m_{t+1} *satisfies*

$$m_{t+1} - E_t\left(m_{t+1}\right) = -\left(\Delta c_{t+1} - E_t\left(\Delta c_{t+1}\right)\right) + \frac{(1-\gamma)\beta}{1-\beta}\left(\sum_{i=1}^{2}\alpha_i\Delta\mu_{it}\right), \tag{1.30}$$

where α_i *is expenditure share of each type of good.*

The last term characterizes effects of variation in investors' beliefs on the pricing kernel. As shown in equation (1.29), investors update their beliefs more actively when parameter uncertainty is higher. Across different goods, more attention is allocated to goods of higher parameter uncertainty whose signals contain more information about expected consumption growth. As a result, the pricing kernel is tilted toward types of goods with higher parameter uncertainty. To sharpen the intuition, equation (1.30) can be reformulated to

$$\begin{aligned} m_{t+1} - E_t\left(m_{t+1}\right) = &- \underbrace{\left(1 + \frac{(\gamma-1)\beta}{1-\beta}\cdot\frac{\Sigma_{1t}+\Sigma_{2t}}{2}\right)\left(\Delta c_{t+1} - E_t\left(\Delta c_{t+1}\right)\right)}_{\text{CCAPM}} \\ &- \underbrace{\frac{(\gamma-1)\beta}{1-\beta}\frac{\Sigma_{1t}-\Sigma_{2t}}{2}\left[\alpha_1\left(\Delta c_{1t} - \mu_{1t-1}\right) - \alpha_2\left(\Delta c_{2t} - \mu_{2t-1}\right)\right]}_{\text{Consumption fragility}}. \end{aligned} \tag{1.31}$$

The first term in equation (1.31) is the aggregate consumption factor that features in the conventional CCAPM proposed by Breeden (1979). When $\Sigma_{1t} > \Sigma_{2t}$, the second factor longs goods of higher parameter uncertainty and shorts goods of lower parameter uncertainty. It can be viewed as a factor of consumption fragility which over-weights types of goods that are more sensitive to breaks. The model implies the fragility factor has positive price of risk.

1.5.1 Empirical test of cross-sectional implications

In this section, we test the implication that the consumption fragility factor have a positive price of risk. We use the standard deviation of g_{it} conditional on data up to time t to measure parameter uncertainty. Figure 1.10 presents the heat map of standard deviations of g_{it} over time, starting in 1969Q2. On average, consumption growth in clothing, energy, transportation, and financial service have higher parameter uncertainty, while food, housing and utility service, and health care have lower parameter uncertainty. Consistent with estimates of locations of breaks in Section 1.3.1, parameter uncertainty in energy goods is high the 1970s, while parameter uncertainty in financial services is high in the early 2000s.

To test the implication, at each time t, we separate the 11 types of goods into 3 groups based on parameter uncertainty. The top group contains 3 consumer goods that have the largest uncertainty about g_{it}; the bottom group contains 3 consumer goods that have the least uncertainty about g_{it}; while the medium group contains the remaining 5 goods. Let $e_{H,t}$, $e_{M,t}$ and $e_{L,t}$ denote the consumption growth innovations of goods of high, medium, and low parameter uncertainty groups: $e_{(\cdot),t+1} = \Delta c_{(\cdot)t+1} - E_t\left(\Delta c_{(\cdot)t+1}\right)$, where $\Delta c_{(\cdot)t}$ is the expenditure-share weighted average of log consumption growth in each group. We use expenditure shares of the three groups $\alpha_{H,t}$, $\alpha_{M,t}$, $\alpha_{L,t}$ as proxy for corresponding α_is in (1.33).

Equation (1.31) suggests a two-factor model of aggregate consumption innovation e_t and relative consumption innovation $e_{fragile,t+1} \equiv \alpha_{H,t} e_{H,t} - \alpha_{L,t} e_{L,t}$. $e_{fragile,t+1}$ emphasizes the components in aggregate consumption that are more subject to breaks. Investors' learning and

their aversion to long-run uncertainty implies its price of risk is positive and is increasing in the difference of parameter uncertainty across the two groups. The fragility factor has a correlation of -0.1 with the aggregate consumption factor.

Let R_j be the return of asset j and r_j be the log return. Under the approximation in Campbell et al. (2018), the pricing equation is

$$E_t\left[R_{j,t+1} - R_{0,t+1} + \left(r_{j,t+1} - r_{ft+1}\right)\left(m_{t+1} - E_t\left(m_{t+1}\right)\right)\right] = 0. \tag{1.32}$$

Based on equation (1.31), our model implies positive λ_c and $\lambda_{fragile}$ in the pricing equation:

$$E_t\left[R_{j,t+1} - R_{0,t+1} - \left(r_{j,t+1} - r_{0t+1}\right)\left(\lambda_c e_t + \lambda_{fragile} e_{fragile}\right)\right] = 0. \tag{1.33}$$

We can further test the implication that price of risk of $e_{fragile}$ is increasing in the difference of parameter uncertainty across the two groups by adding a facor consisting of interactions of $e_{fragile}$ and a proxy for uncertainty difference between the top and bottom group. Let $\sigma_{diff,t}$ denotes the difference between average parameter uncertainty across the two groups and M_{diff} denotes its sample median. We use the indicator function $\mathbb{1}(\sigma_{diff,t} > M_{diff})$ as the proxy for uncertainty difference. Hence, the model implies positive values of λ_c, $\underline{\lambda}_{fragile}$ and $\bar{\lambda}_{fragile}$ in the pricing equation

$$E_t\left[R_{j,t+1} - R_{0,t+1} - \left(r_{j,t+1} - r_{0t+1}\right)\left(\lambda_c e_t + \underline{\lambda}_{fragile} e_{fragile} + \bar{\lambda}_{fragile}\mathbb{1}(\sigma_{diff,t} > M_{diff}) e_{fragile}\right)\right] = 0. \tag{1.34}$$

Our test assets consist of the 25 Fama-French portfolios sorted by size and book-to-market equity, 10 momentum portfolios sorted by prior 12 months returns, and 10 industry portfolios. The data is from Kenneth French's website. Our reference asset return $R_{0,t}$ is the risk-free rate. Table 1.8 reports estimates of the price of risk of the CCAPM, the consumption fragility factor $e_{fragile}$, and the interaction factor. Estimation is by two-step GMM with HAC standard errors of lag 1. The price of risk of aggregate consumption growth is positive and significant for both CCAPM and extended CCAPM. The magnitude is reduced to 121 from 183 after including

$e_{fragile}$. The price of risk of $e_{fragile}$ is 682 and is significantly positive. The third row of Table 1.8 presents the conditional price of risk for $e_{fragile}$. The conditional price of risk of the fragility factor is roughly 40% higher when $\sigma_{diff,t}$ is above its median relative to other periods. The mean absolute pricing error is 1.30% for CCAPM and 1.29% for CAPM, while the value is reduced to 0.66% for the extended CAPM with the consumption fragility factor. Similar improvement is found in the R^2 of pricing errors benchmarked against CAPM. The CCAPM generates R^2 of -0.11 while the extended CAPM generates R^2 of 0.71, a significant improvement relative to CCAPM and CAPM.

To investigate the source of the fragility factor's price of risk, we examine factor loadings of portfolios sorted by size, book-to-market equity ratio, and returns in past 12 months as reported in Table 1.9. The three panels from left to right reports size, value, and momentum portfolios respectively. The first column of each panel reports loadings of the aggregate consumption factor, and the second column of each panel reports loadings of the consumption fragility factor. The third column of each panel reports average excess returns. Based on our sample from 1969Q1 to 2019Q4, the size, value and momentum premium are -0.2%, 0.8%, and 3.0% at quarterly rates. The left panel shows that average excess returns are increasing in book to market equity ratios and past 12 months returns, while the size premium is small in our sample. Factor loadings on aggregate consumption growth are increasing in book-to-market equity ratios, while loadings on the fragility factor is not monotonic. The right panel shows that portfolios of higher momentum have higher exposure to the consumption fragility factor: the the fifth portfolio has beta in relative consumption growth of 12.6, while the first portfolio has beta 1.5 and their difference is statistically significant.

In general, the aggregate consumption beta accounts for value premium, while the consumption fragility beta accounts for momentum premium. The empirical evidence is consistent with findings of Liu and Zhang (2008) that macroeconomic factors like industrial production can explain part of momentum premium.

1.6 Conclusion

Using a panel of disaggregate consumption goods, we identify infrequent and persistent breaks to consumption growth dynamics. The impact of the breaks is very heterogeneous across consumption goods, leading to unevenly distributed parameter uncertainty across different goods.

Having demonstrated that consumption growth is neither stable nor continuous, we build a Lucas tree model with breaks in consumption growth dynamics. The resulting parameter uncertainty combined with investors' learning leads to persistent changes in investors' beliefs. Because investors with recursive preferences are averse to persistent sources of consumption growth risk, breaks generate high price of risk embedded in the market portfolio. Parameter uncertainty plays a key role in driving variation in price-dividend ratios, which in turn helps to forecast excess returns on the market portfolio.

Differences in parameter uncertainty across different consumption goods imply that investors learn more about expected growth in the current regime from consumption goods whose parameters are most sensitive to shifts in the underlying economic state. The resulting pricing kernel is tilted away from the aggregate consumption growth that features in the conventional CCAPM. We account for the deviation from the CCAPM by a consumption fragility factor which is the consumption growth of goods with high parameter sensitivity relative to goods with low parameter sensitivity. Consistent with our hypothesis, we identify a significantly positive risk premium in the consumption fragility factor using a cross-section of returns on portfolios sorted on different attributes. The price of risk of the new factor is higher when the difference in parameter sensitivity between the two types of goods is higher. Exposures to the fragility factor is increasing in the portfolios' prior 12-month returns suggesting that the consumption fragility factor helps explain the momentum premium.

In practice, investors learn from an abundance of signals they observe in addition to data on consumption growth. In future work, we intend to use the approach introduced in this paper

to explore how pervasive breaks across different economic variables can help explain how asset prices respond to news on macroeconomic and financial variables.

1.7 Acknowledgements

Chapters 1, in full, is currently being prepared for submission for publication of the material. Ritong Qu, the dissertation author, is the primary investigator and author of this material.

Table 1.1: Summary statistics of log consumption growth

Quarterly data: 1959Q2 to 2020Q3						
	Mean(%)	Standard Deviation(%)	Skewness	Kurtosis	Autocorrelation	Weight(%)
Aggregate Consumption	1.81	1.94	-5.22	87.92	-0.16	
Food	0.66	1.89	1.47	9.35	-0.01	13.51
Clothing	2.42	4.84	1.67	45.49	-0.12	6.14
Energy	-0.05	4.67	-2.12	36.79	-0.17	4.28
Other nondurable	2.37	1.94	-0.19	2.65	0.24	9.27
Housing & Utility	1.81	1.15	-0.29	-0.25	0.11	21.13
Health care	2.43	3.75	-2.56	70.82	-0.15	13.89
Transportation	1.65	7.01	-7.01	113.23	-0.17	3.77
Recreation	2.43	9.49	-8.24	131.93	-0.25	3.57
Food & Accommodation	1.14	6.80	-4.98	101.93	-0.25	7.59
Financial Service	2.64	3.01	0.42	0.91	0.21	7.30
Other Service	1.41	3.45	-6.28	77.08	-0.02	9.56
Quarterly data: 1959Q2 to 2019Q4						
	Mean(%)	Standard Deviation(%)	Skewness	Kurtosis	Autocorrelation	Weight(%)
Aggregate Consumption	1.92	0.87	-0.22	1.27	0.45	
Food	0.57	1.69	0.33	2.22	0.09	13.57
Clothing	2.51	2.76	-0.19	0.58	0.08	6.17
Energy	0.11	3.16	-1.88	12.23	0.02	4.31
Other nondurable	2.28	1.85	-0.56	2.04	0.30	9.26
Housing & Utility	1.82	1.15	-0.30	-0.25	0.12	21.11
Health care	2.59	1.76	0.30	2.22	0.43	13.83
Transportation	2.11	2.44	-0.78	1.54	0.57	3.77
Recreation	3.11	2.55	-0.71	5.07	-0.10	3.56
Food & Accommodation	1.52	2.16	0.10	0.32	0.11	7.59
Financial Service	2.67	3.02	0.41	0.90	0.21	7.27
Other Service	1.67	2.01	-0.17	0.43	0.29	9.56
Annual data: 1929 to 2019						
	Mean(%)	Standard Deviation(%)	Skewness	Kurtosis	Autocorrelation	Weight(%)
Aggregate Consumption	1.78	2.11	-1.39	5.06	0.48	
Food	0.84	2.70	1.01	5.05	0.30	17.53
Clothing	1.67	3.93	-1.31	2.74	0.31	8.09
Energy	0.79	5.69	1.12	13.56	0.40	4.69
Other nondurable	2.29	3.31	-1.44	5.07	0.34	9.13
Housing & Utility	2.17	1.86	0.11	0.26	0.70	19.98
Health care	2.45	2.98	0.11	4.23	0.30	10.54
Transportation	1.80	5.46	-0.25	1.94	0.53	3.63
Recreation	2.28	4.39	-1.80	8.03	0.36	3.19
Food & Accommodation	1.87	4.62	-0.15	2.81	0.43	7.62
Financial Service	2.16	3.94	-0.40	1.01	0.18	6.13
Other Service	1.31	3.28	-1.60	4.58	0.48	9.47

Notes: This table presents mean, standard deviation, skewness, kurtosis, and autocorrelation of aggregate log-consumption growth and its components. The top panel is calculated based on sample period from 1959Q2 to 2020Q3. The middle panel is calculated based on sample period from 1959Q2 to 2019Q4, excluding the Covid-19 period. The bottom panel summarizes annual data from 1929 to 2019. For quarterly data, the mean and standard deviation are transformed to annual percentage points. The last column presents the sample average of the expenditure share of each type of good.

Table 1.2: Estimates of Intercept Coefficient g_{ik}

Starts	1959Q2	1972Q4	1981Q2	2000Q1	2020Q1
Ends	1972Q3	1981Q1	1999Q4	2019Q4	2020Q3
Food	0.47		0.65		2.29
	(0.40)		(0.18)		(1.66)
Clothing	0.80	2.90			−2.71
	(0.75)	(0.53)			(11.20)
Energy	2.75	−2.06	0.36	−1.01	−2.62
	(0.96)	(2.36)	(0.86)	(0.73)	(15.14)
Other nondurable	2.44		2.12		5.77
	(0.44)		(0.23)		(2.21)
Housing & Utility	2.60	2.16	1.44		
	(0.17)	(0.27)	(0.10)		
Health care	4.38	3.09	1.50	2.36	−3.19
	(0.60)	(0.42)	(0.21)	(0.27)	(8.57)
Transportation	3.36	1.13	4.17	1.28	−28.80
	(0.65)	(0.98)	(0.58)	(0.58)	(36.34)
Recreation	3.01	4.81		1.35	−25.97
	(0.98)	(0.45)		(0.52)	(29.55)
Food & Accommodation	1.36	2.30	0.99	1.59	−8.36
	(0.61)	(1.12)	(0.50)	(0.38)	(14.68)
Financial Service	4.21			−0.65	
	(0.81)			(0.49)	
Other Service	1.73			0.99	−2.64
	(0.26)			(0.39)	(4.67)
Summary Statistics					
Mean absolute change		1.94	1.37	2.11	9.89
Most-affected goods		Energy	Transportation	Financial Service	Transportation
		Transportation	Energy	Recreation	Recreation
		Clothing	Health care	Transportation	Food & Accommodation

Notes: This table presents estimates of g_{ik}. The header of the table marks the beginning and ending quarters of each regime. The reported coefficients are in annualized percentage points (the original intercept times 4). The bottom panel presents the mean absolute change of g_{ik} after each break. The most-affect goods list the three most-affected types of good in terms of the absolute value of change in parameters. Coefficients that are not affected by the break (with posterior probability below 0.5) are left blank. The standard errors are in parenthesis.

Table 1.3: Estimates of Idiosyncratic Volatility σ_{ik}

Starts	1959Q2	1972Q4	1981Q2	2000Q1	2020Q1
Ends	1972Q3	1981Q1	1999Q4	2019Q4	2020Q3
Food	1.92		0.97		2.42
	(0.14)		(0.05)		(0.43)
Clothing	2.45	3.48			17.36
	(0.23)	(0.19)			(2.97)
Energy	3.70	7.09	3.72	3.09	23.03
	(0.36)	(0.73)	(0.37)	(0.22)	(4.31)
Other nondurable	1.71		1.21		3.42
	(0.15)		(0.09)		(0.77)
Housing & Utility	0.67	0.97	0.62		
	(0.06)	(0.12)	(0.04)		
Health care	2.35	1.35	1.01	1.27	11.49
	(0.22)	(0.14)	(0.07)	(0.09)	(2.01)
Transportation	1.78	2.51	2.37	2.36	48.84
	(0.17)	(0.29)	(0.20)	(0.19)	(8.37)
Recreation	3.92	2.42		2.05	39.64
	(0.37)	(0.16)		(0.18)	(7.96)
Food & Accommodation	2.10	3.15	2.01	1.48	18.38
	(0.21)	(0.29)	(0.17)	(0.12)	(3.17)
Financial Service	4.57			2.41	
	(0.24)			(0.22)	
Other Service	1.90			1.50	7.66
	(0.11)			(0.12)	(1.42)
Common Component	1.57				2.05
	(0.19)				(0.50)
			Summary Statistics		
Mean absolute change		1.29	0.97	0.62	15.53
Most affected goods		Energy	Energy	Financial Service	Transportation
		Recreation	Food & Accommodation	Energy	Recreation
		Food & Accommodation	Food	Food & Accommodation	Energy

Notes: This table presents estimates of σ_{ik}. The header of the table marks the beginning and ending quarters of each regime. The reported coefficients are in annualized percentage points (the original quarterly times 2). The bottom panel presents the mean absolute change of σ_{ik} after each break. The most-affect goods list the three most-affected types of good in terms of the absolute value of change in parameters. Coefficients that are not affected by the break (with posterior probability below 0.5) are left blank. The standard errors are in parenthesis.

Table 1.4: Parameter choices of benchmark break model

Parameter	Description	Value
β	Subjective discount factor	0.994
γ	Coefficient of risk aversion	6
ψ	Coefficient of IES $(1/\rho)$	2
λ	The probability that a break happens	0.0167
$\bar{g}$	Unconditional mean of log consumption growth	0.45%
$\bar{G}$	Upper bound of truncated normal distribution of $\frac{g_{K_t}-\bar{g}}{\sigma_{K_t}}$	1
$\underline{G}$	Lower bound of truncated normal distribution $\frac{g_{K_t}-\bar{g}}{\sigma_{K_t}}$	-1
Σ_0	Variance of $\frac{g_{K_t}-\bar{g}}{\sigma_{K_t}}$	0.1
L	Leverage of dividend growth to consumption growth	3
$\bar{\sigma}_c$	Unconditional idiosyncratic volatility in consumption growth	1.0%
$\bar{\sigma}_d$	Unconditional idiosyncratic volatility in dividend growth	7.9%
σ_{Low}	$\sigma_{dK_t}/\bar{\sigma}_d, \sigma_{K_t}/\bar{\sigma}_c$ in low volatility regime	0.91
σ_{High}	$\sigma_{dK_t}/\bar{\sigma}_d, \sigma_{K_t}/\bar{\sigma}_c$ in high volatility regime	1.30
π_{Low}	Probability that new regime has low volatility	0.66

Notes: This table presents parameter values used in the benchmark calibration. All parameters are calibrated at quarterly frequency except for γ and ψ.

Table 1.5: Model-implied moments of asset returns with different parameters in data-generating process

	$E(r_m - r_f)$	$\sigma(r_m - r_f)$	$E(r_f)$	$\sigma(r_f)$	SR	$\sigma(M)/E(M)$
Data						
-	5.32	21.04	0.60	1.40	0.25	-
Benchmark: $\lambda = 0.017$, $\sqrt{\Sigma_0} = 0.3$						
-	5.41	17.82	1.56	1.32	0.30	0.51
$\lambda = 0.017$						
$\sqrt{\Sigma_0} = 0.2$	2.60	16.23	2.17	0.62	0.16	0.30
$\sqrt{\Sigma_0} = 0.4$	7.41	18.13	1.03	1.69	0.41	0.68
$\sqrt{\Sigma_0} = 0.3$						
$\lambda = 0.050$	2.80	16.06	2.23	0.54	0.17	0.28
$\lambda = 0.012$	5.46	16.86	1.48	1.24	0.32	0.57

Notes: This table presents moments of historical and model implied asset returns with diferent parameters of data generating process. The first panel presents sample moments estimated from data. The second panel presents the result when $\sqrt{\Sigma_0}$ ranges from 0.2 to 0.4, holding other parameters fixed at $1/\rho = 2$, $\gamma = 6$ and $\lambda = 0.017$. The third panel presents similar set of result when λ varies from 0.05 to 0.0125, with $\sqrt{\Sigma_0}$ fixed at 0.3. Columns $E(r_m - r_f)$, $\sigma(r_m - r_f)$, $E(r_f)$, $\sigma(r_f)$ are unconditional mean and volatility of excess return and risk-free rate, in annualized percentage points. Column *SR* denotes Sharpe ratio and $\sigma(M)/E(M)$ denotes price of risk.

Table 1.6: Model-implied moments of asset returns with different parameters of investors' preference

γ	$E(r_m - r_f)$	$\sigma(r_m - r_f)$	$E(r_f)$	$\sigma(r_f)$	SR	$\sigma(M)/E(M)$
			Data			
-	5.32	21.04	0.60	1.40	0.34	-
			$1/\rho = 2$			
4	2.78	19.40	2.19	0.91	0.14	0.25
6	5.41	17.82	1.56	1.32	0.30	0.51
8	7.44	15.88	1.06	1.43	0.47	0.86
			$1/\rho = 1.5$			
4	2.31	18.71	2.57	0.89	0.12	0.23
6	4.86	17.51	2.18	1.19	0.28	0.48
8	7.63	15.60	1.62	1.31	0.49	0.92

Notes: This table presents moments of historical and model implied asset returns with different preference parameters. The first panel presents sample moments estimated from data. The second panel presents the result when EIS, $1/\rho = 2$ and risk aversion γ ranges from 4 to 8. The third panel presents similar set of result with $1/\rho = 1.5$. Columns $E(r_m - r_f)$, $\sigma(r_m - r_f)$, $E(r_f)$, $\sigma(r_f)$ are unconditional mean and volatility of excess return and risk-free rate, in annualized percentage terms, respectively. Column *SR* denotes Sharpe ratio and $\sigma(M)/E(M)$ denotes price of risk.

Table 1.7: Predictive regressions of excess returns

Excess return					
			Data		
	1 Quarter	1 Year	2 Years	3 Years	5 Years
β	-0.021	-0.087	-0.172	-0.233	-0.370
	(0.016)	(0.048)	(0.065)	(0.070)	(0.092)
R^2	0.008	0.035	0.079	0.129	0.226
			Model		
	1 Quarter	1 Year	2 Years	3 Years	5 Years
β	-0.059	-0.210	-0.361	-0.486	-0.679
R^2	0.016	0.052	0.080	0.096	0.110
Aggregate consumption growth					
			Data		
	1 Quarter	1 Year	2 Years	3 Years	5 Years
β	0.000	-0.002	-0.008	-0.013	-0.024
	(0.001)	(0.003)	(0.007)	(0.010)	(0.016)
R^2	0.000	0.004	0.024	0.040	0.070
			Model		
	1 Quarter	1 Year	2 Years	3 Years	5 Years
β	0.001	0.004	0.008	0.011	0.017
R^2	0.000	0.001	0.001	0.002	0.002

Notes: This table presents the slope coefficient and R^2 of predictive regressions of excess return (top panel) and aggregate consumption growth (bottom panel). The y variables are cumulative quarterly excess return or consumption growth over the next 1 quarter, 1 year, 2 years, 3 years, and 5 years. For each panel, the top sub-panel presents estimates from data and the sub-bottom panel presents model implication from the benchmark break model. The HAC standard errors are in parenthesis.

Table 1.8: Estimation of linear factor models

Aggregate consumption	Fragility factor	Uncertainty diff x Fragility	Market	MAE(%)	R^2
182.70				1.30	-0.11
(47.25)					
121.58	681.90			0.66	0.71
(48.15)	(98.92)				
126.29	480.63	364.70		0.65	0.74
(43.19)	(107.85)	(181.71)			
			15.08	1.29	0.00
			(4.19)		

Notes: This table presents the estimated price of risk for the CAPM, CCAPM, extended CCAPM the consumption fragility factor. The first column presents price of risk of the aggregation consumption, the second column is price of risk of the fragility factor. The third column is the price of risk of the interaction of the fragility factor and the uncertain difference proxy. The forth column is the market factor. The test assets are the 25 Fama-French portfolios sorted by size and book-to-market equity and 10 momentum portfolios sorted by returns of the past 12 months. Estimation is by two-step GMM. HAC standard errors are in parentheses. The last two columns report absolute pricing error (MAE) and R^2 which are based on first-stage estimate. The R^2 is benchmarked against CAPM such that R^2 of CAPM is 0 by construction.

Table 1.9: Consumption growth and relative consumption growth betas

Size				Value				Momentum			
	Consumption	Fragility	R_{ex} (%)		Consumption	Fragility	R_{ex} (%)		Consumption	Fragility	R_{ex}(%)
ME 1	0.4	3.7	2.0	BM 1	1.2	4.2	1.6	Prior 1	8.6	0.2	-0.1
ME 2	0.6	4.3	2.1	BM 2	1.1	3.5	1.9	Prior 2	4.0	1.2	1.7
ME 3	0.7	4.6	2.1	BM 3	1.3	3.3	1.9	Prior 3	4.1	1.1	1.6
ME 4	1.6	5.2	2.1	BM 4	2.0	4.4	1.9	Prior 4	3.2	4.5	1.9
ME 5	1.9	4.6	1.6	BM 5	1.1	5.6	2.6	Prior 5	6.6	11.2	2.8
Diff	1.5	0.9	-0.3	Diff	-0.1	1.4	0.9	Diff	-2.0	11.0	2.9
	(1.5)	(2.7)			(1.6)	(2.7)			(2.3)	(3.9)	

Notes: This table presents regression coefficients of aggregate consumption innovations and the fragility factor on excess returns of five size-sorted portfolios, five value-sorted portfolios and five momentum-sorted portfolios. For each panel, the first and second columns reports factor loadings and the last column reports average excess returns. The sixth row is the difference of the fifth and the first row. The numbers in parenthesis are standard errors of the difference in factor loadings.

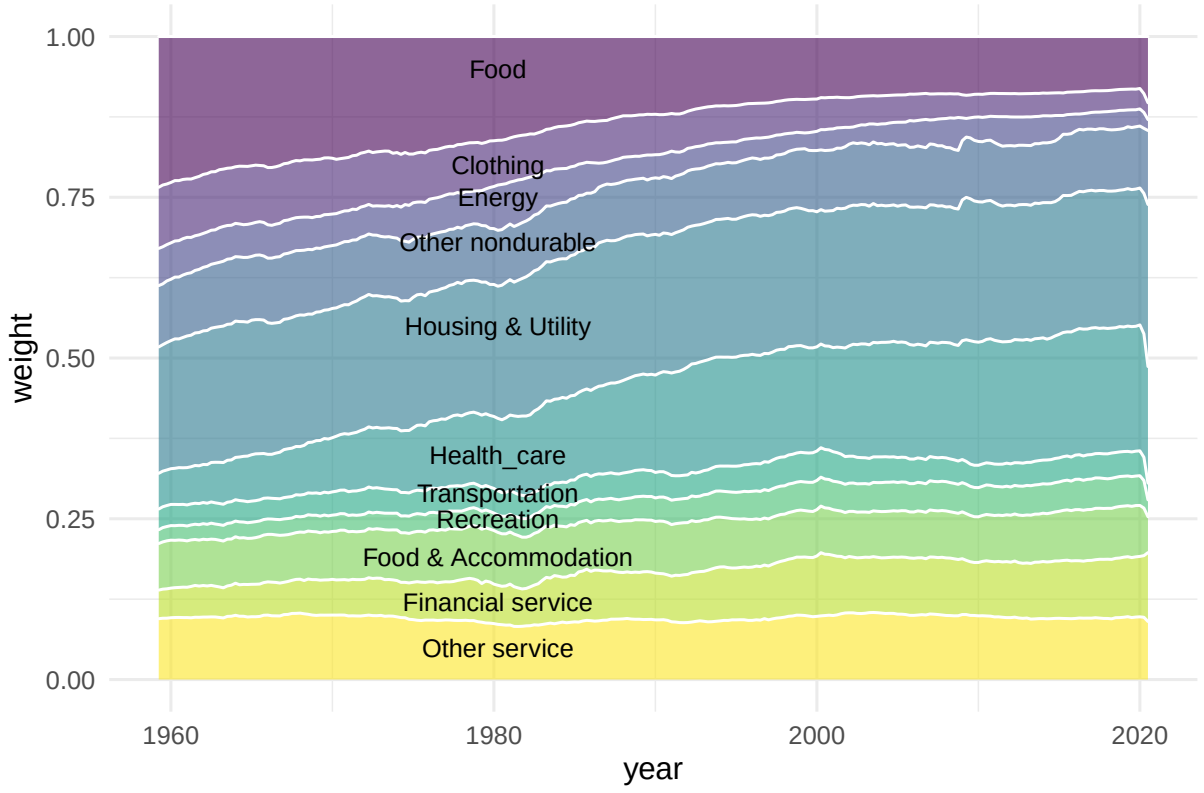
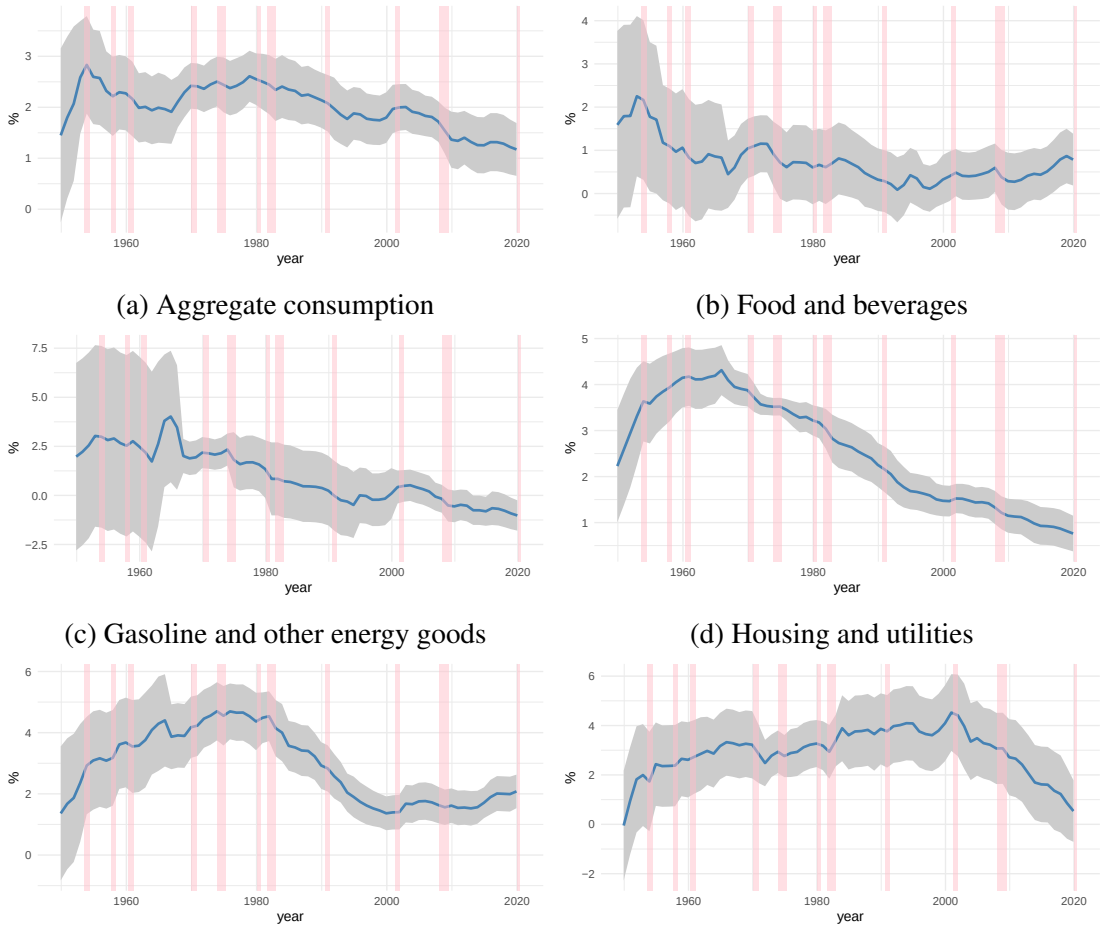


Figure 1.1: Weights of consumption expenditures in each categories over time



(a) Aggregate consumption

(b) Food and beverages

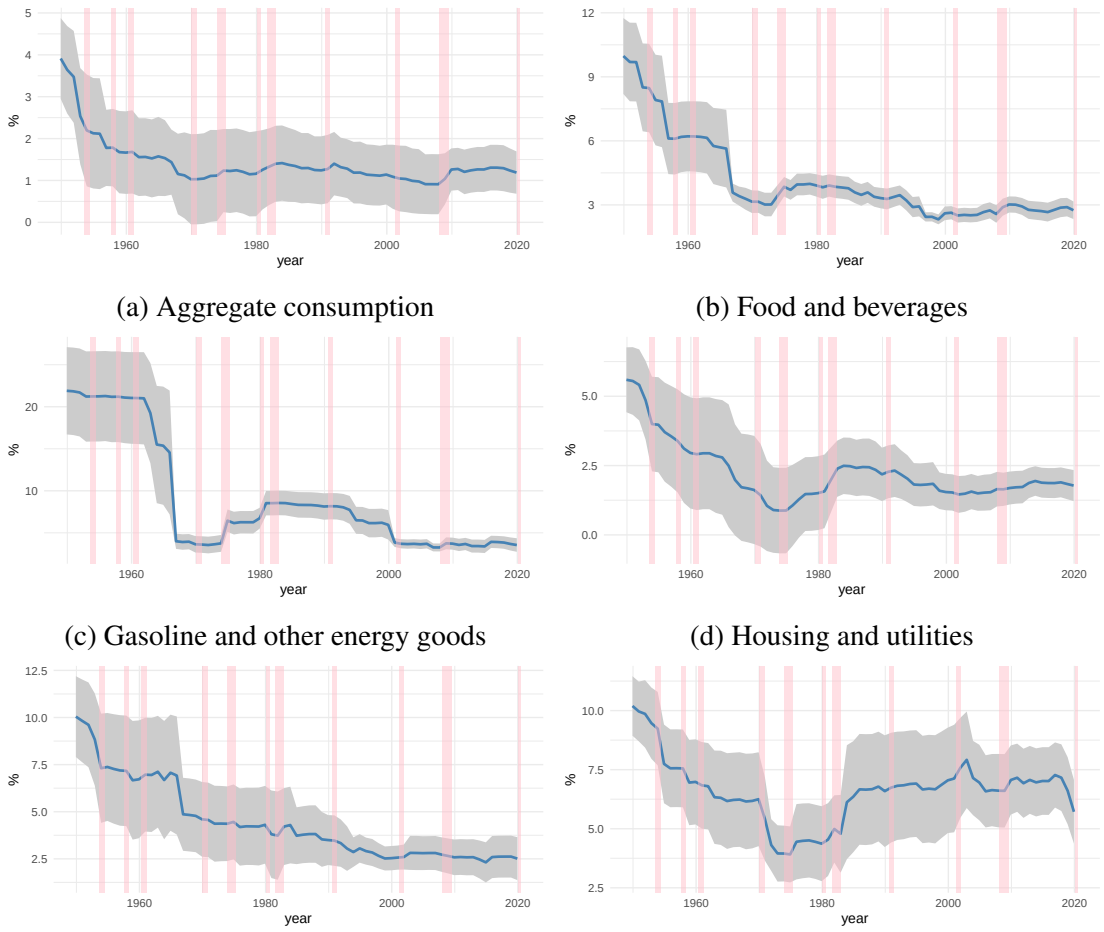
(c) Gasoline and other energy goods

(d) Housing and utilities

(e) Health care

(f) Financial services and insurance

Figure 1.2: Time series of annualized 20-year rolling average of quarterly consumption growth for selected series
The shaded area are 95% confidence intervals.



(a) Aggregate consumption

(b) Food and beverages

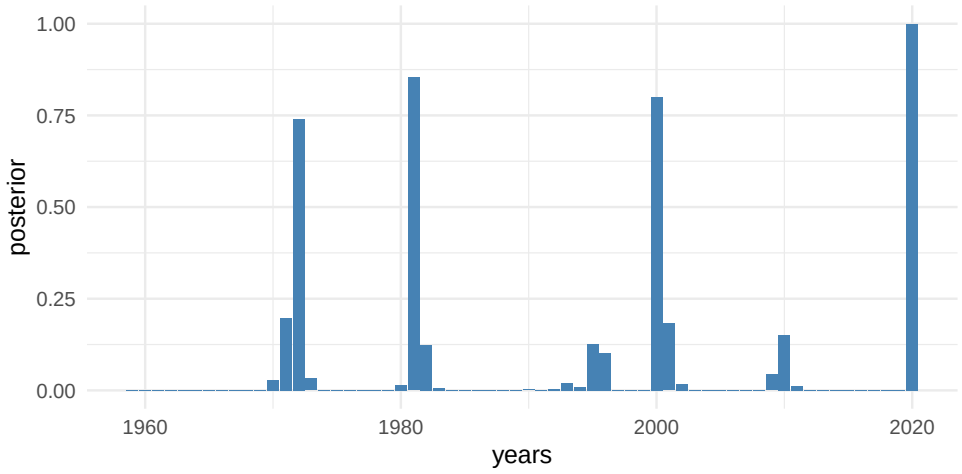
(c) Gasoline and other energy goods

(d) Housing and utilities

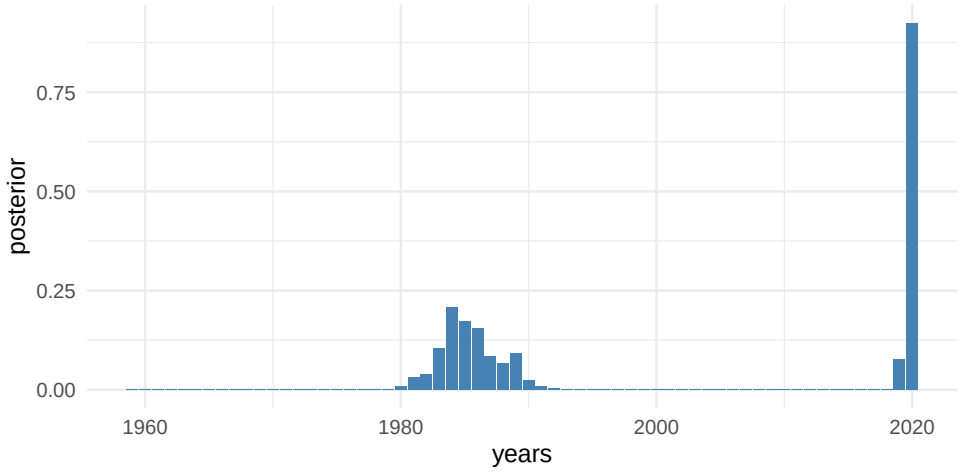
(e) Health care

(f) Financial services and insurance

Figure 1.3: Time series of annualized 20-year rolling volatility of quarterly consumption growth for selected series
The shaded area are 95% confidence intervals.

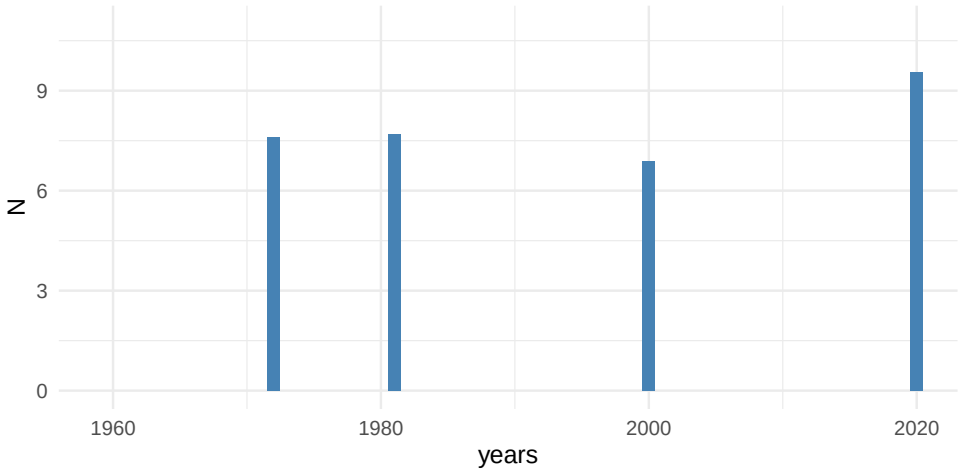


(a) 11 consumer goods

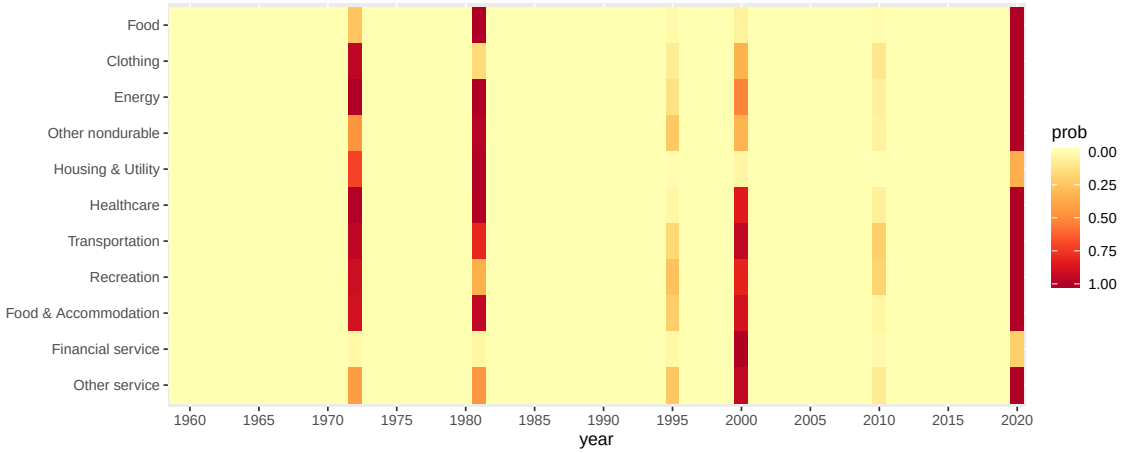


(b) Aggregate consumption

Figure 1.4: Posterior locations of breaks
The top panel presents the posterior probability $P_{t|T}$ that a break happened between period t and $t+1$ based on 11 consumer goods. The bottom panel presents the posterior probability $P_{t|T}$ that a break happened between period t and $t+1$ based on the aggregate consumption.



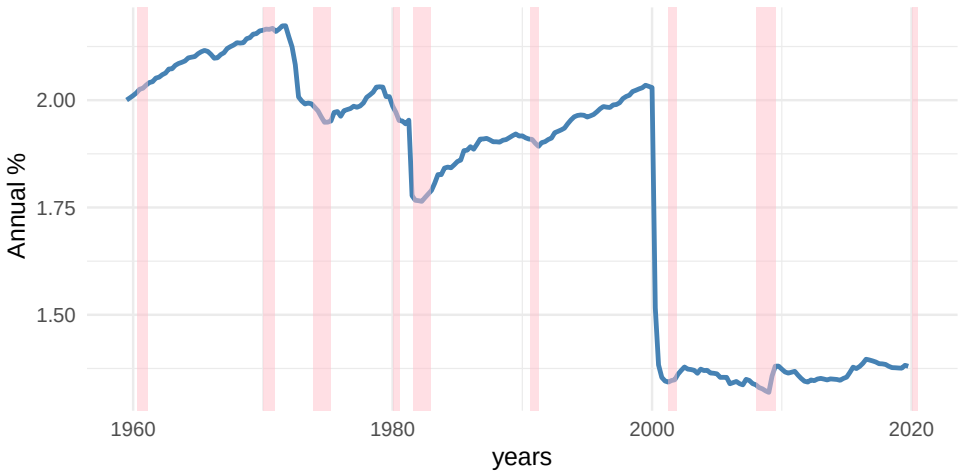
(a) Expected number of goods affected by breaks



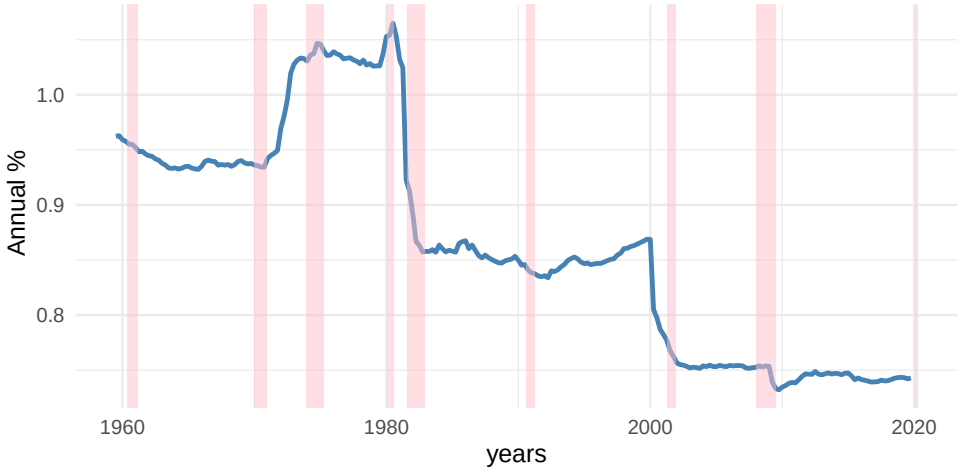
(b) Posterior of location of breaks affecting each good

Figure 1.5: Expected number of goods affected by breaks and posterior of location of breaks affecting each good

The top panel presents expected number of goods affected by breaks conditional on a break happens at t. The second panel presents the posterior probability $P_{it|T}$ that a break happens to good i between period t and $t+1$.



(a) Conditional expected growth rate



(b) Conditional volatility

Figure 1.6: Posterior mean of the expected growth rate and volatility of aggregate consumption
The expected growth of aggregate consumption is the expenditure-share weighted average of g_{it} of each good. The volatility of aggregate consumption growth is calculated as the combined effect of idiosyncratic components and the common component.

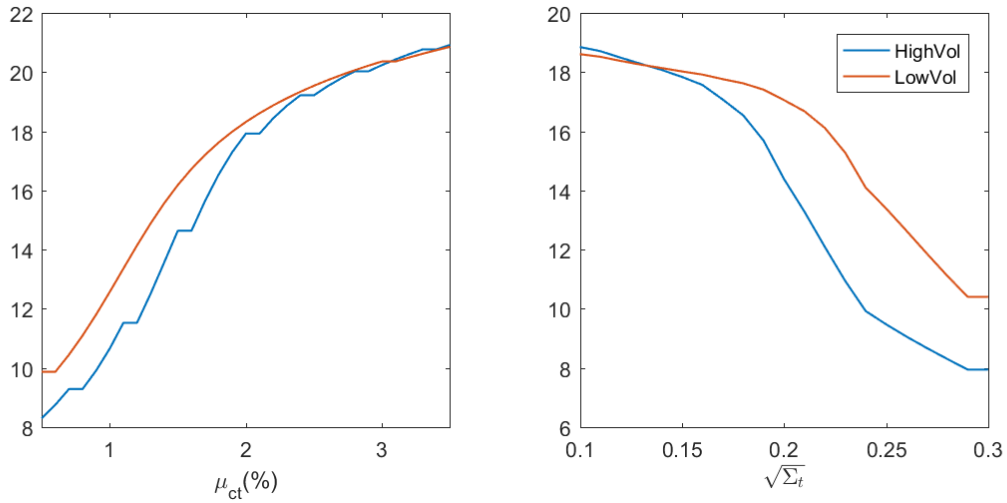
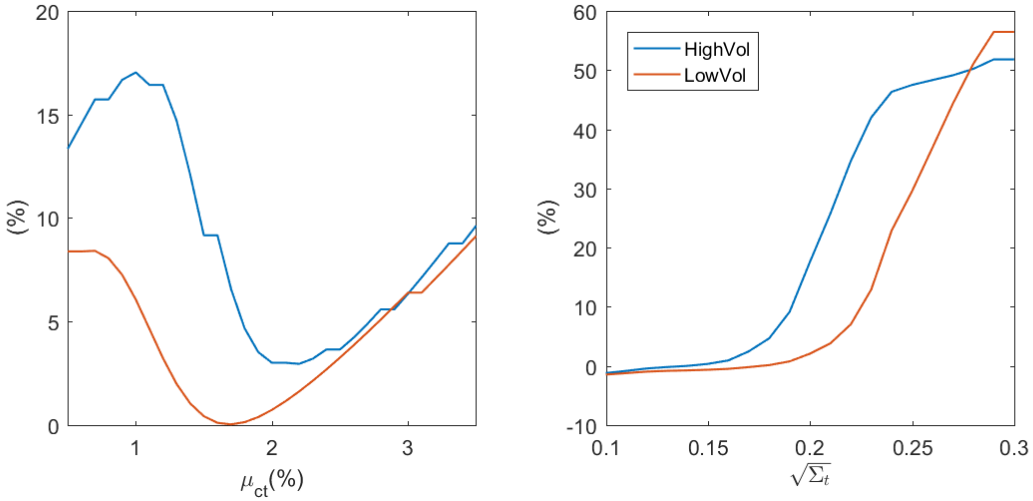
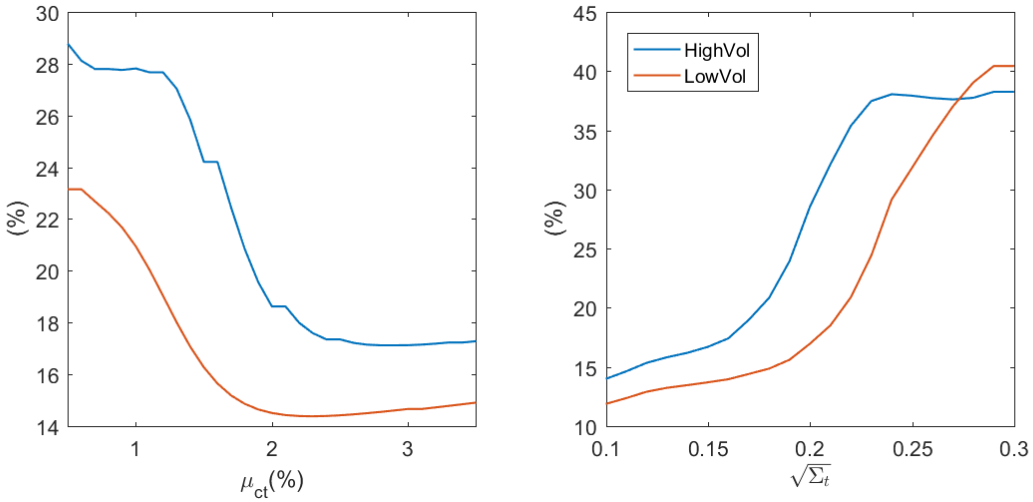


Figure 1.7: Price-dividend ratio and state variables
The left panel plot price-dividend ratio against μ_{ct} holding $\Sigma_t = 3.5\%$. The right panel plot price-dividend ratio against $\sqrt{\Sigma_t}$ holding $\mu_{ct} = 1.8\%$.

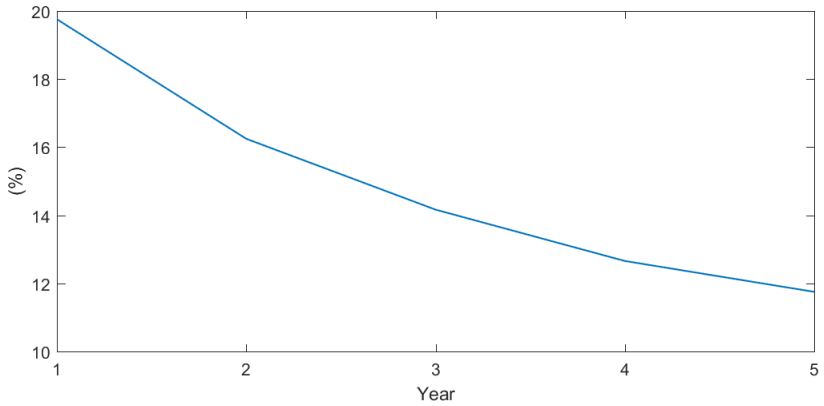


(a) $E_t(R_{t+1} - Rf_{t+1})$

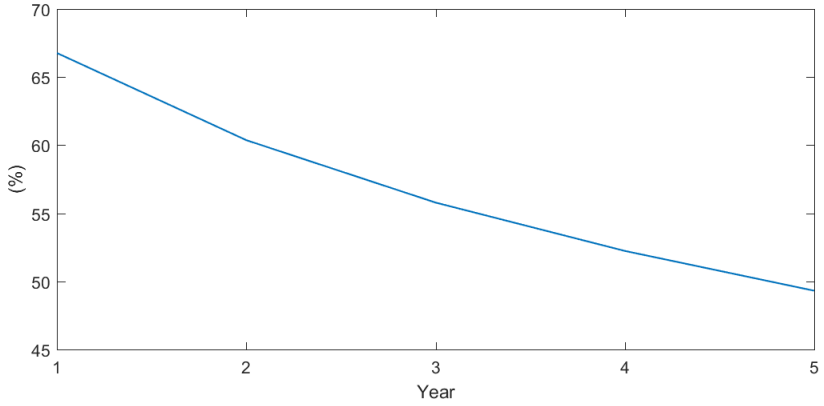


(b) $Vol_t(R_{t+1} - Rf_{t+1})$

Figure 1.8: Risk premium, volatility and state variables
The top panel plots annualized risk premium against state variables. The bottom panel plots annualized volatility of market return against state variables. For the figures on the left side of top and bottom panels, the x-axis is μ_{ct} holding $\Sigma_t = 3.5\%$. For the figures on the right side of top and bottom panels, the x-axis is $\sqrt{\Sigma_t}$ holding $\mu_{ct} = 1.8\%$.



(a) Excess return



(b) Volatility

Figure 1.9: Forecasts of market risk premium and volatility after the Covid shock
The forecasts are based on the Lucas tree model. The forecasts are made at 2020Q3, assuming that a break happened on 2020Q1 and the mean of investors' beliefs of consumption growth is -1.5% at an annual rate.

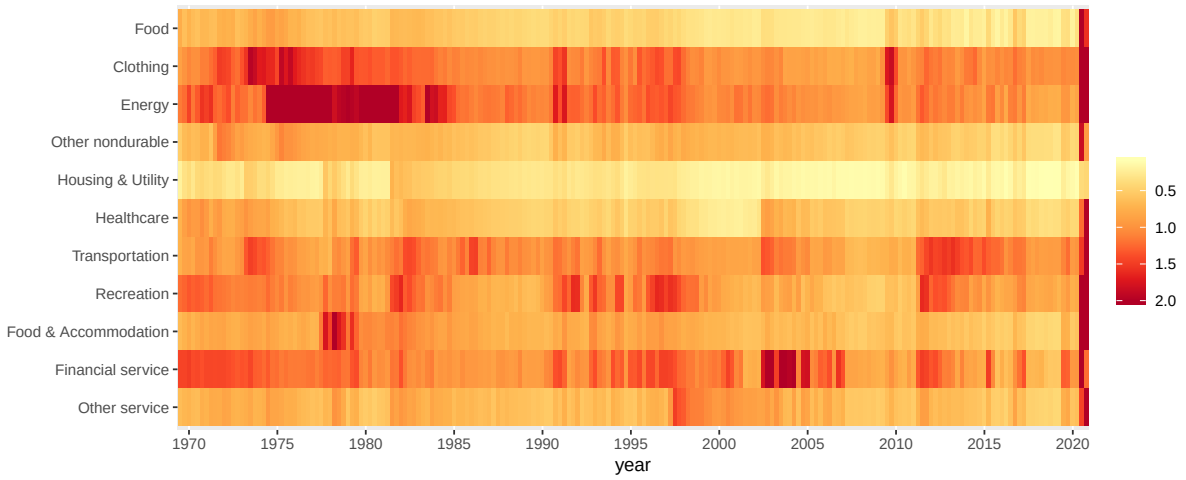


Figure 1.10: Parameter uncertainty in real-time estimation of intercept $g_{it|t}$ of major types of goods

The figure presents the standard deviations of posteriors of $g_{it|t}$ based on data up to time t. The units are in annual percentage points.

Chapter 2

Identifying Forecasting Skills: A Bootstrap Test for Comparing Predictive Accuracy with Panel Data

2.1 Introduction

Panels of forecasts of outcomes recorded over multiple periods of time for many different variables and forecasters are increasingly widespread in economics and finance. For example, in their biannual World Economic Outlook publication, the IMF reports forecasts of economic indicators such as real GDP growth and inflation for more than 180 countries and at multiple horizons. Financial analysts predict company earnings growth, profits, and other outcomes for hundreds of firms spanning multiple industries and countries. The travel and tourism industry forecasts daily demand for hotel rooms and occupancy rates across countries and cities. Multiple forecasts also arise regularly in comparisons of the predictive accuracy of different econometric modeling approaches such as univariate autoregressive, multivariate and machine learning methods.

Comparisons of predictive accuracy across what is typically very large sets of forecasts is complicated by the sheer number of possible pair-wise test statistics that needs to be conducted in order to identify superior predictive performance. Conventional critical values that apply to a single pair-wise test of the null of equal predictive accuracy no longer remain valid when multiple test statistics are examined. To deal with this multiple hypothesis testing issue, we develop in this paper a set of "Sup" tests that allow us to conduct robust inference on whether one or more forecasts is significantly more accurate than some benchmark forecast. Our analysis exploits the panel data structure to conduct tests for the existence of superior forecasting skills for *any* forecast(er), *any* outcome variable, at *any* horizon, or at *any* point in time. The existence of several dimensions, including cross-sectional and time-series dimensions, for an arbitrarily large set of individual forecasts introduces a high-dimensional multiple hypothesis testing problem. Our approach handles situations with multiple moment inequalities by controlling the family-wise error rate.

Outcomes in any individual period are likely to be strongly cross-sectionally correlated, so

accurate forecasting performance across many variables in a particular period could simply be due to luck, i.e., the result of a forecaster essentially making one good judgment about the realization of a highly influential common factor. Provided that the cross-sectional dimension of the data set is large enough–or, equivalently, that the forecasts of sufficiently many variables are being compared–and there is sufficiently independent variation across forecast errors, we demonstrate that valid comparisons of predictive accuracy can be performed on a *single* cross-section, i.e., for a *single* time period, after controlling for common factors. This result requires us to apply a cross-sectional central limit theorem. We establish conditions under which this can be justified in the context of a model that decomposes the forecast errors of individual variables into a correlated component that is driven by exposures to a set of common factors and an uncorrelated, idiosyncratic component.

Our paper makes several contributions to the existing literature on multiple comparisons of predictive performance.[1] The seminal paper of White (2000) and subsequent work by Hansen (2005), Romano and Wolf (2005), Hansen, Lunde and Nason (2011) address the multiple hypothesis problem in settings with a low-dimensional model space and a single dependent outcome variable. We generalize this to a setting with multiple dimensions, including cases in which the dimensions of both the number of forecasters and the cross-section are large, while retaining the ability to identify which forecasters and for which variables we find superior predictive performance. Further, we generalize results in the extant literature to a setting that uses a single cross-section to conduct inference on the average forecasting performance across a large number of (cross-sectional) units in a single time period.

To handle cases in which the number of pair-wise comparisons increases with the length of the time series–and even can be much larger than this–we use the approach developed by

[1] An earlier literature develops methods for conducting inference on the relative accuracy of a pair of forecasting models. For example, Chong and Hendry (1986) develop tests of forecast encompassing, while Diebold and Mariano (1995) and West (1996) develop distribution theory and propose statistics for testing the null of equal predictive accuracy for the non-nested case. Clark and McCracken (2001) and Giacomini and White (2006) develop methods for testing equal predictive accuracy for forecasts generated by nested models.

Chernozhukov, Chetverikov and Kato (2018).[2] This approach implements a version of the high-dimensional bootstrap from Chernozhukov, Chetverikov and Kato (2013, 2017) which accounts for serial dependence using a blocking technique. We use the bootstrap tests to identify superior predictive skills based on a potentially large set of forecast comparisons or, equivalently, multiple moment inequalities. The resulting bootstrap is easy to implement and, in addition to testing the null that no forecast is more accurate than a given benchmark, identifies the variables, forecasts, horizons, or time periods for which the benchmark is beaten. To the best of our knowledge, our approach provides the first tests for superior predictive accuracy conducted over multiple units in a panel setting.

Although our analysis builds on Chernozhukov, Chetverikov and Kato (2018), there are also some important technical differences between our method and theirs. In particular, we develop studentized test statistics that apply to dependent data, whereas Chernozhukov, Chetverikov and Kato (2018) study a non-studentized test statistic (see their Appendix B.1). In many empirical applications, the scale of forecast errors can differ drastically across units and/or time and so normalizing the test statistics to have unit variance under the null hypotheses can improve the power of the test (Hansen (2005)). In practice, this really matters as we demonstrate through Monte Carlo simulations.

Our paper also contributes to the literature on panel forecasting. Papers that use panel data to evaluate predictive accuracy such as Keane and Runkle (1990) and Davies, Lahiri et al. (1995) do not address the multiple hypothesis testing issue and can therefore not be used to identify the dimensions in which forecasting performance is genuinely superior (or inferior) relative to some benchmark.[3]

We use our approach in an empirical analysis of the "term structure" of forecast errors which examines the rate at which macroeconomic forecasts improve as the distance to the target date shrinks. Specifically, we analyze forecasts from the IMF's World Economic Outlook (WEO)

[2]The approach in White (2000) assumes that the number of pair-wise comparisons is fixed.

[3]Baltagi (2013) provides an extensive review of forecast applications that use panel data.

publication of five variables–inflation, GDP, import, and export growth, and the current account balance–across up to 185 countries recorded over the 30-year period from 1990 to 2019. WEO forecasts are updated in April and October every year and we focus on current-year and next-year forecasts, giving us four different points on the term structure representing forecast horizons of 21, 15, 9, and 3 months prior to the end of the target year.

Comparing predictive accuracy across these horizons, we test if individual country-level forecasts become significantly more accurate at the shorter horizons and, if so, at which revision points the improvements are largest. Our most extensive comparisons of predictive accuracy across countries, variables, and forecast horizons involve more than 2,700 pair-wise tests. Empirically, we find little evidence of significant improvements in the accuracy of next-year forecasts between the Spring and Fall WEO issues, suggesting that the IMF's learning curve is flat between 21 and 15 months prior to the end of the year being predicted. Accuracy gains for the GDP growth, inflation, and current account forecasts materialize between the Fall next-year and Spring current-year WEO issues (horizons of 15 and 9 months) and further accelerate between the Spring and Fall current-year issues (horizons of 9 and 3 months). The IMF's learning curve for these variables is, thus, considerably steeper during the target year.[4]

Applying our Sup tests to clusters of countries defined according to developmental stage and geographic region, we find not only that predictive accuracy is higher for advanced economies than for emerging market and developing economies, but also that predictive accuracy improves by a significantly larger margin for the advanced economies as the forecast horizon shrinks. The steeper learning curve is consistent with the flow of data and data quality being better for advanced economies than for emerging market and developing economies, making it more difficult to track the state of the economy and improve forecast accuracy for the latter group.

Our cross-sectional individual-year tests for improvements in predictive accuracy at

[4]For the two remaining variables–import and export–we find only limited evidence that forecast accuracy improves significantly in the 18-month period lasting from 21 to 3 months prior to the end of the predicted year. This finding is consistent with a flat learning curve for these variables.

shorter horizons show that forecasts of import and export fail to improve significantly in many years, particularly prior to 2002. For GDP growth and inflation, forecast accuracy improves significantly at the shorter horizons in almost all years, consistent with the IMF incorporating new information to improve forecasts of these variables.

The outline of the paper is as follows. Section 2.2 describes methods for identifying predictive skills with panels of forecasts, while the next two sections describe our test methodologies in panel (Section 2.3) and cross-sectional (Section 2.4) settings. Section 2.5 conducts our empirical analysis of the WEO forecasts and Section 2.6 concludes. Monte Carlo simulation results and technical proofs are contained in Appendices.

2.2 Identifying Predictive Skills in Panels of Forecast

This section introduces our framework for comparing forecasting performance in a panel setting with an arbitrary number of forecasts and outcome variables observed across multiple time periods. Our analysis explicitly accounts for the multiple hypothesis testing problem that arises when many test statistics are simultaneously evaluated.

2.2.1 Setup

Consider a set of h-step-ahead forecasts of a panel of variables (units) $i = 1, ..., N$ observed over T time periods $t + h = 1, ..., T$. We refer to the outcome of variable i at time $t + h$ as $y_{i,t+h}$ and to the associated $h-$step-ahead forecast at time t as $\hat{y}_{i,t+h|t,m}$, where $m = 1, ..., M$ refers to an individual survey participant or a model used to generate the forecast. In all cases, the forecast horizon, $h \geq 0$, is a non-negative integer.

This type of panel data gives us four dimensions over which to compute averages and/or conduct inference: variables/units ($i = 1, ..., N$), forecasters/models ($m = 1, ..., M$), forecast horizons ($h = 1, ..., H$) and time periods ($t + h = 1, ..., T$).

Following standard practice (e.g., Granger (1999)), forecasting performance is evaluated through a loss function, $L(y_{i,t+h}, \hat{y}_{i,t+h|t,m})$, which takes as its inputs the outcome and the forecast and maps these values to the real line. By far the most common loss function in applied work is squared error loss:

$$L(y_{i,t+h}, \hat{y}_{i,t+h|t,m}) = e^2_{i,t+h,m}, \tag{2.1}$$

where $e_{i,t+h,m} = y_{i,t+h} - \hat{y}_{i,t+h|t,m}$ is the forecast error.

Forecasting performance is usually measured relative to some benchmark, m_0, which might be an incumbent forecasting approach while the M alternative forecasts represent competitors that could replace the benchmark.

The resulting *loss differential* of forecast m, measured relative to the benchmark m_0, is given by[5]

$$\Delta L_{i,t+h,m} = L(y_{i,t+h}, \hat{y}_{i,t+h|t,m_0}) - L(y_{i,t+h}, \hat{y}_{i,t+h|t,m}). \tag{2.2}$$

Under squared error loss, $\Delta L_{i,t+h,m} = e^2_{i,t+h,m_0} - e^2_{i,t+h,m}$. Positive values of $\Delta L_{i,t+h,m}$ show that forecast m generated a lower loss than the benchmark, m_0, in period $t+h$, while negative values show the reverse.

2.2.2 Single Pairwise Comparison of Forecast Accuracy

We begin with a simple setting that compares the predictive accuracy of two forecasts for a single variable over a single horizon ($M = N = H = 1$) so that the two forecasts used in the comparison (m_0 and m) as well as the identity of the variable (i) and the horizon (h) used in the comparison are fixed ex-ante (predetermined). Under this assumption, the comparison does *not* involve a multiple hypothesis testing problem and so inference about the hypothesis that the two forecasts are equally accurate, on average, can be conducted using the approach proposed by

[5]For simplicity, we drop the reference to t and m_0 in the subscripts of ΔL.

Diebold and Mariano (1995):

$$H_0^{DM} : E[\Delta L_{i,t+h,m}] = 0. \tag{2.3}$$

Assuming that a time series of outcomes and forecasts $\{y_{i,t+h}, \hat{y}_{i,t+h|t,m}\}$ $t+h = 1,...,T$ is observed, the Diebold-Mariano null in (2.3) can readily be tested using a t-test on the time-series average of the loss differential $\overline{\Delta L}_{i,m} = T^{-1} \sum_{t+h=1}^{T} \Delta L_{i,t+h,m}$.[6] Provided that parameter estimation error ("learning") can be ignored or is incorporated as part of the null hypothesis, the test statistic will be asymptotically normally distributed.[7]

Whenever we do not fix the variable, i, the forecast, m, and the horizon, h, and instead conduct tests either across multiple variables, forecasts, or horizons, a multiple hypothesis testing problem arises and we cannot rely on the conventional distribution results that underpin tests of H_0^{DM}. We next discuss how to deal with this issue.

2.2.3 Multiple Comparisons of Forecast Accuracy

Assuming initially that the time-series dimension of the data is used to compute sample averages, the remaining three dimensions of our four-dimensional panel can be used to identify whether *any* of the forecasts outperforms the associated benchmark for *any* of the variables or at *any* horizon. This corresponds to considering the "sup" loss differential across $i \in \{1,...,N\}$, $m \in \{1,..,M\}$ and $h \in \{1,..,H\}$:

$$H_0 : \max_{h\in\{1,...,H\}} \max_{i\in\{1,...,N\}} \max_{m\in\{1,..,M\}} E[\Delta L_{i,t+h,m}] \leq 0. \tag{2.4}$$

[6] It is common to use heteroskedasticity and autocorrelation consistent standard errors when conducting this test, see Diebold and Mariano (1995).

[7] Giacomini and White (2006) discuss conditions under which this type of test is valid even for forecasts generated by nested models while Clark and West (2007) derive distributional results that account for parameter estimation error and nested models. Clark and McCracken (2001) consider the case with recursive updating in the parameters of nested forecasting models and show that this gives rise to a non-standard distribution of the resulting test statistic.

It is common to study individual forecast horizons. For a given forecast horizon (fixed h), under the null of "no outperformance" for any variable or any forecast, we have

$$H_0 : \max_{i\in\{1,...,N\}} \max_{m\in\{1,..,M\}} E[\Delta L_{i,t+h,m}] \leq 0. \tag{2.5}$$

This null nests, as a special case, the "reality check" null considered by White (2000), Hansen (2005), and Romano and Wolf (2005) that, for a particular variable, i, no forecast $m = 1,...,M$ can beat the benchmark m_0:

$$H_0^{RC} : \max_{m\in\{1,...,M\}} E[\Delta L_{i,t+h,m}] \leq 0. \tag{2.6}$$

The null in (2.6) cannot be used to identify whether a particular forecast outperforms the benchmark for some variables but not for others.

To address the possibility that some forecasts are superior to the benchmark for a subset of variables–including just a single variable–we can test whether any of the M forecasts beats the benchmark among variables within some cluster C_k comprising $N_k < N$ of the variables, where both C_k and N_k are assumed to be predetermined:

$$H_0^{C} : \max_{i\in C_k} \max_{m\in\{1,..,M\}} E[\Delta L_{i,t+h,m}] \leq 0. \tag{2.7}$$

2.3 Test Statistics and Bootstrap

We next introduce our test statistic and bootstrap methods. To handle cases in which N increases with T and even can be much larger than T, we use the approach developed by Chernozhukov, Chetverikov and Kato (2018).[8] This approach implements a version of the high-dimensional bootstrap from Chernozhukov, Chetverikov and Kato (2013, 2017) which accounts

[8]The approach in White (2000) assumes that N is fixed.

for serial dependence using a blocking technique.

2.3.1 Bootstrap

Suppose we only compare the performance of a single forecast (m) to that of the benchmark (m_0) so that, without risk of confusion, we can drop the forecast subscript, m, from (2.2) and define $\Delta L_{i,t+h} = L(y_{i,t+h}, \hat{y}_{i,t+h|t,m_0}) - L(y_{i,t+h}, \hat{y}_{i,t+h|t,m})$ and $\hat{\mu}_i = T^{-1}\sum_{t+h=1}^{T} \Delta L_{i,t+h}$. Appendix B.1 of Chernozhukov, Chetverikov and Kato (2018) considers the test statistic $J_T = \max_{1\leq i\leq N} \sqrt{T}\hat{\mu}_i$. We depart from the analysis in their paper by introducing a studentized test statistic. As suggested by Hansen (2005), studentization can improve the power in tests of predictive performance in many empirical applications where $\hat{\mu}_i$ displays strong forms of heteroskedasticity. Such heteroskedasticity may arise due to differences in sample lengths used to compute the test statistics or due to differences in the degree of variability in the loss differentials across different variables.

Consider the following test statistic for the maximum value of the average loss differential, computed across the $i = 1, ..., N$ cross-sectional units:

$$R_T = \max_{1\leq i\leq N} \frac{T^{-1/2}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}}{\hat{a}_i}, \tag{2.8}$$

where $I_{i,t+h} = \mathbf{1}\{\Delta L_{i,t+h} \text{ is observed}\}$ is an indicator for whether the loss differential for unit i is observed in period $t+h$ and $\hat{a}_i > 0$ is a normalizing quantity that is either deterministic or estimated from the data. Ideally, we observe all the loss differentials and so $I_{i,t+h} = 1$ for all i and all $t+h$. In practice, it is common that not all of the $\Delta L_{i,t+h}$'s are available.

To account for serial dependency in loss differentials, let B_T be an integer that measures the average block length used in the bootstrap and define the number of blocks $K := K_T = \lfloor T/B_T \rfloor$. For $j \in \{1, ..., K-1\}$, let $H_j = \{(j-1)B_T + 1, ..., jB_T\}$ and $H_K = \{(K-1)B_T + 1, ..., T\}$ denote the jth and Kth time-series blocks, respectively.

Example 2.3.1. We consider a variety of possible normalizations of the test statistic, R_T:

- No normalization: $\hat{a}_i = 1$ for $1 \leq i \leq N$. This choice does not attempt to balance differences in $Var(T^{-1/2}\sum_{t=1}^{T} I_{i,t+h}\Delta L_{i,t+h})$ across i. Hence, the behavior of R_T will tend to be dominated by those units i with the largest values of $Var(T^{-1/2}\sum_{t=1}^{T} I_{i,t+h}\Delta L_{i,t+h})$ which are likely to produce the most extreme values of the numerator in (2.8).

- Full normalization: $\hat{a}_i = \sqrt{K^{-1}\sum_{j=1}^{K}\left(B_T^{-1/2}\sum_{t+h\in H_j} I_{i,t+h}(\Delta L_{i,t+h} - \hat{\mu}_i)\right)^2}$ with $\hat{\mu}_i = T^{-1}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}$. This normalization is an estimate of the long-run variance and hence can correct the cross-sectional differences in scale of $T^{-1/2}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}$. However, this could be a rather noisy estimate as it is essentially computed from K observations, each being the sum of data in a block. In small samples, the noise in this estimate could create substantial size distortions.

- Partial normalization: $\hat{a}_i = \sqrt{T^{-1}\sum_{t+h=1}^{T} I_{i,t+h}(\Delta L_{i,t+h} - \hat{\mu}_i)^2}$. This choice of normalization corrects for different scales in the unconditional variance of $Var(I_{i,t+h}\Delta L_{i,t+h})$. This is a sensible choice when the variability of $I_{i,t+h}\Delta L_{i,t+h}$ differs significantly across i but does not guarantee that the variance of $T^{-1/2}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}/\hat{a}_i$ stays approximately constant across i.[9]

- Sample-sized normalization: $\hat{a}_i = \sqrt{T_i/T}$, where $T_i = \sum_{t+h=1}^{T} I_{i,t+h}$. This choice is sensible when T_i/T varies significantly across i and the variance of $T^{-1/2}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}$ is driven by the number of observations in each series.

- Double normalization: $\hat{a}_i = \sqrt{T_i/T} \times \sqrt{K^{-1}\sum_{j=1}^{K}\left(\sum_{t+h\in H_j} I_{i,t+h}(\Delta L_{i,t+h} - \hat{\mu}_i)\right)^2}$. This choice normalizes both by the number of observations T_i and the long-run variance of the observed $\Delta L_{i,t+h}$.

[9] The reason is that the unconditional variance is not the same as the long-run variance of the partial sum $T^{-1/2}\sum_{t+h=1}^{T} I_{i,t+h}\Delta L_{i,t+h}$ since the latter also depends on any serial correlation in $\Delta L_{i,t+h}$.

Critical values for R_T in (2.8) can be based on the following multiplier bootstrap procedure. Let $\{\xi_j\}_{j=1}^K$ be a set of i.i.d $N(0,1)$ variables used to construct the test statistic

$$R_T^* = \max_{1\leq i\leq N} R_{i,T}^*, \tag{2.9}$$

where

$$R_{i,T}^* = \frac{K^{-1/2}\sum_{j=1}^K \xi_j \left(B_T^{-1/2}\sum_{t+h\in H_j} I_{i,t+h}\Delta L_{i,t+h}\right)}{\hat{a}_i}.$$

To cover different hypotheses and test statistics encountered in empirical analysis, we consider a general setting in which the number of forecasts of $y_{i,t+h}$ can be large. Suppose that for each $1\leq i\leq N$, we have a set of $\mathcal{D}_i$ models generating $|\mathcal{D}_i|$ forecasts for all $1\leq t+h\leq T$, namely $\hat{y}_{i,t+h|t,m}$ for $m\in\mathcal{D}_i$, in addition to the benchmark $\hat{y}_{i,t+h|t,m_0}$. Hence, we can allow the number of forecasts to vary across variables, although for simplicity we assume that this number does not depend on time.[10]

The following general setup covers as special cases the earlier null hypotheses:[11]

$$H_0: \max_{1\leq i\leq N}\max_{m\in\mathcal{D}_i} E\left[\Delta L_{i,t+h,m}\right]\leq 0. \tag{2.10}$$

To test this null, define

$$U_{t+h} = \left(\{\Delta L_{1,t+h,m}\}_{m\in\mathcal{D}_1}, \{\Delta L_{2,t+h,m}\}_{m\in\mathcal{D}_2}, ..., \{\Delta L_{N,t+h,m}\}_{m\in\mathcal{D}_N}\right)$$

so U_{t+h} is a column vector of dimension $\mathcal{N}=\sum_{i=1}^N|\mathcal{D}_i|$ with kth component denoted by $U_{k,t+h}$. Consider the test statistic

$$\tilde{R}_T = \max_{1\leq k\leq\mathcal{N}} \frac{T^{-1/2}\sum_{t+h=1}^T U_{k,t+h}}{\hat{a}_k}, \tag{2.11}$$

[10]Extension to the case where $\mathcal{D}_i$ is time-varying is conceptually trivial but makes the notation more cumbersome without offering additional insights.

[11]For simplicity we ignore the dimension referring to the forecast horizon, but the analysis is easily extended to also cover this.

where $\hat{a}_k$ is computed using any of the schemes described in Example 2.3.1.

Bootstrap critical value are constructed analogously

$$\tilde{R}^*_T = \max_{1\leq k\leq \mathcal{N}} \tilde{R}^*_{k,T}, \tag{2.12}$$

where

$$\tilde{R}^*_{k,T} = \frac{K^{-1/2}\sum_{j=1}^{K}\xi_j\left(B_T^{-1/2}\sum_{t+h\in H_j}U_{k,t+h}\right)}{\hat{a}_k}.$$

To establish the distributional properties of the test statistic in (2.11), we require a set of regularity conditions. To this end, let $W_{k,t+h} = U_{k,t+h} - E(U_{k,t+h})$, while $W_{t+h} = \left(W_{1t+h},...,W_{\mathcal{N}(t+h)}\right)$. We summarize our assumptions as follows:

Assumption 1. *Suppose that the following conditions hold:*

(1) The distribution of W_{t+h} does not depend on $t+h$.

(2) $P(\max_{1\leq t+h\leq T}\|W_{t+h}\|_\infty \leq D_T) = 1$ for some $D_T \geq 1$.

(3) $\{W_{t+h}\}_{t+h=1}^{T}$ is β-mixing with mixing coefficient $\beta_{\text{mixing}}(\cdot)$.

(4) $c_1 \leq E\left(k^{-1/2}\sum_{t+h=s+1}^{s+k}W_{j,t+h}\right)^2$, $E\left(k^{-1/2}\sum_{t+h=s+1}^{s+k}W_{j,t+h}\right)^2 \leq C_1$ for any j, s and k.

(5) $T^{1/2+b}D_T\log^{5/2}(\mathcal{N}T) \lesssim B_T \lesssim T^{1-b}/(\log\mathcal{N})^2$ and $\beta_{\text{mixing}}(s) \lesssim \exp(-b_1 s^{b_2})$ for some constant $b, b_1, b_2 > 0$.

(6) There exist a nonrandom vector $a = (a_1,...,a_{\mathcal{N}})' \in \mathbb{R}^{\mathcal{N}}$ and constants $\kappa_1, \kappa_2 > 0$ such that $\kappa_1 \leq a_j \leq \kappa_2$ for all $1 \leq j \leq \mathcal{N}$ and $\max_{1\leq j\leq\mathcal{N}}|\hat{a}_j - a_j| = o_P(1/\log\mathcal{N})$.

Part (1) of Assumption 1 requires strict stationarity and can be relaxed at the expense of more technicalities in the proof. Part (2) imposes a bound on the tail behavior of the loss difference. When the loss difference is bounded, we can choose D_T to be a constant; when the loss difference is sub-Gaussian, we can choose $D_T \asymp \sqrt{\log(\mathcal{N}T)}$ and adapt the proof to handle $P(\max_{1\leq t\leq T}\|W_{t+h}\|_\infty \leq D_T) \to 1$. This bound on the variables is needed for the high-dimensional bootstrap and Gaussian approximation even in the i.i.d case; see Chernozhukov, Chetverikov

and Kato (2013, 2017, 2018).[12] The β-mixing condition in part (3) is routinely imposed in the literature and holds for many stochastic processes. Part (4) requires the loss differences for all variables to be of roughly the same order of magnitude. Part (5) imposes rate conditions; notice that we allow $\mathcal{N} \gg T$. Finally, part (6) states that $\hat{a}_j$ needs to be uniformly consistent for some non-random quantity that is bounded away from zero and infinity.

We can verify that part (6) of Assumption 1 holds for the normalization schemes listed above as we next formalize:

Lemma 2.3.1. *Let Assumption 1(1)-(5) hold. Then all the normalizations in Example 2.3.1 satisfy part (6) of Assumption 1.*

Using Assumption 1, we have the following result:

Theorem 2.3.1. *Suppose Assumption 1 holds. Under*

$$H_0 : \max_{1 \leq i \leq N} \max_{m \in \mathcal{D}_i} E\left[\Delta L_{i,t+h,m}\right] \leq 0,$$

we have

$$\limsup_{T \to \infty} P\left(\tilde{R}_T > \tilde{Q}^*_{T,1-\alpha}\right) \leq \alpha,$$

where $\tilde{Q}^*_{T,1-\alpha}$ is the $(1-\alpha)$ quantile of $\tilde{R}^*_T$ conditional on the data. Moreover, if $E(\Delta L_{i,t+h,m}) = 0$ for all $1 \leq i \leq N$ and $m \in \mathcal{D}_i$, then

$$\limsup_{T \to \infty} P\left(\tilde{R}_T > \tilde{Q}^*_{T,1-\alpha}\right) = \alpha.$$

[12] One way to relax part (2) of Assumption 1 is to use the union bound together with moderate deviation inequalities for self-normalized sums, but this might lead to more conservative procedures; see Chernozhukov, Chetverikov and Kato (2018). Notice that although other parts of Assumption 1 require $D_T \ll T^{1/2-2b}$, it is possible to relax D_T to be larger than $\sqrt{\log(\mathcal{N}T)}$. To do so, we only need components of W_{t+h} to have bounded m-th moment with m satisfying $\mathcal{N} \ll T^{(1/2-2b)m-1}$. Hence, certain "heavy-tailed" processes such as GARCH processes can be allowed, provided that they have a sufficient number of moments. For conditions ensuring that this hold for a GARCH(1,1) process, see Bollerslev (1986).

Theorem 2.3.1 establishes the asymptotic validity of the bootstrap procedure. Under the null of equal expected loss for all variables, the multiplier bootstrap test is asymptotically exact and, hence, not conservative.

The result readily applies to comparisons across multiple forecast horizons as described in Section 2.2.3; for this case we only need to replace U_{t+h} with

$$U_t = \left(\{\Delta L_{1,t+h,m}\}_{m\in\mathcal{D}_1,1\leq h\leq H}, \{\Delta L_{2,t+h,m}\}_{m\in\mathcal{D}_2,1\leq h\leq H}, \ldots, \{\Delta L_{N,t+h,m}\}_{m\in\mathcal{D}_N,1\leq h\leq H}\right).$$

The studentization used for $\tilde{R}_T$ serves a similar role as the self-normalization in Chernozhukov, Chetverikov and Kato (2018) for the independent case and can improve the power of the test. By arguments similar to those in Chernozhukov, Chetverikov and Kato (2018), we expect the test to have non-trivial power against alternatives of order $\max_{1\leq i\leq N}\max_{m\in\mathcal{D}_i} E(\Delta L_{i,t+h,m}) = O(\sqrt{T^{-1}\log\mathcal{N}})$ with a rate that is minimax optimal. Since the number of hypotheses tested only enters through a logarithmic factor, the proposed test has consistency against fixed alternatives even if this number grows exponentially with T.

It is important to note that the dimension $\mathcal{N}$ only has a very small impact on the requirements that guarantee the validity of the procedure. This holds because in the regularity conditions (Assumption 1), only the rate $\log(\mathcal{N})$ matters, which means that $\mathcal{N}$ can increase at the rate T^c for any constant $c > 0$.

Theorem 2.3.1 provides a transparent technical result on the nature of the normalization. The main concern is whether the estimation errors (each of order $O(T^{-1/2})$) in the vast number ($\mathcal{N}$) of normalizations $\hat{a}_j$ could create problems for the asymptotic behavior of the test. From Chernozhukov, Chetverikov and Kato (2018), we know that the sample variance can be used as valid normalizations in the independent case. However, it is not clear whether the specific structure of the sample variance plays an important role or how to obtain a meaningful "sample variance" in the dependent case. Theorem 2.3.1 answers this question and states that the only

requirement for the normalization terms is that they are consistent for *some* non-random quantities a_j at a very slow rate. We do not need to specify what these non-random quantities are and they do not strictly have to be variances.

2.3.2 Family-wise Error Rate

We next show how to use Theorem 2.3.1 to construct confidence sets for under- and overperforming units. For notational simplicity, we consider $|\mathcal{D}_i| = 1$ so $\mathcal{N} = N$. Define $A = \{i : \mu_i > 0\}$, where $\mu_i = T^{-1}\sum_{t+h=1}^{T} E\Delta L_{i,t+h}$, so that A is the set of units, i, for which an alternative forecast, m, is genuinely better than the benchmark, m_0.

To estimate this set, consider

$$\hat{A} = \left\{ i : \frac{T^{-1/2}\sum_{t+h=1}^{T} \Delta L_{i,t+h}}{\hat{a}_i} > Q^*_{T,1-\alpha} \right\}.$$

If $\hat{A}$ contains a unit that is not in A, i.e., $\hat{A}\backslash A \neq \emptyset$, $\hat{A}$ makes a false discovery since it includes units for which the alternative forecast performs no better than m_0.

A consequence of Theorem 2.3.1 is that the probability of a false discovery is asymptotically at most α. To see this, notice that

$$\begin{aligned}
&P\left(\hat{A}\backslash A \neq \emptyset\right) \\
&= P\left(\frac{T^{-1/2}\sum_{t=1}^{T} \Delta L_{i_0,t+h}}{\hat{a}_{i_0}} > Q^*_{T,1-\alpha} \quad \text{for some } i_0 \in \hat{A}\backslash A \right) \\
&\leq P\left(\max_{i \in A^c} \frac{T^{-1/2}\sum_{t=1}^{T} \Delta L_{i,t+h}}{\hat{a}_i} > Q^*_{T,1-\alpha} \right) \\
&\leq \alpha + o(1),
\end{aligned}$$

where the last inequality follows by Theorem 2.3.1 applied to A^c (instead of $\{1,...,N\}$). By construction, $\max_{i \in A^c} E\mu_i \leq 0$. We summarize this result as follows:

Corollary 2.3.1. *Suppose Assumption 1 holds. Consider A and $\hat{A}$ defined above. Then*

$$\limsup_{T\to\infty} P\left(\hat{A} \subseteq A\right) \geq 1-\alpha.$$

Hence, with probability at least $1-\alpha$, $\hat{A}$ only selects cases in which the alternative forecast outperforms the benchmark.

Our approach to bootstrapping the distribution of the maximum value chosen from a large set of test statistics is related to the reality check methodology pioneered by White (2000), though there are also important differences. Most notably, White (2000) tests hypotheses about the population parameter value.[13] Moreover, he assumes that the forecasts are generated by parametric models and thus take the form $f_{t+h|t} = f(Z_t, \hat{\beta}_h)$, using the parameter updating scheme discussed in West (1996).[14] Finally, White (2000) assumes that the number of forecasts each time period is fixed, whereas we allow it to be expanding with the sample size, T. As pointed out by White (2000) (page 1111) and Chernozhukov, Chetverikov and Kato (2018) (Comment 4.7), assuming a fixed number of forecasts, models or moment conditions is an important limitation in many empirical applications. Here we allow the number of forecasts to be much larger than T which can be quite important for panel forecasts with large N.

2.3.3 Moment Selection

The literature on testing moment inequalities suggests that test power can be improved by reducing the number of inequalities via moment selection; see e.g., Hansen (2005); Andrews and Soares (2010); Romano, Shaikh and Wolf (2014). To see how this works, we start with the goal of testing moment inequalities in $A = \{1,...,N\}$.[15] We would like to use the data to find a

[13] See, e.g., the discussion on page 1099 in White (2000).

[14] See Assumption A.2 in the Appendix to West (1996).

[15] This can be generalized to $A = \{1,...,\mathcal{N}\}$, where $\mathcal{N}$ varies depending on which null hypothesis is being tested. For example, $\mathcal{N} = N$ in H_0^S, whereas $\mathcal{N} = N \times M$ in H_0^{NSS}. Again, for simplicity, we focus on the case of $|\mathcal{D}_i| = 1$ (so $\mathcal{N} = N$).

set A_0 such that with high probability, say $1-\beta$, the moment inequalities contained in $A\backslash A_0$ are satisfied. Provided that this holds, we only need to test the moment inequalities in A_0. When $|A_0| \ll |A|$, excluding the moment inequalities in $A\backslash A_0$ can be expected to improve the power of the test, although we need to adjust the size of the test to be $\alpha-\beta$ when testing the moment inequalities in A_0.

Most of the literature on testing moment inequalities focuses on the case where $|A|$ is fixed.[16] Here, we follow the spirit of Romano, Shaikh and Wolf (2014) and use a bootstrapped threshold. We summarize the details in Algorithm 1.

Algorithm 1. *Implement the following steps:*

1. *Choose* $\beta \in (0,\alpha)$ *to be either a constant or a sequence tending to zero.*

2. *Compute*
$$R_{i,T} = \frac{T^{-1/2}\sum_{t=1}^{T}\Delta L_{i,t+h}}{\hat{a}_i} \qquad \forall 1 \leq i \leq N.$$

3. *Compute the bootstrapped threshold* C_β*, which is the* $1-\beta$ *quantile of* $\|R_T^*\|_\infty$ *conditional on the data, where* R_T^* *is defined in (2.9). In other words,* $P(\|R_T^*\|_\infty > C_\beta \mid data) = \beta$.

4. *Select* $A_0 = \{i:\ R_{i,T} > -C_\beta\}$.

5. *Compute the test statistic* $\max_{i\in A_0} R_{i,T}$.

6. *Compute the bootstrap critical value* $C_{\alpha-\beta,A_0}$ *satisfying* $P(\max_{i\in A_0} R_{i,T}^* > C_{\alpha-\beta,A_0} \mid data) = \alpha-\beta$*, where* $R_{i,T}^*$ *is defined in (2.9).*

Although this procedure requires us to decrease the size of the test from α to $\alpha-\beta$ for small β, the test statistic and the bootstrap critical value are computed as the maximum over indices in A_0 rather than over the original set $\{1,...,N\}$. When $|A_0|$ is much smaller than N, the

[16] Hansen (2005) proposes a threshold of $\sqrt{\log\log N}$ based on the law of iterated logarithm so that A_0 contains moments whose sample counterpart is larger than $-\sqrt{T^{-1}\log\log N}$.

price we pay for using a reduced nominal size is small and the procedure can result in improved power.[17]

2.3.4 Monte Carlo Simulations

Appendix A reports the results from a set of Monte Carlo simulations which we use to study the finite sample properties of our test statistics. We draw the following conclusions from these simulations. Both studentized and non-studentized test statistics have reasonable size properties when N and M are small, regardless of the time-series dimension, T. For small N, M and T, the test statistics are slightly oversized. However, as N and M grow bigger, the test statistics tend to become under-sized. Undersizing is particularly pronounced for the studentized test statistic when $\alpha = 0.05$ but is less of a concern for $\alpha = 0.10$. Using a critical level of $\alpha = 0.10$ for the studentized test statistic in many cases gets us close to a size of 5-10%.

The Monte Carlo simulations also show that the power of the studentized test statistic is far better than that of the non-studentized test statistic, even when size-adjusted critical values are used in the power calculations. This is an important consideration because accounting for the multiple hypothesis testing problem easily leads to procedures with weak power and, hence, conservative inference. For this reason, we use studentized test statistics with a size of $\alpha = 0.10$ throughout our empirical applications.

2.4 Comparing Forecasting Performance in Individual Cross-sections

In situations with a large number of variables, N, we can attempt to exploit the cross-sectional dimension of the data to address whether the performance of any individual forecaster, averaged cross-sectionally, is better than the benchmark in a *single* period or over a short time

[17] The high-dimensional testing problem is further discussed by Chernozhukov, Chetverikov and Kato (2018).

span. In this section we introduce a set of assumptions about a common-factor structure in the individual forecast errors that allows us to conduct inference on predictive accuracy using individual cross-sections.

2.4.1 Common Factors in Forecast Errors

Tests for superior predictive skills in a single cross-section can be based on the distribution of the average cross-sectional loss differential of forecaster m, $\hat{\mu}_{m,t+h} = N^{-1}\sum_{i=1}^{N}\Delta L_{i,t+h,m}$. For inference to be valid, we require the use of a cross-sectional central limit theorem for the resulting test statistic which means that the cross-sectional dependency in the loss differentials cannot be too strong. To establish conditions under which this holds, consider the following factor structure for the forecast errors

$$e_{i,t+h,m} = \lambda_{i,m}' f_{t+h} + u_{i,t+h,m}, \tag{2.13}$$

for $1 \leq i \leq N$ and $1 \leq t+h \leq T$, where $f_{t+h} \in \mathbb{R}^k$ is a set of latent factors common to the forecast errors. Economic variables often contain common components that none of the forecasters anticipated and these can make forecast errors highly correlated. The factor structure assumed in (2.13) is a natural representation of this situation.

Under the factor structure in (2.13), the squared error loss differential takes the form

$$\begin{aligned}\Delta L_{i,t+h,m} &= \left(\lambda_{i,m_0}' f_{t+h} + u_{i,t+h,m_0}\right)^2 - \left(\lambda_{i,m}' f_{t+h} + u_{i,t+h,m}\right)^2 \\ &= f_{t+h}'(\lambda_{i,m_0}\lambda_{i,m_0}' - \lambda_{i,m}\lambda_{i,m}')f_{t+h} + u_{i,t+h,m_0}^2 - u_{i,t+h,m}^2 \\ &\quad + 2f_{t+h}'(\lambda_{i,m_0}u_{i,t+h,m_0} - \lambda_{i,m}u_{i,t+h,m}).\end{aligned} \tag{2.14}$$

To rule out that the cross-sectional dependencies are so strong as to prevent us from establishing distributional results for the cross-sectional average loss differentials, we assume that the idiosyncratic terms are independent conditional on the factor structure:

Assumption 2. *Let $\mathcal{F}$ be the σ-algebra generated by $\{f_{t+h}\}_{1\leq t+h\leq T}$ and $\{\lambda_{i,m}\}_{1\leq i\leq N,\ 0\leq m\leq M}$. Conditional on $\mathcal{F}$, $\{u_i\}_{1\leq i\leq N}$ is independent across i and $E(u_i \mid \mathcal{F}) = 0$, where $u_i = \{u_{i,t+h,m}\}_{1\leq t+h\leq T, 1\leq m\leq M} \in \mathbb{R}^{T\times M}$.*

Using Assumption 2, we have

$$\frac{1}{N}\sum_{i=1}^{N}\Delta L_{i,t+h,m} - E\left(\frac{1}{N}\sum_{i=1}^{N}\Delta L_{i,t+h,m} \mid \mathcal{F}\right) = \frac{1}{N}\sum_{i=1}^{N}\xi_{i,t+h,m},$$

where $\xi_{i,t+h,m} = 2f'_{t+h}(\lambda_{i,m_0}u_{i,t+h,m_0} - \lambda_{i,m}u_{i,t+h,m}) + \left(u^2_{i,t+h,m_0} - u^2_{i,t+h,m}\right) - E\left(u^2_{i,t+h,m_0} - u^2_{i,t+h,m} \mid \mathcal{F}\right)$. Under Assumption 2, $\{\xi_{i,t+h,m}\}_{i=1}^{N}$ has mean zero and is independent across i conditional on $\mathcal{F}$. Therefore, we can use a central limit theorem to show that $\frac{1}{N}\sum_{i=1}^{N}\Delta L_{i,t+h,m}$ is an asymptotically normal estimator for $E\left(\frac{1}{N}\sum_{i=1}^{N}\Delta L_{i,t+h,m} \mid \mathcal{F}\right)$. By virtue of a high-dimensional Gaussian approximation, we can extend this intuition to a simultaneous test across many periods, $t+h$, and/or forecasts, m.

2.4.2 Hypotheses about Performance in Individual Cross-sections

The conditional null that, given $\mathcal{F}$, no forecaster, $m = 1,....,M$, is better, on average across all units, than the benchmark in a particular time period, $t+h$, can be tested by considering the maximum of the expected value of the cross-sectionally averaged loss differentials in period $t+h$, $\overline{\Delta L}_{t+h,m} = \frac{1}{N}\sum_{i=1}^{N}\Delta L_{i,t+h,m}$:

$$H_0^{ES}: \max_{(t+h,m)\in\mathcal{A}} E\left(\overline{\Delta L}_{t+h,m} \mid \mathcal{F}\right) \leq 0, \tag{2.15}$$

where $\mathcal{A}$ is the set defined by $\mathcal{A} = \{t+h\}\times\{m = 1,....,M\}$. The hypothesis in (2.15) is strictly about performance in period $t+h$ so we refer to this null as characterizing "event skills" (ES). Equivalently, the null in (2.15) is concerned with whether the average predictive accuracy in period $t+h$ of *any* of the M forecasters is better than that of the benchmark.

We can also test whether, across all periods $t+h=1,...,T$ and all forecasters, $m=1,...,M$, *any* of the forecasters were more accurate, on average across all units, than the benchmark in *any* time period (given $\mathcal{F}$):

$$H_0^{ESall}: \max_{t+h\in\{1,...,T\}} \max_{m\in\{1,...,M\}} E\left(\overline{\Delta L}_{t+h,m} \mid \mathcal{F}\right) \leq 0, \tag{2.16}$$

where now $\mathcal{A}=\{t+h=1,...,T\}\times\{m=1,...,M\}$ in (2.16). This null can be used to test whether any forecaster's cross-sectional average performance beats the benchmark during any period in the sample.

Finally, we can also consider the average performance over small subsamples of time, e.g., during individual calendar years or during some periods of time characterized, e.g., by high volatility. Denoting the subset of dates as T_c, we can accommodate this case by re-defining $\mathcal{A}=\{t+h\in T_c\}\times\{m=1,...,M\}$, and considering the null hypothesis

$$H_0^{ES_c}: \max_{t+h\in T_c} \max_{m\in\{1,...,M\}} E\left(\overline{\Delta L}_{t+h,m} \mid \mathcal{F}\right) \leq 0, \tag{2.17}$$

2.4.3 Test statistics

The test statistic we propose for testing (2.15), (2.16) or (2.17) is given by

$$Z = \max_{(t+h,m)\in\mathcal{A}} \frac{\sqrt{N}\overline{\Delta L}_{t+h,m}}{\sqrt{N^{-1}\sum_{i=1}^{N}\widetilde{\Delta L}_{i,t+h,m}^{2}}}, \tag{2.18}$$

where $\widetilde{\Delta L}_{i,t+h,m}=\Delta L_{i,t+h,m}-\overline{\Delta L}_{t+h,m}$ is the demeaned loss differential of variable i for forecaster m. Critical values for this test statistic can be obtained from a bootstrap

$$Z_* = \max_{(t+h,m)\in\mathcal{A}} \frac{N^{-1/2}\sum_{i=1}^{N}\varepsilon_i\widetilde{\Delta L}_{i,t+h,m}}{\sqrt{N^{-1}\sum_{i=1}^{N}\widetilde{\Delta L}_{i,t+h,m}^{2}}}, \tag{2.19}$$

where the multipliers $\varepsilon_i \sim N(0,1)$ are generated independently of the data. Note that we assume cross-sectional conditional independence for the idiosyncratic terms. Moreover, we assume that the multipliers ε_i are i.i.d, rather than having the block structure needed to handle serial dependence in the test statistics which use data from multiple time periods.

Using these assumptions, we can establish the validity of the above procedure:

Theorem 2.4.1. *Let Assumption 2 hold. Suppose that* $(\kappa_{N,3}^3 \vee \kappa_{N,4}^2 \vee B_N)^2 \log^{7/2}(TMN) \lesssim N^{1/2-c}$ *for some* $c \in (0,1/2)$*, where* $B_N = (E \max_{t,m,i} |\xi_{i,t+h,m}|^4)^{1/4}$*,* $\kappa_{N,3} = (\max_{i,t,m} E|\xi_{i,t+h,m}|^3)^{1/3}$ *and* $\kappa_{N,4} = (\max_{i,t,m} E|\xi_{i,t+h,m}|^4)^{1/4}$*. Then under* H_0 *in (2.15) we have*

$$\limsup_{N\to\infty} P\left(Z > Q^*_{N,1-\alpha,Z}\right) \leq \alpha,$$

where $Q^*_{N,1-\alpha,Z}$ *is the* $(1-\alpha)$ *quantile of* Z_* *conditional on the data. Moreover, if* $E\left(\frac{1}{N}\sum_{i=1}^N \Delta L_{i,t+h,m} \mid \mathcal{F}\right) = 0$ *for all* $(t+h,m) \in \mathcal{A}$*, then*

$$\limsup_{N\to\infty} P\left(Z > Q^*_{N,1-\alpha,Z}\right) = \alpha.$$

Here, B_N, $\kappa_{N,3}$ and $\kappa_{N,4}$ measure the tail of $\xi_{i,t+h,m}$, which is the deviation of the loss differential $\Delta L_{i,t+h,m}$ from its conditional mean. When deviations are bounded, $B_N, \kappa_{N,3}$ and $\kappa_{N,4}$ are positive constants. If $\xi_{i,t+h,m}$ has sub-Gaussian tails, then $B_N = O(\log(TMN))$ and $\kappa_{N,3}$ and $M_{N,4}$ are constants. The proof of Theorem 2.4.1 follows almost exactly the same lines as the proof of Theorem 4.3 of Chernozhukov, Chetverikov and Kato (2018) with two exceptions: (1) the independence assumption is replaced by conditional independence given $\mathcal{F}$ and (2) the assumption of identical distributions is changed and can be handled by slight changes to the definition of B_N, $\kappa_{N,3}$ and $\kappa_{N,4}$. We omit the details of the proof for this reason.

Theorem 2.4.1 is stated for the null in (2.15), but the null hypotheses in (2.16) or (2.17) can be tested in the same way by replacing $\max_{(t+h,m)\in\mathcal{A}}$ with either $\max_{t+h\in\{1,\ldots,T\}} \max_{m\in\{1,\ldots,M\}}$ or with $\max_{t+h\in T_c} \max_{m\in\{1,\ldots,M\}}$ in (2.18) and (2.19).

2.5 Evaluating the Term Structure of Forecast Errors for the IMF

Do macroeconomic forecasts become significantly more accurate over time as the distance to the target date is reduced so that, on average, short-term forecasts are be more precise than long-run forecasts? We would expect this property to hold as a result of flows of news which should enable forecasters to update their predictions by conditioning on an expanded information set. However, whether such improvements really do occur is likely to depend on the quality and relevance of the information flow and on whether the forecaster exploits such information in a reasonably efficient manner. For example, in cases with noisy and irregular statistical data, short-term forecasts may not be much more accurate than long-term forecasts. How far out in time (prior to the target date) any improvements occur is also likely to reflect the persistence of the underlying variable: Improvements in predictive accuracy can be expected to occur at longer horizons for variables that are highly persistent–since news shocks affect outcomes further ahead in time–compared to variables with little persistence.

The term structure of forecast errors–the mapping between predictive accuracy and the forecast horizon–therefore contains information about the data generating process of the variable being predicted, the arrival rate and quality of new information, as well as on forecasters' ability to use such information to improve forecast accuracy, i.e., the forecaster's "learning curve".

Macroeconomic surveys often ask participants to predict outcomes across across multiple forecast horizons. For example, the Survey of Professional Forecasters and the Blue Chip forecasts request that survey participants forecast outcomes several quarters or even years ahead in time. This enables comparisons of forecast accuracy across multiple horizons and makes this type of data ideally suited for analysis by means of our new Sup tests.

This section provides an empirical application that uses our methods to compare the accuracy at long and short horizons of the International Monetary Fund's (IMF) World Economic

Outlook (WEO) forecasts of five variables across up to 185 countries over a 30-year period. The WEO publication contains the "flagship" forecasts published by the IMF which receive more attention perhaps than any other global forecasts and is widely covered by public media.[18]

To be able to pinpoint for which variables and at which horizons any improvements in predictive accuracy occurs, we conduct our initial analysis separately for each of the five variables and at particular forecast horizons. Instead we take advantage of the large cross-sectional (country-level) dimension of the data which allows us to analyze forecasting performance for particular clusters or subsets of countries and identify for which types of economies the IMF forecasts become significantly more accurate as the forecast horizon is reduced. We also conduct analyses that consider performance across all dimensions of the data (horizons, variables, and countries). Using the richness of this information, we can better understand the IMF's learning curve as well as determinants of any improvements in predictive accuracy as a function of the forecast horizon.

2.5.1 Predictive Accuracy Across Different Horizons

WEO forecasts are published twice each year, namely in April (labeled Spring, or S) and October (Fall, or F). The WEO publication contains forecasts for the current year ($h = 0$), next year ($h = 1$), and up to five years ahead in time ($h = 5$). Since the time-series dimension of our sample is relatively short, we focus our benchmark analysis on the current-year and next-year forecast horizons in order to have a sufficient number of observations to compute sample averages. This gives us four horizons, listed in decreasing order: $\{h = 1, S;\ h = 1, F;\ h = 0, S;\ h = 0, F\}$. However, we also show results for the 2-5 year horizons at the end of this section.

For a subset of mostly advanced countries, current-year forecasts go back to 1990, while next-year forecasts start in 1991. For other countries, the forecasts start later, providing us with a somewhat shorter data sample. In all cases, the last outcome for our data is recorded for 2019, giving us a maximum sample of 30 years. Our analysis covers forecasts of five variables, namely

[18] The WEO forecasts have been the subject of a number of academic studies summarized in Timmermann (2007).

real GDP growth, inflation, import growth, export growth, and the current account balance as a percentage of GDP.

We expect predictive accuracy to improve as the forecast horizon is reduced and more information about the outcome becomes available. Because we observe the WEO forecasts of a particular outcome at different horizons, we can test if this holds. Ordering the WEO forecasts from the longest ($h = 1, S$) to the shortest ($h = 0, F$) horizon, under the squared error loss in (2.1) we have, from Patton and Timmermann (2011)

$$E[e^2_{i,h=0,F}] \leq E[e^2_{i,h=0,S}] \leq E[e^2_{i,h=1,F}] \leq E[e^2_{i,h=1,S}]. \tag{2.20}$$

Using these inequalities, we compare predictive accuracy across four pairs of long and short forecast horizons (h_L and h_S, respectively), namely (i) Spring versus fall next year ($h_L = 1, S; h_S = 1, F$); (ii) Fall next year versus spring current-year ($h_L = 1, F; h_S = 0, S$); (iii) Spring versus fall current-year ($h_L = 0, S; h_S = 0, F$); and (iv) Spring next year versus fall current-year ($h_L = 1, S; h_S = 0, F$).[19] The fourth comparison is concerned with cumulative improvements in predictive accuracy across the three shorter six-month intervals and so summarizes the lessons learned over the 18-month period preceding the Fall current year forecast.

Our comparisons use the squared forecast error loss differential for variable i in period t given forecasts generated at short and long horizons, $t - h_S$ and $t - h_L$ for $h_L > h_S$:

$$\Delta L_{i,t,h_L \to h_S} = (y_{i,t} - \hat{y}_{i,t|t-h_S})^2 - (y_{i,t} - \hat{y}_{i,t|t-h_L})^2. \tag{2.21}$$

Similarly, define the change in the squared forecast error from reversing the order and going from the short to the long horizon:

$$\Delta L_{i,t,h_S \to h_L} = (y_{i,t} - \hat{y}_{i,t|t-h_L})^2 - (y_{i,t} - \hat{y}_{i,t|t-h_S})^2. \tag{2.22}$$

[19] Current-year forecasts ($h = 0$) can be viewed as a mixture of nowcasts and forecasts.

To understand the IMF's learning curve, we initially apply our approach to test for improvements in squared error loss accuracy across all countries for a given pair of forecast horizons and a given variable. For example, to test the null that, for each country, i, the short-term forecast is at least as accurate as the long-term forecast, we test the null[20]

$$H_0 : \max_{i \in \{1,\ldots,N\}} \left(E[\Delta L_{i,t,h_L \to h_S}] \right) \leq 0. \tag{2.23}$$

To test the converse proposition that, for each country, i, forecast accuracy does not improve as the forecast horizon is reduced, we test the null

$$H_0 : \max_{i \in \{1,\ldots,N\}} \left(E[\Delta L_{i,t,h_S \to h_L}] \right) \leq 0. \tag{2.24}$$

2.5.2 Results for Individual Countries

Rejections of the null in (2.24) imply that forecasts are improving as the forecast horizon gets shorter. To get a sense of whether the accuracy of the WEO forecasts improves across different horizons, Figure 2.1 shows a heat diagram depicting how the accuracy of the WEO country-level inflation forecasts evolves as we move from $h = 1,S$ to $h = 1,F$ (top left panel), from $h = 1,F$ to $h = 0,S$ (top right panel), from $h = 0,S$ to $h = 0,F$ (bottom left panel) and on a cumulative basis ($h = 1,S$ versus $h = 0,F$, bottom right). The last comparison measures whether the WEO current-year Fall forecasts ($h = 0,F$) are more accurate than the prior-year Spring forecasts ($h = 1,S$) and thus accumulates any gains in accuracy over the three preceding six-month intervals. Colors applied to each country are based on the p-values for testing the null in (2.24). Red color corresponds to small p-values, indicating that short horizon forecasts are significantly more accurate than long-horizon forecasts. Green color indicates weak evidence

[20]For simplicity, we suppress references to the variable in the subscript.

against significant improvements in accuracy as the forecast horizon gets shorter.[21]

We find no evidence that inflation forecasts at the longest horizon ($h = 1, S$) are significantly less accurate than forecasts at the shorter horizon ($h = 1, F$), indicating that little useful information arrives between 15 and 21 months prior to the end of the year whose inflation is predicted–or at least that such information is not incorporated in the IMF's forecasts. Significant improvements in forecast accuracy start showing up as we move from the prior-year fall to the current-year spring (top right, $h = 1, F$ vs $h = 0, S$) and from current-year spring to current-year fall (bottom left, $h = 0, S$ vs $h = 0, F$) inflation forecasts, with notable improvements for many European countries, United States, and Australia. Finally, on a cumulative basis, we identify significant improvements in the accuracy of the inflation forecasts for most of the aforementioned countries in addition to countries such as Canada, Chile, and India.[22]

Table 2.1 supplements Figure 2.1 by reporting the outcome of comparisons of the accuracy of the WEO forecasts of GDP growth and inflation across the four different forecast horizons. Panels A and C set up the test statistic so that rejections (small p-values) indicate significant improvements as the forecast horizon gets *longer*, thus testing the null in (2.23). Reassuringly, we fail to find a single country for which the GDP growth forecasts (Panel A) or inflation forecasts (Panel C) are significantly *less* accurate at the shorter forecast horizons than at the longer horizons.

Testing the reverse null in (2.24) we find four instances–Brazil, Mexico, Italy, and Portugal–for which the Sup test identifies significant improvements in the accuracy of next-year GDP growth forecasts as we move from the spring to the fall WEO (Panel B). Evidence of significant improvements in the accuracy of the GDP growth forecasts gets stronger for the current-year forecasts ($h = 0, S$ vs. $h = 0, F$) for which the null is rejected for ten countries. On a cumulative basis (last column), we identify significant improvements in short-term GDP growth forecasts

[21] When implementing sup tests on current-year and next-year forecasts, we use a block length $B_T = 1$ while we set $B_T = 2, 3, 4, 5$ when analyzing forecasts at horizons 2, 3, 4 and 5 years. To increase the power of tests, we only include countries with at least 20 forecasts.

[22] Note that a significant improvement for one of the shorter six-month incremental horizons is no guarantee that a country experiences a significant improvement on a cumulative basis, mainly because the cumulative 18-month forecast revisions are more volatile than the shorter six-month revisions.

($h = 0, F$) relative to long-term forecasts ($h = 1, S$) for 21 countries.

For the inflation forecasts (Panel D), consistent with Figure 2.1 we find no evidence of significant improvements in predictive accuracy as we move between the two longest forecast horizons ($h = 1, S$ vs. $h = 1, F$). Conversely, at the shorter horizons there are now more countries with significant improvements than for GDP growth, namely 10 ($h = 1, F$ vs. $h = 0, S$) and 15 ($h = 0, S$ vs. $h = 0, F$) cases, respectively. Improvements in inflation forecasts are, thus, concentrated in the revisions between the next-year fall and current-year fall periods. On a cumulative basis, we identify significant improvements in inflation forecasts for 32 countries.

Table 2.2 conducts the same set of tests for forecasts of import and export growth and the current account balance. For all three variables, we fail to identify a single case in which forecasts are significantly more accurate at the longer horizons than at the shorter ones (Panels A, C, and E). For import and export, our tests only identify a combined total of six cases in which forecast accuracy improves significantly between the two fall issues of WEO ($h = 1, F$ vs. $h = 0, F$). Even on a cumulative basis ($h = 1, S$ vs. $h = 0, F$), the Sup tests only detect six countries with significant improvements in predictive accuracy for each of the import and export growth variables.

Improvements in predictive accuracy at shorter horizons is notably stronger for the current account forecasts. In fact, we identify significant improvements for all three six-month reductions in forecast horizon, including 23 countries for which the cumulative forecast revision leads to significant improvements.

These results show that there is surprisingly weak evidence that IMF forecasts of imports and exports improve significantly as the forecast horizon gets reduced from 21 to 3 months prior to the end of the target year. The learning curve is, thus, quite flat for these variables. Conversely, the learning curve for the current account forecasts behaves more in line with the forecasts of GDP growth and inflation with many more cases showing significant improvements in predictive accuracy for this variable as the forecast horizon is reduced.

2.5.3 Heterogeneity Across Regions and Types of Economies

To examine whether improvements in the accuracy of the WEO forecasts vary across geographical regions and types of economies we next test the hypothesis in (2.7). To this end, we group the countries into nine categories adopted by the IMF, namely (i) advanced economies, (ii) emerging and developing economies, (iii) emerging and developing Europe, (iv) low income developing countries, (v) Latin America and Caribbean, (vi) Commonwealth of Independent States, (vii) Middle East, North Africa, Afghanistan and Pakistan, (viii) Emerging and developing Asia, and (ix) Sub-Sahara Africa.

The results, presented in Tables 2.3 and 2.4, show large variation across types of economies, regions, and variables. For reference, the first column (world) summarizes the evidence from Tables 2.1 and 2.2, confirming that we identify most countries with significant improvements in predictive accuracy for the inflation and current account series and fewest for the import and export growth variables, with GDP growth in the middle.

The remaining columns reveal a great deal of heterogeneity in rejection rates across different clusters of countries and also demonstrate that the number of rejections in fact can be lower for tests undertaken on a larger, more heterogeneous set of countries compared to a smaller, more homogeneous set. This shows up most evidently for the advanced economy cluster in the second column. For example, for the cumulative revisions ($h = 1, S$ vs. $h = 0, F$), we identify significant improvements in the accuracy of the inflation forecasts for 27 out of 36 advanced economies as compared to only 33 out of 182 world economies. The corresponding numbers are 16 out of 36 advanced economies for GDP growth versus 17 out of 149 emerging market and developing economies and 23 out of 185 world economies. Across regional clusters, the number of rejections tends to be high for Latin America and the Caribbean and low for emerging and developing Europe.

For the import and export growth forecasts (Table 2.4), the number of rejections is twice as high when the Sup tests are undertaken on the cluster of advanced economies compared

to conducting the tests on the full set of world economies. The number of rejections is correspondingly lower among emerging market and developing economies (column 3) compared to what we would expect if rejection rates were homogeneous across different clusters.

We conclude from these findings that there is stronger evidence for the advanced economies than for any of the other groups that the WEO forecasts become significantly more accurate as the forecast horizon shrinks. Our empirical results thus suggest that conducting the Sup tests on a more homogeneous sample of units can lead to increased power consistent with the analysis in Hansen (2005).

Advanced Economies

The higher power of the Sup tests among the set of advanced economies allows us to identify the individual economies in this group for which forecasts improve significantly as the horizon shrinks. Table 2.5 therefore applies the Sup test associated with (2.7) to GDP growth and inflation forecasts for this group.

Once again, we find no case for which the short-horizon forecasts are significantly less accurate than the corresponding forecasts generated at longer horizons (Panels A, C). Conversely, we identify significant reductions in squared error loss after accounting for the multiple comparison problem not only for the current-year forecasts ($h = 0, S$ versus $h = 0, F$) but also for the one-year-ahead GDP growth forecasts (Panel B). For example, for next-year GDP growth forecasts, we identify significant improvements in predictive accuracy between the Spring and Fall WEO issues for Italy, Japan and Portugal. At the shorter horizons, we also see improvements in predictive accuracy for major economies such as United States, United Kingdom, Japan, France, and Spain. Portugal is the only country with improvements in predictive accuracy across all six-month increments to the forecast horizon.

For the inflation forecasts (Panel D), we identify 12 countries for which the predictive accuracy improves significantly between the three shortest horizons and no country with improved

accuracy between the two next-year horizons ($h = 1, S$ vs. $h = 1, F$). On a cumulative basis, we see significant improvements for all major advanced economies.

2.5.4 Joint Tests Across Variables and Forecast Horizons

The top part of panel A in Table 2.6 reports results from Sup tests conducted, at each horizon, across all five variables ($j = 1, ..., 5$):

$$H_0 : \max_{j\in\{1,...,J\}} \max_{i\in\{1,...,N\}} \max_{m\in\{1,..,M\}} E[\Delta L_{i,j,t+h,m}] \leq 0. \tag{2.25}$$

These comparisons result in up to 878 pairwise tests and produce up to 47 rejections for the world economies. Our tests again identify a much larger number of rejections at the shortest ($h = 0, S$ vs. $h = 0, F$) horizons and for advanced economies compared to emerging markets and developing economies (44 vs. 27).

The bottom part of panel A in Table 2.6 further pools the Sup tests across the three pairings of non-overlapping horizons ($H = 3$):

$$H_0 : \max_{h\in\{1,...,H\}} \max_{j\in\{1,...,J\}} \max_{i\in\{1,...,N\}} \max_{m\in\{1,..,M\}} E[\Delta L_{i,j,t+h,m}] \leq 0. \tag{2.26}$$

This generates a maximum of 2,751 pair-wise forecast comparisons for the world economies. For this extended list of pairwise comparisons we identify only 14 rejections of the null that long-run forecasts are at least as accurate as the short-run forecasts for every variable, country, and horizon. Once again, this illustrates how adding more moment inequalities can in fact lead to weaker power and fewer rejections.

Our results are based on the studentized test statistic in (2.8). To examine if it makes a difference whether we use a studentized test, in unreported results we also conduct tests based on the non-standardized test statistic with $\hat{a}_i = 1$. Using this test statistic, we find no rejections of the

null that long-horizon forecasts are as accurate as short-horizon forecasts. This finding reflects the dominance of outliers from a few emerging market and developing economies whose test statistics tend to be far more volatile than those from other economies.

Panel B of Table 2.6 reports results from implementing our moment selection procedure described in Algorithm 1 on the same set of moments as in Panel A. In fact, the moment selection procedure produces fewer rejections than what we find for the regular approach (Panel A). This happens because our data contains only few countries whose long-horizon forecasts are significantly better than their short-horizon forecasts. As a result, the moment selection procedure fails to exclude many hypotheses and its use of a smaller nominal size leads to fewer rejections.

2.5.5 Improvements in Predictive Accuracy for Individual Years

So far, our empirical analysis has focused on the panel-based Sup tests from Section 2.3. We next consider the tests conducted on individual cross-sections as described in Section 2.4.

Figure 2.2 shows results from cross-sectional comparisons of GDP growth and inflation forecasting performance in individual years, averaged across the roughly 180 individual countries in our sample. Each row tracks a particular pairing of long and short forecast horizons, with circles indicating individual years with rejections of the null that forecasts at the longer horizon are at least as accurate as forecasts at the shorter horizon. Open circles indicate years with rejections of the null in (2.15), while closed circles show years in which the joint null in (2.16) gets rejected.

Comparing the two longest forecast horizons ($h = 1,S$ versus $h = 1,F$) for the GDP growth and inflation forecasts, we fail to reject the null (2.15) for around half of the years in our 30-year sample. Conversely, at the two shortest horizons ($h = 0,S$ vs. $h = 0,F$), we only fail to reject this null for two years. GDP growth and inflation forecasts thus improve significantly, on average, in almost all years at the current-year horizons, as well as on a cumulative basis. There is much weaker evidence of improvements in predictive accuracy every year at the longer next-year

horizons. Overall, short-horizon forecasts of GDP growth and inflation are significantly more accurate than long-horizon forecasts for a majority of the years during our sample even after accounting for the multiple hypothesis testing problem associated with testing the joint null in (2.16).

Figure 2.3 detects far fewer rejections of the null in (2.15) and (2.16), particularly for the import and export growth forecasts (top panels). For these, the rate at which we identify individual years with significantly improved forecasting performance increases over time with few rejections prior to 2002 even at the two shortest horizons ($h = 0, S$ vs. $h = 0, F$). Compared with this, the rejection rates for the current account forecasts (bottom panel) are both more numerous and also display a more even pattern through our sample.

These results demonstrate that our cross-sectional tests can be used to identify individual years with improvements in forecasting performance as the forecast horizon gets shorter.

2.5.6 IMF's Learning Curve at Longer Horizons

Up to this point our analysis focused on current-year and next-year forecasts. To get a sense of the IMF's learning curve at the longer horizons, we also consider the accuracy of WEO forecasts spanning 2-5 year forecast horizons.

Figure 2.4 illustrates the IMF's learning curve in the form of box-and-whisker plots depicting the distribution of MSE ratios MSE_{h_S}/MSE_{h_L} which summarize the relative MSE accuracy of short-horizon (h_S) vs. long-horizon (h_L) forecasts. In each panel, the top four rows use Spring WEO forecasts at horizons ranging from one through five years while the rows below track the shorter horizons. An MSE ratio of unity indicates no improvement in predictive accuracy, while ratios below unity measure the degree of improvement in predictive accuracy between the horizons in the numerator and denominator.

The plots show only modest evidence of improved forecast accuracy between the two-year and five-year horizons except, perhaps, for the current account forecasts. Notable improvements in

predictive accuracy show up at the shorter horizons, however, particularly for the GDP growth and inflation series whose median values and 90th percentiles fall below 0.7 and unity, respectively, for the two shortest forecast horizons.

Table 2.7 presents formal tests for improvements in predictive accuracy, again comparing Spring WEO forecasts at forecast horizons spaced one year apart from one through five years.[23] We identify very few cases with significant improvements in predictive accuracy for the one-year increments to the forecast horizon–essentially zero cases between the three and five-year horizons and one or two cases at the shorter two or three-year horizons. On a cumulative basis, for the four-year horizons from $h = 1,S$ to $h = 5,S$, we find significant improvements in GDP growth forecasts for about 10 of the world economies with even fewer cases for inflation (one instance), import and export growth (both zero cases), and the current account balance (six cases).

This evidence is consistent with our empirical findings that significant improvements in the predictive accuracy of the WEO forecasts are most common for GDP growth, inflation, and the current account balance, and rarer for import and export forecasts. They also show that the IMF's learning curve is quite flat at the longer horizons–extended here to five years–and that improvements in forecast accuracy accelerate notably at the shortest forecast horizons.

2.6 Conclusion

We develop new methods for comparing the accuracy of panels of forecasts and testing if individual forecasts are significantly more accurate than some benchmark for at least one outcome variable, one forecaster (model), one horizon, or one period. Our tests control the family wise error rate and thus account for the multiple hypothesis testing problem that arises when forecasting performance is compared across large numbers of pairings. Building on Chernozhukov, Chetverikov and Kato (2018), we show that a bootstrap approach can be used

[23] To increase the power of our tests, we use the full normalization scheme from Example 2.3.1 with $B_T = 2,3,4,5$ for forecast horizons of 2, 3, 4 and 5 years.

to test hypotheses about predictive accuracy in high-dimensional settings such as those that are increasingly encountered in practice.

Our empirical application to the IMF's World Economic Outlook forecasts of five variables for more than 180 countries recorded at several forecast horizons over a 30-year sample suggests that the term structure of forecast errors is quite flat at longer 2-5 year horizons with little evidence of significant improvements in predictive accuracy. Improvements in predictive accuracy are stronger at shorter horizons spanning nine to three months, for GDP growth, inflation and the current account balance and in advanced economies. Improvements in predictive accuracy are notably weaker at longer horizons, for import and export growth and for emerging markets and developing economies. These findings are confirmed across a range of tests conducted both in a panel setting and for individual cross-sections (years). The results suggest that information that helps significantly improve the accuracy of economic forecasts tends to be relatively short-lived.

2.7 Acknowledgements

Chapter 2 is currently being prepared for submission for publication and is couathored with Allan Timmermann and Yinchu Zhu. Ritong Qu, the dissertation author, is the primary investigator and author of this material.

Table 2.1: Sup tests comparing predictive accuracy of GDP growth and inflation across different horizons

Panel A: GDP, m_0 = short horizon, m_1 = long horizon
H_0: Short-horizon forecasts are at least as accurate as long-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F
0.995	0.996	1.000	1.000

Panel B: GDP, m_0 = long horizon, m_1 = short horizon
H_0: Long-horizon forecasts are at least as accurate as short-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F	
0.015	0.025	0.006	0.001	
Brazil	India	Canada	Argentina	Lebanon
Italy	Switzerland	Chile	Brazil	Liberia
Mexico	Zimbabwe	Israel	Comoros	Malta
Portugal		Italy	Congo, Democratic	Panama
		Japan	Congo, Republic of	Peru
		Mongolia	Guyana	Portugal
		Spain	Haiti	Sudan
		Switzerland	Israel	Switzerland
		Ukraine	Italy	Tunisia
		United Kingdom	Kenya	United States
				Zimbabwe

Panel C: Inflation, m_0 = short horizon, m_1 = long horizon
H_0: Short-horizon forecasts are at least as accurate as long-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F
0.316	0.944	1.000	0.998

Panel D: Inflation, m_0 = long horizon, m_1 = short horizon
H_0: Long-horizon forecasts are at least as accurate as short-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F	
0.127	0.000	0.000	0.000	
	Angola	Belgium	Angola	Guatemala
	Australia	Dominican Republic	Austria	India
	France	Finland	Bangladesh	Indonesia
	Germany	Georgia	Belarus	Italy
	Hungary	Indonesia	Belgium	Japan
	Luxembourg	Japan	Cyprus	Kenya
	New Zealand	Lithuania	Denmark	Lithuania
	Slovak Republic	Nepal	Dominican Republic	Luxembourg
	Slovenia	Norway	Egypt	Malaysia
	Switzerland	Panama	Estonia	Mongolia
		Peru	Ethiopia	New Zealand
		Poland	Finland	Norway
		Singapore	France	Sweden
		United Kingdom	Germany	Switzerland
		United States	Ghana	Thailand
				United States
				Zambia

Notes: The first row in each panel reports the p-value of the Sup test for the null that the benchmark forecasts m_0 are at least as accurate as the alternative forecasts m_1 for all countries included in the comparison. Small p-values indicate rejections of the null. For cases where the null is rejected, we list the countries for which the alternative forecast is significantly more accurate than the benchmark forecast, i.e., countries whose t-statistics are higher than the 90% quantile of the maximum value of the bootstrapped t-statistic. Panels A and B examine GDP growth forecasts while Panels C and D examine inflation forecasts. Columns 1-3 compare WEO forecasts at consecutive 6-month revision points, while column 4 evaluates the cumulative revision from the one-year-ahead spring forecasts ($h = 1, S$) to the current-year fall forecasts ($h = 0, F$).

Table 2.2: Sup tests comparing predictive accuracy of export growth, import growth and the current account-GDP ratio across different horizons

Panel A: Import, m_0 = short horizon, m_1 = long horizon
H_0: Short-horizon forecasts are at least as accurate as long-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F
0.999	0.779	1.000	0.909

Panel B: Import, m_0 = long horizon, m_1 = short horizon
H_0: Long-horizon forecasts are at least as accurate as short-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F
0.190	0.137	0.006	0.001
		Italy	Australia
		Japan	Chile
		Venezuela	Ireland
			Switzerland
			United States
			Venezuela

Panel C: Export, m_0 = short horizon, m_1 = long horizon
H_0: Short-horizon forecasts are at least as accurate as long-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F
0.316	0.944	1.000	0.998

Panel D: Export, m_0 = long horizon, m_1 = short horizon
H_0: Long-horizon forecasts are at least as accurate as short-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F
0.127	0.000	0.000	0.000
	Egypt	Japan	Canada
		Uruguay	Egypt
			India
			Ireland
			Korea
			Myanmar

Panel E: Current account, m_0 = short horizon, m_1 = long horizon
H_0: Short-horizon forecasts at least as accurate as long-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F
0.538	0.990	0.940	1.000

Panel F: Current account, m_0 = long horizon, m_1 = short horizon
H_0: Long-horizon forecasts at least as accurate as short-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F	
0.034	0.009	0.009	0.000	
Algeria	Italy	Algeria	Algeria	Italy
	Japan	Colombia	Australia	Japan
	Malta	Japan	Bangladesh	New Zealand
	Pakistan	South Africa	Canada	Pakistan
	Poland	Venezuela	Chile	Saudi Arabia
	Saudi Arabia		China	Slovenia
	Spain		Egypt	South Africa
			France	Spain
			Greece	Trinidad and Tobago
			Guyana	Turkey
			Israel	Uruguay
				Venezuela

Notes: The first row in each panel reports the p-value of the Sup test for the null that the benchmark forecasts m_0 are at least as accurate as the alternative forecasts m_1 for all countries included in the comparison. Small p-values indicate rejections of the null. For cases where the null is rejected, we list the countries for which the alternative forecast is significantly more accurate than the benchmark forecast, i.e., countries whose t-statistics are higher than the 90% quantile of the maximum value of the bootstrapped t-statistic. Panels A and B examine import growth forecasts; Panels C and D examine export growth forecasts; Panels E and F examine current account-GDP ratio forecasts. Columns 1-3 compare WEO forecasts at consecutive six-month revision points, while column 4 evaluates the cumulative revision from the one-year-ahead spring forecasts ($h = 1, S$) to the current-year fall forecasts ($h = 0, F$).

Table 2.3: Sup tests comparing the accuracy of GDP growth and inflation rate forecasts across clusters of economies

Panel A: GDP Growth

	world	ae	emde	lics	eeur	dasia	lac	menap	cis	ssa
h=1,S vs. h=1,F	0.01	0.01	0.01	0.07	0.15	0.41	0.00	0.13	0.05	0.05
Rejections	4	3	2	1	0	0	3	0	1	1
h=1,F vs. h=0,S	0.02	0.00	0.05	0.03	0.20	0.02	0.06	0.12	0.04	0.02
Rejections	3	6	2	1	0	1	2	0	4	2
h=0,S vs. h=0,F	0.01	0.00	0.02	0.08	0.03	0.01	0.01	0.02	0.01	0.06
Rejections	10	15	3	3	2	2	8	3	2	3
h=1,S vs. h=0,F	0.00	0.00	0.00	0.00	0.02	0.03	0.00	0.00	0.01	0.00
Rejections	23	16	17	9	2	4	12	5	3	12
Countries	185	36	149	58	12	27	32	23	12	43

Panel B: Inflation

	world	ae	emde	lics	eeur	dasia	lac	menap	cis	ssa
h=1,S vs. h=1,F	0.12	0.13	0.10	0.05	0.14	0.11	0.45	0.45	0.21	0.04
Rejections	0	0	0	2	0	0	0	0	0	2
h=1,F vs. h=0,S	0.00	0.00	0.03	0.03	0.00	0.04	0.03	0.01	0.04	0.01
Rejections	10	12	3	2	5	1	3	2	2	4
h=0,S vs. h=0,F	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.06	0.02	0.14
Rejections	15	12	7	3	1	5	5	3	3	0
h=1,S vs. h=0,F	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
Rejections	33	27	18	11	4	11	8	6	6	9
Countries	182	36	146	57	12	26	31	22	12	43

Notes: This table reports p-values for Sup tests comparing long-horizon to short-horizon WEO forecasts of GDP growth (Panel A) or inflation (Panel B). The null hypothesis is that none of the long-horizon WEO forecasts are less accurate than the corresponding short-horizon forecasts for each of the countries within a particular group. Small p-values indicate that the null is rejected and some short-horizon WEO forecasts are significantly more accurate than their long-horizon counterparts. Each panel also shows the number of countries for which the null hypothesis is rejected using a nominal size of $\alpha = 0.1$. 'ae' refers to advanced economies, 'emde' is emerging and developing economies, 'eeur' is emerging and developing Europe, 'lics' is low income developing countries, 'lac' is Latin America and Caribbean, 'cis' is Commonwealth of Independent States, 'menap' is Middle East, North Africa, Afghanistan, and Pakistan, 'dasia' is emerging and developing Asia, and 'ssa' is Sub-Sahara Africa.

Table 2.4: Sup tests comparing the accuracy of import, export, and current account forecasts across clusters of economies

Panel A: Import										
	world	ae	emde	lics	eeur	dasia	lac	menap	cis	ssa
h=1,S vs. h=1,F	0.19	0.28	0.17	0.22	0.17	0.18	0.05	0.55	0.20	0.19
Rejections	0	0	0	0	0	0	1	0	0	0
h=1,F vs. h=0,S	0.13	0.02	0.25	0.19	0.06	0.07	0.07	0.54	0.31	0.56
Rejections	0	4	0	0	1	2	2	0	0	0
h=0,S vs. h=0,F	0.00	0.00	0.07	0.20	0.04	0.10	0.02	0.02	0.22	0.10
Rejections	3	6	2	0	1	0	3	1	0	0
h=1,S vs. h=0,F	0.01	0.00	0.06	0.05	0.02	0.05	0.01	0.02	0.12	0.03
Rejections	5	12	3	1	1	1	7	1	0	1
Countries	180	35	145	57	11	24	32	23	12	43
Panel B: Export										
	world	ae	emde	lics	eeur	dasia	lac	menap	cis	ssa
h=1,S vs. h=1,F	0.46	0.15	0.43	0.38	0.46	0.12	0.44	0.54	0.17	0.34
Rejections	0	0	0	0	0	0	0	0	0	0
h=1,F vs. h=0,S	0.05	0.04	0.04	0.37	0.45	0.21	0.12	0.01	0.09	0.65
Rejections	1	3	1	0	0	0	0	1	1	0
h=0,S vs. h=0,F	0.05	0.02	0.04	0.43	0.14	0.03	0.01	0.23	0.17	0.36
Rejections	2	3	1	0	0	3	1	0	0	0
h=1,S vs. h=0,F	0.00	0.00	0.00	0.03	0.08	0.00	0.08	0.00	0.04	0.18
Rejections	6	10	3	1	2	6	1	2	1	0
Countries	180	35	145	57	11	24	32	23	12	43
Panel C: Current account										
	world	ae	emde	lics	eeur	dasia	lac	menap	cis	ssa
h=1,S vs. h=1,F	0.03	0.09	0.03	0.27	0.29	0.34	0.21	0.01	0.51	0.22
Rejections	1	2	1	0	0	0	0	1	0	0
h=1,F vs. h=0,S	0.01	0.00	0.04	0.07	0.00	0.11	0.09	0.01	0.08	0.03
Rejections	5	6	4	2	3	0	1	4	3	3
h=0,S vs. h=0,F	0.01	0.02	0.01	0.05	0.01	0.02	0.00	0.00	0.22	0.01
Rejections	5	3	5	1	2	4	5	3	0	1
h=1,S vs. h=0,F	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00
Rejections	23	16	16	4	6	4	13	8	2	5
Countries	180	35	145	57	11	24	32	23	12	43

Notes: This table reports p-values for Sup tests comparing the accuracy of long-horizon vs. short-horizon WEO forecasts of import growth (Panel A), export growth (Panel B), and the current account balance (Panel C). The null hypothesis is that none of the long-horizon WEO forecasts are less accurate than the corresponding short-horizon forecasts for each of the countries within a particular group. Small p-values indicate that the null is rejected and some short-horizon WEO forecasts are significantly more accurate than their long-horizon counterparts. Each panel also shows the number of countries for which the null hypothesis is rejected using a nominal size of $\alpha = 0.1$. 'ae' refers to advanced economies, 'emde' is emerging and developing economies, 'eeur' is emerging and developing Europe, 'lics' is low income developing countries, 'lac' is Latin America and Caribbean, 'cis' is Commonwealth of Independent States, 'menap' is Middle East, North Africa, Afghanistan, and Pakistan, 'dasia' is emerging and developing Asia, and 'ssa' is Sub-Sahara Africa.

Table 2.5: Sup tests comparing the accuracy of forecasts of GDP growth and inflation across different forecast horizons for advanced economies

Panel A: GDP, m_0 = short horizon, m_1 = long horizon
H_0: Short-horizon forecasts are at least as accurate as long-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F
0.975	0.755	1.000	1.000

Panel B: GDP, m_0 = long horizon, m_1 = short horizon
H_0: Long-horizon forecasts are at least as accurate as short-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F
0.007	0.007	0.003	0.002
Italy	Canada	Austria	Belgium
Japan	Hong Kong SAR	Belgium	Canada
Portugal	Luxembourg	Canada	Cyprus
	Portugal	Cyprus	Finland
	Switzerland	Estonia	France
	United States	France	Germany
		Israel	Greece
		Italy	Hong Kong SAR
		Japan	Ireland
		Malta	Israel
		New Zealand	Italy
		Portugal	Japan
		Spain	Malta
		Switzerland	Portugal
		United Kingdom	Switzerland
			United States

Panel C: Inflation, m_0 = short horizon, m_1 = long horizon
H_0: Short-horizon forecasts are at least as accurate as long-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F
0.888	1.000	1.000	1.000

Panel D: Inflation, m_0 = long horizon, m_1 = short horizon
H_0: Long-horizon forecasts are at least as accurate as short-horizon forecasts

h=1, S vs. h=1, F	h=1, F vs. h=0, S	h=0, S vs. h=0, F	h=1, S vs. h=0, F	
0.151	0.000	0.001	0.000	
	Australia	Belgium	Austria	Japan
	Cyprus	Denmark	Belgium	Korea
	Finland	Finland	Canada	Lithuania
	France	France	Cyprus	Luxembourg
	Germany	Germany	Czech Republic	Netherlands
	Italy	Italy	Denmark	New Zealand
	Luxembourg	Japan	Estonia	Norway
	New Zealand	Lithuania	Finland	Singapore
	Slovak Republic	Norway	France	Slovak Republic
	Slovenia	Singapore	Germany	Slovenia
	Spain	United Kingdom	Ireland	Spain
	Switzerland	United States	Italy	Sweden
				Switzerland
				United Kingdom
				United States

Notes: The first row in each panel reports the p-value of the Sup test for the null that the benchmark forecasts m_0 are at least as accurate as the forecasts in the alternative set m_1 for all advanced economies. Small p-values indicate rejections of the null. For cases where the null is rejected, we list the countries for which the alternative forecast is significantly more accurate than the benchmark forecast, i.e., countries whose t-statistics are higher than the 90% quantile of the maximum value of the bootstrapped t-statistic. Panels A and B examine GDP forecasts while Panels C and D examine inflation forecasts. Columns 1-3 compare WEO forecasts at consecutive six-month revision points, while column 4 evaluates the cumulative revision from the one-year-ahead spring forecasts ($h = 1, S$) to the current-year fall forecasts ($h = 0, F$).

Table 2.6: Sup tests comparing predictive accuracy across clusters of economies with pooling across variables and horizons

	world	ae	emde	lics	eeur	dasia	lac	menap	cis	ssa
Panel A: Sup tests pooled across variables										
h=1,S vs. h=1,F	0.05	0.04	0.04	0.21	0.45	0.35	0.01	0.03	0.15	0.18
Rejections	1	3	1	0	0	0	2	1	0	0
h=1,F vs. h=0,S	0.00	0.00	0.13	0.10	0.04	0.08	0.05	0.04	0.14	0.08
Rejections	9	15	0	1	2	1	1	4	0	1
h=0,S vs. h=0,F	0.01	0.00	0.01	0.01	0.06	0.00	0.00	0.01	0.02	0.03
Rejections	14	21	4	1	0	2	6	4	1	1
h=1,S vs. h=0,F	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.00
Rejections	47	44	27	12	8	10	15	12	6	8
Country-variable pairs	878	177	701	276	51	121	155	110	55	209
Sup tests pooled across variables and horizons										
Pool horizons	0.00	0.00	0.03	0.03	0.11	0.01	0.01	0.02	0.04	0.07
Rejections	14	21	4	1	0	2	6	4	1	1
Country-variable-horizon pairs	2751	534	2217	864	169	383	480	345	180	660
Panel B: Sup tests with moment selection										
	world	ae	emde	lics	eeur	dasia	lac	menap	cis	ssa
h=1,S vs. h=1,F	0.10	0.09	0.09	0.24	0.50	0.39	0.06	0.08	0.21	0.23
Rejections	1	1	1	0	0	0	1	1	0	0
h=1,F vs. h=0,S	0.05	0.05	0.19	0.14	0.10	0.13	0.10	0.09	0.19	0.13
Rejections	6	10	0	0	1	0	1	3	0	0
h=0,S vs. h=0,F	0.05	0.05	0.07	0.06	0.11	0.06	0.05	0.06	0.07	0.08
Rejections	8	10	3	1	0	2	5	2	1	1
h=1,S vs. h=0,F	0.05	0.05	0.05	0.05	0.06	0.06	0.05	0.05	0.07	0.05
Rejections	26	31	15	9	2	7	10	7	2	8
Country-variable pairs	878	177	701	276	51	121	155	110	55	209
Pool 3 horizons										
Pool horizons	0.05	0.05	0.09	0.07	0.17	0.06	0.06	0.07	0.09	0.11
Rejections	8	14	2	1	0	1	3	1	1	0
Country-variable-horizon pairs	2751	534	2217	864	169	383	480	345	180	660

Notes: This table reports p-values for Sup tests comparing long-horizon to short-horizon WEO forecasts. The null hypothesis is that none of the long-horizon WEO forecasts are less accurate than the corresponding short-horizon forecasts for each of the countries. The first set of tests pools results across our five variables but keeps forecast horizons separate while the second set of results pools across both variables and horizons. Small p-values indicate that the null is rejected and some short-horizon WEO forecasts are significantly more accurate than their long-horizon counterparts. Each panel also shows the number of countries for which the null hypothesis is rejected using a nominal size of $\alpha = 0.1$. The bottom panel reports outcomes from the same set of tests but uses moment selection with $\alpha = \beta = 0.05$.

Table 2.7: Sup tests comparing predictive accuracy across longer forecast horizons

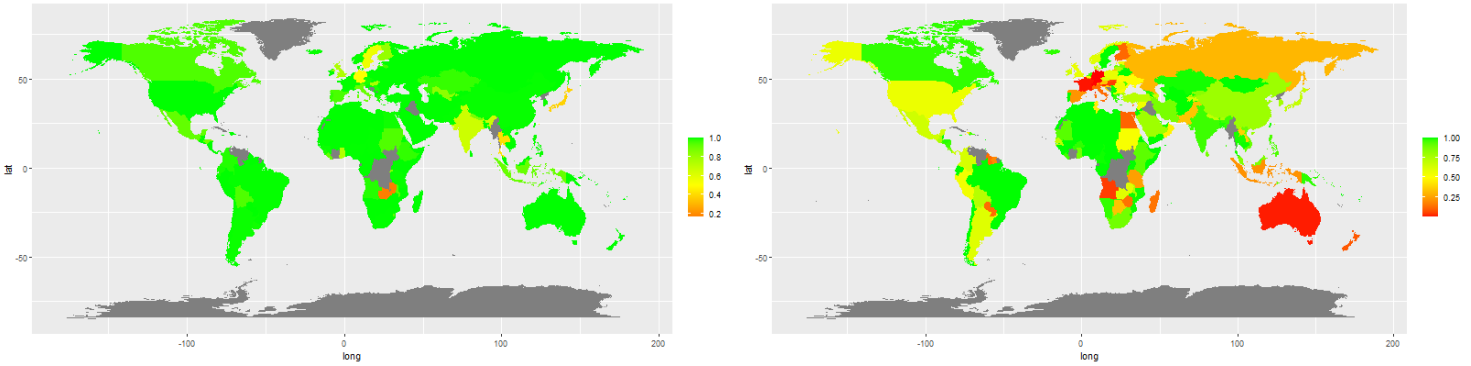
	GDP	Inflation	Import	Export	Current Account
h=5,S vs. h=4,S	0.42	0.28	0.49	0.66	0.57
Rejections	0	0	0	0	0
h=4,S vs. h=3,S	0.22	0.31	0.11	0.30	0.54
Rejections	0	0	0	0	0
h=3,S vs. h=2,S	0.05	0.02	0.41	0.11	0.28
Rejections	2	1	0	0	0
h=2,S vs. h=1,S	0.04	0.41	0.16	0.47	0.36
Rejections	2	0	0	0	0
h=5,S vs. h=1,S	0.00	0.00	0.58	0.45	0.01
Rejections	10	1	0	0	6
Countries	175	154	155	155	158

Notes: This table reports p-values for Sup tests comparing long-horizon to short-horizon WEO forecasts with forecast horizons separated by one year (first four comparisons) or four years (final comparison). The null hypothesis is that none of the long-horizon WEO forecasts are less accurate than the corresponding short-horizon forecasts. Small p-values indicate that the null is rejected and some short-horizon WEO forecasts are significantly more accurate than their long-horizon counterparts. Each panel also shows the number of countries for which the null hypothesis is rejected using a nominal size of $\alpha = 0.1$.

Table 2.8: Sup tests comparing predictive accuracy across longer forecast horizons

	GDP	Inflation	Import	Export	Current Account
h=5,S vs. h=4,S	0.42	0.28	0.49	0.66	0.57
Rejections	0	0	0	0	0
h=4,S vs. h=3,S	0.22	0.31	0.11	0.30	0.54
Rejections	0	0	0	0	0
h=3,S vs. h=2,S	0.05	0.02	0.41	0.11	0.28
Rejections	2	1	0	0	0
h=2,S vs. h=1,S	0.04	0.41	0.16	0.47	0.36
Rejections	2	0	0	0	0
h=5,S vs. h=1,S	0.00	0.00	0.58	0.45	0.01
Rejections	10	1	0	0	6
Countries	175	154	155	155	158

Notes: This table reports p-values for Sup tests comparing long-horizon to short-horizon WEO forecasts with forecast horizons separated by one year (first four comparisons) or four years (final comparison). The null hypothesis is that none of the long-horizon WEO forecasts are less accurate than the corresponding short-horizon forecasts. Small p-values indicate that the null is rejected and some short-horizon WEO forecasts are significantly more accurate than their long-horizon counterparts. Each panel also shows the number of countries for which the null hypothesis is rejected using a nominal size of $\alpha = 0.1$.



(a) h=1, S vs. h=1, F

(b) h=1, F vs. h=0, S

(c) h=0, S vs. h=0, F

(d) h=1, S vs. h=0, F

Figure 2.1: Sup tests comparing the accuracy of WEO inflation forecasts across different forecast horizons

The figures shows outcomes of Sup tests for the null that long-horizon inflation forecasts are at least as accurate as short-horizon forecasts. Red color (a small p-value) indicates that long-horizon forecasts are significantly less accurate for a particular country than the short-horizon forecasts. Green color (large p-values) indicates weak evidence against long-horizon forecasts being at least as accurate as the short-horizon forecasts.

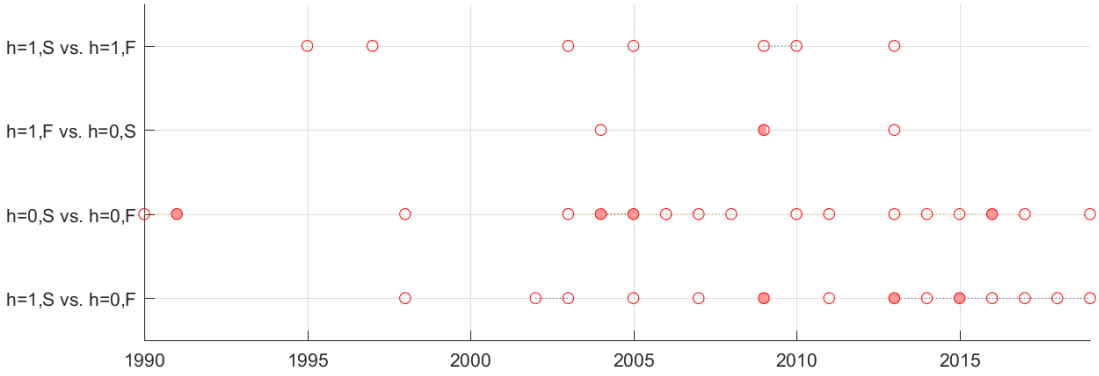
(a) GDP growth

(b) Inflation

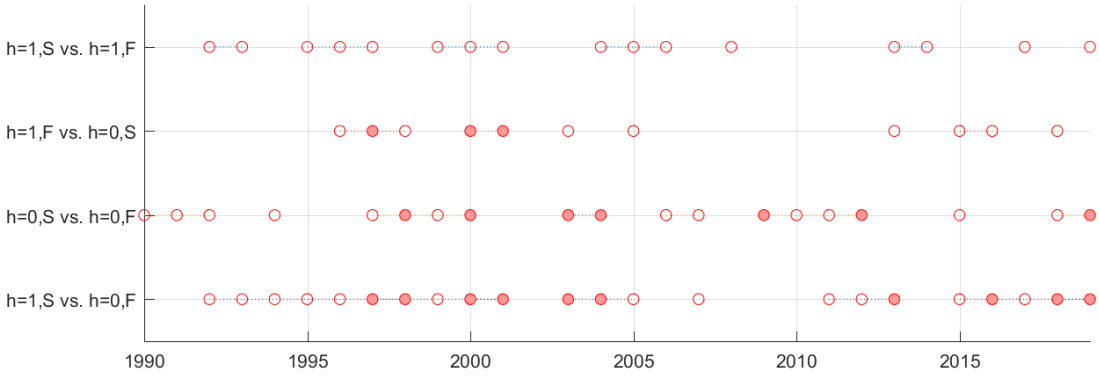
Figure 2.2: Sup test comparing average performance of long- and short-horizon forecasts of GDP growth and inflation in individual calendar years

Each panel shows the outcome of tests for whether the benchmark forecasts (m_0) are at least as accurate as all forecasts in the alternative set (m_1) in individual years (with rejections marked by open circles). We also show results from tests of the null that the benchmark forecasts are at least as accurate as all forecasts in the alternative set during every single year in the sample (marked by filled circles). Circles indicate years in which the null is rejected at the 10% significance level, suggesting that the short-horizon forecast is significantly more accurate than the long-horizon forecast. Panel A shows results for GDP growth while Panel B shows results for inflation rate forecasts.

(a) Import growth



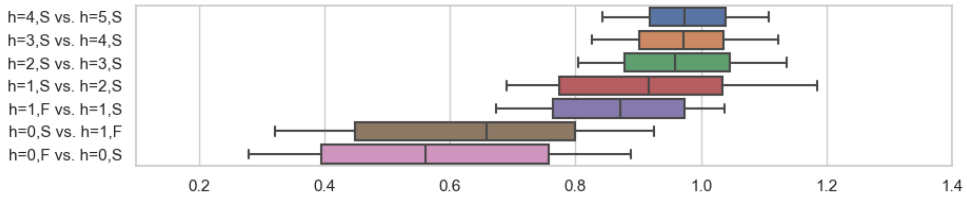
(b) Export growth



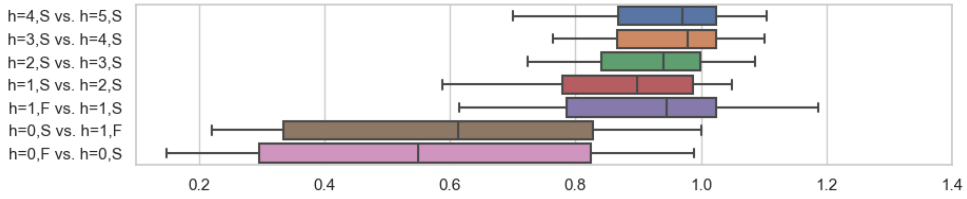
(c) Current account-GDP ratio

Figure 2.3: Sup test comparing average performance of long- and short-horizon forecasts of GDP growth and inflation in individual calendar years

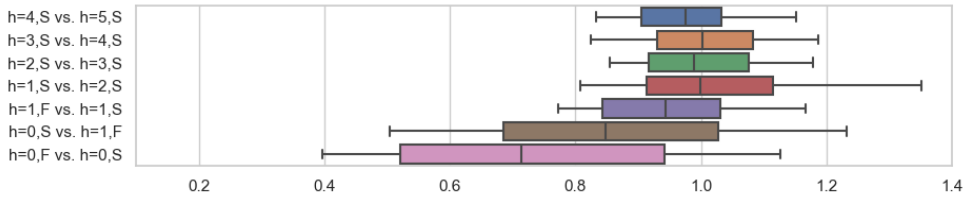
Each panel shows the outcome of tests for whether the benchmark forecasts (m_0) are at least as accurate as all forecasts in the alternative set (m_1) in individual years (with rejections marked by open circles). We also show results from tests of the null that the benchmark forecasts are at least as accurate as all forecasts in the alternative set during every single year in the sample (marked by filled circles). Circles indicate years in which the null is rejected at the 10% significance level, suggesting that the short-horizon forecast is significantly more accurate than the long-horizon forecast. Panel A shows results for import growth, Panel B shows results for export growth, while Panel C shows results for the current account balance.



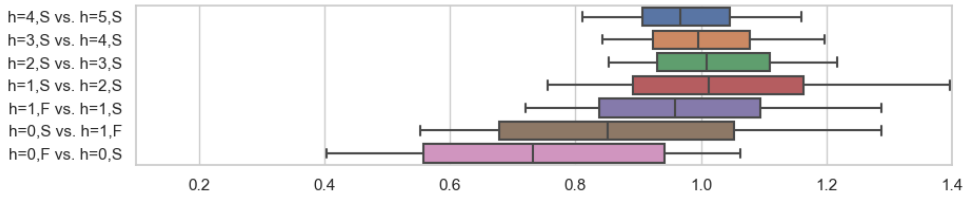
(a) GDP growth



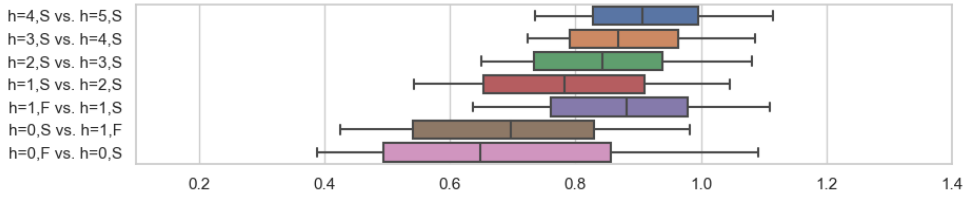
(b) Inflation



(c) Import growth



(d) Export growth



(e) Current account-GDP ratio

Figure 2.4: Distribution of ratios of short-horizon MSE values over long-horizon MSE-values These box-and-whisker diagrams show the median, interquartile, and 10% and 90% quantiles for the MSE ratios of short- versus long-horizon MSE values recorded for different pairings of horizons and for different variables.

Chapter 3

Comparing Forecasting Performance in Cross-sections

3.1 Introduction

What, if anything, can we learn about forecasting performance from a *single* cross-section of data? This question is becoming highly relevant as large cross-sections of forecasts are now routinely recorded for numerous economic and financial outcomes: financial analysts predict company earnings and revenues for hundreds of firms covering multiple industries; credit card companies conduct billions of forecasts for real-time transactions to guard against fraud; banks and international organizations forecast macroeconomic outcomes across many countries and sectors.

Comparisons of forecasting performance conducted on a single cross-section has the potential for yielding important economic insights that easily get masked by averaging performance over longer spans of time. First, forecasting performance may be state- and time-dependent. A test conducted on a single cross-section might find that model-based forecasts are inferior to survey forecasts during, say, the Covid-19 epidemic although the two forecasts are equally accurate when their performance gets averaged over a longer sample. Such a finding could indicate that survey participants possessed important forward-looking information about the impact of this event that was not reflected in past data. Second, when conducted on individual time periods, cross-sectional tests can be used to identify points in time during which one forecast performs relatively well or to identify shifts over time in forecasting performance. Third, performance evaluations conducted on individual cross-sections facilitate faster real-time comparisons of predictive accuracy than conventional methods that require calculating often lengthy time-series averages which tends to slow down discovery of deterioration or breakdown in forecasting performance. Fourth, inference conducted on a single cross-section dispenses with time-series stationarity assumptions that are unlikely to be valid in many situations.

From an inferential perspective, the key challenge for cross-sectional comparisons of forecasting performance is the likely presence of common components in forecast errors. Such

common components can invalidate the use of a cross-sectional central limit theorem (CLT) to derive distributional results for test statistics based on cross-sectional averages. To address this challenge, we develop a common factor framework for capturing cross-sectional dependencies in forecast errors and separately consider the cases with homogeneous and heterogeneous factor loadings. The case with homogeneous factor loadings gives rise to tests of equal unconditionally expected squared error loss, while heterogeneous factor loadings lead to tests that condition on factor realizations. Although these tests are fundamentally different we show that, in practice, they lead to very similar inference. Forecast comparisons conducted on individual cross-sections are robust to changes in both the number of factors and in the factor loadings which can be an important concern in empirical work, see Cheng, Liao and Schorfheide (2016).

Common components in the forecast errors contain important economic information about the underlying models used by forecasters and the extent to which shocks are fundamentally unpredictable. Large shocks to outcomes that were unanticipated by *all* forecasters and, thus, are common, cancel out from pairwise comparisons of squared forecast error *differences* to the extent that they affect individual forecasters by the same amount. Conversely, idiosyncratic error components that are specific to individual forecasters do not cancel out from squared error loss differentials.

To get a better sense of the commonality and predictability of economic shocks, we propose a new decomposition of the squared forecast error differential into a squared bias component, which tracks differences in forecast exposures to common factors, and an idiosyncratic error variance component. Only the total squared forecast error differential is observed, so we develop three approaches to estimate the common factors in forecast errors, namely (i) a cluster method that imposes homogeneity restrictions on factor loadings within clusters of variables and can be computed on a single cross-section; (ii) a common correlated effects estimator based on Pesaran (2006); and (iii) a principal components approach. Unlike the cluster approach, the second and third approach require the availability of time-series data to estimate factor loadings.

Moreover, these approaches work under different assumptions about the number of factors and patterns in factor loadings and cover many of the situations encountered by applied researchers.

We illustrate our new tests in an empirical application to financial analysts' short-term forecasts of individual firms' quarterly earnings. We compare the predictive accuracy across six brokerages covering a total of between 1,400 and 1,800 different firms during a sample that spans twenty years. We find evidence of highly significant correlation across brokerage firms' earnings forecast errors, most of which can be captured through their loadings on a single common factor. Empirically, we find that our cross-sectional tests of equal predictive accuracy across brokerage firms are highly robust regardless of whether factor loadings are assumed to be homogeneous or heterogeneous and so yield similar results for the conditional and unconditional cases. For the vast majority of quarters, brokerage firms produce similarly accurate earnings forecasts, but we also identify some quarters with rejections of the null of equal predictive accuracy.

Using our decompositions we find that, in general, differences in idiosyncratic error variances account for more of the variation in squared error loss differences in brokerage firms' earnings forecasts than the squared bias. Differences in the accuracy of earnings forecasts in individual quarters thus appear to be mostly driven by differences in brokerage firms' ability to reduce uncertainty about the idiosyncratic earnings component and is less a reflection of differences in exposures to common factor shocks.

Our paper expands to a cross-sectional setting a large literature that compares the predictive accuracy of time-series forecasts. Chong and Hendry (1986) propose tests of forecast encompassing. More recently, Diebold and Mariano (1995) and West (1996) develop tests for comparing the null of equal predictive accuracy. Clark and McCracken (2001) and McCracken (2007) focus on comparisons of predictive accuracy for forecasts that are generated by nested models, while accounting for the effect of recursive updating in the parameter estimates used to generate forecasts. Giacomini and White (2006) propose a test of equal predictive accuracy that accounts for the presence of non-vanishing parameter estimation error and develop methods

for conditional forecast comparisons. We build on these earlier contributions, but show how the presence of a cross-sectional dimension can enrich the set of economic hypotheses that can be tested and dispenses with the need for restrictive assumptions on time-series stationarity for the underlying data generating process.

A related literature evaluates the efficiency of forecasts with panel data; see, e.g., Keane and Runkle (1990), Davies and Lahiri (1995), and Patton and Timmermann (2012). However, this literature does not provide methods for comparing the relative accuracy of different forecasts or for conducting tests of the null of equal predictive accuracy across different forecasts. An advantage of our new tests is that they can be computed using only a single cross-section-provided that cross-sectional dependencies are properly accounted for. This makes the tests particularly useful in microeconomic forecast applications which often have short time-series dimensions since such surveys are conducted infrequently or due to the attrition of individual households that enter and exit.[1]

The outline of the paper is as follows. Section 3.2 presents our new tests for comparing predictive accuracy with individual cross-sections, while Section 3.3 develops our decomposition of the mean squared forecast errors into a squared bias and an idiosyncratic error variance component and derives statistics for testing the null that these two components are of the same magnitude across different forecasts. Section 3.4 conducts an empirical analysis that compares the predictive accuracy of firm-level short-term earnings forecasts across six brokerage firms. Section 3.5 uses Monte Carlo simulations to explore the finite-sample size and power properties of our tests in a variety of settings and Section 3.6 concludes. Technical proofs are in an Appendix.

[1] Giacomini, Lee and Sarpietro (2019) discuss micro forecasting approaches for annual PSID panels while Liu, Moon and Schorfheide (2018) and Liu, Moon and Schorfheide (2019) develop ways to forecast in panels with very short time-series dimensions.

3.2 Tests for Cross-sectional Comparisons of Predictive Accuracy

Formal tests used in comparisons of forecasting performance such as the well-known Diebold-Mariano test (Diebold and Mariano (1995)) rely on time-series averages. While these tests have proven useful in many economic applications, an important limitation of their usage is that sample sizes (T) are often short and so their statistical power can be quite low.[2] Conversely, in situations with long samples, non-stationarities in the underlying data generating process becomes an issue for inference. Moreover, new time-series observations arrive only slowly when outcomes are measured at a monthly, quarterly, or annual frequency, reducing the usefulness of real-time comparisons of predictive accuracy. These points highlight shortcomings of inference on predictive accuracy based on time-series averages.

In contrast, individual forecasting models can often be used to generate hundreds or even thousands of cross-sectional forecasts each period, as in the case of forecasts for individual customers, market places, product categories, or firms. Data with small T and large n can be used to compare the accuracy of pairs of forecasts in a particular time period or over a short period of time. Conducting such tests requires, however, an understanding of the assumptions under which it is possible to establish the distribution of cross-sectional averages underlying the test statistics. Most obviously, the loss differentials cannot be too strongly cross-sectionally dependent–otherwise a CLT will not apply to the cross-sectional test statistics.

We next develop a framework and a set of tests that allow us to conduct inference about relative predictive accuracy on single cross-sections.

[2] This is particularly relevant for microeconomic applications that often rely on short surveys, see, e.g., Giacomini, Lee and Sarpietro (2019) and Liu, Moon and Schorfheide (2019).

3.2.1 Setup

Let y_{it+h} denote the realized value of unit i at time $t+h$, where $i = 1, ..., n$ refers to the cross-sectional dimension and $t+h$ refers to the "target date", i.e., the point in time at which we observe the outcome. Further, suppose we observe the h-step-ahead forecast of y_{it+h} generated conditional on information available to the forecaster at time t. We denote these by $\hat{y}_{it+h|t,m}$, where $m = 1, ..., M$ indexes the individual forecasts (e.g., forecasting models) and $h \geq 0$ is the forecast horizon.

To compare the predictive accuracy of different forecasts we use a loss function that quantifies the cost of different forecast errors. Following Diebold and Mariano (1995), define the loss associated with forecast m as $L_{it+h|t,m} = L(y_{it+h}, \hat{y}_{it+h|t,m})$. Consistent with most empirical work, we assume that the loss is a quadratic function of the forecast error, $e_{it+h,m} = y_{it+h} - \hat{y}_{it+h|t,m}$, and thus takes the form[3]

$$L(y_{it+h}, \hat{y}_{it+h|t,m}) \equiv L_{it+h|t,m} = e^2_{it+h,m}. \tag{3.1}$$

Similarly, the squared-error loss differential between forecasts m_1 and m_2 for unit i at time $t+h$ is given by (dropping the reference to $m_1$and m_2)

$$\Delta L_{i,t+h|t} = e^2_{it+h,m_1} - e^2_{it+h,m_2}. \tag{3.2}$$

Following Diebold and Mariano (1995) and Giacomini and White (2006), we treat the forecasts as given and make high-level assumptions on the distribution of the forecast errors or, more generally, the losses $L_{it+h|t}$. Hence, we do not consider the effect of estimation error on the distribution of the test statistics which we derive.[4]

[3] See Elliott, Komunjer and Timmermann (2005) for a more general loss function that nests squared error loss.

[4] Estimation error and its effect on tests for equal predictive accuracy features prominently in the analysis of West (1996), Clark and McCracken (2001), McCracken (2007), and Hansen and Timmermann (2015).

To keep our analysis simple, we focus on pair-wise comparisons of forecasting performance ($M = 2$). Often, empirical researchers have access to a large number of forecasts, e.g. from surveys with large numbers of participants, from different forecasting models, or even from several cross-sections spanning different time periods. This introduces a multiple hypothesis testing problem when analyzing outcomes of several (pair-wise) test statistics. Dealing with this issue is beyond the scope of the present paper, but Qu, Timmermann and Zhu (2019) propose a Sup test procedure that allows for multiple comparisons while controlling the family-wise error rate.

3.2.2 Factor Structure

To capture cross-sectional dependencies in forecast errors, suppose we can decompose the forecast error of model m, $e_{i,t+h,m} = y_{i,t+h} - \hat{y}_{i,t+h|t,m}$, into a common component, f_{t+h}, with factor loadings λ_{im}, and an idiosyncratic component, $u_{it+h,m}$, so that, for $m = 1, 2$,

$$e_{i,t+h,m} = \lambda'_{im} f_{t+h} + u_{i,t+h,m}. \tag{3.3}$$

Under this setup, forecast errors are allowed to be affected by the same common factors, f_{t+h}, but we allow for differences in the factor loadings (λ_{im}) across units, i, and forecasts, m. Factor loadings, λ_{im}, can be either random or fixed as we make clear in the analysis below.

The assumed factor structure in (3.3) is typically well-motivated in economic forecast applications. Outcomes of economic variables such as GDP growth and inflation are likely to contain an important common unpredictable component reflecting large unanticipated supply shocks (e.g., commodity price shocks) or crises in financial markets. Common factors can be either global or regional in nature and are likely to have a very different impact on, e.g., advanced versus developing economies. The presence of common and idiosyncratic shocks is also consistent with macroeconomic models such as Mackowiak and Wiederholt (2009). A distinct advantage

of the setup, which we demonstrate in our empirical analysis, is that the presence of common factors in the forecast errors is empirically testable through simple econometric tests.

We next consider how to conduct cross-sectional tests of equal predictive accuracy using the squared error loss function in (3.2) and the factor structure in (3.3).

3.2.3 Null Hypotheses

The assumed common factor structure in (3.3) introduces a common component that does not disappear asymptotically even as $n \to \infty$. To address this issue, we consider two different approaches for testing the null of equal predictive accuracy in a single cross-section.

First, we can test the unconditional null that the cross-sectional average loss differential at time $t+h$, $\overline{\Delta L}_{t+h} = n^{-1}\sum_{i=1}^{n} \Delta L_{i,t+h|t}$, equals zero in expectation:

$$H_{0,t+h}^{unc}: \; E(\overline{\Delta L}_{t+h}) = 0. \tag{3.4}$$

While the forecasts are only expected to be equally accurate at a single point in time, $t+h$, differences in predictive accuracy at that time are hypothesized to balance out across units, $i = 1, ..., n$. As we show below, this requires that the common factor component that introduces dependence in forecast errors cancels out in the loss differentials.

Second, we can test whether two forecasts are expected to be equally accurate, at time $t+h$, *conditional* on a particular outcome of the factor realizations, f_{t+h}, and factor loadings $\{\lambda_{i1}, \lambda_{i2}\}_{i=1}^{n}$ so that, for $\mathcal{F} = \sigma(f_{t+h}, \{\lambda_{i1}, \lambda_{i2}\}_{i=1}^{n})$,

$$H_{0,t+h}^{cond}: \; E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = 0. \tag{3.5}$$

This approach is valid provided that, conditional on the realized factor, a cross-sectional CLT applies to the idiosyncratic error components.

The conditional null in (3.5) is different from the unconditional null in (3.4) but is often

of separate economic interest. For example, we can use (3.5) to test whether, conditional on the unusual realizations of the factors that occurred during the Global Financial Crisis, the accuracy of a set of alternative forecasts was the same. Or, as the complement to this, we can test whether the forecasts were equally accurate during more "normal" years.

If, in fact, factor realizations were the main driver of differences in the predictive accuracy of a pair of forecasts, we can imagine situations in which we reject the null in (3.4) without rejecting (3.5). Conversely, two forecasts could be equally accurate "on average" in a given period because one forecast is more strongly affected by shocks to the common factors and less affected by idiosyncratic error shocks, while the reverse holds for the other forecast and the effects balance out. In this case, we do not reject the null in (3.4), whereas the conditional null in (3.5) is rejected.

We next discuss settings under which the hypotheses in (3.5) and (3.4) hold along with how they can be tested.

3.2.4 Homogeneous Factor Loadings

Suppose loadings on the common factors affecting the individual forecast errors in (3.3) are the same across the two forecasts so $\lambda_{i1} = \lambda_{i2} = \lambda_i$. Under quadratic error loss,

$$\Delta L_{i,t+h|t} = (u^2_{i,t+h,1} - u^2_{i,t+h,2}) + 2(u_{i,t+h,1} - u_{i,t+h,2})\lambda_i' f_{t+h}. \tag{3.6}$$

Common unpredictable shocks that are not picked up by any of the forecasts can be thought of as satisfying the assumption of homogeneous factor loadings since they can have a different effect on different units ($\lambda_{i1} \neq \lambda_{j1}$ for $i \neq j$), but will affect the forecasts in the same way($\lambda_{i1}' = \lambda_{i2}'$ for all i). These shocks will, therefore, cancel out from the forecast error differentials. For example, if the effects of a major event such as the Global Financial Crisis were unanticipated by both forecasts and affected them by the same amount, they cancel out from the loss differential.

Under homogeneous factor loadings, the cross-sectional dependence arising from the forecasts' exposure to the common factors, f_{t+h}, does not play an important role in deriving the asymptotics of tests of the null in (3.4) since $\lambda_i' f_{t+h}$ in (3.6) is multiplied by $(u_{i,t+h,1} - u_{i,t+h,2})$. This is assured under the following assumption which requires (conditionally) independent idiosyncratic errors as well as a Lyapounov condition:

Assumption 3. *Suppose that the loadings are homogeneous,* $\lambda_{i1} = \lambda_{i2} = \lambda_i$ *for* $i = 1, ..., n$. *Conditional on* $\mathcal{F} = \sigma(f_{t+h}, \{\lambda_{i1}, \lambda_{i2}\}_{i=1}^n)$, $\{(u_{i,t+h,1}, u_{i,t+h,2})\}_{i=1}^n$ *is independent across* i *with mean zero and bounded* $(4+\delta)$ *moments for some* $\delta > 0$. *Moreover,* $\min_{1 \leq i \leq n} \text{Var}[(u_{i,t+h,1} - u_{i,t+h,2}) \mid \mathcal{F}] \geq c$ *for some constant* $c > 0$ *and*

$$\frac{\left(\sum_{i=1}^n |\lambda_i' f_{t+h}|^{2+\delta}\right)^{1/(2+\delta)}}{\left(\sum_{i=1}^n |\lambda_i' f_{t+h}|^2\right)^{1/2}} = o_P(1).$$

To test the null of equal expected loss for the cross-sectional average in (3.4), consider the test statistic

$$Q_{t+h} = \frac{n^{1/2} \overline{\Delta L}_{t+h|t}}{\sqrt{n^{-1} \sum_{i=1}^n \left(\Delta L_{i,t+h|t}\right)^2}}. \tag{3.7}$$

Under the assumption of pair-wise homogeneous factor loadings, (3.6) shows that testing the null of equal predictive accuracy in period $t+h$ amounts to testing that $E(u_{i,t+h,1}^2 - u_{i,t+h,2}^2) = 0$. This is easily accomplished under Assumption 3 which ensures independence across i for $(u_{i,t+h,1}, u_{i,t+h,2})$ so that asymptotic normality can be established for Q_{t+h} in (3.7) as we next show:[5]

Theorem 3.2.1. *Suppose Assumption 3 holds. Then under the null of equal expected cross-*

[5] Alternatively, we can test this null under assumptions of stationarity which allows us to exploit time-series variation in the factors.

sectional predictive accuracy, $H^{unc}_{0,t+h} : E(\overline{\Delta L}_{t+h}) = 0$*, we have*

$$\limsup_{n\to\infty} P\left(|Q_{t+h}| > z_{1-\alpha/2}\right) \leq \alpha,$$

where $z_{1-\alpha/2}$ *is the* $(1-\alpha/2)$ *quantile of a* $N(0,1)$ *variable.*

Theorem 3.2.1 shows that homogeneous factor loadings lead to a simple test of the null of equal expected loss for the pooled average using data only on a single cross-section. Moreover, the test statistic follows a Gaussian distribution in large cross-sections.

For now, we do not go into details of how the assumption of homogeneous loadings can be tested. However, as we show below, our approach for testing the null in (3.4) remains valid as long as $n^{-1}\sum_{i=1}^{n}[(\lambda'_{i,1}f_{t+h})^2 - (\lambda'_{i,2}f_{t+h})^2] = 0$. Moreover, this condition can be tested empirically and we propose ways to do so later on.

3.2.5 Heterogeneous Factor Loadings

Next, consider the case with heterogeneous factor loadings for the forecast errors, i.e., $\lambda_{i,1} \neq \lambda_{i,2}$. For this case, the loss differential in (3.6) is generalized to

$$\begin{aligned}\Delta L_{i,t+h|t} = &\left[(\lambda'_{i,1}f_{t+h})^2 - (\lambda'_{i,2}f_{t+h})^2\right] \\ &+ \left[u^2_{i,t+h,1} - u^2_{i,t+h,2} + 2(\lambda'_{i,1}f_{t+h}u_{i,t+h,1} - \lambda'_{i,2}f_{t+h}u_{i,t+h,2})\right]. \qquad (3.8)\end{aligned}$$

When the factor loadings differ for the forecasts, equation (3.8) shows that the relative predictive accuracy in period $t+h$ contains a systematic component, $E\left[(\lambda'_{i,1}f_{t+h})^2 - (\lambda'_{i,2}f_{t+h})^2\right]$. Even if f_{t+h} is independent of the factor loadings, $\{(\lambda_{i,1},\lambda_{i,2})\}_{i=1}^{n}$, and these loadings are independent across i, $n^{-1/2}\sum_{i=1}^{n}\left[(\lambda'_{i,1}f_{t+h})^2 - (\lambda'_{i,2}f_{t+h})^2\right]$ is asymptotically normal only conditional on f_{t+h}. This suggests conducting a test of equal expected predictive accuracy conditional on the factor realization as is done in (3.5).

To test the conditional null in (3.5), let $\mathcal{F} = \sigma(f_{t+h}, \{\lambda_{i1}, \lambda_{i2}\}_{i=1}^n)$ and assume that $E(u_{i,t+h,1} \mid \mathcal{F}) = E(u_{i,t+h,2} \mid \mathcal{F}) = 0$. Define

$$\xi_{i,t+h} = \left(u_{i,t+h,1}^2 - u_{i,t+h,2}^2\right) - E\left(u_{i,t+h,1}^2 - u_{i,t+h,2}^2 \mid \mathcal{F}\right) + 2(\lambda_{i,1}' f_{t+h} u_{i,t+h,1} - \lambda_{i,2}' f_{t+h} u_{i,+ht,2}).$$

Using equation (3.8), we have

$$\overline{\Delta L}_{t+h} - E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = n^{-1} \sum_{i=1}^n \xi_{i,t+h}. \tag{3.9}$$

The ideal variance estimate for the object in (3.9) is $n^{-1}\sum_{i=1}^n \xi_{i,t+h}^2$. However, at the unit level, we only observe $e_{i,t+h,m}$ and hence are restricted to computing $n^{-1}\sum_{i=1}^n (\Delta L_{i,t+h|t} - \overline{\Delta L}_{t+h})^2$. Consider the following test statistic

$$\widetilde{Q}_{t+h} = \frac{n^{1/2}\overline{\Delta L}_{t+h}}{\sqrt{n^{-1}\sum_{i=1}^n (\Delta L_{i,t+h|t} - \overline{\Delta L}_{t+h})^2}}. \tag{3.10}$$

To establish properties of the test statistic in (3.10), we need a set of regularity conditions which we summarize in the following assumption:

Assumption 4. *Conditional on $\mathcal{F} = (f_{t+h}, \{\lambda_{i1}, \lambda_{i2}\}_{i=1}^n)$, $\{(u_{i,t+h,1}, u_{i,t+h,2})\}_{i=1}^n$ is independent across i with mean zero and bounded $(4+\delta)$ moments for some $\delta > 0$. Moreover, $\min_{1 \le i \le n} \mathrm{Var}[\xi_{i,t+h} \mid \mathcal{F}] \ge c$ for some constant $c > 0$.*

Using this assumption, we can now test the null $E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = 0$ or, equivalently, establish a confidence interval for $E(\overline{\Delta L}_{t+h} \mid \mathcal{F})$:

Theorem 3.2.2. *Suppose Assumption 4 holds. Then, under the conditional null $H_{0,t+h}^{cond} : E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = 0$, the following result holds for the test statistic in (3.10)*

$$\limsup_{n\to\infty} P\left(|\widetilde{Q}_{t+h}| > z_{1-\alpha/2}\right) \le \alpha,$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of a $N(0,1)$ variable.

Results based on the test statistic in (3.10) can be interpreted in two ways. First, as explained above, they can be viewed as tests of the conditional null $E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = 0$. Second, if we assume that the factor loadings $\{(\lambda_{i,1}, \lambda_{i,2})\}_{i=1}^{n}$ are random, independent across i and independent of f_{t+h}, we can use the test statistic in (3.10) to test $E(\overline{\Delta L}_{t+h} \mid f_{t+h}) = 0$ without also conditioning on the factor loadings ($\lambda_{i,1}$ and $\lambda_{i,2}$). Testing the latter hypothesis introduces an additional term in the numerator of (3.10)

$$E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) - E(\overline{\Delta L}_{t+h|t} \mid f_{t+h}) = f'_{t+h} \left(n^{-1} \sum_{i=1}^{n} \left[\lambda_{i,1}\lambda'_{i,1} - \lambda_{i,2}\lambda'_{i,2} - E(\lambda_{i,1}\lambda'_{i,1} - \lambda_{i,2}\lambda'_{i,2}) \right] \right) f_{t+h}.$$

However, the denominator in (3.10) still overestimates the variance of the numerator of the test statistic under the null. As a result, Theorem 3.2.2 remains valid for testing the null $E(\overline{\Delta L}_{t+h} \mid f_{t+h}) = 0$ and the critical values remain the same.

Under either interpretation, it follows from (3.8) that the variance estimate in (3.10) is conservative. Under the first interpretation, this follows because the variance estimate takes into account variation in the factor structure and in $E(u^2_{i,t+h,1} - u^2_{i,t+h,2})$. Under the second interpretation, the variance estimate still includes cross-sectional variations in $E(u^2_{i,t+h,1} - u^2_{i,t+h,2})$. This seems unavoidable without introducing additional modeling assumptions that impose structure on this variation.[6]

3.2.6 Other loss functions

In practice, applied researchers might consider loss functions other than the squared error loss in (3.1), including linex, absolute error or piece-wise linear loss, see, e.g., Elliott, Komunjer

[6]Essentially, we have a CLT for independent but non-identically distributed variables, $\Delta L_{i,t+h|t} - E\left[\Delta L_{i,t+h|t} | f_{t+h}\right]$, but the exact variance is difficult to estimate because $E\left[\Delta L_{i,t+h|t} | f_{t+h}\right]$ cannot be estimated from the observed data.

and Timmermann (2005). Fortunately, the methodology in Section 3.2.5 can readily be extended to such loss functions.

To see this, suppose we replace Assumption 4 with the assumption that, conditional on $\mathcal{F}$, $\{\Delta L_{i,t+h|t}\}_{i=1}^{n}$ is independent, where $\Delta L_{i,t+h|t} = L(y_{it+h}, \hat{y}_{it+h|t,1}) - L(y_{it+h}, \hat{y}_{it+h|t,2})$ for a general loss function $L(\cdot,\cdot)$. Under the conditional null hypothesis $E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = 0$, $\overline{\Delta L}_{t+h} - E(\overline{\Delta L}_{t+h} \mid \mathcal{F})$ will be the average of terms that, conditional on $\mathcal{F}$, have mean zero and are independent. Therefore, with moment conditions similar to those in Assumption 4, Theorem 3.2.2 remains valid. In Section 3.5, we demonstrate this point using Monte Carlo simulations for the test of equal conditionally expected loss applied to the linex loss function. We find results that are very similar to those obtained under squared error loss.

For the unconditional test, we can consider a linear factor structure as a series approximation. For example, suppose that $e_{i,t+h,m} = \lambda_{im}' f_{t+h} + u_{i,t+h,m}$ and $L_{i,t+h,m} = \phi(e_{i,t+h,m})$ for some function $\phi(\cdot)$. Provided that $\phi(\cdot)$ is smooth enough and $e_{i,t+h,m}$ is bounded, standard approximation results can be used to give a polynomial approximation $\phi(x) \approx \sum_{j=0}^{k} a_j x^j$, where k grows slowly with the sample size. Because $(\lambda_{im}' f_{t+h} + u_{i,t+h,m})^j$ contains powers of $\lambda_{im}' f_{t+h}$, polynomials of factors and factor loadings become new factors in an augmented linear factor structure. Clearly, the details of this approach (e.g., approximation rate and strong factor conditions) require serious theoretical analysis, which we leave for future research.

Cross-sectional comparisons of forecast errors sometimes involve variables that are measured in very different units. This can mean that the comparisons are dominated by a few variables, possibly impairing the finite-sample behavior of the test statistics. To address this point, one can use squared percentage errors which tend to be more comparable across variables. Alternatively, individual variables' forecast errors can be scaled by their standard errors prior to calculating the test statistics.

3.3 Decomposing Differences in Forecasting Performance

Equation (3.3) decomposes the forecast errors into a common factor component and an uncorrelated idiosyncratic error component. In some economic applications, it is important to be able to attribute differences in forecasting performance to these two sources. For example, Mackowiak and Wiederholt (2009) develop a rational inattention model in which firms acquire and process information subject to a constraint on their total attention budget. Consistent with the setup in (3.3), Mackowiak and Wiederholt (2009) partition firms' information set into signals about a common (aggregate) factor and an idiosyncratic term. The constraint on each forecaster's attention introduces a trade-off between reducing the uncertainty about the common factor versus reducing the variance of the idiosyncratic error. Similarly, the finance literature on performance of investment managers distinguishes between generalists who possess market timing skills that require an ability to predict pervasive (common) factors affecting a broad set of asset returns versus stock pickers with security selection skills which require specialist firm-level knowledge akin to more precise signals on the idiosyncratic error terms (see, e.g., Blake et al. (2013)).

The importance of these types of skills is likely to vary over time as a result of common factor volatility being higher during recessions or in financial crises (favoring market timers) and lower during expansions and calmer periods (favoring stock pickers), see, e.g., Kacperczyk, Nieuwerburgh and Veldkamp (2014). By conducting tests on individual cross-sections, our approach can help identify periods in which forecasters with a comparative advantage at predicting the common factors (generalists) perform relatively better than the forecasters who focus instead on the idiosyncratic error component (specialists).

Decomposing forecast errors into common factors and uncorrelated idiosyncratic terms is also important in applications of forecast combination since these terms matter for calculating optimal combination weights which depend on both the overall error variance and on the covariance between forecast errors. The larger the contribution to forecast errors from the

common factors and the more homogeneous the factor loadings are, the closer the optimal combination weights will be to equal-weighting. Related to this, the scope for achieving gains in predictive accuracy from forecast combination is likely to be highest during times when the correlation in forecast errors is weakest, i.e., less driven by common factors with similar loadings and more by idiosyncratic errors.

We next discuss how to conduct inference on the squared conditional bias and idiosyncratic variance components.

3.3.1 Decomposing the Conditional Squared Error Loss

Using equation (3.8), we can express the (cross-sectional) average conditional squared error loss difference as the sum of the average difference in squared conditional bias and the average difference in the conditional idiosyncratic error variance:

$$\underbrace{n^{-1}\sum_{i=1}^{n} E\left(\Delta L_{i,t+h|t} \mid \mathcal{F}\right)}_{E(\overline{\Delta L}_{t+h} \mid \mathcal{F})} = \underbrace{n^{-1}\sum_{i=1}^{n}\left[(\lambda'_{i,1} f_{t+h})^2 - (\lambda'_{i,2} f_{t+h})^2\right]}_{bias^2_{t+h}} + \underbrace{n^{-1}\sum_{i=1}^{n} E\left(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}\right)}_{E\left(\Delta u^2_{t+h} \mid \mathcal{F}\right)}. \tag{3.11}$$

The terms on the right hand side of the decomposition in (3.11) are unobserved. However, note that

$$\overline{\Delta L}_{t+h} - bias^2_{t+h} = \overline{\Delta u}^2_{t+h} + \frac{2}{n}\sum_{i=1}^{n}\left[\lambda'_{i,1} f_{t+h} u_{i,t+h,1} - \lambda'_{i,2} f_{t+h} u_{i,t+h,2}\right], \tag{3.12}$$

where $\overline{\Delta u}^2_{t+h} = n^{-1}\sum_{i=1}^{n}\left(u^2_{i,t+h,1} - u^2_{i,t+h,2}\right)$. Provided that n is relatively large so the last term on the right side of (3.12) is small, the bias-adjusted average loss differential on the left hand side of (3.12) can be expected to be a good estimate of the difference in the two forecasts' idiosyncratic variance at time t, $E\left(\Delta u^2_{t+h} \mid \mathcal{F}\right)$.[7]

[7] Of course, we do not directly observe the idiosyncratic errors and factors. However, since $\overline{\Delta L}_{t+h}$ is observed, from (3.12) we only need to estimate the factor-induced squared bias term, $bias^2_{t+h}$.

We next discuss three strategies for computing $(\lambda_{i,1}' f_{t+h})^2 - (\lambda_{i,2}' f_{t+h})^2$. The first exploits clusters in factor loadings and so is applicable when factor loadings are homogeneous within certain groups of units. This approach can be computed on a single cross-section and poses no limit on the number of factors affecting the forecast errors but requires that clusters can be identified within which there is little or no heterogeneity in the factor loadings. The second approach uses the common correlated effects (CCE) method of Pesaran (2006) and so requires the availability of panel data to estimate factor loadings from time series data. This approach does not impose tight restrictions on factor loadings but, in practice, limits the number of common factors driving the forecast errors. The third approach, principal components (PCA), again requires the availability of panel data and is similar to the CCE approach. However, it does not impose tight bounds on the number of common factors in the forecast error differentials.

3.3.2 Clustering in Factor Loadings

It is common in empirical applications to have data on units that share certain observable characteristics or features which make them more similar than randomly selected units. For example, advanced economies may react in a broadly similar way to supply shocks which, in turn, affect emerging or developing economies very differently. Or, the effect of an interest rate increase on the default probability of credit card holders may be quite different across high, medium, and low income households, yet be broadly similar within these three categories.

In this section we develop a class of estimators using the identifying assumption that clusters of cross-sectional units share the same factor loadings, while allowing factor loadings to differ across clusters. Formally, suppose that a set of K clusters $\bigcup_{k=1}^{K} H_k = \{1, ..., n\}$ form a partition of all n units so that each unit belongs to a unique cluster, H_k, i.e., $H_j \cap H_l = \emptyset$ with $n_k = |H_k|$ elements in the kth cluster. We assume that the cluster membership for each unit is known ex ante and so is not determined endogenously from the data. Moreover, suppose that the factor loadings $(\lambda_{i,1}, \lambda_{i,2})$ can differ across clusters $(\lambda_{i,1}, \lambda_{i,2}) \neq (\lambda_{j1}, \lambda_{j2})$ for $i \in H_k$ and $j \in H_l$,

but are homogeneous within clusters

$$(\lambda_{i,1}, \lambda_{i,2}) = (\lambda_{1,(k)}, \lambda_{2,(k)}) \text{ for all } i \in H_k. \tag{3.13}$$

Testing Equal Idiosyncratic Error Variances

We first discuss how to test the conditional null of equal average idiosyncratic error variance for the two forecasts given $\mathcal{F}$ for all units in cluster k:

$$H_0^{idio}: \; n_k^{-1} \sum_{i \in H_k} E(u_{i,t+h,1}^2 - u_{i,t+h,2}^2 \mid \mathcal{F}) = 0. \tag{3.14}$$

To test this null, we need to construct an estimate of the idiosyncratic variance within each cluster. To see how group patterns in factor loadings allow us to identify the idiosyncratic variance component, $\overline{\Delta u}_{t+h}^2$, define the errors from the two forecasts, averaged within each cluster, as

$$\overline{e}_{1,k,t+h} \equiv n_k^{-1} \sum_{i \in H_k} \left(y_{i,t+h} - \hat{y}_{i,t+h|t,1} \right) = \lambda_{1,(k)}' f_{t+h} + n_k^{-1} \sum_{i \in H_k} u_{i,t+h,1},$$

and

$$\overline{e}_{2,k,t+h} \equiv n_k^{-1} \sum_{i \in H_k} \left(y_{i,t+h} - \hat{y}_{i,t+h|t,2} \right) = \lambda_{2,(k)}' f_{t+h} + n_k^{-1} \sum_{i \in H_k} u_{i,t+h,2}.$$

Squaring these within-cluster average forecast errors, we have

$$\overline{e}_{1,k,t+h}^2 - \overline{e}_{2,k,t+h}^2 = (\lambda_{1,(k)}' f_{t+h})^2 - (\lambda_{2,(k)}' f_{t+h})^2 + \left(n_k^{-1} \sum_{i \in H_k} u_{i,t+h,1} \right)^2 - \left(n_k^{-1} \sum_{i \in H_k} u_{i,t+h,2} \right)^2$$
$$+ 2\lambda_{1,(k)}' f_{t+h} n_k^{-1} \sum_{i \in H_k} u_{i,t+h,1} - 2\lambda_{2,(k)}' f_{t+h} n_k^{-1} \sum_{i \in H_k} u_{i,t+h,2}. \tag{3.15}$$

Define $\overline{\Delta L}_{t+h,k} \equiv n_k^{-1} \sum_{i \in H_k} \Delta L_{i,t+h}$ and let $\overline{\Delta u}_{t+h,k}^2$ be the average loss differential for

cluster k adjusted for the difference $(\overline{e}^2_{1,k,t+h} - \overline{e}^2_{1,k,t+h})$:

$$\overline{\Delta u}^2_{t+h,k} = \overline{\Delta L}_{t+h,k} - (\overline{e}^2_{1,k,t+h} - \overline{e}^2_{1,k,t+h}). \tag{3.16}$$

This suggests using the following statistic to test H_0^{idio} in (3.14):

$$S_k = \frac{\sqrt{n_k}\overline{\Delta u}^2_{t+h,k}}{\sqrt{n_k^{-1}\sum_{i\in H_k}(\Delta L_{i,t+h} - \overline{\Delta L}_{t+h,k})^2}}. \tag{3.17}$$

Theorem 3.3.1. *Suppose Assumption 4 holds. Then under the null hypothesis* H_0^{idio} : $n_k^{-1}\sum_{i\in H_k} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) = 0$*, we have*

$$\limsup_{n_k\to\infty} P\left(|S_k| > z_{1-\alpha/2}\right) \leq \alpha,$$

where $z_{1-\alpha/2}$ *is the* $(1-\alpha/2)$ *quantile of a* $N(0,1)$ *variable.*

Alternatively, we can test the weaker null of equal expected squared idiosyncratic forecast errors holding on average, i.e., across all units though not necessarily within each cluster:

$$H_0^{idio-av}:\ n^{-1}\sum_{i=1}^{n} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) = 0. \tag{3.18}$$

To this end, let $\overline{\Delta u}^2_{t+h} = \sum_{k=1}^K \frac{n_k}{n}\overline{\Delta u}^2_{t+h,k}$ be the cluster-weighted average difference in squared idiosyncratic forecast errors, and consider the test statistic

$$S_c = \frac{\sqrt{n}\overline{\Delta u}^2_{t+h}}{\sqrt{n^{-1}\sum_{k=1}^K\sum_{i\in H_k}(\Delta L_{i,t+h} - \overline{\Delta L}_{t+h,k})^2}}. \tag{3.19}$$

We use S_c to test the null in (3.18) of equal average idiosyncratic forecast error variance:

Corollary 3.3.1. *Suppose Assumption 4 holds and assume that* $\lim_{n\to\infty} n_k/n > 0$ *for all* $1 \leq k \leq K$.

Then under the null $H_0^{idio-av}: \; n^{-1}\sum_{i=1}^{n} E(u_{i,t+h,1}^2 - u_{i,t+h,2}^2 \mid \mathcal{F}) = 0$*, we have*

$$\limsup_{n\to\infty} P\left(|S_c| > z_{1-\alpha/2}\right) \leq \alpha,$$

where $z_{1-\alpha/2}$ *is the* $(1-\alpha/2)$ *quantile of a* $N(0,1)$ *variable.*

Using Corollary 3.3.1, we can compute a $1-\alpha$ confidence interval for the squared idiosyncratic forecast errors $\overline{\Delta u}_{t+h}^2$ as

$$\overline{\Delta u}_{t+h}^2 \pm \frac{z_{1-\alpha/2}}{\sqrt{n}} \sqrt{n^{-1}\sum_{k=1}^{K}\sum_{i\in H_k}(\Delta L_{i,t+h} - \overline{\Delta L}_{t+h,k})^2}. \tag{3.20}$$

Testing Equal Squared Biases

Next, consider the squared bias component of the expected loss differential in (3.11). Under the assumed homogeneous factor loadings within clusters in (3.13), we have

$$n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}' f_{t+h})^2 - (\lambda_{i,2}' f_{t+h})^2\right] = \sum_{k=1}^{K}\frac{n_k}{n}\left((\lambda_{1,(k)}' f_{t+h})^2 - (\lambda_{2,(k)}' f_{t+h})^2\right).$$

We can estimate $(\lambda_{1,(k)}' f_{t+h})^2 - (\lambda_{2,(k)}' f_{t+h})^2$ by $\overline{e}_{1,k,t+h}^2 - \overline{e}_{2,k,t+h}^2$. By (3.15), we have

$$\begin{aligned}\overline{e}_{1,k,t+h}^2 - \overline{e}_{2,k,t+h}^2 = {} & (\lambda_{1,(k)}' f_{t+h})^2 - (\lambda_{2,(k)}' f_{t+h})^2 \\ & + 2\lambda_{1,(k)}' f_{t+h} n_k^{-1}\sum_{i\in H_k} u_{i,t+h,1} - 2\lambda_{2,(k)}' f_{t+h} n_k^{-1}\sum_{i\in H_k} u_{i,t+h,2} + O_P(n_k^{-1}).\end{aligned}$$

To test the null of equal squared bias, we use the following test statistic:

$$B_{n,1} = \frac{\sqrt{n}\sum_{k=1}^{K}\frac{n_k}{n}(\overline{e}_{1,k,t+h}^2 - \overline{e}_{2,k,t+h}^2)}{2\sqrt{n^{-1}\sum_{k=1}^{K}\sum_{i\in H_k}(\overline{e}_{1,k,t+h}\hat{u}_{i,t+h,1} - \overline{e}_{2,k,t+h}\hat{u}_{i,t+h,2})^2}}, \tag{3.21}$$

where $\hat{u}_{i,t+h,1} = y_{i,t+h} - \hat{y}_{i,t+h|t,m_1} - \overline{e}_{1,k,t+h}$ and $\hat{u}_{i,t+h,2} = y_{i,t+h} - \hat{y}_{i,t+h|t,m_2} - \overline{e}_{2,k,t+h}$. We can

show that $B_{n,1}\left(n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}'f_{t+h})^2-(\lambda_{i,2}'f_{t+h})^2\right]\right)\to^d N(0,1)$, and so:

Theorem 3.3.2. *Suppose Assumptions 3 holds and assume that* $\lim_{n\to\infty} n_k/n > 0$ *for all* $1 \le k \le K$. *Then under* $H_0:\ n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}'f_{t+h})^2-(\lambda_{i,2}'f_{t+h})^2\right]=0$, *we have*

$$\limsup_{n\to\infty} P\left(|B_{n,1}|>z_{1-\alpha/2}\right)\le\alpha.$$

The null of equal squared bias relates to our earlier discussion of homogeneous versus heterogeneous factor loadings: If factor loadings are the same across two sets of forecasts, their squared bias differential should also be close to zero.

Theorem 3.3.2 yields a $1-\alpha$ confidence interval for $n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}'f_{t+h})^2-(\lambda_{i,2}'f_{t+h})^2\right]$:

$$\sum_{k=1}^{K}\frac{n_k}{n}(\overline{e}_{1,k,t+h}^2-\overline{e}_{2,k,t+h}^2)\pm\frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{n^{-1}\sum_{k=1}^{K}\sum_{i\in H_k}(\overline{e}_{1,k,t+h}\hat{u}_{i,t+h,1}-\overline{e}_{2,k,t+h}\hat{u}_{i,t+h,2})^2}. \quad (3.22)$$

Note that because $B_{n,1}\left(n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}'f_{t+h})^2-(\lambda_{i,2}'f_{t+h})^2\right]\right)\to^d N(0,1)$, the confidence interval is asymptotically exact.

3.3.3 Factor Structure Estimated by CCE

In many empirical applications, a cluster structure may not be suitable either because units are not easily assigned to individual clusters or because factor loadings are not homogeneous within clusters. For such applications, a more traditional factor setting may be more appropriate. To this end, suppose we observe a panel of forecast errors $\{e_{i,s+h,m}\}_{1\le i\le n,\ 1\le s\le T}$ generated according to the factor model in (3.3), $e_{i,s+h,m}=\lambda_{i,m}'f_{s+h}+u_{i,s+h,m}$, where $m=1,2$, $\lambda_{i,m}\in\mathbb{R}^{r\times v}$ and $f_{s+h}\in\mathbb{R}^r$ with $v\ge r$, so the number of observables, v, is at least equal to the number of factors, r. The requirement that $v\ge r$ implies that if we do not include observables other than the two sets of forecast errors, we can allow for at most two factors. Conversely, including more

observable variables that are driven by the same factors lets us relax this restriction and allow for additional factors.

Difference in Idiosyncratic Error Variances

Let $e_{i,s+h} = (e_{i,s+h,1}, e_{i,s+h,2})' \in \mathbb{R}^2$ and $u_{i,s+h} = (u_{i,s+h,1}, u_{i,s+h,2})' \in \mathbb{R}^2$ be 2×1 vectors of forecast errors and idiosyncratic residuals and define the cross-sectional averages $\bar{e}_{s+h} = n^{-1}\sum_{i=1}^n e_{i,s+h}$, $\bar{u}_{s+h} = n^{-1}\sum_{i=1}^n u_{i,s+h}$ and $\bar{\lambda} = n^{-1}\sum_{i=1}^n \lambda_i$ with $\lambda_i = (\lambda_{i,1}, \lambda_{i,2}) \in \mathbb{R}^{r\times 2}$. Assuming that we can invoke a CLT for the cross-sectional average of the idiosyncratic shocks, $\bar{u}_{s+h}$ will be small and $\bar{e}_{s+h} \approx \bar{\lambda}' f_{s+h}$ can be used as a proxy for the unobserved factors. This is the common correlated effects (CCE) idea proposed in Pesaran (2006). In turn, we can estimate the individual factor loadings, λ_{im}, from a time-series regression

$$\hat{\lambda}_i' = \left(\sum_{s=1}^T e_{i,s+h}\bar{e}_{s+h}'\right)\left(\sum_{s=1}^T \bar{e}_{s+h}\bar{e}_{s+h}'\right)^{-1}.$$

Let $\lambda_{i,1}$ denote the first column of λ_i, with similar notations used for $\hat{\lambda}_{i,1}$ and $\hat{\lambda}_{i,2}$. Consider the following regularity conditions:

Assumption 5. *The following conditions hold for $m = 1, 2$:*

(1) the smallest eigenvalue of $\bar{\lambda}\bar{\lambda}'$ is bounded away from zero.

(2) conditional on $\{f_{s+h}\}_{s+h=1}^T$ and $\{\lambda_i\}_{i=1}^n$, $\{u_{i,t+h,m}\}_{i=1}^n$ has mean zero with bounded variance and is independent across i.

The first part of Assumption 5 implies that the number of factors cannot exceed the dimension of $e_{i,s+h}$ since otherwise the smallest eigenvalue of $\bar{\lambda}\bar{\lambda}'$ is zero. We also impose additional regularity conditions. These are part of Assumptions A, B and C in Bai (2003) and are routinely imposed in factor analysis.

Assumption 6. *The following conditions hold for $m = 1, 2$:*

(1) $n^{-1}\sum_{i=1}^n \lambda_{i,m}\lambda_{i,m}'$ and $Ef_{s+h}f_{s+h}'$ have eigenvalues bounded away from zero and infinity.

(2) $\sum_{s+h=1}^{T}\sum_{i=1}^{n}\lambda_{i,m}u_{i,s+h,m}f'_{s+h}=O_P(\sqrt{nT})$.

(3) There exists a constant $M>0$ *such that* $\|\gamma_n(s,\tau)\|\leq M$ *and* $T^{-1}\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\|\gamma_n(s,\tau)\|\leq M$, *where* $\gamma_n(s,\tau)=n^{-1}\sum_{i=1}^{n}Eu_{i,s+h}u'_{i,\tau+h}$.

(4) $n/T^2=o(1)$.

Using Assumption 5 and 6, we can characterize the difference between the average squared forecast errors and the average squared factor values, both weighted by the factor loadings, λ'_i :

Lemma 3.3.1. *Under Assumptions 5 and 6, we have*

$$n^{-1/2}\sum_{i=1}^{n}[(\hat{\lambda}'_{i,1}\bar{e}_{t+h})^2-(\lambda'_{i,1}f_{t+h})^2]=2n^{-1/2}\vec{u}'_{t+h}\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left(\sum_{i=1}^{n}\lambda_{i,1}\lambda'_{i,1}\right)f_{t+h}+o_P(1).$$

Next, consider the null that the difference in the squared idiosyncratic variance component of the forecast errors equals zero:

$$H_0:\ n^{-1}\sum_{i=1}^{n}E(u^2_{i,t+h,1}-u^2_{i,t+h,2}\mid\mathcal{F})=0. \tag{3.23}$$

To test this null, we use the following test statistic

$$S_{cce}=\frac{\sqrt{n}\overline{\Delta\hat{u}}^2_{t+h}}{\sqrt{n^{-1}\sum_{i=1}^{n}(\Delta L_{i,t+h|t}-[(\hat{\lambda}'_{i,1}\bar{e}_{t+h})^2-(\hat{\lambda}'_{i,2}\bar{e}_{t+h})^2]-\hat{c}_{t+h}+\hat{u}'_{i,t+h}\hat{D}_{t+h})^2}}, \tag{3.24}$$

where $\hat{c}_{t+h}=n^{-1}\sum_{i=1}^{n}(\Delta L_{i,t+h|t}-[(\hat{\lambda}'_{i,1}\bar{e}_{t+h})^2-(\hat{\lambda}'_{i,2}\bar{e}_{t+h})^2]+\hat{u}'_{i,t+h}\hat{D}_{t+h})$,

$$\overline{\Delta\hat{u}}^2_{t+h}=n^{-1}\sum_{i=1}^{n}\Delta L_{i,t+h}-n^{-1}\sum_{i=1}^{n}[(\hat{\lambda}'_{i,1}\bar{e}_{t+h})^2-(\hat{\lambda}'_{i,2}\bar{e}_{t+h})^2] \tag{3.25}$$

and

$$\hat{D}_{t+h}=n^{-1}\sum_{i=1}^{n}(\hat{\lambda}_{i,1}\hat{\lambda}'_{i,1}-\hat{\lambda}_{i,2}\hat{\lambda}'_{i,2})\bar{e}_{t+h}. \tag{3.26}$$

Using these definitions, we now have the following result:

Theorem 3.3.3. *Suppose that Assumptions 5 and 6 hold. Then under* $H_0 : n^{-1}\sum_{i=1}^{n} E(u_{i,t+h,1}^2 - u_{i,t+h,2}^2 \mid \mathcal{F}) = 0$,

$$S_{cce} \to^d N(0,1).$$

Using that S_{cce} follows a standard Gaussian distribution asymptotically, we can compute a $1-\alpha$ confidence interval for $n^{-1}\sum_{i=1}^{n} E(u_{i,t,1}^2 - u_{i,t,2}^2 \mid \mathcal{F})$ as

$$\overline{\Delta\hat{u}}_{t+h}^2 \pm \frac{z_{1-\alpha/2}}{\sqrt{n}} \sqrt{n^{-1}\sum_{i=1}^{n} (\Delta L_{i,t+h|t} - [(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2 - (\lambda_{i,2}'\bar{e}_{t+h})^2] - \hat{c}_{t+h} + \hat{u}_{i,t+h}'\hat{D}_{t+h})^2} \quad (3.27)$$

Squared Bias Differences

Next, consider the squared bias component of the MSE loss differential. Define

$$D_{t+h} = \bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1} \left(n^{-1}\sum_{i=1}^{n} [\lambda_{i,1}\lambda_{i,1}' - \lambda_{i,2}\lambda_{i,2}'] \right) f_{t+h}.$$

Using

$$\sqrt{n}\big(n^{-1}\textstyle\sum_{i=1}^{n}\big[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2 - (\hat{\lambda}_{i,2}'\bar{e}_{t+h})^2\big] - n^{-1}\sum_{i=1}^{n}\big[(\lambda_{i,1}'f_{t+h})^2 - (\lambda_{i,2}'f_{t+h})^2\big]\big) = 2n^{1/2}\bar{u}_{t+h}'D_{t+h} + o_P(1),$$

it follows that $n^{-1}\sum_{i=1}^{n}\left[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2 - (\hat{\lambda}_{i,2}'\bar{e}_{t+h})^2\right]$ is a $\sqrt{n}$-consistent estimator for the average difference in the squared bias differential, $n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}'f_{t+h})^2 - (\lambda_{i,2}'f_{t+h})^2\right]$, where the estimation error is asymptotically $2\bar{u}_{t+h}'D_{t+h}$. To construct tests for the squared bias difference, consider the following test statistic

$$B_{n,2} = \frac{n^{-1/2}\sum_{i=1}^{n}\left[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2 - (\hat{\lambda}_{i,2}'\bar{e}_{t+h})^2\right]}{2\sqrt{n^{-1}\sum_{i=1}^{n}(\hat{u}_{i,t+h}'\hat{D}_{t+h})^2}}, \quad (3.28)$$

where, again, $\hat{u}_{i,t+h} = e_{i,t+h} - \hat{\lambda}_i'\bar{e}_{t+h}$. The following result characterizes the distribution of this statistic:

Theorem 3.3.4. *Suppose that Assumption 5 holds. Then under* H_0 : $n^{-1}\sum_{i=1}^{n}\left[(\lambda'_{i,1}f_{t+h})^2-(\lambda'_{i,2}f_{t+h})^2\right]=0$,

$$B_{n,2}\rightarrow^{d} N(0,1).$$

Using Theorem 3.3.4, we can construct a confidence interval for the average squared bias differential, $n^{-1}\sum_{i=1}^{n}\left[(\lambda'_{i,1}f_{t+h})^2-(\lambda'_{i,2}f_{t+h})^2\right]$ as

$$n^{-1}\sum_{i=1}^{n}\left[(\hat{\lambda}'_{i,1}\bar{e}_{t+h})^2-(\hat{\lambda}'_{i,2}\bar{e}_{t+h})^2\right]\pm\frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{n^{-1}\sum_{i=1}^{n}(\hat{u}'_{i,t+h}\hat{D}_{t+h})^2}. \tag{3.29}$$

Again, this confidence interval is asymptotically exact.

Comparing (3.27) and (3.29), we note a difference in the asymptotics. Although both variance expressions have $\hat{u}'_{i,t+h}\hat{D}_{t+h}$, the former has the additional term $\Delta L_{i,t+h|t}-[(\hat{\lambda}'_{i,1}\bar{e}_{t+h})^2-(\hat{\lambda}'_{i,2}\bar{e}_{t+h})^2]-\hat{c}_{t+h}$. This difference could well make a difference to the finite-sample performance of the two tests. For example, overfitting could result in a very small $\hat{u}_{i,t+h}$ and thus a small $\hat{u}'_{i,t+h}\hat{D}_{t+h}$. By including the extra term, tests associated with Theorem 3.3.3 might be more robust in small samples.

3.3.4 Factor Structure Estimated by PCA

An alternative to the CCE approach in Section 3.3.3 is to use principal components analysis (PCA) to extract the common factors. A notable advantage of the PCA approach is that, unlike the CCE approach, the number of observed forecast errors does not pose an upper bound on the number of factors. In practice, this means that we can allow for more factors under the PCA approach.

Define the difference in the idiosyncratic forecast error variance

$$\overline{\Delta\hat{u}}^2_{t+h} = n^{-1}\sum_{i=1}^{n}\Delta L_{i,t+h} - n^{-1}\sum_{i=1}^{n}\left[(\hat{\lambda}'_{i,1}\hat{f}_{t+h})^2 - (\hat{\lambda}'_{i,2}\hat{f}_{t+h})^2\right]. \tag{3.30}$$

As before, let $\hat{f}_{t+h}$ and $\hat{\lambda}_i$ be the estimated factors and factor loadings obtained using PCA estimation. Then we have the following results:

Lemma 3.3.2. *Under Assumptions A-F in Bai (2003), we have*

$$\begin{aligned}
&\sqrt{n}\left[\overline{\Delta\hat{u}}^2_{t+h} - n^{-1}\sum_{i=1}^{n}E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})\right] \\
&= n^{-1/2}\sum_{i=1}^{n}\left[(u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) + 2(\lambda'_{i,1}f_{t+h}u_{i,t+h,1} - \lambda'_{i,2}f_{t+h}u_{i,t+h,2})\right] \\
&\quad + o_P(1).
\end{aligned}$$

Notice that we no longer have a term involving $\hat{D}_{t+h}$. Depending on the distribution of the idiosyncratic term, the PCA approach might yield a more efficient estimator than the CCE approach since it does not require us to estimate this term.

From this point, all steps in the inference procedure are exactly the same as those in Section 3.3.3, except that $(\hat{\lambda}'_{i,1}\bar{e}_{t+h}, \hat{\lambda}'_{i,2}\bar{e}_{t+h})$ is replaced by the PCA estimate $(\hat{\lambda}'_{i,1}\hat{f}_{t+h}, \hat{\lambda}'_{i,2}\hat{f}_{t+h})$ and we set $\hat{D}_{t+h} = 0$. Specifically, in Equations (3.24), (3.25) and (3.27), we replace $(\hat{\lambda}'_{i,1}\bar{e}_{t+h}, \hat{\lambda}'_{i,2}\bar{e}_{t+h})$ with the PCA estimate $(\hat{\lambda}'_{i,1}\hat{f}_{t+h}, \hat{\lambda}'_{i,2}\hat{f}_{t+h})$ and set $\hat{D}_{t+h} = 0$. We also replace $B_{n,2}$ in (3.28) with the following

$$\tilde{B}_{n,2} = \frac{n^{-1/2}\sum_{i=1}^{n}\left[(\hat{\lambda}'_{i,1}\hat{f}_{t+h})^2 - (\hat{\lambda}'_{i,2}\hat{f}_{t+h})^2\right]}{2\sqrt{n^{-1}\sum_{i=1}^{n}(\hat{\lambda}'_{i,1}\hat{f}_{t+h}\hat{u}_{i,t+h,1} - \hat{\lambda}'_{i,2}\hat{f}_{t+h}\hat{u}_{i,t+h,2})^2}}, \tag{3.31}$$

where $\hat{u}_{i,t+h,m} = e_{i,t+h,m} - \lambda'_{i,m}f_{t+h}$.

3.4 Empirical Application to Earnings Forecasts

To illustrate the economic insights that can be gained from our new test statistics, we next conduct an empirical analysis that compares the accuracy of analysts' forecasts of quarterly earnings recorded across six large brokerage firms.

3.4.1 Data

Using data from the Institutional Brokers Estimate System (IBES), we examine forecasts of quarterly earnings per share (EPS) generated by analysts at six large brokerage firms, namely Merrill Lynch (MERRILL), JP Morgan Chase (JPMORGAN), Credit Suisse (FBOSTON), Goldman Sachs (GOLDMAN), Morgan Stanley (MORGAN) and Deutsche Bank (LAWRENCE). Analysts' forecasts are not always updated so frequently at long horizons, so we focus on forecasts generated at the two-month horizon to avoid issues caused by stale forecasts.[8]

Our quarterly data span the 20-year period from 2000Q1 to 2020Q1. Table 3.1 presents summary statistics on the number of firms covered by each brokerage firm (Panel A) as well as the average number of firms covered each quarter (Panel B). The total number of firms covered by the brokerage firms in at least one quarter ranges from 1,437 (Lawrence) to 1,825 (Merrill), while the average number of firm-level quarterly EPS estimates reported by the brokerage firms ranges from 239 (Lawrence) to 356 (Merrill).

In addition to inspecting the forecasting performance across all firms, we also use SIC codes to assign individual firms to five industry groupings chosen to match the Fama-French industry classification, namely Consumer, Manufacturing, High Tech, Health, and Other. Firm numbers are highest in the Other category, followed by High Tech, Manufacturing, Consumer, and Health.

[8] We calculate the forecast horizon using daily data on the announcement date (ANNDATS) and forecast period end date (FPEDATS).

3.4.2 Factor Structure in Errors and Loss Differentials

Table 3.2 presents results from testing for the presence of common factors in the EPS forecast errors for the six brokerage firms using the growth ratio (GR) and eigenvalue (ER) statistics of Ahn and Horenstein (2013) as well as the Onatski (2009) test (ED).[9] For three of the brokerage firms (Fboston, Goldman, and Merrill), the three tests identify a single common factor in the forecast errors, while for a fourth (Lawrence), two of the tests suggest a single common factor while the third (ED) uncovers three factors. For the remaining two brokerages, the tests identify either zero (JP Morgan) or two (Morgan) common factors.

Given these findings, we next inspect whether controlling for a common factor in the EPS forecast errors captures their correlation. To this end, Table 3.3 reports the average correlation between the forecast errors without controlling for a common factor (top row labeled 0) as well as after controlling for one or two common factors (second and third rows), along with values of the test statistic of Pesaran (2004). Under the null that the data are uncorrelated, this test statistic is asymptotically normally distributed. The average cross-sectional correlation in forecast errors ranges from 0.07 (Morgan) to 0.12 (Lawrence). Moreover, the underlying correlations are highly statistically significant with test statistics exceeding 25, indicating very strong evidence of cross-correlations in all brokerage firms' EPS forecast errors.[10]

Controlling for exposures to a single common factor, average correlations drop to a much narrower range from -0.01 to 0.03. While some of these test statistics remain statistically significant–notably for Morgan Stanley–the test statistics typically come down by more than an order of magnitude as does the average cross-correlation estimate. Accounting for a second common factor only has a marginal effect on average correlations and test statistics, except for Morgan Stanley whose average correlation declines from 0.03 to 0.01. We conclude that very

[9] We estimate the factors from the subset of firm-brokerage pairings with at least 40 quarterly observations, corresponding to half of the sample period.

[10] Cross-sectional regressions of EPS outcomes on brokerage forecasts yield predictive R^2-values in the range 0.5-1, with an average of 0.91.

little common variation remains in the forecast errors after accounting for a single common factor and, hence, that the setup in (3.3) appears to provide an accurate empirical characterization for our data.

We next explore evidence of heterogeneity in cluster loadings across industries. To the extent that industries differ in how sensitive their earnings are to the economic cycle, we might expect factor loadings to be clustered along industry lines with firms within a particular industry exhibiting more similar factor loadings than firms belonging to different industries. To see if this holds, we estimate a common factor model $\tilde{e}_{it+h} = \lambda_i f_{t+h} + \varepsilon_{it+h}$ on the standardized forecast errors ($\tilde{e}_{it+h}$) subject to the constraint $\sum_{i=1}^{N} \lambda_i^2 = 1$. Specifically, we first demean and scale the forecast errors so they have mean zero and unit standard deviation. Next, we estimate factors and factor loadings by PCA using the EM algorithm.

Table 3.4 shows the standard deviation of the estimated factor loadings across all firms (first column) as well as within the five industry clusters. Factor loadings that are more homogeneous within a particular industry than in the aggregate should give rise to smaller values of the standard deviations than in the first column. We see modest evidence of this: For all but one of the six brokerage firms, the standard deviation of the factor loadings is smaller in three of the five industries compared to in the aggregate. Similarly, for the Consumer, Manufacturing, and High Tech industries, the standard deviation of factor loadings is smaller than the standard deviation of factor loadings in the aggregate for four of the six brokerages. For the "Other" industry, there is typically higher heterogeneity in factor loadings than what we see in the aggregate, indicating that this industry group includes many heterogeneous firms.

3.4.3 Test Results

We next use our new cross-sectional tests of equal predictive accuracy to compare the EPS forecasts. With six brokerage firms, we can conduct a total of 15 pair-wise comparisons. To focus the discussion, we concentrate on four pairs, namely Morgan Stanley vs. Goldman, Morgan

Stanley vs. Merrill, Goldman vs. Merrill, and Lawrence (Deutsche Bank) vs. Merrill.[11]

Figure 3.1 plots time-series of the quarterly values of the cross-sectional average test statistics for the null of equal predictive accuracy. We show separate lines for the test statistics assuming homogeneous factor loadings, (3.7), used to test the unconditional null in (3.4), and heterogeneous factor loadings, (3.10), used to test the conditional null in (3.5). In each panel, positive values of the test statistic indicate that the second forecaster is more accurate than the first forecaster, while negative values suggest the reverse.

The first point to note is that the two sets of test statistics in (3.7) and (3.10) are very similar even though they test different hypotheses and deal with factor-related shocks in different ways. This similarity arises because the tests only differ with respect to the centering of the terms in the denominator which turns out to be of little importance.

Next, consider the pairwise comparisons starting with Morgan Stanley vs. Goldman (top left panel). In most quarters during our sample, the test statistic is not statistically significant, the three exceptions being 2004Q4, 2012Q1 and 2020Q1 where Goldman's forecasts are significantly more accurate than Morgan Stanley's. Comparing Morgan Stanley vs. Merrill (top right corner), Merrill comes out on top in two quarters (2001Q3, 2018Q4). The pairwise comparison of Goldman vs. Merrill (bottom left) only shows one quarter (2004Q3) with significant underperformance for Merrill relative to Goldman, while Lawrence produces significantly more accurate earnings forecasts than Merrill (bottom right) in five quarters (2005Q4, 2011Q1, 2013Q4, 2014Q2 and 2017Q3) and only underperforms significantly during a single quarter (2007Q3).

An important point to bear in mind when interpreting these results is that we are inspecting multiple test statistics–81 in this case–which introduces a multiple hypothesis testing problem. While we do not deal with this issue here, Qu, Timmermann and Zhu (2019) develop a Sup-type bootstrap approach that evaluates the joint statistical significance of individual test statistics.

We conclude the following from these results. First, the empirical results are very robust

[11] For each of the pairwise comparisons of firm-level EPS forecasts, our analysis imposes a requirement of at least five observations.

to whether we assume homogeneous or heterogeneous factor loadings and test the null of equal cross-sectional average predictive accuracy unconditionally or conditional on the factors and factor loadings. Second, our results suggest that the brokerage firms produce short-term earnings forecasts that are equally accurate during the vast majority of quarters but also indicate that there are significant differences in predictive accuracy in a few periods.

3.4.4 Decomposition Results

Figure 3.2 presents a set of heat diagrams displaying the quarterly values of the cross-sectional tests statistics used to test the null of equal idiosyncratic variances (3.23) for pairs of brokerage firms. Each panel corresponds to a particular pair-wise comparison, using the four pairs from Figure 3.1. Red colors indicate quarters in which the first forecaster has a larger idiosyncratic error variance component than the second forecaster, while blue colors indicate the reverse. Asterisks mark quarters in which the test statistic is significant at the 5% level, using a two-sided test. Each diagram contains three rows showing results based on the PCA, CCE, and cluster approaches, respectively.

First consider the comparison of Morgan Stanley vs. Goldman (top panel). The test statistics fluctuate around zero in most quarters without being statistically significant. Using the PCA-based test we see find two quarters in which Morgan Stanley's idiosyncratic error variance was significantly higher than that of Goldman while for the CCE and cluster tests this holds in zero and one quarter, respectively. Given that we are considering 81 quarterly test statistics, this number of rejections is lower than what we would expect by random chance and so does not provide strong evidence that idiosyncratic error variances differ in any significant way across the two brokerage firms. Similar results hold for the Morgan Stanley vs. Merrill Lynch and Goldman vs. Merrill Lynch comparisons. The comparison of the idiosyncratic error variances of Lawrence versus Merrill Lynch (bottom panel) leads to more rejections of the null of equal accuracy–four in total–when based on the PCA method, with one and four rejections using the CCE and cluster

methods, respectively.

In total, across the four pair-wise comparisons in Figure 3.2, the PCA test produces eleven rejections of the null, while the CCE and cluster-based methods record five and nine rejections. These findings suggest that there is little overall evidence of systematic differences in the magnitude of the idiosyncratic error variance component of analysts' EPS estimates.

Figure 3.3 shows the outcome of cross-sectional comparisons of the squared bias component in the errors of the four pairs of brokerage firms' EPS forecasts using the test statistics in (3.21), (3.28) and (3.31). Starting with the Morgan Stanley vs. Goldman comparison, we find nine rejections of the null of equal squared biases based on the PCA test, six rejections based on the CCE test, and a single rejection based on the cluster test. Rejection rates are lower for the three other pairwise comparisons, with five to six rejections for the PCA-based test, two to seven rejections for the CCE-based test, and zero or one rejections for the cluster test.

Overall, across the four pair-wise comparisons, we find 26 rejections of the null based on the PCA test, 22 rejections based on the CCE test, and only two rejections based on the cluster test. Hence, for the PCA and CCE-based tests, the rejection rate is somewhat higher than what we would expect by random chance from applying a test with a 5% size to 324 cross-sectional comparisons (16 rejections), while clearly this is not the case for the cluster-based test. Many of the rejections based on the PCA and CCE tests occur during the Global Financial crisis (2008-09). During this period, factor volatility is likely to have been higher than normal and so this could have boosted the power of the test for equal squared biases.

There are good theoretical reasons why the PCA approach appears to have better power in our empirical application. First, the cluster-based method is likely to be conservative since its asymptotic size is not exact; in fact, in Theorems 3.3.1 and 3.3.2, the asymptotic size is only bounded by the nominal size, rather than being equal to it. Second, the asymptotic variance of the estimates for the difference in squared biases and idiosyncratic variances can be smaller under PCA than under CCE because, as we noted after Lemma 3.3.2, the PCA estimator tends to be

more efficient than the CCE, thus increasing its relative power.

The empirical results displayed in Figures 3.2 and 3.3 show notable differences across the tests of equal idiosyncratic variance versus equal squared biases. The Monte Carlo simulations reported in the next section suggest that the test for equal idiosyncratic error variances can be quite conservative which is consistent with the smaller number of rejections of the null for this test compared to the test for equal squared error bias.

The test statistics plotted in Figures 3.1-3.3 provide evidence on the statistical significance of differences in squared error loss, idiosyncratic variances, and squared biases. They do not show how much of the variation in differences in squared error loss is explained by differences in the idiosyncratic variance and squared bias components. To address this point, Table 3.5 reports the mean and variance of the contributions from these components, both measured relative to the total loss differential. Specifically, defining the cross-sectional sample moments

$$\overline{bias^2_{t+h}} = n^{-1}\sum_{i=1}^{n}\left[(\hat{\lambda}'_{i,1}\hat{f}_{t+h})^2 - (\hat{\lambda}'_{i,2}\hat{f}_{t+h})^2\right],$$
$$\overline{\triangle u^2_{t+h}} = n^{-1}\sum_{i=1}^{n}\left(\hat{u}^2_{i,t+h,1} - \hat{u}^2_{i,t+h,2}\right),$$

the columns labeled mean ratio in Table 3.5 report the time-series averages

$$\frac{100}{81}\left(\frac{\sum_{t=2000Q1}^{2020Q1}\overline{\triangle u^2_{t+h}}}{\sum_{t=2000Q1}^{2020Q1}\overline{\triangle L}_{t+h}}\right), \qquad \frac{100}{81}\sum_{t=2000Q1}^{2020Q1}\left(\frac{\sum_{t=2000Q1}^{2020Q1}\overline{bias^2_{t+h}}}{\sum_{t=2000Q1}^{2020Q1}\overline{\triangle L}_{t+h}}\right),$$

for the idiosyncratic variance (top panel) and squared bias (bottom panel) components, respectively. As before, $\overline{\triangle L}_{t+h} = n^{-1}\sum_{i=1}^{n}\Delta L_{i,t+h|t}$. From (3.11), these measures can be positive or negative but sum to 100.

Similarly, columns labeled variance ratio report the following

$$\frac{Var(\overline{\triangle u^2_{t+h}})}{Var(\overline{\triangle L}_{t+h})}, \qquad \frac{Var(\overline{bias^2_{t+h}})}{Var(\overline{\triangle L}_{t+h})},$$

where

$$Var(\overline{\triangle L}_{t+h}) = \frac{1}{80}\sum_{t=2000Q1}^{2020Q1}\left(\overline{\triangle L}_{t+h} - \overline{\overline{\triangle L}}_{t+h}\right)^2,$$
$$Var(\overline{\triangle u^2_{t+h}}) = \frac{1}{80}\sum_{t=2000Q1}^{2020Q1}\left(\overline{\triangle u^2_{t+h}} - \overline{\overline{\triangle u^2_{t+h}}}\right)^2,$$
$$Var(\overline{bias^2_{t+h}}) = \frac{1}{80}\sum_{t=2000Q1}^{2020Q1}\left(\overline{bias^2_{t+h}} - \overline{\overline{\triangle bias^2_{t+h}}}\right)^2.$$

and $\overline{\overline{\triangle L}}_{t+h} = (1/81)\sum_{t=2000Q1}^{2020Q1}\overline{\triangle L}_{t+h}$, $\overline{\overline{\triangle u^2_{t+h}}} = (1/81)\sum_{t=2000Q1}^{2020Q1}\overline{\triangle u^2_{t+h}}$, and $\overline{\overline{\triangle bias^2_{t+h}}} = (1/81)\sum_{t=2000Q1}^{2020Q1}\overline{\triangle bias^2_{t+h}}$. These variance ratios do not sum to 100 because of the omitted covariance term.

The mean ratios of the four pairwise comparisons reported in Table 3.5 are generally notably higher for differences in the idiosyncratic variances than for differences in squared biases, with the former falling within ranges of 40-87%, 1-95%, and 85-96% for the PCA, CCE, and cluster methods, respectively. Variance ratios are also higher–typically by a large margin–for differences in the idiosyncratic variance component than for differences in the squared biases for all but one pairwise comparison (Morgan Stanley vs. Goldman, CCE method). Variation in the idiosyncratic variance component is thus generally substantially more important to explaining squared error loss differences between brokerage firms' earnings forecasts than variation in the squared bias term.

These results show that, on average, differences in idiosyncratic error variances account for far more of squared error loss differences in brokerage firms' EPS forecasts than the squared bias component. Differences in brokerage firms' quarterly EPS forecast accuracy therefore appear not so much to be driven by differences in their ability to predict common factors, i.e., their skills as "generalists". Rather, differences in predictive accuracy tend to be driven by differences in brokerage firms' ability to reduce uncertainty about the idiosyncratic component of EPS as this relates to their specialist knowledge of individual firm performance. The main exception to this

finding occurs around the Global Financial Crisis (2008-09) during which the squared bias term becomes more important in explaining differences in squared-error losses across brokerages, particularly for the PCA-based test.

3.5 Monte Carlo Simulations

Our final section reports the outcome of a set of Monte Carlo simulations which address the finite-sample properties of our tests.

3.5.1 Setup

Our baseline simulations use a simple setup designed to satisfy the assumptions of the three different estimation procedures (clustering, CCE and PCA) which allows us to more directly compare their performance. First, we generate factors $f_{1,t}$ and $f_{2,t}$ as i.i.d variables from the standard normal distribution. Next, we compute realized outcomes as $y_{it+h} = f_{1,t} + f_{2,t} + \varepsilon_{it+h}$, while forecasts are generated as $\hat{y}_{it+h|t,1} = f_{1,t} + \xi_{it+h,1}$ and $\hat{y}_{it+h|t,2} = f_{2,t} + \xi_{it+h,2}$, where ε_{it+h}, $\xi_{it+h,1}$ and $\xi_{it+h,2}$ are mutually independent i.i.d. $N(0,\sigma^2)$ draws. We calibrate σ^2 to yield a value for the predictive power ρ^2 in a certain range, where for $m \in \{1,2\}$,

$$\rho^2 = 1 - \frac{E(y_{it+h} - \hat{y}_{it+h|t,m})^2}{Ey_{it+h}^2}.$$

Because $\rho^2 = 1/(2+\sigma^2)$, $\rho^2 \in (0,1/2)$. We set $n \in \{10,25,50,100,200,1000\}$ and $\rho^2 \in \{0.05,0.1,0.2,0.25,0.3,0.4,0.45\}$ as well as $T = 80$. All results are based on 2000 random samples.

Initially we consider the performance in a typical time period ($t = 3$) and note that results for other time periods would be similar, given the i.i.d. setting. Section 3.5.4 introduces breaks to the data generating process and so considers both pre- and post-break performance.

Because we are testing a random hypothesis, the hypothesized value is not zero but a random quantity that depends on the realization of the factors and factor loadings. For this reason, and to simplify the presentation of size and power results, we invert our test statistics to form 95% confidence intervals for $E(\overline{\Delta L}_{t+h} \mid \mathcal{F})$ and report the coverage probabilities.

3.5.2 Baseline results

Table 3.6 reports results on the procedure for conducting inference on $E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = 0$ described in Section 3.2.5. Coverage probabilities are generally quite accurate although there is some undercoverage for very small values of n, suggesting that the test might slightly overreject in such cases.

Next, we invert the procedures described in Section 3.3 to construct 95% confidence intervals for the squared error loss decompositions based on the clustering, CCE, and PCA methods. Table 3.7 reports results for the average difference in the squared bias component $n^{-1}\sum_{i=1}^{n}\left[(\lambda_{i,1}'f_{t+h})^2-(\lambda_{i,2}'f_{t+h})^2\right]$ while results for the average difference in the idiosyncratic variance component $n^{-1}\sum E(u_{i,t+h,1}^2-u_{i,t+h,2}^2 \mid \mathcal{F})$ are reported in Table 3.8.

For the tests applied to the squared bias terms (Table 3.7), the coverage probability generally improves with the sample size n, with exception of the PCA method when ρ^2 is very small.[12] In larger samples, the coverage probability for the difference in squared bias is relatively stable as a function of ρ^2, while for smaller values of n, the tests are under-sized for small values of ρ^2 and oversized for large values of ρ^2. Conversely, the confidence intervals for the difference in variances (Table 3.8) tend to be more conservative when ρ^2 is large, with coverage probabilities exceeding 99%. To understand this finding, note that $\rho^2 = 1/(2+\sigma^2)$ and inference on the difference in variance relies on variation in $u_{it+h|m}$ which is $2\sigma^2$ in this case. Large values of ρ^2 are therefore associated with smaller variation in $u_{it+h|m}$ and so the higher-order terms in the

[12] A reason for this finding is that for $n = 1000$ and $T = 80$, the two dimensions of the sample size are not very balanced and the accuracy of the PCA method is determined by $\min\{n, T\}$.

asymptotic expansion tend to be more pronounced which means that the first-order asymptotic approximation underlying the inference procedure is generally less accurate.

Overall, the procedure for testing differences in idiosyncratic variances has better coverage than its counterpart for testing differences in the squared bias component. This might be explained by the greater robustness of the test for equal idiosyncratic variances highlighted earlier. Moreover, the size distortion results suggest that tests for equal squared biases are likely to have more power than tests for equal idiosyncratic error variances.

3.5.3 Decompositions with heterogeneous factor loadings

We next consider various extensions to the baseline simulation setup. To keep the presentation short, all results are reported in a set of appendix tables.

Heterogeneous factor loadings across Clusters

Our first extension allows factor loadings to have a cluster structure. Specifically, we partition the cross-section of n units into five equal-sized clusters and set $\hat{y}_{it+h|t,1} = f_{1,t}\lambda_{k(i)} + \xi_{it+h,1}$, where $\lambda_{k(i)} \in \{0, 0.5, 1, 1.5, 2\}$ and $k(i)$ is the cluster that contains unit i. Similarly, we set $\hat{y}_{it+h|t,2} = f_{2,t}\lambda_{k(i)} + \xi_{it+h,2}$ with $\lambda_{k(i)} \in \{0, 0.5, 1, 1.5, 2\}$. Since each cluster contains $n/5$ units, the clusters are very small for the smallest values of n, i.e., only two and five units per cluster for $n = 10$ and $n = 25$, respectively.

Results from this setup are reported in Appendix Tables C.1 and C.2. For inference on differences in the squared bias (Table C.1), the clustering method has a substantial undercoverage for small values of n but performs notably better with larger sample sizes. This is as expected since the clustering method uses the cluster-wise average to estimate the factor structure and the size of each cluster is $n/5$. The CCE method mostly has sufficient coverage probability while the PCA method tends to be very conservative with coverage probabilities at or above 99%. For inference on differences in the error variance (Table C.2), all three methods perform

reasonably well across various sample sizes, although the clustering and CCE approaches tend to be somewhat conservative while, conversely, the PCA method overrejects if n is very small ($n = 10$).

General heterogeneous factor loadings

Our second extension applies a more general setting in which factor loadings are neither constant, nor have a cluster structure as we generate factor loadings as the absolute value of a standard normal distribution, i.e., $\lambda_{1,i}$ and $\lambda_{2,i}$ are i.i.d $|N(0,1)|$. Using absolute values ensures that $E(\lambda_{1,i}) = E(\lambda_{2,i})$ is positive as required by the CCE method (Assumption 5). Conversely, the clustering method is no longer valid in this setting and so we omit results for this method. Using these heterogeneous factor loadings, we set $y_{it+h} = \lambda_{1,i} f_{1,t} + \lambda_{2,i} f_{2,t} + \varepsilon_{it+h}$ and generate forecasts as $\hat{y}_{it+h|t,1} = \lambda_{1,i} f_{1,t} + \xi_{it+h,1}$ and $\hat{y}_{it+h|t,2} = \lambda_{2,i} f_{2,t} + \xi_{it+h,2}$, where ε_{it+h}, $\xi_{it+h|1}$ and $\xi_{it+h|2}$ are again drawn independently with mean zero and variance σ^2. Appendix Table C.3 shows that the coverage of the CCE method is often better than that of the PCA method for inference on differences in the squared bias with the latter having issues with undercoverage for small values of n; both methods provide sufficient overall coverage but tend to be conservative for inference on differences in variances, particularly when ρ^2 is large.

3.5.4 Variation in the factor structure

Three factors

The key reason for using the PCA method is that once the number of factors exceeds two, PCA is the only valid method for handling the general case with heterogeneous factor loadings.[13] We illustrate this point in a setting with three factors as we set $y_{it+h} = \lambda_{1,i} f_{1,t} + \lambda_{2,i} f_{2,t} + \lambda_{3,i} f_{3,t} + \varepsilon_{it+h}$ and generate the forecasts as $\hat{y}_{it+h|t,1} = \lambda_{1,i} f_{1,t} + \xi_{it+h,1}$ and $\hat{y}_{it+h|t,2} = \lambda_{2,i} f_{2,t} + \xi_{it+h,2}$,

[13] Another reason for using the PCA method is that it remains asymptotically valid even if $E(\lambda_i) = 0$, whereas the CCE method would fail in this setting.

where all variables (including all factors and factor loadings) are generated as before.

As shown in Appendix Table C.4, the 95% coverage probability for the CCE method can be as low as 40% for comparing the squared bias and as low as 59% for comparing variances when $n = 1000$ and $\rho^2 = 0.45$. This phenomenon arises because we only observe two variables (two forecast errors) and the CCE method can handle at most two factors in our setup. With more than two factors, the CCE method does not guarantee consistent estimation of the factor structure. Since we are studying the average across n units, the problem becomes more pronounced as n increases.

Breaks in the number of factors

We next consider a setting in which the number of factors changes as represented by a discrete break to the factor structure: $y_{it+h} = \lambda_{1,i} f_{1,t} + \lambda_{2,i} f_{2,t} + \lambda_{3,i} f_{3,t} \mathbf{1}_{\{t>T/2\}} + \varepsilon_{it+h}$. In this model, the third factor ($f_{3,t}$) only shows up in the second half of the sample. All other details remain the same. Instability in the number of factors is empirically plausible and has been studied in Cheng, Liao and Schorfheide (2016).

Appendix Table C.5 reports coverage probabilities for 95% confidence intervals based on the PCA and CCE methods. We consider two time periods: one before the break ($t = 3$), the other after the break ($t = T - 3$).[14] Overall, the PCA method maintains sufficient coverage probability while the CCE method can suffer from severe undercoverage. Again, the reason is that when there are three factors, CCE cannot consistently estimate the factor structure from two observed variables. The performance of the PCA approach is similar before and after the break. Conversely, the CCE method performs worse after the break than before, most likely because there are only two factors before the break, consistent with a setting in which the CCE approach is valid.

[14] We conduct the PCA analysis for the full sample using three factors because there are three spiked eigenvalues in the data matrix for the full sample.

3.5.5 Linex Loss

Table C.6 reports 95% confidence intervals for testing the null of equal conditional expected loss (3.5) using linex loss:

$$L(e_{it+h}) = \frac{1}{a^2}\left[\exp(ae_{it+h}) - ae_{it+h} - 1\right] \tag{3.32}$$

where $a = 1$. The data generating process is identical to that in the baseline case used to construct Table 3.6. Coverage probabilities are very similar to those in Table 3.6, with a slight undercoverage for small values of n and coverage approximating 95% as n grows larger.

3.5.6 Conditional heteroskedasticity

We now conduct a set of simulations in which the data generating process allows for conditional heteroskedasticity modeled through a simple ARCH process of the form:

$$f_t = \sigma_t \varepsilon_t,$$

where $\sigma_t^2 = (1-r) + r f_{t-1}^2$ with $r = 0.5$. Notice that $E\sigma_t^2 = 1$.

Results are reported in Appendix Tables C.7 and C.8. Compared to the baseline setup in Tables 3.7 and 3.8, the results do not change in any material ways, showing that conditional heteroskedasticity in the innovations of the data generating process need not have a material effect on the performance of our cross-sectional tests for equal predictive accuracy.

3.5.7 Relation to empirical results

In our empirical analysis, the PCA and CCE methods lead to notably more rejections of the null of equal squared biases than the clustering method which rarely rejects the null. To help explain these results, we slightly modify the simulation setup so as to match the high cross-

sectional R^2 values found in our application (0.9 on average) and allow for broad heterogeneity in factor loadings. We accomplish this by adding a third factor to the model and letting factor loadings be random: $y_{it+h} = \lambda_{i1} f_{1,t} + \lambda_{i2} f_{2,t} + \lambda_{i3} f_{3,t} + \varepsilon_{it+h}$, where $f_{1,t}, f_{2,t}, f_{3,t} \sim iidN(0,1)$, $\lambda_{i1} \sim iidN(0,V)$, $\lambda_{i2}, \lambda_{i3} \sim iidN(1,1)$, and $\varepsilon_{i,t+h} \sim iidN(0,\sigma^2)$. The two forecasts are generated as $\hat{y}_{it+h|t,1} = \lambda_{i1} f_{1,t} + \lambda_{i2} f_{2,t}$ and $\hat{y}_{it+h|t,2} = \lambda_{i1} f_{1,t} + \lambda_{i3} f_{3,t}$, respectively, with forecast errors $e_{it+h,1} = \lambda_{i3} f_{3,t} + \varepsilon_{it+h}$ and $e_{it+h,2} = \lambda_{i2} f_{2,t} + \varepsilon_{it+h}$.

Normally distributed factor loadings is likely to cause the biggest problem for the clustering method which approximates heterogeneity in factor loadings by means of a small number of discrete values. Such heterogeneity is particularly important if the fraction of the variation in forecast errors explained by the omitted factors is large. Here, this is given by $\rho_e^2 = 2/(2+\sigma^2)$ and we vary σ to obtain a range of values $\rho_e^2 \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. To mimic the cross-sectional R^2, note that for each t, $1 - E[(y_{it+h} - \hat{y}_{it+h|t,1})^2 | f_{1,t}, f_{2,t}, f_{3,t}] / E[y_{it+h}^2 | f_{1,t}, f_{2,t}, f_{3,t}] = (V f_{1,t}^2 + 2 f_{2,t}^2) / (V f_{1,t}^2 + 2 f_{2,t}^2 + 2 f_{3,t}^2 + \sigma^2)$. Matching the empirical evidence, we set this number to 0.9 for $f_{1,t}^2 = f_{2,t}^2 = f_{3,t}^3 = 1$ by choosing an appropriate value of V.

Simulation results that use this setup are reported in Appendix Table C.9. When the omitted factors matter less for the variation in forecast errors (small ρ_e^2), the PCA and clustering methods generate tighter confidence intervals than the CCE approach. However, as the common factors gain in importance (high ρ_e^2), the PCA approach produces notably narrower confidence intervals than the clustering method, with the CCE approach in the middle. This is consistent with the clustering approach having weaker power and so helps explain the far lower rejection rate observed empirically for this estimator for the equality of squared bias tests in situations with substantial cross-sectional heterogeneity in factor loadings.

3.6 Conclusion

This paper develops new methods for testing the null of equal predictive accuracy on a single cross-section containing pairs of forecasts of multiple outcome variables. In settings where the cross-sectional dependence in forecast errors can be captured by a common factor structure, we show that it is possible to conduct formal inference about equal predictive accuracy and develop a set of test statistics. In particular, we show that the null of equal predictive accuracy can be conducted in settings with a large cross-sectional dimension if either (i) factor loadings are homogeneous across units so that the effect of common factors on forecast errors cancels out in squared error loss differentials; or (ii) we condition on factor realizations and conduct a test of equal predictive accuracy, given these factors.

We illustrate our tests in an empirical application that compares the accuracy of analyst short-term earnings forecasts across six brokerage firms, using a sample covering hundreds of individual firms. While our cross-sectional tests fail to reject the null of equal predictive accuracy for most quarters, we do identify individual quarters with significant differences among pairs of brokers. Moreover, our empirical results suggest that differences in the variances of the idiosyncratic error component tend to be more important than differences in squared biases for explaining variation in differences in brokerage firms' earnings per share squared-error loss performance.

3.7 Acknowledgements

Chapter 3 is currently being prepared for submission for publication and is couathored with Allan Timmermann and Yinchu Zhu. Ritong Qu, the dissertation author, is the primary investigator and author of this material.

Table 3.1: Firm coverage by forecaster

Panel A: Total number of firms covered

	Total	Consumer	Manufacturing	High tech	Health	Other
MERRILL	1825	233	382	409	159	642
JPMORGAN	1796	210	341	455	166	624
FBOSTON	1752	211	375	417	121	628
GOLDMAN	1602	193	358	387	123	541
MORGAN	1473	170	305	352	117	529
LAWRENCE	1437	151	297	372	113	504

Panel B: Average number of firms covered

	Total	Consumer	Manufacturing	High tech	Health	Other
MERRILL	356	45	81	76	29	125
JPMORGAN	311	38	79	72	27	96
FBOSTON	277	34	70	60	17	96
GOLDMAN	283	37	72	64	21	88
MORGAN	243	29	55	53	19	88
LAWRENCE	239	27	57	56	16	82

Note: Panel A reports the number of different firms whose quarterly earnings per share is predicted by each brokerage firm for at least one quarter during our sample. Panel B reports the average number of quarterly earnings per share forecasts generated by each brokerage firm both in the aggregate (first column) and across five industries (columns 2-6).

Table 3.2: Estimated number of common factors in the earnings forecast errors

	GR	*ER*	*ED*
FBOSTON	1	1	1
JPMORGAN	0	0	0
MORGAN	2	2	2
GOLDMAN	1	1	1
LAWRENCE	1	1	3
MERRILL	1	1	1

Note: This table presents estimates of the number of common factors in the earnings forecast errors using the methods in Ahn and Horenstein (2013) and Onatski (2010). Columns labeled "GR" and "ER" report the "'Growth Ratio" and "Eigenvalue Ratio" statistics proposed by Ahn and Horenstein (2013), while the column labeled "ED" reports the Onatski (2010) statistic.

Table 3.3: Correlations across earnings forecast errors

Average correlations in forecast errors						
No. factors	FBOSTON	JPMORGAN	MORGAN	GOLDMAN	LAWRENCE	MERRILL
0	0.09	0.08	0.07	0.08	0.12	0.08
	(40.22)	(43.50)	(25.12)	(37.41)	(39.02)	(57.97)
1	-0.01	0.00	0.03	0.00	-0.01	0.00
	(-2.62)	(0.34)	(9.95)	(1.08)	(-2.24)	(2.82)
2	-0.01	-0.00	0.01	0.01	-0.01	-0.00
	(-2.39)	(-1.87)	(5.07)	(4.41)	(-2.68)	(-0.71)

Note: This table reports estimates of the average pair-wise correlation in earnings forecast errors along with the test statistic for non-zero average correlations proposed by Pesaran (2004) in brackets underneath. Results are presented using raw forecast errors (row labeled "0") as well as residuals from a regression that accounts for one and two common factors in the residuals (rows labeled "1" and "2").

Table 3.4: Heterogeneity in factor loadings within and across industries

	Aggregate	consumer	manufacturing	high tech	health	other
FBOSTON	0.067	0.066	0.075	0.049	0.043	0.070
JPMORGAN	0.062	0.049	0.058	0.052	0.067	0.070
MORGAN	0.096	0.048	0.101	0.049	0.064	0.123
GOLDMAN	0.074	0.086	0.060	0.072	0.048	0.080
LAWRENCE	0.069	0.075	0.061	0.077	0.084	0.065
MERRILL	0.057	0.048	0.051	0.067	0.063	0.051

Note: This table reports the standard deviation of the estimated factor loadings for the earnings forecast errors across all firms (column 1) as well as for different industries (columns 2-6). For each set of forecast errors, we estimate a model with a single common factor on the normalized forecast errors, demeaned and scaled to have a unit sample variance.

$$\tilde{e}_{i,t} = \lambda_{i,1} f_{1,t} + \varepsilon_{i,t},$$

sucject to the constraint: $\sum_i^N \lambda_{i,1}^2 = 1$. The table reports the standard deviation of the factor loadings $\lambda_{i,1}$ within each group of firms.

Table 3.5: Contributions of idiosyncratic error variance and squared bias components

Difference in idiosyncratic variance (%)

	MORGAN vs. GOLDMAN		MORGAN vs. MERRILL		GOLDMAN vs. MERRILL		LAWRENCE vs. MERRILL	
	Mean Ratio	Variance Ratio	Mean Ratio	Variance Ratio	Mean Ratio	Variance Ratio	Mean Ratio	Variance Ratio
PCA	39.85	86.97	55.23	61.75	64.10	51.30	74.82	98.34
CCE	1.31	31.91	51.57	63.01	94.79	105.98	43.34	76.81
Cluster	96.46	93.58	96.35	92.91	93.40	90.74	85.46	89.45

Difference in squared bias (%)

	MORGAN vs. GOLDMAN		MORGAN vs. MERRILL		GOLDMAN vs. MERRILL		LAWRENCE vs. MERRILL	
	Mean Ratio	Variance Ratio	Mean Ratio	Variance Ratio	Mean Ratio	Variance Ratio	Mean Ratio	Variance Ratio
PCA	60.15	7.70	44.77	16.41	35.90	31.67	25.18	6.07
CCE	98.69	64.37	48.43	15.71	5.21	6.94	56.66	7.74
Cluster	3.54	0.35	3.65	0.64	6.60	0.44	14.20	0.42

Note: Columns labeled mean ratio report the sample average of the ratio of the mean contribution to the total loss difference that comes from differences in idiosyncratic variances (top panel) or differences in squared biases (bottom panel) for a given pair of brokerage firms. Columns labeled variance ratio report the ratio of the sample variance of the squared idiosyncratic error differences to the sample variance of the total loss difference (top panel) or the ratio of the variance of the squared bias difference to the variance of the total loss difference (bottom panel), averaged across all quarters in the sample.

Table 3.6: Coverage probabilities for 95% confidence intervals constructed to test the null of equal conditional squared error loss

	Coverage probability						
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	91.6	91.6	91.1	90.8	90.9	90.4	89.8
25	93.8	94.0	93.4	93.4	93.3	93.3	92.9
50	94.5	94.3	94.3	94.1	94.0	94.4	94.2
100	94.5	94.7	94.5	94.5	94.6	94.4	94.4
200	94.8	94.9	94.8	94.8	94.7	94.7	94.8
1000	94.9	95.2	95.1	95.0	95.1	94.8	95.2

Note: This table reports the coverage probability for a 95% confidence interval for the test of equal conditional squared error loss, $E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = 0$ using the Monte Carlo simulation setup described in Section 5.1 and 2,000 random samples. n refers to the number of cross-sectional units used in the pair-wise comparison of loss differences, while ρ^2 measures the predictive power of the underlying forecasts.

Table 3.7: Coverage probabilities for 95% confidence intervals constructed to test the null of equal squared biases

Clustering							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	98.0	96.6	93.8	92.7	92.5	90.8	91.8
25	97.3	96.6	94.4	94.1	93.1	93.2	93.4
50	96.3	95.7	95.6	93.9	95.2	94.1	93.7
100	95.9	95.3	94.9	94.9	94.9	94.1	94.3
200	95.9	94.4	93.8	95.4	95.5	95.3	95.0
1000	95.4	94.8	94.8	95.4	95.1	95.5	95.1

CCE							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	97.7	96.2	93.6	92.1	92.5	90.7	91.4
25	97.1	96.1	94.2	93.9	92.9	93.0	93.2
50	96.3	95.4	95.3	93.6	95.0	93.6	93.1
100	95.7	95.3	94.6	94.8	94.4	93.9	93.8
200	95.3	93.9	93.6	94.9	95.2	95.1	94.5
1000	94.7	94.2	94.7	95.1	94.7	95.2	94.8

PCA							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	99.1	96.9	92.2	91.4	91.9	90.4	91.0
25	97.1	92.6	92.4	93.1	92.5	92.9	93.2
50	95.0	91.7	94.2	92.0	94.6	93.3	93.2
100	90.6	92.7	93.3	94.1	93.6	93.4	93.7
200	90.2	92.5	93.6	93.4	94.8	95.2	94.9
1000	91.3	92.6	93.6	94.3	93.8	94.6	94.9

Note: This table reports the coverage probability for a 95% confidence interval for the test of equal squared biases, using the Monte Carlo simulation setup described in Section 5.1 and 2,000 random samples. n refers to the number of cross-sectional units used in the pair-wise comparison of loss differences, while ρ^2 measures the predictive power of the underlying forecasts. We show coverage probabilities for the clustering, CCE, and PCA methods described in Section 3. The assumed time-series dimension is $T = 80$.

Table 3.8: Coverage probabilities for 95% confidence intervals constructed to test the null of equal idiosyncratic error variances

Coverage probability (clustering)							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	93.7	94.8	95.4	96.4	95.6	98.0	99.1
25	94.8	94.6	97.4	97.3	97.8	98.5	99.4
50	95.3	96.3	97.0	97.3	98.1	99.0	99.6
100	95.3	96.6	97.3	97.8	98.6	99.2	99.8
200	95.8	96.6	97.7	98.6	98.2	99.0	99.4
1000	95.7	96.4	98.0	97.8	98.3	99.5	99.7

Coverage probability (CCE)							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	95.3	96.5	97.0	98.0	97.1	99.2	99.5
25	96.1	96.4	98.4	98.5	98.5	99.3	99.7
50	96.4	97.3	97.9	98.3	99.0	99.6	99.8
100	96.5	97.1	98.4	98.7	99.1	99.7	100.0
200	96.4	97.6	98.5	99.1	98.9	99.4	99.7
1000	96.4	97.1	98.7	98.6	99.1	99.8	99.9

Coverage probability (PCA)							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	91.3	91.0	93.1	94.1	95.4	98.0	99.1
25	94.1	92.8	96.1	96.8	97.4	98.6	99.3
50	94.6	94.5	96.5	97.0	97.9	99.0	99.6
100	93.3	95.5	96.3	97.5	98.4	99.1	99.7
200	95.0	95.8	96.8	98.4	98.0	99.1	99.4
1000	94.8	95.3	97.6	97.1	98.0	99.4	99.7

Note: This table reports the coverage probability for a 95% confidence interval for the test of equal idiosyncratic variances, using the Monte Carlo simulation setup described in Section 5.1 and 2,000 random samples. n refers to the number of cross-sectional units used in the pair-wise comparison of loss differences, while ρ^2 measures the predictive power of the underlying forecasts. We show coverage probabilities for the clustering, CCE, and PCA methods described in Section 3. The assumed time-series dimension is $T = 80$.

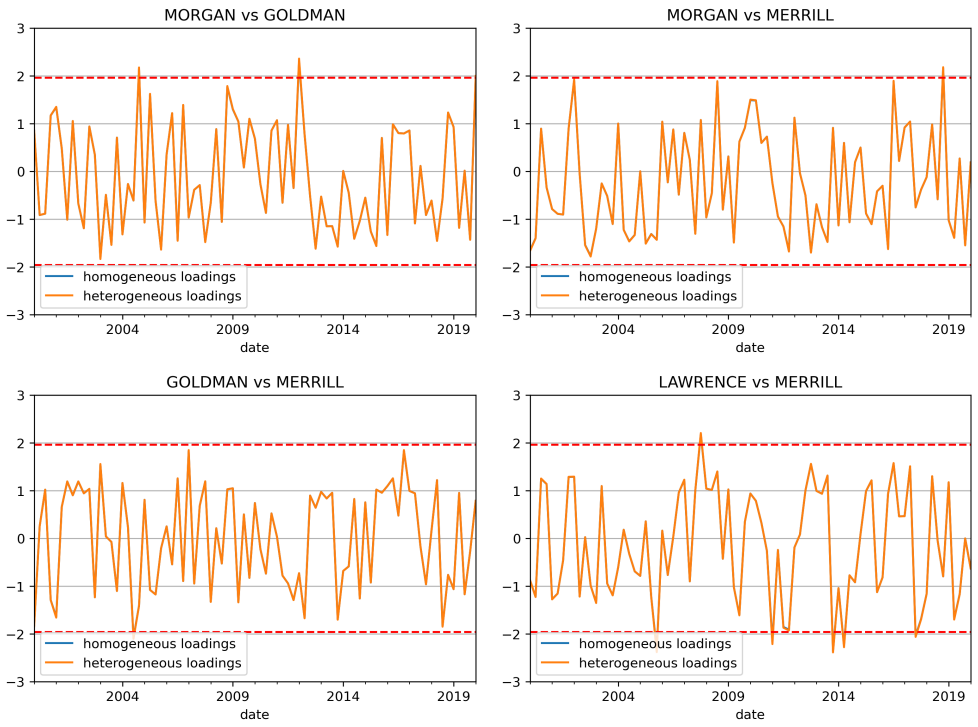


Figure 3.1: Cross-sectional test statistics for comparisons of the null of equal squared error loss conducted on pairs of brokerage firms
Positive values of the test statistics indicate that the second forecaster is more accurate than the first forecaster, while negative values suggest the reverse.

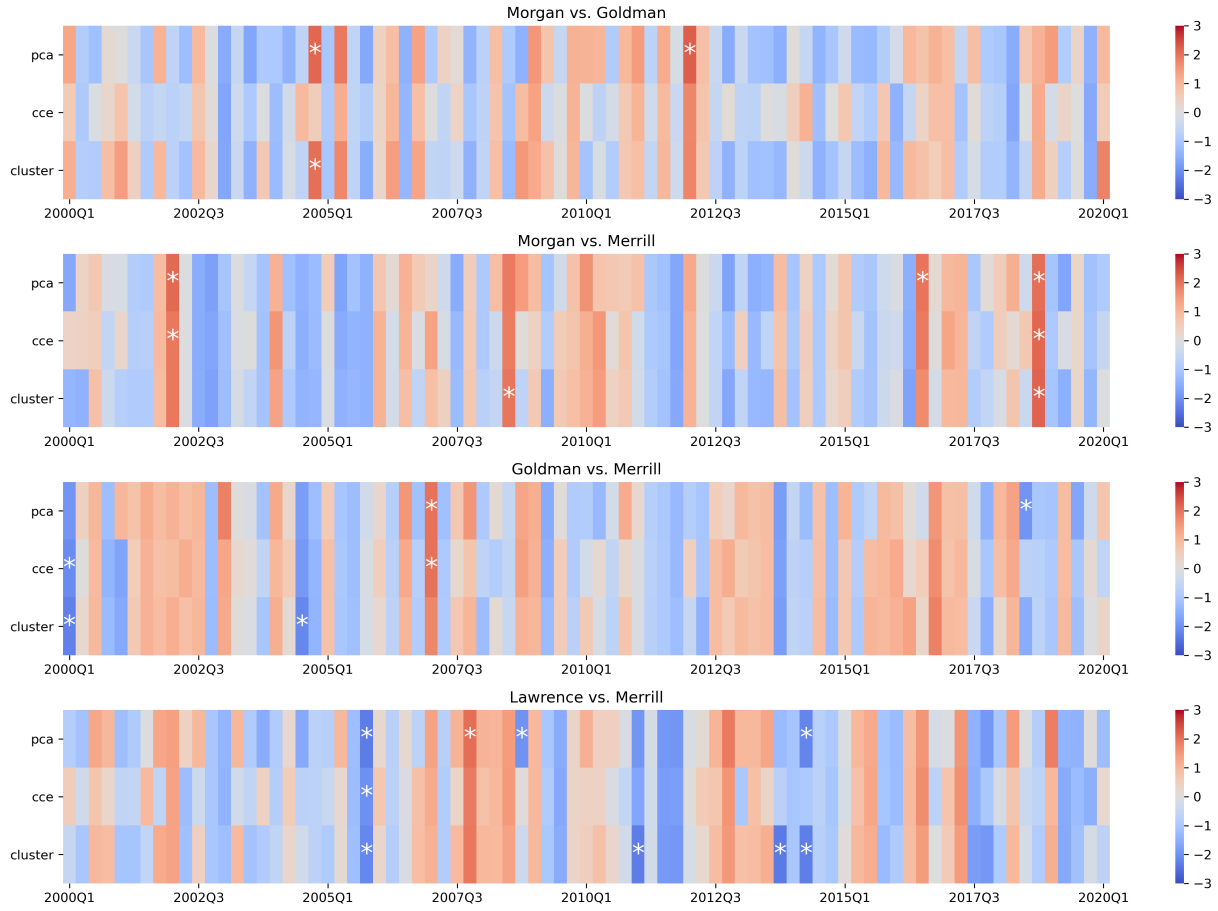


Figure 3.2: Values of the cross-sectional test of equal idiosyncratic variances conducted on individual quarters

Each panel shows the outcome of a cross-sectional test of the null that a pair of forecasters produce the same idiosyncratic error variance in a given quarter. Red color indicates that the idiosyncratic error variance component of the first forecaster is larger than that of the second forecaster. Blue color indicates the reverse. The first and second rows of each panel estimate the factors by PCA and CCE, respectively, while the third row is calculated by assuming identical factor loadings within each cluster. Asterisks represent quarters with test statistics that are statistically significant at the 5% level.

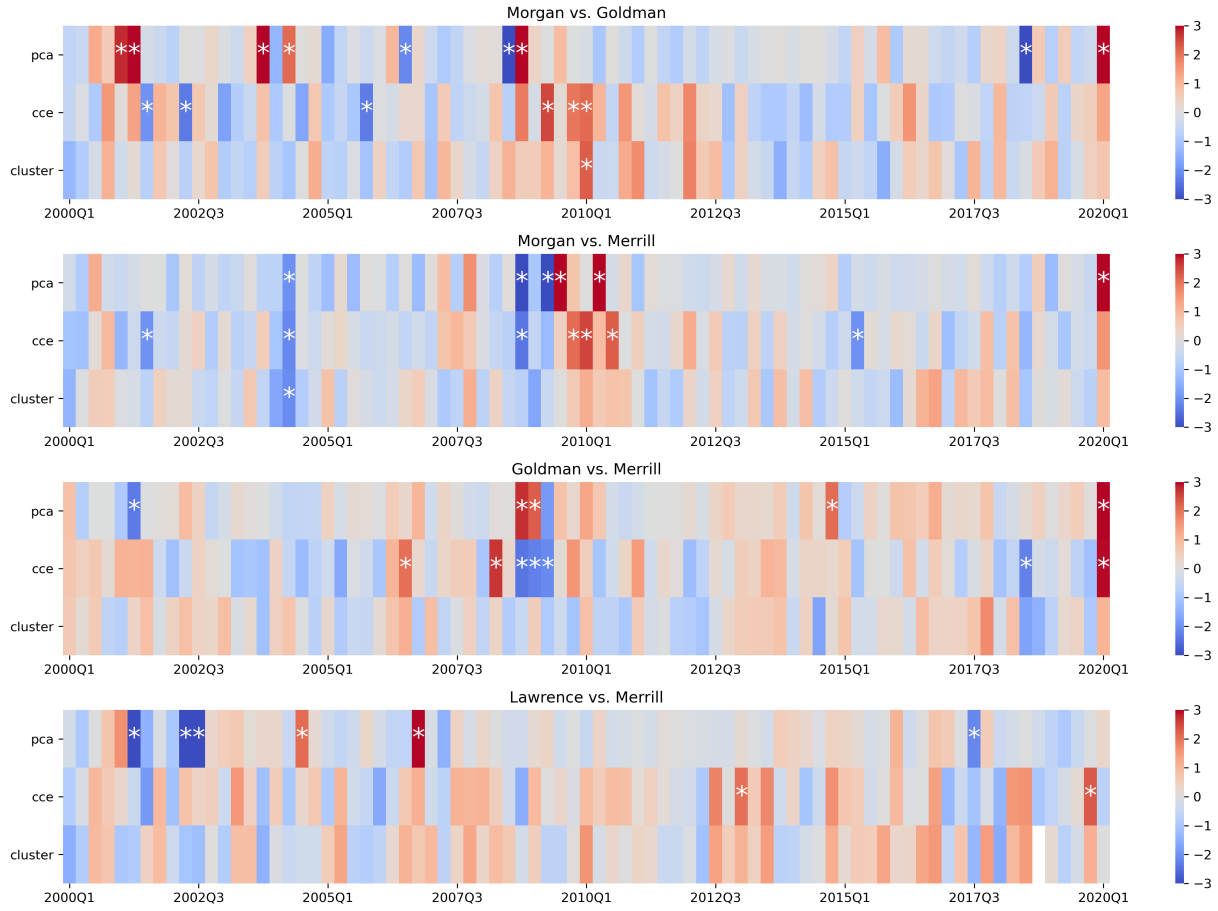


Figure 3.3: Values of the cross-sectional test of equal squared biases conducted on individual quarters

Each panel shows the outcome of a cross-sectional test of the null that a pair of forecasters produce the same squared bias in a given quarter. Red color indicates that the squared bias component of the first forecaster is larger than that of the second forecaster. Blue color indicates the reverse. The first and second rows of each panel estimate the factors by PCA and CCE, respectively, while the third row is calculated by assuming identical factor loadings within each cluster. Asterisks represent quarters with test statistics that are statistically significant at the 5% level.

Appendix A

Appendix for Chapter 1

A.1 Numerical solution of value function and prices

It is more convenient to reparameterize $v_t = v_{H(L)}(\tilde{\mu}_t, \Sigma_t)$, where $\tilde{\mu}_t = \frac{\mu_t - \sum_{i=1}^{N} \alpha_i \bar{g}_i}{\sigma_{ct}}$ and $H(L)$ in the subscripts denote high or low volatility regimes. Our strategy is to solve for v_t backwards from the solutions when Σ_t is close to zero.

A.1.1 Partial solution of v_t and φ_t when $\Sigma_t = 0$

One feature of the model is that investors observe the timing of breaks. When breaks happen, investors' belief is refreshed to $\left(0, \Sigma_0, \sigma_{High}\right)$ or $(0, \Sigma_0, \sigma_{Low})$. This section derived partial close form solution of v_t and φ_t at $\Sigma_0 = 0$ given values of $v_H(0, \Sigma_0), v_L(0, \Sigma_0), \varphi_H(0, \Sigma_0), \varphi_L(0, \Sigma_0)$.

Investors' utility satisfy (1.14), where $Q_t(v_{t+1} + \Delta c_{t+1})$ satisfies

$$Q_t(v_{t+1} + \Delta c_{t+1}) = \frac{1}{1-\gamma} \ln \left\{ \int f_t(g) \left(E_{g_t = g} \exp\left[(1-\gamma)(v_{t+1} + \Delta c_{t+1}) \right] \right) dg \right\}. \tag{A.1}$$

Hence

$$Q_H(\tilde{\mu}_t,0)=\frac{\ln\left\{(1-\lambda)\,e^{(1-\gamma)\left[v_H(\tilde{\mu}_t,0)+\tilde{\mu}_{ct}\sigma_{cH}+\sum_{i=1}^{N}\alpha_i\bar{g}_i+\frac{(1-\gamma)\sigma_{cH}^2}{2}\right]}+\lambda\left[(1-\pi_L)\,e^{(1-\gamma)v_H(0,\Sigma_0)}+\pi_L e^{(1-\gamma)v_L(0,\Sigma_0)}\right]\right\}}{1-\gamma},$$

$$Q_L(\tilde{\mu}_t,0)=\frac{\ln\left\{(1-\lambda)\,e^{(1-\gamma)\left[v_L(\tilde{\mu}_t,0)+\tilde{\mu}_{ct}\sigma_{cL}+\sum_{i=1}^{N}\alpha_i\bar{g}_i+\frac{(1-\gamma)\sigma_{cL}^2}{2}\right]}+\lambda\left[(1-\pi_L)\,e^{(1-\gamma)v_H(0,\Sigma_0)}+\pi_L e^{(1-\gamma)v_L(0,\Sigma_0)}\right]\right\}}{1-\gamma}.$$

Further, we have the solution of $v_{H(L)}$ at $\Sigma_t=0$:

$$v_H(\tilde{\mu}_t,0)=\frac{1}{1-\rho}\ln\left\{(1-\beta)+\beta\exp\left[(1-\rho)\,Q_H(\tilde{\mu}_t,0)\right]\right\},$$
$$v_L(\tilde{\mu}_t,0)=\frac{1}{1-\rho}\ln\left\{(1-\beta)+\beta\exp\left[(1-\rho)\,Q_L(\tilde{\mu}_t,0)\right]\right\}.$$

In the case of no parameter uncertainty and using equations (1.18), (1.21) and (1.23), we have

$$\varphi_t=\frac{\beta\exp\left[g_{dt}+\frac{1}{2}\sigma_{ds_t}^2\right]E_t\left\{(1+\varphi_{t+1})\exp\left[(L-\gamma)\,\Delta c_{t+1}+(\rho-\gamma)\,v_{t+1}\right]\right\}}{\exp\left[(\rho-\gamma)\,Q_t\left(\Delta c_{t+1}+v_{t+1}\right)\right]}.$$

Given $g_{dt}=L(\mu_t-\bar{g})+\bar{g}_d$, φ_t is also a function of $\tilde{\mu}_t$, Σ_t and volatility level. Our strategy is to first solve the case when $\Sigma_t=0$ and move backwards in time with l_t increasing until $l_t=\Sigma_0$. Looking at $\varphi(\tilde{\mu}_{ct},0)$, we have

$$\varphi_H(\tilde{\mu}_t,0)=\frac{\beta\exp\left[g_{dt}+\frac{1}{2}\sigma_{dH}^2\right]E_t\left\{(1+\varphi_{t+1})\exp\left[(L-\gamma)\,\Delta c_{t+1}+(\rho-\gamma)\,v_{t+1}\right]\right\}}{\exp\left[(\rho-\gamma)\,Q_H(\tilde{\mu}_{ct},0)\right]}.$$

$$\varphi_L(\tilde{\mu}_t,0)=\frac{\beta\exp\left[g_{dt}+\frac{1}{2}\sigma_{dL}^2\right]E_t\left\{(1+\varphi_{t+1})\exp\left[(L-\gamma)\,\Delta c_{t+1}+(\rho-\gamma)\,v_{t+1}\right]\right\}}{\exp\left[(\rho-\gamma)\,Q_L(\tilde{\mu}_{ct},0)\right]}.$$

Let us define

$$\bar{\varphi}_0 = (1-\pi_L)(1+\varphi_H(0,\Sigma_0))\, e^{(\rho-\gamma)v_H(0,\Sigma_0)} + \pi_L(1+\varphi_L(0,\Sigma_0))\, e^{(\rho-\gamma)v_L(0,\Sigma_0)}.$$

We have the partial solution of $\varphi_{H(L)}$ at $\Sigma_t = 0$:

$$\varphi_H(\tilde{\mu}_t,0) = \frac{\beta \exp\left[g_{dt} + \frac{1}{2}\sigma_{dH}^2\right]\left\{(1-\lambda)(1+\varphi_H(\tilde{\mu}_t,0))\, e^{(\rho-\gamma)v_H(\tilde{\mu}_t,0)+(L-\gamma)\left[\tilde{\mu}_t\sigma_{cH}+\bar{g}+\frac{(L-\gamma)\sigma_{cH}^2}{2}\right]} + \lambda\bar{\varphi}_0\right\}}{\exp\left[(\rho-\gamma)Q_H(\tilde{\mu}_t,0)\right]}.$$

$$\varphi_L(\tilde{\mu}_t,0) = \frac{\beta \exp\left[g_{dt} + \frac{1}{2}\sigma_{dL}^2\right]\left\{(1-\lambda)(1+\varphi_L(\tilde{\mu}_t,0))\, e^{(\rho-\gamma)v_L(\tilde{\mu}_t,0)+(L-\gamma)\left[\tilde{\mu}_t\sigma_{cL}+\bar{g}+\frac{(L-\gamma)\sigma_{cL}^2}{2}\right]} + \lambda\bar{\varphi}_0\right\}}{\exp\left[(\rho-\gamma)Q_L(\tilde{\mu}_t,0)\right]}.$$

A.1.2 Compute v_t and φ_t by backward iteration

We solve the v_t and φ_t through backward value function iteration. When the time interval between two breaks tends to infinity, parameter uncertainty Σ_t will converge to zero and v_t and φ_t will converge to the case of $\Sigma_t = 0$. Note there is a one to one mapping between Σ_t and the length of time period to the most recent break as shown in equation (1.19). Let us assume the most recent break happens at time 0, hence,

$$\frac{1}{\Sigma_t} = \frac{1}{\Sigma_0} + t.$$

We use $T_{max} = 400$ quarters as the maximum of t. Assume at $t = T_{max}$ the difference between $\left(v_{H(L)}(\tilde{\mu}_t,0), \varphi_{H(L)}(\tilde{\mu}_t,0)\right)$ and $\left(v_{H(L)}(\tilde{\mu}_t,\Sigma_{Tmax}), \varphi_{H(L)}(\tilde{\mu}_t,\Sigma_{Tmax})\right)$ are small enough to be ignored. We further descritize support of $\tilde{\mu}_t$ in to N_{max} points on the interval $[-h\underline{G}, h\bar{G}]$ by separating it evenly. We use $N_{max} = 200$ and $h = 3$. Given an initial specification of $v_{H(L)}(0,\Sigma_0), \varphi_{H(L)}(0,\Sigma_0)$ as $v_{H(L)}^{(0)}(0,\Sigma_0), \varphi_{H(L)}^{(0)}(0,\Sigma_0)$, we solve $v_t^{(0)}, \varphi_t^{(0)}$ using backward iteration:

1. Given $v_t^{(0)}, \varphi_t^{(0)}$ on the descritezed points, solve for $v_{Tmax}^{(0)}, \varphi_{Tmax}^{(0)}$

2. Starting from $t = T_{max}$, solve for $Q_{t-1}^{(0)}$ using equation $(A.1)$.

3. Solve for $v_{t-1}^{(0)}$ using $Q_{t-1}^{(0)}$ and equation (1.20).

4. Solve for $\varphi_t^{(0)}$ using $Q_{t-1}^{(0)}$ and equations (1.18), (1.21) and (1.23)

5. Repeat process 1-3 until t goes to zero and the resulting $\Sigma_t = \Sigma_0$.

6. Set $\left[v_{H(L)}^{(1)}(0,\Sigma_0), \varphi_{H(L)}^{(1)}(0,\Sigma_0)\right] = \left[v_{H(L)}^{(0)}(0,\Sigma_0), \varphi_{H(L)}^{(0)}(0,\Sigma_0)\right]$ where the right hand side is from $v_t^{(0)}, \varphi_t^{(0)}$ at $t = 0$. Repeat the process 1-5.

Finally, we repeat step 1-6 through N iterations with (0) and (1) replaced by n and $n+1$th iterations. N is large enough that the value function converges which can be measured by the distance between $v_{H(L)}^{(N+1)}(0,\Sigma_0)$ and $v_{H(L)}^{(N)}(0,\Sigma_0)$.

A.2 Estimating the model with MCMC algorithm

Prior

We assume the prior of $\Gamma_i = (\gamma_{i0}, \gamma_{i1})'$, $i = 1,...,N$ as i.i.d. normal distribution, $\Gamma_i \sim N\left(\mu_G, \sigma_G^2 \cdot I_2\right)$. We assume $\mu_G = 0$ and $\sigma_G^2 = 1$.

We assume the prior of d^{-1} that characterizes the scale parameter of Gamma distribution (1.5) where λ_k is drawn is Gamma:

$$d \sim \text{Gamma}\left(u_d^*, v_d^*\right).$$

The expected length of duration is c/d. Our goal is to examine regime of long duration so we truncate the support of d to $[0, \bar{d}]$. We assume $c = 20$, $\mu_d^* = 20$, $v_d^* = 160$ and $\bar{d} = 1/6$ such that the expectation of c/d before truncation is 40 years and the minimum of $c/d = c/\bar{d}$ is 30 years.

We assume the prior of $1/\sigma_\beta^2$ that characterizes the dispersion of regression coefficients (1.7) of the new regime is Gamma:

$$\frac{1}{\sigma_\beta^2} \sim \text{Gamma}\left(u_\beta, v_\beta\right).$$

We assume $\mu_\beta = 5$ and $v_\beta = 0.5$.

MCMC procedure

We use MCMC methods to estimate model parameters. The following is notation preparations. Let Γ denote

$$\Gamma = \begin{pmatrix} \gamma_{10} & \cdots & \gamma_{N0} \\ \gamma_{11} & \cdots & \gamma_{N1} \end{pmatrix}'$$

$\beta = (A_{ik}|i = 1,...,N, k \in K_i)$, $\sigma^2 = \left(\sigma_{ik}^2|i = 1,...,N, k \in K_i\right)$, $\sigma_f^2 = \left(\sigma_{fk}^2|k \in K_f\right)$, $\mathbb{1} = (\mathbb{1}_1,...\mathbb{1}_K)$, $F = (f_1,...,f_T)$, $\Delta c_t = (\Delta c_{1t},...,\Delta c_{Nt})$, $\Delta c = (\Delta c_1',...,\Delta c_T')'$. Let $\tau = (0,\tau_1,...,\tau_K)$ where τ_k is the timing of break k. Our MCMC algorithm involves a marginal Gibbs sampler with Metropolis proposal distributions in some blocks. We separate the parameters into several blocks: Block 1 consists of parameters related to the timing and number of breaks $\{K,\tau,\mathbb{1}\}$, block 2 involves the common component $\{\Gamma,F\}$, block 3 involves hyperparameters parameters governing regime length and distributions where parameters of new regimes are drawn $\left\{d,\sigma_\beta^2\right\}$. We sample the number and timing of breaks using Metropolis-Hastings algorithm in a manner similar to Smith (2017) and Geweke and Jiang (2011). We can improve efficiency in sampling $\{K,\tau,\mathbb{1}\}$ by marginalizing β and σ. The general steps below:

1. Sample starting value of parameters: Sample Γ, d, σ_β^2 from their priors. Use cross-sectional mean of Δc_{it} as starting value of f_t. Set the starting value $K = 1$ and $\tau_K = \lfloor T/2 \rfloor$.

2. Sample the number and location of the breaks: Conditional on Γ, F, d, σ_β^2, sample K, τ, $\mathbb{1}$

using Metropolis-Hastings algorithm.

3. Conditional on d, σ_β^2, K, τ and $\mathbb{1}$, sample Γ and F.

4. Sample hyperparameters governing distributions of regime duration, regression coefficients and volatility of regimes: Conditional on Γ , K, τ and $\mathbb{1}$, sample d and σ_β^2 using Metropolis-Hastings algorithm or Gibbs sampler.

5. Repeat step 2.

A.2.1 Sampling the location, pervasiveness and number of breaks

This section elaborate how to implement step 2: sampling K, τ, $\mathbb{1}$ conditional on Γ, F, d, σ_β^2. We sample the location, the pervasiveness of breaks, and the number of breaks in a sequential manner using Metropolis-Hastings sampler. Before we proceed, we lay out useful density expressions. Let K_{it} denote the indices of the last non common break that hits consumption good i at time t: $K_{it} = \max\{k|\mathbb{1}_{ik} = 1, k \leq K_{it}\}$. K_i denote The conditional density of Δc is derived using similar tricks in Proposition 1 in Smith and Timmermann (2018),

$$\begin{aligned} p\left(\Delta c|F,\Gamma,K,\tau,\mathbb{1},d,\sigma_\beta^2\right) &= p\left(\Delta c|F,\Gamma,K,\tau,\mathbb{1},\sigma_\beta^2\right) \\ &= (2\pi)^{\frac{-TN}{2}} \prod_{i=1}^{N}\prod_{k\in K_i} \frac{v^u}{\Gamma(u)} \frac{\Gamma(\tilde{u}_{ik})}{\tilde{v}_{ik}^{\tilde{u}_{ik}}} \frac{|\Sigma_{ik}|^{1/2}}{|V_\beta|^{1/2}} \end{aligned} \tag{A.2}$$

where

$$\begin{aligned} \Sigma_{ik}^{-1} &= V_\beta^{-1} + |\{t|K_{it} = k\}|, \\ \mu_{ik} &= \Sigma_{ik}\left(\sum_{t\in\{\tau|K_{i\tau}=k\}} \Delta c_{it}\right), \\ \tilde{u}_{ik} &= u + |\{t|K_{it} = k\}|/2, \end{aligned}$$

$$\tilde{v}_{ik} = \frac{1}{2}\left(2v - \mu_{ik}'\Sigma_{ik}^{-1}\mu_{ik} + \sum_{t\in\{\tau|K_{i\tau}=k\}} \Delta c_{it}^2\right).$$

The following is the density of F conditional on $\{K, \tau, \mathbb{1}\}$ and other parameters. It can be shown that σ_{ft}^2 can be marginalized. Let K_f denote the set of breaks that hit the common component

$$p\left(F|K,\tau,\mathbb{1},d,\sigma_\beta^2\right) = p\left(F|K,\tau,\mathbb{1},\sigma_\beta^2\right) = (2\pi)^{\frac{-T}{2}} \prod_{k\in K_f} \frac{v_f^{u_f}\Gamma\left(\tilde{u}_f\right)}{\tilde{v}_f^{\tilde{u}_f}\Gamma\left(u_f\right)}, \tag{A.3}$$

where

$$\tilde{u}_{fk} = u + \left|\left\{t|K_{ft} = k\right\}\right|/2,$$

$$\tilde{v}_{fk} = \frac{1}{2}\left(2v + \sum_{t\in\left\{\tau|K_{f\tau}=k\right\}} f_t^2\right).$$

The probability of τ conditional on $K, \mathbb{1}, d, \sigma_\beta^2$ can be expressed as

$$p\left(\tau|K,\mathbb{1},d,\sigma_\beta^2\right) = p\left(\tau|d\right) = \prod_{k=1}^{K} \frac{1}{l_k!}\frac{\Gamma(c+l_k)}{(d+1)^{c+l_k}}\frac{d^c}{\Gamma(c)}, \tag{A.4}$$

where $l_k = \tau_{k+1} - \tau_k$. In practice, the exact value of l_K is not known given τ: it is only known that $l_K \geq T - \tau_K$. The last factor of equation $(A.4)$ can be replaced by the right tail of CDF of negative Bernoulli distribution NB $\left(r = c, p = \frac{1}{d+1}\right)$.

The following is the probability of $\mathbb{1}$ conditional on $\tau, K, d, \sigma_\beta^2$:

$$p\left(\mathbb{1}|\tau,K,d,\sigma_\beta^2\right) = \prod_{k=1}^{K} \binom{N+1}{N_k} \mathrm{B}\left(N_k+1, N+2-N_k\right) \tag{A.5}$$

which can be derived from the data generating process.

Sample timing of breaks

For each of the $k = 1, ..., K$ breakpoints we perturb τ_k by an integer j that is sampled uniformly from the interval $[-s, s]$, such that $\hat{\tau}_k = \tau_k + j$. The proposed $\hat{\tau}$ has its kth element replaced by $\hat{\tau}_k$. This is a Metropolis-Hastings algorithm of random walk proposition function. The proposal is accepted with probability $\min(1, \alpha)$ where

$$\alpha = \frac{p\left(\Delta c | F, \hat{\tau}, \Gamma, K, \mathbb{1}, \sigma_\beta^2\right) p\left(F | \hat{\tau}, K, \mathbb{1}\right) p\left(\hat{\tau} | d\right)}{p\left(\Delta c | F, \tau, , \Gamma, K, \mathbb{1}, \sigma_\beta^2\right) p\left(F | \tau, K, \mathbb{1}\right) p\left(\tau | d\right)}$$

which can be calculated using equations $(A.2)$, $(A.3)$, $(A.4)$.

Sample pervasiveness of breaks

Let $\mathbb{1}_i$ be the ith row of $\mathbb{1}$ indicating which breaks hit the ith series. Let $\mathbb{1}_f$ be the last row of $\mathbb{1}$ indicating which breaks hit the common component. For each of the $i = 1, ..., N$ series and the common component, we propose $\hat{\mathbb{1}}_{i(f)}$ where each element is sampled independently from a Bernoulli distribution of P. The whole sequence of elements in $\hat{\mathbb{1}}_{i(f)}$ is sampled in one block to increase efficiency because of dependence of breaks affecting the same series. The proposal is accepted with probability $\min(1, \alpha)$ where

$$\alpha = \frac{p\left(\Delta c | F, \tau, \Gamma, K, \hat{\mathbb{1}}, \sigma_\beta^2\right) p\left(F | \tau, K, \hat{\mathbb{1}}\right) p\left(\hat{\mathbb{1}} | \tau, \pi\right) P^{|K_i|} (1-P)^{K-|K_i|}}{p\left(\Delta c | F, \tau, , \Gamma, K, \mathbb{1}, \sigma_\beta^2\right) p\left(F | \tau, K, \mathbb{1}\right) p\left(\mathbb{1} | \tau, \pi\right) P^{|\hat{K}_i|} (1-P)^{K-|\hat{K}_i|}}$$

which can be calculated using equations $(A.2)$, $(A.3)$, $(A.5)$.

Sample number of breaks

We adopt the reversible jump MCMC approach of Green (1995) which is also used in Geweke and Jiang (2011) and Smith (2017). The computational burdern is alleviated by

marginalizing regime specific parameters β, σ^2, σ_f^2. The proposal is a mixture of birth move and death move.

1. Birth move: With equal probability select $\hat{\tau}$ among periods $\{1,...,T\}\setminus\tau$. At time $\hat{\tau}$, we propose with probability P which of the N series and common component got hit. The proposed number of breaks $\hat{K} = K+1$. Let $\hat{N}_{birth}$ denotes the number of series that is hit by the new break. The proposed $\hat{\tau}$ and resulting $\hat{\mathbb{1}}$ and $\hat{K}$ are accepted with probability $\min(1,\alpha)$ where

$$\alpha = \frac{p\left(\Delta c|F,\hat{\tau},,\Gamma,\hat{K},\hat{\mathbb{1}},\sigma_\beta^2\right) p\left(F|\hat{K},\hat{\tau},\hat{\mathbb{1}},d,\sigma_\beta^2\right) p(\hat{\tau}|d)\, p\left(\hat{\mathbb{1}}|\tau,\pi\right) T}{p\left(\Delta c|F,\tau,,\Gamma,K,\mathbb{1},\sigma_\beta^2\right) p\left(F|K,\tau,\mathbb{1},d,\sigma_\beta^2\right) p(\tau|d)\, p\left(\hat{\mathbb{1}}|\tau,\pi\right) P^{\hat{N}_{birth}} (1-P)^{N+1-\hat{N}_{birth}} (K+1)}.$$

   The last factor $\frac{T}{P^{\hat{N}_{birth}}(1-P)^{N+1-\hat{N}_{birth}}(K+1)}$ equals to $\frac{1/(K+1)}{P^{\hat{N}_{birth}}(1-P)^{N+1-\hat{N}_{birth}}/T}$ the ratio of proposal probability of death move over birth move.

2. Death move: With equal probability, delete one element among τ and the corresponding column in $\hat{\mathbb{1}}$ to get $\hat{\tau}$ and $\hat{\mathbb{1}}$. The proposed number of breaks $\hat{K} = K-1$. Let $\hat{N}_{death}$ denotes the number of series that is hit by the break deleted. The proposed $\hat{\tau}$ and resulting $\hat{\mathbb{1}}$ and $\hat{K}$ are accepted with probability $\min(1,\alpha)$ where

$$\alpha = \frac{p\left(\Delta c|F,\hat{\tau},,\Gamma,\hat{K},\hat{\mathbb{1}},\sigma_\beta^2\right) p\left(F|\hat{K},\hat{\tau},\hat{\mathbb{1}},d,\sigma_\beta^2\right) p(\hat{\tau}|d)\, p\left(\hat{\mathbb{1}}|\tau,\pi\right) P^{\hat{N}_{death}} (1-P)^{N+1-\hat{N}_{death}} K}{p\left(\Delta c|F,\tau,,\Gamma,K,\mathbb{1},\sigma_\beta^2\right) p\left(F|K,\tau,\mathbb{1},d,\sigma_\beta^2\right) p(\tau|d)\, p\left(\hat{\mathbb{1}}|\tau,\pi\right) T}.$$

A.2.2 Sample the common component and its parameters

Given the regime specific coefficients β, σ^2 and σ_f^2 , equations (1.1) specifies a state-space model of latent factor $F_t = (f_t, f_{t-1})'$ with state equation

$$F_t = BF_{t-1} + \left(\sigma_{ft}\varepsilon_{ft}, 0\right)' \tag{A.6}$$

where

$$B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Let $\Gamma_i = (\gamma_{i0}, \gamma_{i1})'$ observation equations

$$y_{it} = \Gamma_i' F_t + \sigma_{it}\varepsilon_{it},\ i = 1,...,N \tag{A.7}$$

where $y_{it} = \Delta c_{it} - A_{it}$. The factor component F and parameters B_f and Γ can be sampled by modifying algorithms proposed in Carter and Kohn (1994). The general steps are the following.

1. Sample β, σ^2 and σ_f^2 conditional on $\left\{F, \Gamma, \Delta c, K, \tau, \mathbb{1}, d, \sigma_\beta^2\right\}$.
2. Sample F conditional on $\left\{F, \Gamma, \Delta c, K, \tau, \mathbb{1}, d, \sigma_\beta^2, \beta, \sigma^2, \sigma_f^2\right\}$.
3. Sample Γ conditional on $\left\{F, , \Delta c, K, \tau, \mathbb{1}, d, \sigma_\beta^2, \beta, \sigma^2, \sigma_f^2\right\}$.

The posterior of elements in σ^2 follow inverse Gamma distribution of density which can be sampled directly:

$$p\left(\sigma_{ik}^2 | F, \Gamma, \Delta c, K, \tau, \mathbb{1}, d, \sigma_\beta^2\right) = \frac{\tilde{v}_{ik}^{\tilde{u}_{ik}}}{\Gamma(\tilde{u}_{ik})}\left(\sigma_{ik}^2\right)^{-(\tilde{u}_{ik}+1)} \exp\left(-\frac{\tilde{v}_{ik}}{\sigma_{ik}^2}\right),\ k \in K_i,\ i = 1,...,N.$$

where $\tilde{u}_{ik}$ and $\tilde{v}_{ik}$ are defined in $(A.2)$. The posterior of β conditional on σ^2 is Gaussian:

$$p\left(A_{ik}\right) = N\left(\mu_{ik}, \Sigma_{ik}\right),\ k \in K_i,\ i = 1,...,N.$$

where μ_{ik} and Σ_{ik} are defined in $(A.2)$. The posterior of elements in σ_f^2 follow inverse Gamma distribution of density which can be sampled directly:

$$p\left(\sigma_{fk}^2 | F, \Gamma, \Delta c, K, \tau, \mathbb{1}, d, \sigma_\beta^2\right) = \frac{\tilde{v}_{fk}^{\tilde{u}_{fk}}}{\Gamma(\tilde{u}_{fk})}\left(\sigma_{fk}^2\right)^{-(\tilde{u}_{fk}+1)} \exp\left(-\frac{\tilde{v}_{fk}}{\sigma_{fk}^2}\right),\ k \in K_f, \text{where } \tilde{u}_{fk}$$

and $\tilde{v}_{fk}$ are defined in $(A.3)$.

The posterior of F is normal of which the mean and variance can be computed recursively using Kalman filter. Let $\mu_{ft|t}$ and $R_{ft|t}$ denote the posterior mean and covariance matrix of F_t conditional on $\left\{\beta, \sigma^2, \sigma_f^2, \Gamma, \Delta c^t, K, \tau, \mathbb{1}, d, \sigma_\beta^2\right\}$ where Δc^t is subset of . Let $\mu_{ft+1|t}$ and $R_{ft+1|t}$ denote the posterior mean and variance of F_t conditional on $\left\{\beta, \sigma^2, \sigma_f^2, \Gamma, \Delta c^t, K, \tau, \mathbb{1}, d, \sigma_\beta^2\right\}$. Let

σ_t^2 denotes vector of variance of idiosyncratic innovations $\left(\sigma_{1t}^2, ..., \sigma_{Nt}^2\right)$. The parameters $\mu_{ft|t}$, $R_{ft|t}$, $\mu_{ft+1|t}$ and $R_{ft+1|t}$ are updated following the Kalman filters:

$$\begin{aligned}
\mu_{ft+1|t} &= B\mu_{ft|t}, \\
R_{ft+1|t} &= BR_{ft|t}B' + \text{diag}\left(\sigma_{ft}^2, 0\right), \\
e_{t|t} &= \left(y_{1t} - \Gamma_1\mu_{ft|t-1}, ..., y_{Nt} - \Gamma_N\mu_{ft|t-1}\right)', \\
\mu_{ft|t} &= \mu_{ft|t-1} + R_{ft|t-1}\Gamma'\left(\Gamma R_{ft|t-1}\Gamma' + \text{diag}\left(\sigma_t^2\right)\right)^{-1} e_{t|t}, \\
R_{ft|t} &= R_{ft|t-1} - R_{ft|t-1}\Gamma'\left(\Gamma R_{ft|t-1}\Gamma' + \text{diag}\left(\sigma_t^2\right)\right)^{-1}\Gamma R_{ft|t-1}
\end{aligned}$$

We can then sample the entire set of factor observations conditional on the parameters starting from the latest period T and move backward. First, sample F_T from $N\left(\mu_{fT|T}, R_{fT|T}\right)$. Let $\mu_{ft|t,f_{t+1}}$ and $R_{ft|t,f_{t+1}}$ denote the posterior mean and variance of F_t conditional on $\left\{F_{t+1}, \beta, \sigma^2, \sigma_f^2, \Gamma, \Delta c, K, \tau, \mathbb{1}, \sigma_\beta^2\right\}$.

$$\begin{aligned}
\mu_{ft|t,f_{t+1}} &= \mu_{ft|t} + R_{ft|t}B'R_{ft+1|t}^{-1}\left(f_{t+1} - \mu_{ft+1|t}\right), \\
R_{ft|t,f_{t+1}} &= R_{ft|t} - R_{ft|t}B'R_{ft+1|t}^{-1}BR_{ft|t}.
\end{aligned}$$

Conditional on $\left\{\beta, \sigma^2, \Gamma, \Delta c, F\right\}$, the posterior of Γ_i, $i = 1, ..., N$ is $N\left(\tilde{\mu}_{Gi}, \tilde{\Sigma}_{Gi}\right)$:

$$\begin{aligned}
\tilde{\Sigma}_{Gi}^{-1} &= \sigma_G^{-2}I_2 + \sum_{t=1}^{T} F_tF_t'/\sigma_{it}^2, \\
\tilde{\mu}_{Gi} &= \tilde{\Sigma}_{Gi}\left(\sigma_G^{-2}\mu_G + \sum_{t=1}^{T} F_ty_t/\sigma_{it}^2\right).
\end{aligned}$$

A.2.3 Sample parameters governing regime length and break distributions

We sample regime d governing the distribution where regime length is drawn using Metropolis-Hasting algorithm. The proposal $\hat{d} = d\exp(s\cdot\varepsilon)$, where ε is drawn from a standard normal distribution and s is features average step length. $\hat{d}$ is accepted with probability $\min(1,\alpha)$ where

$$\alpha = \frac{p\left(\tau|\hat{d}\right)p\left(\hat{d}|u_d^*,v_d^*\right)}{p\left(\tau|d\right)p\left(d|u_d^*,v_d^*\right)}.$$

$p\left(d|u_d^*,v_d^*\right)$ is the prior density of d of Gamma $\left(u_d^*,v_d^*\right)$.

We sample $1/\sigma_\beta^2$ governing the variance of new coefficients using Metropolis-Hasting algorithm. The proposal $1/\hat{\sigma}_\beta^2 = 1/\sigma_\beta^2\exp(s\cdot\varepsilon)$ where ε is drawn from a standard normal distribution and s is features average step length. $\hat{\sigma}_\beta^2$ is accepted with probability $\min(1,\alpha)$ where

$$\alpha = \frac{p\left(\Delta c|F,\tau,\Gamma,K,\mathbb{1},\hat{\sigma}_\beta^2\right)p\left(1/\hat{\sigma}_\beta^2|u_\beta,v_\beta\right)}{p\left(\Delta c|F,\tau,\Gamma,K,\mathbb{1},\sigma_\beta^2\right)p\left(1/\sigma_\beta^2|u_\beta,v_\beta\right)}.$$

$p\left(1/\sigma_\beta^2|u_\beta,v_\beta\right)$ is the prior density of $1/\sigma_\beta^2$ of Gamma $\left(u_\beta,v_\beta\right)$.

Appendix B

Appendix for Chapter 2

B.1 Monte Carlo Simulations

To explore the finite-sample performance of our bootstrap procedure for identifying superior forecasting skills, this section conducts a series of Monte Carlo simulations addressing both the size and power of our bootstrap. We adopt the following setup: for forecasters $m = 1, ..., M$, variables $i = 1, ..., N$ and time periods $t + h = 1, ..., T$, the forecast errors are assumed to obey the factor structure

$$e_{i,t+h,m} = \lambda_{i,m} f_{t+h} + u_{i,t+h,m},$$

where f_{t+h} is a mean-zero Gaussian AR(1) process with autoregressive coefficient ρ and variance σ_f^2. We generate $\lambda_{i,m}$ as i.i.d random variables from a $N(0, \sigma_\lambda^2)$ distribution truncated such that $\lambda_{i,m}^2 \sigma_f^2 \leq 0.9$; we then set $u_{i,t+h,m}$ as a mean-zero Gaussian AR(1) process with AR coefficient ρ and variance $1 - \lambda_{i,m}^2 \sigma_f^2$. Here, $\{f_{t+h}\}_{t+h=1}^{T}$, $\{\lambda_{i,m}\}_{1 \leq i \leq N, 1 \leq m \leq M}$ and $\{u_{i,t+h,m}\}_{1 \leq i \leq N, 1 \leq m \leq M, 1 \leq t+h \leq T}$ are mutually independent. We set $(\sigma_f, \sigma_\lambda) = (2, 1.2)$. When $T > 30$, we use $\rho = 0.5$ and a block size $B_T = T^{0.6}$; otherwise, we set $\rho = 0$ and $B_T = 1$. We consider a no normalization and a partial normalization scheme, both of which are described in Example 2.3.1. Under these schemes, all forecast errors have MSE values equal to one so the null

hypothesis that no forecasts underperform the benchmark model, m_0, holds.

Table B.1 reports size results from 1,200 Monte Carlo simulations using a variety of combinations for the sample size, $T = \{25, 50, 100, 200\}$, the number of forecasters, $M = \{2, 10, 100\}$ and the number of outcome variables $N = \{1, 10, 25, 50, 100\}$. Each MC simulation uses 250 bootstraps. We report results both with and without studentizing the Sup test statistic and use critical values of $\alpha = 0.05, 0.10$.

In general, the size of the non-studentized test statistic is reasonably closely aligned with the true size although it tends to be undersized for large values of N and T, particularly when M is also large. The size properties of the studentized test statistic are quiteR good for small-to-modest values of N,M, and T, but this test statistic tends to be severely undersized when N, T, M are large. The undersizing is particularly pronounced for $\alpha = 0.05$. Interestingly, when the time-series dimension is small ($T = 25$), the studentized test statistic is actually over-sized and the rejection rate increases in the number of variables, N. This pattern reverses in the tests that use larger sample sizes, i.e., $T = 50, 100, 200$.

The size simulations can be used to compute size-adjusted critical values that deliver more accurate finite-sample performance. In particular, for each value of (N, M, T), we can compute size-adjusted critical values for the p-value such that the rejection probability for this sample size under the null hypothesis is made to be exactly α. Whenever the rejection rate in Table B.1 exceeds α, the corresponding size-adjusted p-value, displayed in Table B.2, will be adjusted downward (below α), whereas the reverse holds when the rejection rate in Table B.1 falls below α. An interesting observation from Table B.2 is that using a critical level of $\alpha = 0.10$ for the studentized test statistic in many cases gets us close to a size of 5%. This is the chief reason we use a 10% size throughout the empirical analysis.

To explore the power properties of the Sup test statistics with and without studentization, consider the following setup. For each of the N outcome variables, we use one forecast as the benchmark while the remaining $(M - 1)$ forecasts are competitors. In other words, we split the

NM forecasts into N benchmarks and $(N-1)M$ competing forecasts. Next, we randomly select 20% of these competing forecasts and add $\left(2T^{-1}\log(MN)\right)^{1/8}$ to the selected forecast errors, which then have larger MSE than the baseline forecasts.

This design for the power experiments is in line with that in Chernozhukov, Chetverikov and Kato (2018). In their simulations, 5% of the moments violate the null hypothesis while this figure is 20% in ours. We choose a larger percentage of moments that violate the null hypothesis due to the smaller sample size in our experiments: their sample size is always 400 and our sample size ranges from 25 to 200.[1]

Table B.3 reports the power of the Sup test statistics with and without studentization. To facilitate comparisons across the two test statistics, we use size-adjusted critical values. With exception of a few instances when $N = 1$, the power of the Sup test statistic is generally much higher with studentization than without, e.g., the power can be 10-20% for the non-studentized test statistic but 70-80% for the studentized test statistic. The general conclusion is, thus, that using the studentized rather than the non-studentized test statistic yields far better power.

B.2 Proofs

This appendix provides proofs for the theoretical results in our paper.

B.2.1 Preliminary results

Before proving our theorem, we start by recalling some results from Chernozhukov, Chetverikov and Kato (2018), re-stated here using our notation so that these results can be readily used in our analysis. Let $q_T > r_T$ with $q_T + r_T \leq T/2$. Further, let $B_T = q_T + r_T$ and $K = K_T = \lfloor T/(q_T + r_T) \rfloor$ (the integer part of $T/(q_T + r_T)$). For $1 \leq k \leq K$, define $A_k = \{t :$

[1] Because of our smaller sample sizes, it is necessary to let the magnitude of departures from the null depend on the sample size in order to obtain meaningful comparisons; if we change $\left(2T^{-1}\log(MN)\right)^{1/8}$ to a fixed value, we would likely find that the methods either have power close to one or close to the nominal size.

$(k-1)(q_T+r_T)+1 \leq t \leq (k-1)(q_T+r_T)+q_T\}$. Let $\{\varepsilon_k\}_{k=1}^K$ be i.i.d $N(0,1)$ random variables that are independent of the data. In the proofs, we use W_t instead of W_{t+h} for notational simplicity; changing $t+h$ to t does not affect the theoretical arguments.

Theorem B.2.1 (Theorem B.1 of Chernozhukov, Chetverikov and Kato (2018))**.** *Let Assumption 1 hold. Then there exist constants $C, c > 0$ depending only on c_1, c_2 and C_1 such that*

$$E \sup_{x \in \mathbb{R}} \left| P\left(\max_{1 \leq j \leq \mathcal{N}} T^{-1/2} \sum_{t=1}^{T} W_{jt} \leq x \right) - P\left(\max_{1 \leq j \leq \mathcal{N}} \frac{1}{\sqrt{K q_T}} \sum_{k=1}^{K} \sum_{t \in A_k} (W_{j,t} - \hat{\mu}_j)\varepsilon_k \leq x \mid \{W_s\}_{s=1}^T \right) \right| \leq C T^{-c},$$

where $\hat{\mu}_j = T^{-1}\sum_{t=1}^T W_{jt}$. Moreover,

$$\sup_{x \in \mathbb{R}} \left| P\left(\max_{1 \leq j \leq \mathcal{N}} T^{-1/2} \sum_{t=1}^{T} W_{jt} \leq x \right) - P\left(\max_{1 \leq j \leq \mathcal{N}} Z_j \leq x \right) \right| \leq C T^{-c},$$

where $Z = (Z_1, ..., Z_{\mathcal{N}})' \in \mathbb{R}^{\mathcal{N}}$ is a centered Gaussian vector with variance matrix $EZZ' = (Kq_T)^{-1} \sum_{k=1}^K E\left[\left(\sum_{t \in A_k} W_t\right) \left(\sum_{t \in A_k} W_t\right)' \right]$.

Proof. The first claim follows from the statement of Theorem B.1 of Chernozhukov, Chetverikov and Kato (2018). The second statement is from the proof of Theorem B.1 of Chernozhukov, Chetverikov and Kato (2018); see Equation (94) therein. □

Bootstrap approximation of normalized test statistic

The following theorem is a general result on the bootstrap approximation of the normalized test statistic, assuming a good approximation of the normalization. In Appendix B.2.1, we provide further results on the approximation of the normalization. We first state the following Theorem B.2.2, present its proof, and then prove the auxiliary lemmas used.

Theorem B.2.2. *Let Assumption 1 hold. Suppose that* $T^{-1}\sqrt{K}r_T(\log\mathcal{N})^{3/2} = o(1)$, $T^{-1}q_T(\log\mathcal{N})^{3/2} = o(1)$, $T^{-1}Kr_T\log^2\mathcal{N} = o(1)$ *and* $T^{-1}r_T^2D_T^2\log^3\mathcal{N} = o(1)$. *Let* $a = (a_1,...,a_{\mathcal{N}})' \in \mathbb{R}^{\mathcal{N}}$ *be nonrandom with* $\kappa_1 \le a_j \le \kappa_2$ *for all* $1 \le j \le \mathcal{N}$ *and* $\kappa_1, \kappa_2 > 0$. *Suppose that* $\hat{a} = (\hat{a}_1,...,\hat{a}_{\mathcal{N}})$ *satisfies* $\min_{1\le j\le\mathcal{N}}\hat{a}_j > 0$ *and* $\|\hat{a}-a\|_\infty = o_P(1/\log\mathcal{N})$. *Then*

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}\frac{T^{-1/2}\sum_{t=1}^T W_{jt}}{\hat{a}_j}\le x\right)\right.$$
$$\left. - P\left(\max_{1\le j\le\mathcal{N}}\frac{T^{-1/2}\sum_{k=1}^K\sum_{t\in\bar{A}_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k}{\hat{a}_j}\le x \mid \{W_s\}_{s=1}^T\right)\right| = o_P(1),$$

where $\bar{A}_k = \{t:\ (k-1)(q_T+r_T)+1 \le t \le k(q_T+r_T)\}$.

Proof. We first apply Theorem B.2.1 to $\{(a_1^{-1}W_{1t},...,a_{\mathcal{N}}^{-1}W_{\mathcal{N}t})\}_{t=1}^T$, obtaining

$$E\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}a_j^{-1}T^{-1/2}\sum_{t=1}^T W_{jt}\le x\right)\right.$$
$$\left. - P\left(\max_{1\le j\le\mathcal{N}}a_j^{-1}(Kq_T)^{-1/2}\sum_{k=1}^K\sum_{t\in A_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k\le x \mid \{W_s\}_{s=1}^T\right)\right| \le CT^{-c},$$

where $C, c > 0$ are constants that only depend on c_1, c_2, C_1, κ_1 and κ_2. By Lemma B.2.5,

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}a_j^{-1}(Kq_T)^{-1/2}\sum_{k=1}^K\sum_{t\in A_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k\le x \mid \{W_s\}_{s=1}^T\right)\right.$$
$$\left. - P\left(\max_{1\le j\le\mathcal{N}}\hat{a}_j^{-1}T^{-1/2}\sum_{k=1}^K\sum_{t\in\bar{A}_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k\le x \mid \{W_s\}_{s=1}^T\right)\right| = o_P(1).$$

Therefore, we have

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}a_j^{-1}T^{-1/2}\sum_{t=1}^T W_{jt}\le x\right)\right.$$
$$\left. - P\left(\max_{1\le j\le\mathcal{N}}\hat{a}_j^{-1}T^{-1/2}\sum_{k=1}^K\sum_{t\in\bar{A}_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k\le x \mid \{W_s\}_{s=1}^T\right)\right| = o_P(1).$$

It follows from Lemma B.2.3 that

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq\mathcal{N}}a_j^{-1}T^{-1/2}\sum_{t=1}^{T}W_{jt}\leq x\right)\right.$$
$$\left.-P\left(\max_{1\leq j\leq\mathcal{N}}\hat{a}_j^{-1}T^{-1/2}\sum_{t=1}^{T}W_{jt}\leq x\right)\right|=o_P(1).$$

The desired result follows from this. □

Lemma B.2.1. *Let $R=(R_1,...,R_{\mathcal{N}})'$, $\hat{R}=(\hat{R}_1,...,\hat{R}_{\mathcal{N}})'$, $\zeta=(\zeta_1,...,\zeta_{\mathcal{N}})'$, $\hat{\zeta}=(\hat{\zeta}_1,...,\hat{\zeta}_{\mathcal{N}})'$ and $Z=(Z_1,...,Z_{\mathcal{N}})'$ be random vectors in $\mathbb{R}^{\mathcal{N}}$. Suppose that ζ and $\hat{\zeta}$ are $\mathcal{F}$-measurable for some σ-algebra $\mathcal{F}$. Also assume that Z is a centered Gaussian vector with $\min_{1\leq j\leq\mathcal{N}}E(Z_j^2)\geq b$ almost surely for some constant $b>0$. If $\max_{1\leq j\leq\mathcal{N}}|\hat{R}_j-R_j|=o_P(1/\sqrt{\log\mathcal{N}})$ as $N\to\infty$ (or other dimensions tend to infinity), then the following holds:*

$$E\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq\mathcal{N}}Z_j\leq x\right)-P(\max_{1\leq j\leq\mathcal{N}}\hat{R}_j\leq x\mid\mathcal{F})\right|$$
$$\leq 3E\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq\mathcal{N}}Z_j\leq x\right)-P\left(\max_{1\leq j\leq\mathcal{N}}R_j\leq x\mid\mathcal{F}\right)\right|+o(1).$$

Moreover, if $\|\hat{\zeta}-\zeta\|_\infty=o_P(1/\sqrt{\log\mathcal{N}})$, the following holds:

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq\mathcal{N}}Z_j\leq x\right)-P\left(\max_{1\leq j\leq\mathcal{N}}\hat{\zeta}_j\leq x\right)\right|$$
$$\leq 3\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq\mathcal{N}}Z_j\leq x\right)-P\left(\max_{1\leq j\leq\mathcal{N}}\zeta_j\leq x\right)\right|+o(1).$$

Proof. **Step 1:** show the first claim.

For an arbitrary $\eta>0$, let $c=\eta/\sqrt{\log\mathcal{N}}$. Define the event $\mathcal{M}=\{\max_{1\leq j\leq\mathcal{N}}|\hat{R}_j-R_j|\leq c\}$ and variables $\xi=\max_{1\leq j\leq\mathcal{N}}R_j$ and $\hat{\xi}=\max_{1\leq j\leq\mathcal{N}}\hat{R}_j$. Let $a_N=\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\leq j\leq\mathcal{N}}Z_j\leq x\right)-P\left(\xi\leq x\mid\mathcal{F}\right)\right|$.

We first notice that, given the event $\mathcal{M}$, $|\hat{\xi}-\xi| \leq c$, and thus

$$\begin{aligned}
&\left|\mathbf{1}\{\xi \leq x\} - \mathbf{1}\left\{\hat{\xi} \leq x\right\}\right| \\
&= \mathbf{1}\left\{\hat{\xi} \leq x \text{ and } \xi > x\right\} + \mathbf{1}\left\{\hat{\xi} > x \text{ and } \xi \leq x\right\} \\
&= \mathbf{1}\left\{\xi - \hat{\xi} \geq \xi - x \text{ and } \xi - x > 0\right\} + \mathbf{1}\left\{\xi - \hat{\xi} < \xi - x \text{ and } \xi - x \leq 0\right\} \\
&\leq \mathbf{1}\left\{|\xi - x| \leq |\hat{\xi} - \xi|\right\} \leq \mathbf{1}\{|\xi - x| \leq c\}. \qquad \text{(B.1)}
\end{aligned}$$

Hence,

$$\begin{aligned}
&\left|P(\xi \leq x \mid \mathcal{F}) - P(\hat{\xi} \leq x \mid \mathcal{F})\right| \\
&\leq E\left[\left|\mathbf{1}\{\xi \leq x\} - \mathbf{1}\left\{\hat{\xi} \leq x\right\}\right| \mid \mathcal{F}\right] \\
&\leq P(|\xi - x| \leq c \mid \mathcal{F}) + P(\mathcal{M}^c \mid \mathcal{F}) \\
&\leq P(\xi \leq x + c \mid \mathcal{F}) - P(\xi \leq x - 2c) + P(\mathcal{M}^c \mid \mathcal{F}) \\
&\leq P\left(\max_{1 \leq j \leq \mathcal{N}} Z_j \leq x + c \mid \mathcal{F}\right) - P\left(\max_{1 \leq j \leq \mathcal{N}} Z_j \leq x - 2c\right) + P(\mathcal{M}^c \mid \mathcal{F}) + 2a_N. \qquad \text{(B.2)}
\end{aligned}$$

Let $\iota = (1, ..., 1)' \in \mathbb{R}^{\mathcal{N}}$. Then, by Lemma A.1 of Chernozhukov, Chetverikov and Kato (2017), it follows that almost surely, for any $x \in \mathbb{R}$,

$$P(Z \leq (x + c)\iota \mid \mathcal{F}) - P(Z \leq (x - 2c)\iota \mid \mathcal{F}) \leq 3cC_b\sqrt{\log \mathcal{N}},$$

where $C_b > 0$ is a constant that only depends on b. Here, $Z \leq (x + c)\iota$ means $Z_j \leq (x + c)$ for all $1 \leq j \leq \mathcal{N}$; similarly, $Z \leq (x - 2c)\iota$ means that $Z_j \leq (x - 2c)$ for all $1 \leq j \leq \mathcal{N}$. Hence, for any $z \in \mathbb{R}$, $P(Z \leq z\iota) = P(\max_{1 \leq j \leq \mathcal{N}} Z_j \leq z)$. Therefore, the above display implies that for any $x \in \mathbb{R}$,

$$P\left(\max_{1 \leq j \leq \mathcal{N}} Z_j \leq x + c \mid \mathcal{F}\right) - P\left(\max_{1 \leq j \leq \mathcal{N}} Z_j \leq x - 2c\right) \leq 3cC_b\sqrt{\log \mathcal{N}}. \qquad \text{(B.3)}$$

By (B.2), we have

$$\left|P(\xi \le x \mid \mathcal{F}) - P(\hat{\xi} \le x \mid \mathcal{F})\right| \le 3cC_b\sqrt{\log \mathcal{N}} + 2a_N + P(\mathcal{M}^c \mid \mathcal{F}).$$

Since the above display holds for any $x \in \mathbb{R}$, we have

$$\sup_{x\in\mathbb{R}}\left|P(\xi \le x \mid \mathcal{F}) - P(\hat{\xi} \le x \mid \mathcal{F})\right| \le 3cC_b\sqrt{\log \mathcal{N}} + 2a_N + P(\mathcal{M}^c \mid \mathcal{F}),$$

and, thus,

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x\right) - P(\hat{\xi} \le x \mid \mathcal{F})\right| \le 3a_N + 3cC_b\sqrt{\log \mathcal{N}} + P(\mathcal{M}^c \mid \mathcal{F}).$$

Taking expectations on both sides, we obtain

$$E\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x\right) - P(\hat{\xi} \le x \mid \mathcal{F})\right| \le 3Ea_N + 3cC_b\sqrt{\log \mathcal{N}} + P(\mathcal{M}^c) = 3Ea_N + 3\eta C_b + P(\mathcal{M}^c).$$

Since $\max_{1\le j\le \mathcal{N}} |\hat{R}_j - R_j| = o_P(1/\sqrt{\log \mathcal{N}})$, $P(\mathcal{M}^c) = o(1)$, and so

$$E\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x\right) - P(\hat{\xi} \le x \mid \mathcal{F})\right| \le 3Ea_N + 3\eta C_b + o(1).$$

Because $\eta > 0$ is arbitrary, it follows that

$$E\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x\right) - P(\hat{\xi} \le x \mid \mathcal{F})\right| = Ea_N + o(1).$$

Step 2: show the second claim.

The argument is similar to Step 1, but we include the details for completeness.

Fix an arbitrary $\eta > 0$. Let $c_1 = \eta/\sqrt{\log \mathcal{N}}$, $\psi = \max_{1\le j\le \mathcal{N}} \zeta_j$ and $\hat{\psi} = \max_{1\le j\le \mathcal{N}} \hat{\zeta}_j$. Define $d_N = \sup_{x\in\mathbb{R}} \left|P\left(\psi \le x\right) - P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x\right)\right|$. Define the event $\mathcal{M}_1 = \{\|\hat{\zeta} - \zeta\|_\infty \le$

$c_1\}$. As in (B.1), we notice that, given the event $\mathcal{M}_1$,

$$|\mathbf{1}\{\psi \le x\} - \mathbf{1}\{\hat{\psi} \le x\}| \le \mathbf{1}\{|\psi - x| \le c_1\}.$$

Thus,

$$\begin{aligned}
&|P(\psi \le x) - P(\hat{\psi} \le x)| \\
&\le P(|\psi - x| \le c_1) + P(\mathcal{M}_1^c) \\
&= P(x - c_1 \le \psi \le x + c_1) + P(\mathcal{M}_1^c) \\
&\le P(\psi \le x + c_1) - P(\psi \le x - 2c_1) + P(\mathcal{M}_1^c) \\
&\le P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x + c_1\right) - P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x - 2c_1\right) + 2d_N + P(\mathcal{M}_1^c) \\
&\overset{\text{(i)}}{\le} 3c_1 C_b \sqrt{\log \mathcal{N}} + 2d_N + P(\mathcal{M}_1^c),
\end{aligned}$$

where (i) follows by (B.3) (with c replaced by c_1). Since the above bound holds for any $x \in \mathbb{R}$, we have that, given the event $\mathcal{M}_1$,

$$\sup_{x\in\mathbb{R}} |P(\psi \le x) - P(\hat{\psi} \le x)| \le 3c_1 C_b \sqrt{\log \mathcal{N}} + 2d_N = 6c_1 C_b \eta + 2d_N + P(\mathcal{M}_1^c).$$

Therefore,

$$\sup_{x\in\mathbb{R}} \left| P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x\right) - P(\hat{\psi} \le x) \right| \le 6c_1 C_b \eta + 3d_N + P(\mathcal{M}_1^c).$$

Notice that $P(\mathcal{M}_1^c) = o(1)$ due to $\|\hat{\zeta} - \zeta\|_\infty = o_P(1/\sqrt{\log \mathcal{N}})$. Since $\eta > 0$ is arbitrary, we have

$$\sup_{x\in\mathbb{R}} |P(\psi \le x) - P(\hat{\psi} \le x)| \le 3d_N + o(1).$$

This completes the proof. □

Lemma B.2.2. *Suppose that* $Z = (Z_1, ..., Z_{\mathcal{N}})'$ *is a centered Gaussian vector with* $\max_{1\le j\le \mathcal{N}} E(Z_j^2) \le b$ *for some constant* $b > 0$. *Then for any* $z \in (0, \mathcal{N}/5)$, $P\left(\|Z\|_\infty \ge \sqrt{2b\log(\mathcal{N}/z)}\right) \le 2z$.

Proof. Clearly, $Z_j/\sqrt{EZ_j^2} \sim N(0,1)$. Thus, $P\left(|Z_j|/\sqrt{EZ_j^2} > x\right) = 2\Phi(-x) = 2 - 2\Phi(x)$, where $\Phi(\cdot)$ denotes the cdf of a $N(0,1)$ variable. Since $EZ_j^2 \le b$, we have that for any $x > 0$, $P\left(|Z_j| > \sqrt{b}x\right) \le 2(1-\Phi(x))$. By the union bound, it follows that for any $x > 0$,

$$P\left(\max_{1\le j\le \mathcal{N}} |Z_j| > \sqrt{b}x\right) \le 2\mathcal{N}(1-\Phi(x)).$$

Taking $x = \Phi^{-1}(1-a)$ for $a \in (0,1)$, we have $P(\|Z\|_\infty > \sqrt{b}\Phi^{-1}(1-a)) \le 2\mathcal{N}a$. By Lemma 1 of Zhu and Bradic (2018), for $a \le 1/5$, $\Phi^{-1}(1-a) \le \sqrt{2\log(1/a)}$. This means that $P(\|Z\|_\infty > \sqrt{2b\log(1/a)}) \le 2\mathcal{N}a$. The desired result follows by setting $a = z/\mathcal{N}$. □

Lemma B.2.3. *Let the assumptions of Theorem B.2.2 hold. Then*

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le \mathcal{N}} a_j^{-1}T^{-1/2}\sum_{t=1}^{T} W_{jt} \le x\right) - P\left(\max_{1\le j\le \mathcal{N}} \hat{a}_j^{-1}T^{-1/2}\sum_{t=1}^{T} W_{jt} \le x\right)\right| = o_P(1).$$

Proof. Define $\zeta = (\zeta_1, ..., \zeta_{\mathcal{N}})'$ and $\hat{\zeta} = (\hat{\zeta}_1, ..., \hat{\zeta}_{\mathcal{N}})'$, where $\zeta_j = a_j^{-1}T^{-1/2}\sum_{t=1}^{T} W_{jt}$ and $\hat{\zeta}_j = \hat{a}_j^{-1}T^{-1/2}\sum_{t=1}^{T} W_{jt}$. Also, let $Z = (Z_1, ..., Z_{\mathcal{N}})' \in \mathbb{R}^{\mathcal{N}}$ be a centered Gaussian vector with variance matrix $EZZ' = (Kq_T)^{-1}\sum_{k=1}^{K} E\left[\left(\sum_{t\in A_k} W_t\right)\left(\sum_{t\in A_k} W_t\right)'\right]$.

By Theorem B.2.1, we have

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x\right) - P\left(\max_{1\le j\le \mathcal{N}} \zeta_j \le x\right)\right| = o(1). \tag{B.4}$$

Applying the same argument with (Z_j, ζ_j) replaced by $(-Z_j, -\zeta_j)$, we obtain

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}(-Z_j)\le x\right)-P\left(\max_{1\le j\le\mathcal{N}}(-\zeta_j)\le x\right)\right|=o(1).$$

By assumption, $\max_{1\le j\le\mathcal{N}}EZ_j^2\le C_1$. Hence, by Lemma B.2.2, we have that, for any $\eta\in(0,1/5)$,

$$P\left(\|Z\|_\infty\ge\sqrt{2C_1\log(\eta\mathcal{N})}\right)\le 2\eta.$$

It follows that

$$P\left(\max_{1\le j\le\mathcal{N}}\zeta_j>\sqrt{2C_1\log(\eta\mathcal{N})}\right)\le P\left(\max_{1\le j\le\mathcal{N}}Z_j>\sqrt{2C_1\log(\eta\mathcal{N})}\right)+o(1)\le 2\eta+o(1)$$

and

$$P\left(\max_{1\le j\le\mathcal{N}}(-\zeta_j)>\sqrt{2C_1\log(\eta\mathcal{N})}\right)\le P\left(\max_{1\le j\le\mathcal{N}}(-Z_j)>\sqrt{2C_1\log(\eta\mathcal{N})}\right)+o(1)\le 2\eta+o(1).$$

Therefore,

$$\begin{aligned}&P\left(\|\zeta\|_\infty>\sqrt{2C_1\log(\eta\mathcal{N})}\right)\\&\quad\le P\left(\max_{1\le j\le\mathcal{N}}\zeta_j>\sqrt{2C_1\log(\eta\mathcal{N})}\right)+P\left(\max_{1\le j\le\mathcal{N}}(-\zeta_j)>\sqrt{2C_1\log(\eta\mathcal{N})}\right)\le 4\eta+o(1).\end{aligned}$$

Since η is arbitrary, we have

$$\|\zeta\|_\infty=O_P\left(\sqrt{\log\mathcal{N}}\right). \tag{B.5}$$

By Assumption 1, $\min_{1\le j\le\mathcal{N}}E(Z_j^2)\ge c_1$. Next, we verify

$$\|\hat{\zeta}-\zeta\|_\infty=o_P(1/\sqrt{\log\mathcal{N}}). \tag{B.6}$$

Notice that $\hat{\zeta}_j = \hat{a}_j^{-1} a_j \zeta_j$. Thus,

$$\|\hat{\zeta} - \zeta\|_\infty = \max_{1\le j\le \mathcal{N}} |\hat{\zeta}_j - \zeta_j| \le \|\zeta\|_\infty \max_{1\le j\le \mathcal{N}} \left|\hat{a}_j^{-1} a_j - 1\right|.$$

Since $\min_{1\le j\le \mathcal{N}} a_j \ge \kappa_1$ and $\|\hat{a} - a\|_\infty = o_P(1)$, we have $\max_{1\le j\le \mathcal{N}} \left|\hat{a}_j^{-1} a_j - 1\right| = O_P(\|\hat{a} - a\|_\infty)$. Since $\|\zeta\|_\infty = O_P(\sqrt{\log \mathcal{N}})$, we have (B.6) by the assumption on $\|\hat{a} - a\|_\infty$.

Therefore, from Lemma B.2.1, we have

$$\begin{aligned}\sup_{x\in\mathbb{R}} &\left| P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x\right) - P\left(\max_{1\le j\le \mathcal{N}} \hat{\zeta}_j \le x\right)\right| \\ &\le 3 \sup_{x\in\mathbb{R}} \left| P\left(\max_{1\le j\le \mathcal{N}} Z_j \le x\right) - P\left(\max_{1\le j\le \mathcal{N}} \zeta_j \le x\right)\right| + o_P(1) \overset{\text{(i)}}{=} o_P(1),\end{aligned}$$

where (i) follows by (B.4). Now by the triangular inequality, (B.4) implies

$$\sup_{x\in\mathbb{R}} \left| P\left(\max_{1\le j\le \mathcal{N}} \zeta_j \le x\right) - P\left(\max_{1\le j\le \mathcal{N}} \hat{\zeta}_j \le x\right)\right| = o_P(1).$$

This completes the proof. □

Lemma B.2.4. *Let the assumptions of Theorem B.2.2 hold. Then*

$$\max_{1\le j\le \mathcal{N}} \left| \sum_{k=1}^{K} \sum_{t\in \bar{A}_k \setminus A_k} (W_{j,t} - \hat{\mu}_j)\varepsilon_k \right| = O_P\left(\sqrt{K r_T \log \mathcal{N}} + r_T D_T \log \mathcal{N} + r_T \sqrt{T^{-1} K \log \mathcal{N}} \right),$$

where $\bar{A}_k = \{t : \ (k-1)(q_T + r_T) + 1 \le t \le k(q_T + r_T)\}$ *for* $1 \le k \le K-1$ *and* $\bar{A}_K = \{(K-1)(q_T + r_T) + q_T + 1, ..., T\}$.

Proof. For $1 \le k \le K$, let $u_{j,k} = \sum_{t\in \bar{A}_k \setminus A_k} W_{j,t}\varepsilon_k$. By Berbee's coupling (e.g., Lemma 7.1 of Chen et al. (2016)), there exist variables $\{v_k\}_{k=1}^K$ with $v_k = (v_{1,k}, ..., v_{\mathcal{N},k})'$ such that (1) $\{v_k\}_{k=1}^K$ is independent across k and independent of $\{\varepsilon_k\}_{k=1}^K$; (2) v_k has the same distribution as $\sum_{t\in \bar{A}_k \setminus A_k} W_t$ and (3) $P(\bigcap_{k=1}^K \{v_k = \sum_{t\in \bar{A}_k \setminus A_k} W_t\}) \ge 1 - K\beta_{\text{mixing}}(q_T)$.

Since $\varepsilon_k \sim N(0,1)$, Assumption 1 implies that $\max_{1\le j\le\mathcal{N}}\sum_{k=1}^{K} E(v_{j,k}\varepsilon_k)^2 \le Kr_TC_1$. Also notice that $\|\sum_{t\in\bar{A}_k\setminus A_k} W_t\|_\infty \le r_T D_T$. It follows by Lemma D.3 of Chernozhukov, Chetverikov and Kato (2018) that

$$E\left(\max_{1\le j\le\mathcal{N}}\left|\sum_{k=1}^{K} v_{j,k}\varepsilon_k\right|\right) \le M\left(\sqrt{Kr_TC_1\log\mathcal{N}} + r_TD_T\log\mathcal{N}\right),$$

where $M>0$ is a universal constant. Since $\max_{1\le j\le\mathcal{N}}\left|\sum_{k=1}^{K-1} v_{j,k}\varepsilon_k\right| = \max_{1\le j\le\mathcal{N}}|\sum_{k=1}^{K-1} u_{j,k}|$ with a probability of at least $1-K\beta_{\text{mixing}}(q_T)$ and $K\beta_{\text{mixing}}(q_T) \le T\beta_{\text{mixing}}(r_T) = o(1)$, we have that

$$\max_{1\le j\le\mathcal{N}}\left|\sum_{k=1}^{K} u_{j,k}\right| = O_P\left(\sqrt{Kr_T\log\mathcal{N}} + r_TD_T\log\mathcal{N}\right). \tag{B.7}$$

In the proof of Lemma B.2.3, we showed that $\max_{1\le j\le\mathcal{N}}\hat{\mu}_j = O_P(\sqrt{T^{-1}\log\mathcal{N}})$; see (B.5). By a similar argument, we have $\max_{1\le j\le\mathcal{N}}(-\hat{\mu}_j) = O_P(\sqrt{T^{-1}\log\mathcal{N}})$. Hence,

$$\|\hat{\mu}\|_\infty = O_P(\sqrt{T^{-1}\log\mathcal{N}}).$$

It follows that

$$\begin{aligned}\max_{1\le j\le\mathcal{N}}\left|\sum_{k=1}^{K}\sum_{t\in\bar{A}_k\setminus A_k}\hat{\mu}_j\varepsilon_k\right| &= r_T\|\hat{\mu}\|_\infty\left|\sum_{k=1}^{K-1}\varepsilon_k\right| \\ &= O_P(r_T\sqrt{T^{-1}\log\mathcal{N}}\times\sqrt{K-1}) = O_P(r_T\sqrt{T^{-1}K\log\mathcal{N}}).\end{aligned}$$

The above display and (B.7) imply that

$$\max_{1\le j\le\mathcal{N}}\left|\sum_{k=1}^{K}\sum_{t\in\bar{A}_k\setminus A_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k\right| = O_P\left(\sqrt{Kr_T\log\mathcal{N}} + r_TD_T\log\mathcal{N} + r_T\sqrt{T^{-1}K\log\mathcal{N}}\right).$$

This completes the proof. □

Lemma B.2.5. *Let the assumptions of Theorem B.2.2 hold. Then*

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}} a_j^{-1}(Kq_T)^{-1/2}\sum_{k=1}^{K}\sum_{t\in A_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k\le x\mid\{W_s\}_{s=1}^{T}\right)\right.$$
$$\left.-P\left(\max_{1\le j\le\mathcal{N}} \hat{a}_j^{-1}T^{-1/2}\sum_{k=1}^{K}\sum_{t\in\bar{A}_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k\le x\mid\{W_s\}_{s=1}^{T}\right)\right|=o_P(1),$$

where $\{\bar{A}_k\}_{k=1}^{K}$ *is defined in the statement of Lemma B.2.4.*

Proof. Define $R=(R_1,...,R_{\mathcal{N}})'$ and $\hat{R}=(\hat{R}_1,...,\hat{R}_{\mathcal{N}})'$, where $R_j=a_j^{-1}(Kq_T)^{-1/2}\sum_{k=1}^{K}\sum_{t\in A_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k$ and $\hat{R}_j=\hat{a}_j^{-1}T^{-1/2}\sum_{k=1}^{K}\sum_{t\in\bar{A}_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k$. Let $\mathcal{F}$ denote the σ-algebra generated by $\{W_s\}_{s=1}^{T}$.

Also, let $Z=(Z_1,...,Z_{\mathcal{N}})'\in\mathbb{R}^{\mathcal{N}}$ be a centered Gaussian vector with variance matrix $EZZ'=(Kq_T)^{-1}\sum_{k=1}^{K}E\left[\left(\sum_{t\in A_k}W_t\right)\left(\sum_{t\in A_k}W_t\right)'\right]$.

By Theorem B.2.1, we have

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}Z_j\le x\right)-P\left(\max_{1\le j\le\mathcal{N}}R_j\le x\mid\mathcal{F}\right)\right|=o_P(1). \tag{B.8}$$

Applying the same argument with (Z_j,R_j) replaced by $(-Z_j,-R_j)$, we obtain

$$\sup_{x\in\mathbb{R}}\left|P\left(\max_{1\le j\le\mathcal{N}}(-Z_j)\le x\right)-P\left(\max_{1\le j\le\mathcal{N}}(-R_j)\le x\mid\mathcal{F}\right)\right|=o_P(1).$$

By assumption, $\max_{1\le j\le\mathcal{N}}EZ_j^2\le C_1$. Hence, by Lemma B.2.2, we have that, for any $\eta\in(0,1/5)$,

$$P\left(\|Z\|_\infty\ge\sqrt{2C_1\log(\eta\mathcal{N})}\right)\le 2\eta.$$

It follows that

$$P\left(\max_{1\le j\le\mathcal{N}}R_j>\sqrt{2C_1\log(\eta\mathcal{N})}\mid\mathcal{F}\right)\le P\left(\max_{1\le j\le\mathcal{N}}Z_j>\sqrt{2C_1\log(\eta\mathcal{N})}\right)+o_P(1)\le 2\eta+o_P(1)$$

and

$$P\Big(\max_{1\le j\le\mathcal{N}}(-R_j)>\sqrt{2C_1\log(\eta\mathcal{N})}|\mathcal{F}\Big)\le P\Big(\max_{1\le j\le\mathcal{N}}(-Z_j)>\sqrt{2C_1\log(\eta\mathcal{N})}\Big)+o_P(1)\le 2\eta+o_P(1).$$

In turn, we have

$$\begin{aligned}&P\Big(\|R\|_\infty>\sqrt{2C_1\log(\eta\mathcal{N})}|\mathcal{F}\Big)\\ &\le P\left(\max_{1\le j\le\mathcal{N}}R_j>\sqrt{2C_1\log(\eta\mathcal{N})}\mid\mathcal{F}\right)+P\left(\max_{1\le j\le\mathcal{N}}(-R_j)>\sqrt{2C_1\log(\eta\mathcal{N})}\mid\mathcal{F}\right)\le 4\eta+o_P(1).\end{aligned}$$

Since η is arbitrary, we have

$$\|R\|_\infty=O_P(\sqrt{\log\mathcal{N}}). \tag{B.9}$$

By Assumption 1, $\min_{1\le j\le\mathcal{N}}EZ_j^2\ge c_1$. Hence, by Lemma B.2.1, it suffices to verify that

$$\max_{1\le j\le\mathcal{N}}|\hat{R}_j-R_j|=o_P(1/\sqrt{\log\mathcal{N}}). \tag{B.10}$$

We notice that

$$\hat{a}_j\sqrt{T}\hat{R}_j-a_j\sqrt{Kq_T}R_j=\sum_{k=1}^{K}\sum_{t\in\bar{A}_k\setminus A_k}(W_{j,t}-\hat{\mu}_j)\varepsilon_k.$$

Hence, by Lemma B.2.4, we have

$$\max_{1\le j\le\mathcal{N}}\left|\hat{a}_j\sqrt{T}\hat{R}_j-a_j\sqrt{Kq_T}R_j\right|=O_P\left(\sqrt{Kr_T\log\mathcal{N}}+r_TD_T\log\mathcal{N}+r_T\sqrt{T^{-1}K\log\mathcal{N}}\right).$$

Since $\|\hat{a}-a\|_\infty = o_P(1)$ and $\min_{1\le j\le \mathcal{N}} a_j \ge \kappa_1 > 0$, we have

$$\max_{1\le j\le \mathcal{N}}\left|\hat{R}_j - a_j\hat{a}_j^{-1}\sqrt{Kq_T/T}R_j\right| = O_P\left(\sqrt{T^{-1}Kr_T\log\mathcal{N}} + T^{-1/2}r_T D_T\log\mathcal{N} + r_T T^{-1}\sqrt{K\log\mathcal{N}}\right).$$

By (B.9), it follows that

$$\begin{aligned}
&\max_{1\le j\le \mathcal{N}}\left|\hat{R}_j - R_j\right| \\
&\le \max_{1\le j\le \mathcal{N}}\left|\hat{R}_j - a_j\hat{a}_j^{-1}\sqrt{Kq_T/T}R_j\right| + \max_{1\le j\le \mathcal{N}}\left|a_j\hat{a}_j^{-1}\sqrt{Kq_T/T} - 1\right| \times \max_{1\le j\le \mathcal{N}}\left|R_j\right| \\
&= O_P\left(\sqrt{T^{-1}Kr_T\log\mathcal{N}} + T^{-1/2}r_T D_T\log\mathcal{N} + r_T T^{-1}\sqrt{K\log\mathcal{N}}\right) \\
&\quad + \max_{1\le j\le \mathcal{N}}\left|a_j\hat{a}_j^{-1}\sqrt{Kq_T/T} - 1\right| O_P(\sqrt{\log\mathcal{N}}) \\
&\le O_P\left(\sqrt{T^{-1}Kr_T\log\mathcal{N}} + T^{-1/2}r_T D_T\log\mathcal{N} + r_T T^{-1}\sqrt{K\log\mathcal{N}}\right) \\
&\quad + \max_{1\le j\le \mathcal{N}}\left|a_j\hat{a}_j^{-1} - 1\right|\sqrt{Kq_T/T}O_P(\sqrt{\log\mathcal{N}}) + \max_{1\le j\le \mathcal{N}}\left|\sqrt{Kq_T/T} - 1\right| O_P(\sqrt{\log\mathcal{N}}).
\end{aligned} \tag{B.11}$$

By assumption, we have $\|\hat{a}-a\|_\infty = o_P(1/\log\mathcal{N})$ and $\min_{1\le j\le \mathcal{N}} a_j \ge \kappa_1 > 0$. Observe that

$$\max_{1\le j\le \mathcal{N}}\left|a_j\hat{a}_j^{-1} - 1\right|\sqrt{Kq_T/T} \le O_P\left(\max_{1\le j\le \mathcal{N}}|\hat{a}_j - a_j|\right)$$

and

$$\begin{aligned}
\left|\sqrt{Kq_T/T} - 1\right| = \frac{1 - Kq_T/T}{1+\sqrt{Kq_T/T}} &< \frac{T - Kq_T}{T} \\
&< \frac{((K+1)(q_T+r_T) - Kq_T)}{T} = \frac{(K+1)r_T + q_T}{T}.
\end{aligned}$$

By the assumptions on the rates in the statement of Theorem B.2.2, we obtain (B.10) from (B.11). This completes the proof. □

We next show the following result.

Preliminary results on variance approximation

Lemma B.2.6. *Let the assumptions of Theorem B.2.2 hold. Moreover, suppose that* $\beta_{\text{mixing}}(i) \lesssim \exp(-b_1 i^{b_2})$. *Then*

$$\max_{1\le j\le \mathcal{N}} |T^{-1}\sum_{t=1}^{T}(W_{j,t}-E(W_{j,t}))| = O_P\left(D_T T^{-1/2}\left(\log\mathcal{N}+(\log T)^{1/(2b_2)}\right)\right).$$

Proof. We apply Bernstein's blocking technique with the same block structure as A_k and $\bar{A}_k$, but the choice of K and q_T is only specific to the proof of this lemma. Let $R_{k,j}=\sum_{t\in A_k}(W_{j,t}-E(W_{j,t}))$. We choose r_T such that $K\beta_{\text{mixing}}(r_T)=o(1)$. This means that $Kq_T\exp(-b_1 r_T^{b_2})=o(1)$.

By Berbee's coupling and Lemma 8 of Chernozhukov, Chetverikov and Kato (2015), it follows that

$$\begin{aligned}
&\max_{1\le j\le\mathcal{N}}\left|\sum_{k=1}^{K}R_{k,j}\right| \\
&= O_P\left(\max_{1\le j\le\mathcal{N},\ 1\le k\le K}|R_{k,j}|\sqrt{\log\mathcal{N}}+\sqrt{\max_{1\le j\le\mathcal{N}}\sum_{k=1}^{K}ER_{k,j}^2\times\log\mathcal{N}}\right) \\
&\le O_P\left(\max_{1\le k\le K}|A_k|\max_{1\le j\le\mathcal{N},\ 1\le t\le T}|W_{j,t}-E(W_{j,t})|\sqrt{\log\mathcal{N}}+\sqrt{\max_{1\le j\le\mathcal{N}}\sum_{k=1}^{K}ER_{k,j}^2\times\log\mathcal{N}}\right) \\
&\overset{\text{(i)}}{=} O_P\left(D_T q_T\sqrt{\log\mathcal{N}}+D_T\sqrt{Kq_T}\log\mathcal{N}\right),
\end{aligned}$$

where (i) follows from the definition of A_k and $ER_{k,j}^2\lesssim D_T^2 q_T$ (due to Lemma 7.2 of Chen et al. (2016)). Moreover,

$$\max_{1\le j\le\mathcal{N}}\left|\sum_{t=1}^{T}(W_{j,t}-E(W_{j,t}))-\sum_{k=1}^{K}R_{k,j}\right| = \max_{1\le j\le\mathcal{N}}\left|\sum_{k=1}^{K}\sum_{t\in\bar{A}_k/A_k}(W_{j,t}-E(W_{j,t}))\right|$$

$$\leq 2D_T \sum_{k=1}^{K} |\bar{A}_k / A_k| \leq 2D_T (Kr_T + q_T).$$

Therefore,

$$\max_{1\leq j\leq \mathcal{N}} \left| \sum_{t=1}^{T} (W_{j,t} - E(W_{j,t})) \right| = O_P \left(D_T q_T \sqrt{\log \mathcal{N}} + D_T \sqrt{K q_T} \log \mathcal{N} + D_T K r_T \right).$$

Using $K \asymp T/(q_T + r_T)$, we choose $q_T \asymp \sqrt{T r_T / \sqrt{\log \mathcal{N}}}$ to get

$$\max_{1\leq j\leq \mathcal{N}} \left| \sum_{t=1}^{T} (W_{j,t} - E(W_{j,t})) \right| = O_P \left(D_T \sqrt{T} \log \mathcal{N} + D_T \sqrt{T r_T} \right).$$

Now the requirement of $K q_T \exp(-b_1 r_T^{b_2}) = o(1)$ implies that we can choose $r_T \asymp (\log T)^{1/b_2}$, which means that

$$\max_{1\leq j\leq \mathcal{N}} \left| \sum_{t=1}^{T} (W_{j,t} - E(W_{j,t})) \right| = O_P \left(D_T \sqrt{T} \left(\log \mathcal{N} + (\log T)^{1/(2b_2)} \right) \right).$$

The desired result follows from this. □

Lemma B.2.7. *Let the assumptions of Theorem B.2.2 hold. Moreover, suppose that* $K\beta_{\text{mixing}}(q_T + r_T) = o(1)$. *Then*

$$\max_{1\leq j\leq \mathcal{N}} \left| T^{-1} \sum_{k=1}^{K} \left[\left(\sum_{t\in \bar{A}_k} (W_{j,t} - \mu_j) \right)^2 - E \left(\sum_{t\in \bar{A}_k} (W_{j,t} - \mu_j) \right)^2 \right] \right|$$
$$= O_P \left(\sqrt{T^{-1}(q_T + r_T) D_T^4 \log \mathcal{N}} + T^{-1} (q_T + r_T)^2 D_T^2 \log \mathcal{N} \right).$$

Proof. For simplicity, we assume that K is an even number and denote $L_T = K/2$. Since $K\beta_{\text{mixing}}(q_T + r_T) = o(1)$, we can use Berbee's coupling and Lemma 8 of Chernozhukov,

Chetverikov and Kato (2015), obtaining

$$\begin{aligned}
&\max_{1\le j\le \mathcal{N}} \left| \sum_{l=1}^{L_T} \left[\left(\sum_{t\in \bar{A}_{2l-1}} (W_{j,t}-\mu_j) \right)^2 - E\left(\sum_{t\in \bar{A}_{2l-1}} (W_{j,t}-\mu_j) \right)^2 \right] \right| \\
&= O_P\left(\sqrt{ \sum_{l=1}^{L_T} E\left(\sum_{t\in \bar{A}_{2l-1}} (W_{j,t}-\mu_j) \right)^4 \log \mathcal{N} } + \max_{1\le l\le L_T,\ 1\le j\le \mathcal{N}} \left(\sum_{t\in \bar{A}_{2l-1}} (W_{j,t}-\mu_j) \right)^2 \times \log \mathcal{N} \right) \\
&\overset{\text{(i)}}{=} O_P\left(\sqrt{ \sum_{l=1}^{L_T} |\bar{A}_{2l-1}|^2 (2D_T)^4 \times \log \mathcal{N} } + \max_{1\le l\le L_T} |\bar{A}_{2l-1}|^2 (2D_T)^2 \times \log \mathcal{N} \right) \\
&= O_P\left(\sqrt{ T(q_T+r_T) D_T^4 \log \mathcal{N} } + (q_T+r_T)^2 D_T^2 \log \mathcal{N} \right),
\end{aligned}$$

where (i) follows by $E\left(\sum_{t\in \bar{A}_{2l-1}} (W_{j,t}-\mu_j)\right)^4 \lesssim |\bar{A}_{2l-1}|^2 (2D_T)^4$ (due to Lemma 7.2 of Chen et al. (2016)). Similarly, we can show

$$\begin{aligned}
&\max_{1\le j\le \mathcal{N}} \left| \sum_{l=1}^{L_T} \left[\left(\sum_{t\in \bar{A}_{2l}} (W_{j,t}-\mu_j) \right)^2 - E\left(\sum_{t\in \bar{A}_{2l}} (W_{j,t}-\mu_j) \right)^2 \right] \right| \\
&\qquad = O_P\left(\sqrt{ T(q_T+r_T) D_T^4 \log \mathcal{N} } + (q_T+r_T)^2 D_T^2 \log \mathcal{N} \right).
\end{aligned}$$

The desired result follows from this. □

B.2.2 Proof of Theorem 2.3.1

We apply Theorem B.2.2. We separate $K_T = q_T + r_T$ with $q_T > r_T$. Specifically, we choose $r_T = \kappa_1 (\log T)^{1/b_2}$ with $\kappa_1 = (2/b_1)^{1/b_2}$ and $q_T = K_T - r_T$. It suffices to show that the conditions of Theorem B.2.2 can be satisfied by this choice of (q_T, r_T).

Since $K \asymp T/(q_T+r_T)$ and $(r_T/q_T)\log^2 \mathcal{N} = o(1)$, we have $T^{-1}\sqrt{K} r_T (\log \mathcal{N})^{3/2} = o(1)$ and $T^{-1} K r_T \log^2 \mathcal{N} = o(1)$. By $q_T D_T \log^{5/2}(\mathcal{N}T) \le C_1 T^{1/2-c_2}$ and $r_T < q_T$, we have $T^{-1} q_T (\log \mathcal{N})^{3/2} = o(1)$ and $T^{-1} r_T^2 D_T^2 \log^3 \mathcal{N} = o(1)$. It remains to show that Assumption 1 is

satisfied by this choice of (q_T, r_T). In particular, with $\mathcal{N} = N$, we need to verify the following

$$\max\{K\beta_{\text{mixing}}(r_T),\ (r_T/q_T)\log^2\mathcal{N}\} \leq C_1 T^{-c_2} \tag{B.12}$$

and

$$q_T D_T \log^{5/2}(\mathcal{N}T) \leq C_1 T^{1/2-c_2} \tag{B.13}$$

for some $0 < c_2 < 1/4$.

Since $\beta_{\text{mixing}}(r_T) \lesssim \exp(-b_1 r_T^{b_2})$, $K \leq T$, it follows that $K\beta_{\text{mixing}}(r_T) \lesssim T\exp(-b_1 r_T^{b_2})$. Hence, we have $K\beta_{\text{mixing}}(r_T) \lesssim T^{-1}$.

Since $K \asymp T/(q_T + r_T)$ and $q_T > r_T$, we have $q_T \asymp T/K$. Therefore,

$$(r_T/q_T)\log^2\mathcal{N} \lesssim (\log T)^{1/b_2} K T^{-1}\log^2\mathcal{N}.$$

By Assumption 1, $KT^{-1}\log^2\mathcal{N} \lesssim T^{-b}$ for some $b \in (0, 1/4)$. Thus, we only need to choose $c_2 = b/2$ to obtain $(r_T/q_T)\log^2\mathcal{N} \lesssim T^{-c_2}$. This proves (B.12).

By $q_T \asymp T/K$ and Assumption 1, we have

$$q_T D_T \log^{5/2}(\mathcal{N}T) \asymp K^{-1} T D_T \log^{5/2}(\mathcal{N}T) \lesssim T^{1/2-b} \lesssim T^{1/2-c_2},$$

which proves (B.13). Now we have verified all the conditions of Theorem B.2.2, which implies that

$$\sup_{x\in\mathbb{R}}\left| P\left(\max_{1\leq j\leq\mathcal{N}} \frac{T^{-1/2}\sum_{t=1}^T W_{jt}}{\hat{a}_j} \leq x\right) - P\left(\max_{1\leq j\leq\mathcal{N}} \frac{T^{-1/2}\sum_{k=1}^K \sum_{t\in\bar{A}_k}(W_{j,t} - \hat{\mu}_j)\varepsilon_k}{\hat{a}_j} \leq x \mid \{W_s\}_{s=1}^T\right)\right| = o_P(1).$$

This means that

$$\lim_{T\to\infty} P\left(\max_{1\le j\le \mathcal{N}} \frac{T^{-1/2}\sum_{t=1}^{T}(U_{j,t}-EU_{j,t})}{\hat{a}_j} > \tilde{Q}^{*}_{T,1-\alpha}\right) = \alpha.$$

Under the null hypothesis of $EU_{j,t} \le 0$, we have that

$$\max_{1\le j\le \mathcal{N}} \frac{T^{-1/2}\sum_{t=1}^{T}(U_{j,t}-EU_{j,t})}{\hat{a}_j} \ge \max_{1\le j\le \mathcal{N}} \frac{T^{-1/2}\sum_{t=1}^{T}U_{j,t}}{\hat{a}_j} = \tilde{R}_T.$$

In turn, this means that

$$\limsup_{T\to\infty} P\left(\tilde{R}_T > \tilde{Q}^{*}_{T,1-\alpha}\right) \le \alpha.$$

When $EU_{j,t} = 0$, the inequality in the above two equation displays hold with equality. This completes the proof.

B.2.3 Proof of Lemma 2.3.1

We consider two cases.

Case 1: $\hat{a}_j = 1$.

We only need to take $a_j = 1$. Then $\hat{a}_j - a_j = 1$ and the result clearly holds.

Case 2: $\hat{a}_j = \sqrt{K^{-1}\sum_{j=1}^{K}\left(B_T^{-1/2}\sum_{t\in H_j}(\Delta L_{i,t+h} - \hat{\mu}_i)\right)^2}$.

We inherit all the notations from before. Recall $\hat{\mu}_j = T^{-1}\sum_{t=1}^{T} W_{j,t}$. Let $a_j^2 = T^{-1}\sum_{k=1}^{K} E\left(\sum_{t\in\bar{A}_k}(W_{j,t}-\mu_j)\right)^2$ with $\mu_j = T^{-1}\sum_{t=1}^{T}E(W_{j,t})$. Let $\bar{W}_{j,t} = W_{j,t} - \mu_j$ and $\bar{a}_j^2 = T^{-1}\sum_{k=1}^{K}\left(\sum_{t\in\bar{A}_k}\bar{W}_{j,t}\right)^2$. Clearly, $\hat{\mu}_j - \mu_j = T^{-1}\sum_{t=1}^{T}\bar{W}_{j,t}$.

Notice that by triangular inequality for the Euclidean norm in $\mathbb{R}^K$, we have

$$\max_{1\le j\le \mathcal{N}} \left|\hat{a}_j - \bar{a}_j\right| = T^{-1/2} \max_{1\le j\le \mathcal{N}} \left| \sqrt{\sum_{k=1}^{K}\left(\sum_{t\in\bar{A}_k}\bar{W}_{j,t} - |\bar{A}_k|(\hat{\mu}_j-\mu_j)\right)^2} - \sqrt{\sum_{k=1}^{K}\left(\sum_{t\in\bar{A}_k}\bar{W}_{j,t}\right)^2}\right|$$

$$
\begin{aligned}
&\leq T^{-1/2} \max_{1\leq j\leq \mathcal{N}} \sqrt{\sum_{k=1}^{K} |\bar{A}_k|^2 (\hat{\mu}_j - \mu_j)^2} \\
&\leq T^{-1/2} \|\hat{\mu} - \mu\|_\infty \max_{1\leq k\leq K} |\bar{A}_k| \sqrt{K} \\
&\overset{\text{(i)}}{=} O_P\left(D_T T^{-1} \left(\log \mathcal{N} + (\log T)^{1/(2b_2)} \right) \times (q_T + r_T)\sqrt{K} \right) \\
&= O_P\left(D_T K^{-1/2} \left(\log \mathcal{N} + (\log T)^{1/(2b_2)} \right) \right),
\end{aligned}
$$

where (i) follows from Lemma B.2.6. On the other hand, Lemma B.2.7 implies that

$$
\max_{1\leq j\leq \mathcal{N}} \left| \bar{a}_j^2 - a_j^2 \right| = O_P\left(\sqrt{T^{-1}(q_T + r_T) D_T^4 \log \mathcal{N}} + T^{-1}(q_T + r_T)^2 D_T^2 \log \mathcal{N} \right).
$$

Notice that the rate conditions in the assumption imply that the rate in the above two displays are $o_P(1/\log \mathcal{N})$. Since $\min_{1\leq j\leq \mathcal{N}} a_j$ is bounded away from zero, we have $\max_{1\leq j\leq \mathcal{N}} |\hat{a}_j - a_j| = o_P(1/\log \mathcal{N})$.

The proof for the other cases follows by similar arguments as for Case 2.

Table B.1: Finite-sample size of Sup tests computed across multiple variables, forecasters, and time-periods

	$\alpha = 0.05$								$\alpha = 0.1$							
	Without studentization				With studentization				Without studentization				With studentization			
	$M = 2$				$M = 2$				$M = 2$				$M = 2$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.063	0.060	0.050	0.057	0.058	0.063	0.049	0.057	1	0.117	0.135	0.111	0.113	0.117	0.131	0.116	0.117
10	0.054	0.050	0.049	0.052	0.059	0.047	0.027	0.023	10	0.109	0.112	0.105	0.108	0.126	0.113	0.077	0.081
25	0.053	0.048	0.045	0.056	0.073	0.033	0.026	0.014	25	0.115	0.112	0.093	0.112	0.141	0.098	0.076	0.073
50	0.039	0.047	0.040	0.048	0.052	0.022	0.022	0.020	50	0.086	0.117	0.097	0.100	0.122	0.087	0.054	0.059
100	0.033	0.036	0.040	0.039	0.082	0.027	0.011	0.015	100	0.087	0.099	0.109	0.086	0.148	0.074	0.058	0.045
	$M = 10$				$M = 10$				$M = 10$				$M = 10$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.049	0.063	0.059	0.049	0.057	0.044	0.040	0.032	1	0.121	0.158	0.143	0.119	0.113	0.134	0.113	0.088
10	0.068	0.053	0.043	0.047	0.077	0.034	0.021	0.021	10	0.134	0.143	0.121	0.104	0.155	0.101	0.075	0.068
25	0.057	0.067	0.045	0.041	0.087	0.019	0.009	0.008	25	0.127	0.155	0.123	0.130	0.170	0.083	0.043	0.049
50	0.042	0.047	0.056	0.035	0.099	0.020	0.006	0.007	50	0.105	0.135	0.123	0.095	0.196	0.072	0.044	0.033
100	0.036	0.033	0.019	0.026	0.121	0.025	0.004	0.003	100	0.104	0.100	0.077	0.070	0.231	0.079	0.030	0.031
	$M = 100$				$M = 100$				$M = 100$				$M = 100$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.053	0.072	0.051	0.047	0.075	0.039	0.021	0.020	1	0.150	0.165	0.137	0.128	0.135	0.114	0.084	0.070
10	0.059	0.062	0.047	0.036	0.114	0.023	0.005	0.004	10	0.122	0.165	0.131	0.117	0.227	0.098	0.051	0.036
25	0.042	0.034	0.042	0.038	0.130	0.020	0.007	0.002	25	0.113	0.120	0.126	0.100	0.283	0.083	0.033	0.023
50	0.039	0.030	0.024	0.028	0.179	0.016	0.002	0.002	50	0.102	0.129	0.086	0.086	0.369	0.081	0.025	0.021
100	0.031	0.023	0.016	0.022	0.237	0.007	0.001	0.002	100	0.067	0.088	0.070	0.063	0.430	0.063	0.013	0.011

Notes: This table presents the size of Sup tests comparing the finite-sample accuracy of a set of benchmark forecasts m_0 to a set of alternative forecasts, m_1. All numbers are based on 2,000 Monte Carlo simulations conducted under the null of equal predictive accuracy of the forecasts in m_0 and m_1. N denotes the number of variables; M refers to the number of forecasters, while T denotes the number of time-series observations. The two panels on the left present results set the asymptotic size of the test to $\alpha = 0.05$ while the two panels on the right set the asymptotic size of the test to $\alpha = 0.10$. The Monte Carlo simulations generate the forecast errors as $e_{i,t+h,m} = \lambda_{i,m} f_{t+h} + u_{i,t+h,m}$, where f_t is a mean-zero Gaussian AR(1) process with autoregressive coefficient ρ and variance σ_f^2 . We generate $\lambda_{i,m}$ as i.i.d random variables from a $N(0, \sigma_\lambda^2)$ distribution and truncated such that $\lambda_{i,m}^2 \sigma_f^2 \leq 0.9$; we then set $u_{i,t+h,m}$ as a mean-zero Gaussian AR(1) process with AR coefficient ρ and variance $1 - \lambda_{i,m}^2 \sigma_f^2$. $\{f_{t+h}\}_{t+h=1}^{T}$, $\{\lambda_{i,m}\}_{1 \leq i \leq n, 1 \leq m \leq M}$ and $\{u_{i,t+h,m}\}_{1 \leq i \leq n, 1 \leq m \leq M, 1 \leq t+h \leq T}$ are assumed to be mutually independent. We choose $(\sigma_f, \sigma_\lambda) = (2, 1.2)$. When $T > 30$, we use $\rho = 0.5$ and a block size of $BT = T^{0.6}$; otherwise we use $\rho = 0$ and $B_T = 1$. We consider two studentization schemes: no studentization (the first and third panels) and (partial) studentization (the second and fourth panel), which are both described in Example 1. Under this scheme, all forecast errors have an MSE equal to one and thus the null hypothesis that no forecasts underperforms the baseline model holds.

Table B.2: Size-adjusted critical values for the Sup test

	$\alpha = 0.05$								$\alpha = 0.1$							
	Without studentization				With studentization				Without studentization				With studentization			
	$M = 2$				$M = 2$				$M = 2$				$M = 2$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.040	0.048	0.052	0.044	0.048	0.040	0.052	0.044	1	0.084	0.076	0.084	0.088	0.076	0.080	0.080	0.088
10	0.048	0.052	0.052	0.048	0.040	0.052	0.080	0.080	10	0.096	0.088	0.096	0.092	0.076	0.092	0.124	0.120
25	0.048	0.056	0.056	0.044	0.032	0.064	0.080	0.084	25	0.096	0.092	0.108	0.092	0.068	0.108	0.132	0.128
50	0.068	0.056	0.064	0.052	0.048	0.068	0.092	0.092	50	0.112	0.092	0.104	0.104	0.088	0.112	0.140	0.152
100	0.064	0.060	0.060	0.072	0.032	0.076	0.092	0.108	100	0.112	0.104	0.096	0.112	0.064	0.124	0.136	0.168
	$M = 10$				$M = 10$				$M = 10$				$M = 10$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.052	0.040	0.048	0.052	0.044	0.056	0.056	0.068	1	0.084	0.072	0.076	0.084	0.092	0.080	0.092	0.112
10	0.036	0.048	0.056	0.052	0.036	0.068	0.080	0.080	10	0.076	0.080	0.088	0.096	0.060	0.100	0.120	0.128
25	0.044	0.040	0.056	0.052	0.024	0.080	0.108	0.104	25	0.088	0.072	0.088	0.088	0.060	0.112	0.148	0.156
50	0.056	0.056	0.048	0.064	0.020	0.084	0.108	0.132	50	0.096	0.084	0.088	0.104	0.052	0.116	0.156	0.180
100	0.064	0.068	0.076	0.084	0.016	0.076	0.128	0.140	100	0.100	0.104	0.116	0.124	0.040	0.116	0.172	0.188
	$M = 100$				$M = 100$				$M = 100$				$M = 100$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.048	0.036	0.048	0.052	0.036	0.060	0.076	0.084	1	0.072	0.064	0.080	0.084	0.076	0.092	0.116	0.128
10	0.044	0.044	0.052	0.056	0.024	0.076	0.100	0.120	10	0.088	0.068	0.080	0.092	0.044	0.104	0.140	0.168
25	0.056	0.060	0.056	0.064	0.016	0.076	0.112	0.132	25	0.096	0.092	0.084	0.104	0.040	0.112	0.160	0.184
50	0.060	0.064	0.072	0.068	0.012	0.084	0.128	0.148	50	0.100	0.088	0.112	0.112	0.028	0.116	0.172	0.196
100	0.080	0.076	0.084	0.084	0.008	0.096	0.140	0.176	100	0.128	0.108	0.128	0.132	0.016	0.124	0.176	0.224

Notes: This table presents size-adjusted critical values for Sup tests conducted on finite sample data generated in a Monte Carlo simulation. Here, N denotes the number of variables; M is the number of forecasters, and T is the number of time periods. The two panels on the left present results when the asymptotic size of the test is set to $\alpha = 0.05$ while the two panels on the right present result when the asymptotic size of the test is set to $\alpha = 0.10$. For each value of (N, M, T), we compute the critical value for the p-value such that the rejection probability for this sample size under the null hypothesis is set to equal α. We refer to these critical values as size-adjusted critical values. We consider two studentization schemes: no studentization (first and third panels) and (partial) studentization (second and fourth panels), both being described in Example 1. The forecast errors are generated in the same way as those in table A1.

Table B.3: Power of Sup test using size-adjusted critical values

	$\alpha = 0.05$								$\alpha = 0.1$							
	Without studentization				With studentization				Without studentization				With studentization			
	$M = 2$				$M = 2$				$M = 2$				$M = 2$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.050	0.053	0.057	0.035	0.054	0.045	0.064	0.038	1	0.095	0.087	0.105	0.090	0.082	0.083	0.095	0.093
10	0.112	0.129	0.169	0.170	0.351	0.378	0.589	0.702	10	0.218	0.240	0.305	0.304	0.462	0.515	0.682	0.775
25	0.109	0.131	0.121	0.114	0.461	0.576	0.790	0.903	25	0.216	0.208	0.280	0.255	0.609	0.712	0.892	0.950
50	0.140	0.107	0.144	0.124	0.657	0.690	0.884	0.975	50	0.218	0.196	0.225	0.234	0.770	0.834	0.955	0.994
100	0.100	0.105	0.100	0.159	0.621	0.771	0.922	0.993	100	0.197	0.205	0.180	0.244	0.776	0.890	0.976	0.999
	$M = 10$				$M = 10$				$M = 10$				$M = 10$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.473	0.312	0.444	0.610	0.298	0.299	0.384	0.584	1	0.559	0.468	0.551	0.706	0.415	0.393	0.500	0.687
10	0.095	0.112	0.125	0.146	0.637	0.723	0.873	0.978	10	0.182	0.207	0.213	0.274	0.733	0.827	0.945	0.998
25	0.092	0.097	0.128	0.135	0.644	0.829	0.963	0.998	25	0.184	0.204	0.212	0.258	0.818	0.905	0.991	1.000
50	0.097	0.133	0.117	0.159	0.650	0.816	0.967	1.000	50	0.213	0.241	0.218	0.279	0.829	0.904	0.990	1.000
100	0.125	0.150	0.171	0.246	0.645	0.823	0.992	1.000	100	0.241	0.261	0.281	0.365	0.815	0.925	0.999	1.000
	$M = 100$				$M = 100$				$M = 100$				$M = 100$			
$N \setminus T$	25	50	100	200	25	50	100	200	$n \setminus T$	25	50	100	200	25	50	100	200
1	0.721	0.508	0.721	0.857	0.560	0.651	0.822	0.932	1	0.824	0.710	0.864	0.938	0.719	0.763	0.909	0.965
10	0.091	0.089	0.141	0.162	0.668	0.834	0.951	0.998	10	0.227	0.176	0.229	0.296	0.798	0.907	0.982	1.000
25	0.130	0.155	0.144	0.185	0.700	0.843	0.958	0.999	25	0.239	0.252	0.240	0.312	0.843	0.943	0.995	1.000
50	0.158	0.165	0.172	0.186	0.702	0.871	0.986	1.000	50	0.246	0.235	0.319	0.351	0.850	0.928	0.995	1.000
100	0.182	0.155	0.208	0.249	0.701	0.935	0.987	1.000	100	0.319	0.285	0.372	0.443	0.812	0.979	0.998	1.000

Notes: This table reports the finite-sample power of the Sup test conducted across multiple variables, forecasters and time-periods. N denotes the number of variables; M is the number of forecasters and T is the number of time periods. The two panels on the left present results using the 5% size-adjusted critical values from Table A2; the two panels on the right present result using the 10% size-adjusted critical values from Table A2. We consider two studentization schemes: no studentization (first and third panels) and (partial) studentization (second and fourth panels), which are both described in Example 1. To investigate the power properties, we consider the following. Of all the Mn forecasts, N forecasts are assigned to the baseline set m_0 (i.e., each of the N series has one baseline forecasts) while $(N-1)M$ forecasts are assigned to the set of alternatives, m_1. All forecast errors are generated in the same way as the simulation described in Table A1. Then, we randomly select 20% of the competing forecasts and add $(2T^{-1}\log(Mn))^{1/8}$ to their selected forecast errors, which then have larger MSE values than the baseline forecasts.

Appendix C

Appendix for Chapter 3

C.1 Proofs

This section presents proofs of the theoretical results in the main body of our paper.

C.1.1 Theorem 3.2.1

Proof. Using (3.6), we have

$$\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t}) = (u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E[u^2_{i,t+h,1} - u^2_{i,t+h,2}] + 2(u_{i,t+h,1} - u_{i,t+h,2})\lambda_i' f_{t+h}.$$

Hence, conditional on $\mathcal{F}$, $\{\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t})\}_{i=1}^{n}$ is independent across i with mean zero. By Assumption 3, the sequence $\{\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t})\}_{i=1}^{n_t}$ conditional on $\mathcal{F}$ satisfies the Lyapunov condition. Hence, a standard argument yields

$$\frac{n^{-1/2}\sum_{i=1}^{n}[\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t})]}{\sqrt{n^{-1}\sum_{i=1}^{n}[\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t})]^2}} \xrightarrow{d} N(0,1).$$

Under the null that $n^{-1}\sum_{i=1}^{n} E(\Delta L_{i,t+h|t}) = 0$, we have

$$\frac{n^{-1/2}\sum_{i=1}^{n} \Delta L_{i,t+h|t}}{\sqrt{n^{-1}\sum_{i=1}^{n}[\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t})]^2}} \xrightarrow{d} N(0,1).$$

The result now follows by noticing that $n^{-1}\sum_{i=1}^{n}[\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t})]^2 \leq n^{-1}\sum_{i=1}^{n}(\Delta L_{i,t+h|t})^2$. □

C.1.2 Theorem 3.2.2

Proof. Using (3.8), we have

$$\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t}|\mathcal{F}) = (u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E[u^2_{i,t+h,1} - u^2_{i,t+h,2}|\mathcal{F}] + 2(u_{i,t+h,1} - u_{i,t+h,2})\lambda_i' f_{t+h}.$$

Hence, conditional on $\mathcal{F}$, $\{\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F})\}_{i=1}^{n}$ is independent across i with mean zero. By Assumption 4, the sequence $\{\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F})\}_{i=1}^{n_t}$ conditional on $\mathcal{F}$ satisfies the Lyapunov condition. Hence, a standard argument yields

$$\frac{n^{-1/2}\sum_{i=1}^{n}[\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F})]}{\sqrt{n^{-1}\sum_{i=1}^{n}[\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F})]^2}} \xrightarrow{d} N(0,1).$$

Under the null that $n^{-1}\sum_{i=1}^{n} E(\Delta L_{i,t+h|t} \mid \mathcal{F}) = 0$, we have

$$\frac{n^{-1/2}\sum_{i=1}^{n} \Delta L_{i,t+h|t}}{\sqrt{n^{-1}\sum_{i=1}^{n}[\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F})]^2}} \xrightarrow{d} N(0,1).$$

The result now follows by noticing that $n^{-1}\sum_{i=1}^{n}[\Delta L_{i,t+h|t} - E(\Delta L_{i,t+h|t} \mid \mathcal{F})]^2 \leq n^{-1}\sum_{i=1}^{n}(\Delta L_{i,t+h|t})^2$. □

C.1.3 Theorem 3.3.1

Proof. Start by noticing that

$$\sqrt{n_k}\left(\overline{\Delta u}^2_{t+h,k} - n_k^{-1}\sum_{i\in H_k} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})\right)$$
$$= n_k^{-1/2}\sum_{i\in H_k}\left[(u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) + 2(\lambda_1' f_{t+h} u_{i,t+h,1} - \lambda_2' f_{t+h} u_{i,t+h,2})\right]$$
$$+ n_k^{-1/2}\left(n_k^{-1}\sum_{i\in H_k} u_{i,t+h,2}\right)^2 - n_k^{-1/2}\left(n_k^{-1}\sum_{i\in H_k} u_{i,t+h,1}\right)^2.$$

By a CLT,

$$n_k^{-1/2}\left(n_k^{-1}\sum_{i\in H_k} u_{i,t+h,2}\right)^2 - n_k^{-1/2}\left(n_k^{-1}\sum_{i\in H_k} u_{i,t+h,1}\right)^2 = O_P(n_k^{-3/2}) = o_P(1).$$

Therefore, $\overline{\Delta u}^2_{t+h,k}$ is a $\sqrt{n_k}$-consistent estimator for $n_k^{-1}\sum_{i\in H_k} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})$. By the same CLT argument, it follows that

$$\sqrt{n_k}\left(\overline{\Delta u}^2_{t+h,k} - n_k^{-1}\textstyle\sum_{i\in H_k} E(u^2_{i,t+h,1} - u^2_{i,t+h,2}|\mathcal{F})\right)$$
$$= n_k^{-1/2}\textstyle\sum_{i\in H_k}\left[(u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2}|\mathcal{F}) + 2(\lambda_1' f_{t+h} u_{i,t+h,1} - \lambda_2' f_{t+h} u_{i,t+h,2})\right] + o_P(1)$$

is asymptotically normal and that the variance of $\sqrt{n_k}\left(\overline{\Delta u}^2_{t+h,k} - n_k^{-1}\sum_{i\in H_k} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})\right)$ can be estimated by

$$\hat{V} := n_k^{-1}\sum_{i\in H_k}\left[(u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) + 2(\lambda_1' f_{t+h} u_{i,t+h,1} - \lambda_2' f_{t+h} u_{i,t+h,2})\right]^2.$$

Recall our estimate $\bar{V} := n_k^{-1}\sum_{i\in H_k}(\Delta L_{i,t+h} - \overline{\Delta L}_{t+h,k})^2$ with $\overline{\Delta L}_{t+h,k} = n_k^{-1}\sum_{i\in H_k}\Delta L_{i,t+h}$.

It remains to show that $\hat{V} = \bar{V} + o_P(1)$. By (3.8) and (3.9), we have

$$
\begin{aligned}
&\Delta L_{i,t+h} - \overline{\Delta L}_{t+h,k} \\
&= \left[u^2_{i,t+h,1} - u^2_{i,t+h,2} + 2(\lambda'_{i,1} f_{t+h} u_{i,t+h,1} - \lambda'_{i,2} f_{t+h} u_{i,t+h,2})\right] \\
&\quad - n_k^{-1} \sum_{j \in H_k} \left[\left(u^2_{j,t+h,1} - u^2_{j,t+h,2}\right) + 2(\lambda'_{j,1} f_{t+h} u_{j,t+h,1} - \lambda'_{j,2} f_{t+h} u_{j,+ht,2})\right] \\
&= (u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) + 2(\lambda'_1 f_{t+h} u_{i,t+h,1} - \lambda'_2 f_{t+h} u_{i,t+h,2}) + h_{n,1} + h_{n,2},
\end{aligned}
$$

where $h_{n,1} = E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) - n_k^{-1} \sum_{j \in H_k} (u^2_{j,t+h,1} - u^2_{j,t+h,2})$ and $h_{n,2} = -2n_k^{-1} \sum_{j \in H_k} (\lambda'_{j,1} f_{t+h} u_{j,t+h,1} - \lambda'_{j,2} f_{t+h} u_{j,+ht,2})$. Clearly, by a LLN, $h_{n,1} = o_P(1)$ and $h_{n,2} = o_P(1)$. By the elementary inequality $\left|\sqrt{\sum (a_i + b_i)^2} - \sqrt{\sum a_i^2}\right| \leq \sqrt{\sum b_i^2}$, we have that

$$
\left|\sqrt{\hat{V}} - \sqrt{\bar{V}}\right| \leq \sqrt{n_k^{-1} \sum_{i \in H_k} (h_{n,1} + h_{n,2})^2} = |h_{n,1} + h_{n,2}| = o_P(1).
$$

Thus, $\hat{V} = \bar{V} + o_P(1)$. The proof is complete. □

C.1.4 Corollary 3.3.1

Proof. The result follows once we notice that

$$
\begin{aligned}
&\sqrt{n}\left(\sum_{k=1}^{K} \frac{n_k}{n} \overline{\Delta u}^2_{t+h,k} - n^{-1} \sum_{i=1}^{n} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})\right) \\
&= \sqrt{n} \sum_{k=1}^{K} \frac{n_k}{n} \left(\overline{\Delta u}^2_{t+h,k} - n_k^{-1} \sum_{i \in H_k} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})\right) \\
&= \sum_{k=1}^{K} \frac{n_k}{\sqrt{n}} \left\{ n_k^{-1} \sum_{i \in H_k} \left[(u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})\right] + O_P(n_k^{-1}) \right\} \\
&= n^{-1/2} \sum_{i=1}^{n} \left[(u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})\right] + O_P(n^{-1/2}).
\end{aligned}
$$

□

C.1.5 Theorem 3.3.2

Proof. By equation (3.15), we have

$$\begin{aligned}
&\sqrt{n}\left[\sum_{k=1}^{K}\frac{n_k}{n}(\overline{e}^2_{1,k,t+h}-\overline{e}^2_{2,k,t+h})-\sum_{k=1}^{K}\frac{n_k}{n}\left((\lambda'_{1,(k)}f_{t+h})^2-(\lambda'_{2,(k)}f_{t+h})^2\right)\right]\\
&=n^{-1/2}\sum_{k=1}^{K}n_k\left\{\left(n_k^{-1}\sum_{i\in H_k}u_{i,t+h,1}\right)^2-\left(n_k^{-1}\sum_{i\in H_k}u_{i,t+h,2}\right)^2\right\}\\
&\quad+2n^{-1/2}\sum_{i=1}^{n}\left(\lambda'_{i,1}f_{t+h}u_{i,t+h,1}-\lambda'_{i,2}f_{t+h}u_{i,t+h,2}\right).
\end{aligned}$$

Again as in the proof of Theorem 3.3.1, we can show that $n^{-1/2}\sum_{k=1}^{K}n_k\left\{\left(n_k^{-1}\sum_{i\in H_k}u_{i,t+h,1}\right)^2-\left(n_k^{-1}\sum_{i\in H_k}u_{i,t+h,2}\right)^2\right\}=o_P(1)$, and so

$$\begin{aligned}
&\sqrt{n}\left[\sum_{k=1}^{K}\frac{n_k}{n}(\overline{e}^2_{1,k,t+h}-\overline{e}^2_{2,k,t+h})-n^{-1}\sum_{i=1}^{n}\left[(\lambda'_{i,1}f_{t+h})^2-(\lambda'_{i,2}f_{t+h})^2\right]\right]\\
&\qquad=2n^{-1/2}\sum_{i=1}^{n}\left(\lambda'_{i,1}f_{t+h}u_{i,+ht,1}-\lambda'_{i,2}f_{t+h}u_{i,t+h,2}\right)+o_P(1).
\end{aligned}$$

The rest of the proof follows by a CLT as in the proof of Theorem 3.3.1. □

C.1.6 Lemma 3.3.1

Proof. Since we can write $f_{s+h}=(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}(\bar{e}_{s+h}-\bar{u}_{s+h})$, we have $e_{i,s+h}=\lambda'_i(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\bar{e}_{s+h}+u_{i,s+h}-\lambda'_i(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\bar{u}_{s+h}$. It is not difficult to see that

$$\begin{aligned}
\hat{\lambda}'_i&=\left(\sum_{s+h=1}^{T}[\lambda'_i(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\bar{e}_{s+h}+u_{i,s+h}-\lambda'_i(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\bar{u}_{s+h}]\bar{e}'_{s+h}\right)\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}'_{s+h}\right)^{-1}\\
&=\lambda'_i(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}+\left(\sum_{s+h=1}^{T}u_{i,s+h}\bar{e}'_{s+h}\right)\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}'_{s+h}\right)^{-1}\\
&\qquad-\lambda'_i(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\left(\sum_{s+h=1}^{T}\bar{u}_{s+h}\bar{e}'_{s+h}\right)\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}'_{s+h}\right)^{-1}
\end{aligned}$$

and thus

$$\hat{\lambda}_i'\bar{e}_{t+h} = \lambda_i' f_{t+h} + \xi_{i,t+h} + \varepsilon_{i,t+h} + \zeta_{i,t+h},$$

where $\xi_{i,t+h} = \lambda_i'(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\bar{u}_{t+h}$, $\varepsilon_{i,t+h} = \left(\sum_{s+h=1}^{T} u_{i,s+h}\bar{e}_{s+h}'\right)\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}_{s+h}'\right)^{-1}\bar{e}_{t+h}$ and $\zeta_{i,t+h} = -\lambda_i'(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\left(\sum_{s+h=1}^{T}\bar{u}_{s+h}\bar{e}_{s+h}'\right)\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}_{s+h}'\right)^{-1}\bar{e}_{t+h}$.

Next, observe that

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2 - (\lambda_{i,1}'f_{t+h})^2] \\
&= n^{-1/2}\sum_{i=1}^{n}\left(\xi_{i,t+h,1} + \varepsilon_{i,t+h,1} + \zeta_{i,t+h,1}\right)^2 + 2n^{-1/2}\sum_{i=1}^{n}\left(\xi_{i,t+h,1} + \varepsilon_{i,t+h,1} + \zeta_{i,t+h,1}\right)\lambda_{i,1}'f_{t+h}.
\end{aligned}$$

and

$$n^{-1/2}\sum_{i=1}^{n}\xi_{i,t+h,1}\lambda_{i,1}'f_{t+h} = n^{-1/2}\bar{u}_{t+h}'\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left(\sum_{i=1}^{n}\lambda_{i,1}\lambda_{i,1}'\right)f_{t+h}.$$

Therefore, we have

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}[(\hat{\lambda}_{i,1}'\bar{e}_{t+h})^2 - (\lambda_{i,1}'f_{t+h})^2] \\
&= n^{-1/2}\bar{u}_{t+h}'\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left(\sum_{i=1}^{n}\lambda_{i,1}\lambda_{i,1}'\right)f_{t+h} \\
&\quad + n^{-1/2}\sum_{i=1}^{n}\left(\xi_{i,t+h,1} + \varepsilon_{i,t+h,1} + \zeta_{i,t+h,1}\right)^2 + 2n^{-1/2}\sum_{i=1}^{n}\left(\varepsilon_{i,t+h,1} + \zeta_{i,t+h,1}\right)\lambda_{i,1}'f_{t+h}.
\end{aligned}$$

The rest of the proof proceeds in four steps, bounding different components in the above display.

Step 1: show that $n^{-1/2}\sum_{i=1}^{n}\left(\varepsilon_{i,t+h,1} + \zeta_{i,t+h,1}\right)\lambda_{i,1}'f_{t+h} = o_P(1)$.

We observe that

$$n^{-1/2}\sum_{i=1}^{n}\zeta_{i,t+h,1}\lambda_{i,1}'f_{t+h}$$

$$= -n^{-1/2}\bar{e}'_{t+h}\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}'_{s+h}\right)^{-1}\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{u}'_{s+h}\right)\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left(\sum_{i=1}^{n}\lambda_{i,1}\lambda'_{i,1}\right)f_{t+h} \quad \text{(C.1)}$$

and

$$n^{-1/2}\sum_{i=1}^{n}\varepsilon_{i,t+h,1}\lambda'_{i,1}f_{t+h} = n^{-1/2}f'_{t+h}\left(\sum_{s+h=1}^{T}\left(\sum_{i=1}^{n}\lambda_{i,1}u_{i,s+h,1}\right)\bar{e}'_{s+h}\right)\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}'_{s+h}\right)^{-1}\bar{e}_{t+h}. \quad \text{(C.2)}$$

Recall that $\bar{e}_{s+h} = \bar{\lambda}'f_{s+h} + \bar{u}_{s+h}$. Since $\bar{\lambda}\bar{\lambda}'$ has eigenvalues bounded away from zero and infinity and f_{s+h} has non-trivial variance, it follows that $\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}'_{s+h}\right)^{-1} = O_P(T^{-1})$. Moreover,

$$\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{u}'_{s+h} = \bar{\lambda}'\sum_{s+h=1}^{T}f_{s+h}\bar{u}'_{s+h} + \sum_{s+h=1}^{T}\bar{u}_{s+h}\bar{u}'_{s+h}. \quad \text{(C.3)}$$

Notice that $\sum_{s+h=1}^{T}f_{s+h}\bar{u}'_{s+h} = n^{-1}\sum_{i=1}^{n}(\sum_{s+h=1}^{T}f_{s+h}u'_{i,s+h})$ and $T^{-1/2}\sum_{s+h=1}^{T}f_{s+h}u'_{i,s+h}$ has mean zero with bounded variance and is independent across i conditional on $\{f_{s+h}\}_{s+h=1}^{T}$. Therefore, $\sum_{s+h=1}^{T}f_{s+h}\bar{u}'_{s+h} = O_P(\sqrt{T/n})$. Since $E\sum_{s+h=1}^{T}\bar{u}_{s+h}\bar{u}'_{s+h} = O(T/n)$, we have $\sum_{s+h=1}^{T}\bar{u}_{s+h}\bar{u}'_{s+h} = O(T/n)$. By (C.3), we have $\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{u}'_{s+h} = O_P(\sqrt{T/n}+T/n)$. Therefore, (C.1) implies

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\zeta_{i,t+h,1}\lambda'_{i,1}f_{t+h} \\
&= -n^{-1/2}\bar{e}'_{t+h}\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}'_{s+h}\right)^{-1}\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{u}'_{s+h}\right)\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left(\sum_{i=1}^{n}\lambda_{i,1}\lambda'_{i,1}\right)f_{t+h} \\
&= n^{-1/2}O_P(1)\cdot O_P(T^{-1})\cdot O_P(\sqrt{T/n}+T/n)\cdot O_P(1)\cdot O_P(n)\cdot O_P(1) \\
&= O_P(n^{-1/2}+T^{-1/2}) = o_P(1). \quad \text{(C.4)}
\end{aligned}$$

We observe that

$$\sum_{s+h=1}^{T}\left(\sum_{i=1}^{n}\lambda_{i,1}u_{i,s+h,1}\right)\bar{e}'_{s+h}$$

$$
\begin{aligned}
&= \sum_{s+h=1}^{T}\sum_{i=1}^{n}\lambda_{i,1}u_{i,s+h,1}\left(\bar{u}'_{s+h}+f'_{s+h}\bar{\lambda}\right)\\
&= \sum_{s+h=1}^{T}\sum_{i=1}^{n}\lambda_{i,1}u_{i,s+h,1}\bar{u}'_{s+h}+\sum_{s+h=1}^{T}\sum_{i=1}^{n}\lambda_{i,1}u_{i,s+h,1}f'_{s+h}\bar{\lambda}\\
&\overset{\text{(i)}}{=} \sum_{s+h=1}^{T}\sum_{i=1}^{n}\lambda_{i,1}u_{i,s+h,1}\bar{u}'_{s+h}+O_P(\sqrt{nT})\\
&= O_P\left(\sqrt{\sum_{s+h=1}^{T}\left(\sum_{i=1}^{n}\lambda_{i,1}u_{i,s+h,1}\right)^2}\times\sqrt{\sum_{s+h=1}^{T}\|\bar{u}_{s+h}\|^2}\right)+O_P(\sqrt{nT})\\
&\overset{\text{(ii)}}{=} O_P\left(\sqrt{Tn}\times\sqrt{T/n}\right)+O_P(\sqrt{nT}) = O_P\left(\sqrt{nT}(1+\sqrt{T/n})\right),
\end{aligned}
$$

where (i) and (ii) follow by Assumption 5. Hence, by (C.2), we have

$$
\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\varepsilon_{i,t+h,1}\lambda'_{i,1}f_{t+h}\\
&= n^{-1/2}f'_{t+h}\left(\sum_{s+h=1}^{T}\left(\sum_{i=1}^{n}\lambda_{i,1}u_{i,s+h,1}\right)\bar{e}'_{s+h}\right)\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}'_{s+h}\right)^{-1}\bar{e}_{t+h}\\
&= n^{-1/2}\cdot O_P(1)\cdot O_P\left(\sqrt{nT}(1+\sqrt{T/n})\right)\cdot O_P(T^{-1})\cdot O_P(1)\\
&= O_P(T^{-1/2}+n^{-1/2}) = o_P(1).
\end{aligned}
\tag{C.5}
$$

By (C.4) and (C.5), we have proved the claim in Step 1.

Step 2: show that $n^{-1/2}\sum_{i=1}^{n}\xi^2_{i,t+h,1} = o_P(1)$.

Clearly $E\bar{u}_{t+h}\bar{u}'_{t+h} = O(n^{-1})$ and thus $\bar{u}_{t+h}\bar{u}'_{t+h} = O_P(n^{-1})$. It follows that

$$
\begin{aligned}
\sum_{i=1}^{n}\xi_{i,t+h}\xi'_{i,t+h} &= \sum_{i=1}^{n}\lambda'_i(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\bar{u}_{t+h}\bar{u}'_{t+h}\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\lambda_i\\
&= \sum_{i=1}^{n}\text{trace}\left(\bar{u}_{t+h}\bar{u}'_{t+h}\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\lambda_i\lambda'_i(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\right)\\
&= \text{trace}\left(\bar{u}_{t+h}\bar{u}'_{t+h}\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left[\sum_{i=1}^{n}\lambda_i\lambda'_i\right](\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\right)
\end{aligned}
$$

$$= \text{trace } \left(O_P(n^{-1}) \cdot O_P(1) \cdot O_P(n) \cdot O_P(1)\right) = O_P(1).$$

Therefore, $n^{-1/2}\sum_{i=1}^{n} \xi_{i,t+h,1}^2 = O_P(n^{-1/2}) = o_P(1)$.

Step 3: show that $n^{-1/2}\sum_{i=1}^{n} \varepsilon_{i,t+h,1}^2 = o_P(1)$.

Let $q_{n,1} = \left(\sum_{s+h=1}^{T} \bar{e}_{s+h}\bar{e}'_{s+h}\right)^{-1} \bar{e}_{t+h}$. Then

$$\varepsilon_{i,t+h} = \left(\sum_{s+h=1}^{T} u_{i,s+h}\bar{e}'_{s+h}\right) q_{n,1} = \sum_{s+h=1}^{T} u_{i,s+h}\bar{u}'_{s+h}q_{n,1} + \sum_{s+h=1}^{T} u_{i,s+h}f'_{s+h}\bar{\lambda}q_{n,1}.$$

Therefore,

$$n^{-1/2}\sum_{i=1}^{n} \|\varepsilon_{i,t+h}\|^2 \leq 2n^{-1/2}\sum_{i=1}^{n} \left\| \sum_{s+h=1}^{T} u_{i,s+h}\bar{u}'_{s+h}q_{n,1} \right\|^2 + 2n^{-1/2}\sum_{i=1}^{n} \left\| \sum_{s+h=1}^{T} u_{i,s+h}f'_{s+h}\bar{\lambda}q_{n,1} \right\|^2.$$

From the previous argument, $\|q_{n,1}\| = O_P(T^{-1})$. It follows that

$$n^{-1/2}\textstyle\sum_{i=1}^{n} \|\varepsilon_{i,t+h}\|^2 = O_P\left(T^{-2}n^{-1/2}\left[\sum_{i=1}^{n} \left\|\sum_{s+h=1}^{T} u_{i,s+h}\bar{u}'_{s+h}\right\|^2 + \sum_{i=1}^{n} \left\|\sum_{s+h=1}^{T} u_{i,s+h}f'_{s+h}\right\|^2\right]\right). \tag{C.6}$$

We observe that

$$\begin{aligned}
&\text{trace } \left[\sum_{i=1}^{n} \left(\sum_{\tau+h=1}^{T} \bar{u}_{\tau+h}u'_{i,\tau+h}\right)\left(\sum_{s+h=1}^{T} u_{i,s+h}\bar{u}'_{s+h}\right)\right] \\
&= \text{trace } \left[\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T} \bar{u}_{\tau+h}\left(\sum_{i=1}^{n} u'_{i,\tau+h}u_{i,s+h}\right)\bar{u}'_{s+h}\right] \\
&= \sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\left(\sum_{i=1}^{n} u'_{i,\tau+h}u_{i,s+h}\right)\bar{u}'_{s+h}\bar{u}_{\tau+h} \\
&\leq \sqrt{\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\left(\sum_{i=1}^{n} u'_{i,\tau+h}u_{i,s+h}\right)^2} \times \sqrt{\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}(\bar{u}'_{s+h}\bar{u}_{\tau+h})^2}.
\end{aligned}$$

Notice that

$$
\begin{aligned}
&E\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\left(\sum_{i=1}^{n}u'_{i,\tau+h}u_{i,s+h}\right)^2\\
&=\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}Eu'_{i_1,\tau+h}u_{i_1,s+h}u'_{i_2,\tau+h}u_{i_2,s+h}\\
&=\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\sum_{i=1}^{n}E(u'_{i,\tau+h}u_{i,s+h})^2+\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\sum_{i_1\neq i_2}Eu'_{i_1,\tau+h}u_{i_1,s+h}u'_{i_2,\tau+h}u_{i_2,s+h}\\
&\overset{\text{(i)}}{=}\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\sum_{i=1}^{n}E(u'_{i,\tau+h}u_{i,s+h})^2+\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\sum_{i_1\neq i_2}Eu'_{i_1,\tau+h}u_{i_1,s+h}Eu'_{i_2,\tau+h}u_{i_2,s+h}\\
&\leq\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\sum_{i=1}^{n}E(u'_{i,\tau+h}u_{i,s+h})^2+\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\left(\sum_{i=1}^{n}Eu'_{i,\tau+h}u_{i,s+h}\right)^2\\
&=O(nT^2)+O\left(n^2\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\|\gamma_n(s,\tau)\|^2\right)\\
&\overset{\text{(ii)}}{=}O(nT^2)+O\left(n^2\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\|\gamma_n(s,\tau)\|\right)\overset{\text{(iii)}}{=}O(nT^2)+O(n^2T),
\end{aligned}
$$

where (i) follows by the independence of $u_{i,s}$ across i, (ii) follows by $\max_s\max_\tau\|\gamma_n(s,\tau)\|\leq M$ and (iii) follows by Assumption 6. On the other hand, we have

$$
\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}(\bar{u}'_{s+h}\bar{u}_{\tau+h})^2\leq\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\|\bar{u}_{s+h}\|^2\cdot\|\bar{u}_{\tau+h}\|^2=\left(\sum_{s+h=1}^{T}\|\bar{u}_{s+h}\|^2\right)^2=O_P(T^2n^{-1}).
$$

The above three displays imply that

$$
\begin{aligned}
&\text{trace}\left[\sum_{i=1}^{n}\left(\sum_{\tau+h=1}^{T}\bar{u}_{\tau+h}u'_{i,\tau+h}\right)\left(\sum_{s+h=1}^{T}u_{i,s+h}\bar{u}'_{s+h}\right)\right]\\
&\leq\sqrt{\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}\left(\sum_{i=1}^{n}u'_{i,\tau+h}u_{i,s+h}\right)^2}\times\sqrt{\sum_{s+h=1}^{T}\sum_{\tau+h=1}^{T}(\bar{u}'_{s+h}\bar{u}_{\tau+h})^2}\\
&=\sqrt{O_P(nT^2)+O_P(n^2T)}\times\sqrt{O_P(T^2n^{-1})}=O_P\left(T^{3/2}(n^{1/2}+T^{1/2})\right).
\end{aligned}
$$

Since $\sum_{i=1}^{n}\left(\sum_{\tau+h=1}^{T}\bar{e}_{\tau+h}u'_{i,\tau+h}\right)\left(\sum_{s+h=1}^{T}u_{i,s+h}\bar{e}'_{s+h}\right)$ is positive semi-definite, we have

$$\sum_{i=1}^{n}\left(\sum_{\tau+h=1}^{T}\bar{u}_{\tau+h}u'_{i,\tau+h}\right)\left(\sum_{s+h=1}^{T}u_{i,s+h}\bar{u}'_{s+h}\right)=O_P\left(T^{3/2}(n^{1/2}+T^{1/2})\right).$$

By a CLT, we have

$$\sum_{i=1}^{n}\left\|\sum_{s+h=1}^{T}u_{i,s+h}f'_{s+h}\right\|^2=O_P(nT).$$

The above two displays and (C.6) imply

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\|\varepsilon_{i,t+h}\|^2\\
&=O_P\left(T^{-2}n^{-1/2}\left[\sum_{i=1}^{n}\left\|\sum_{s+h=1}^{T}u_{i,s+h}\bar{u}'_{s+h}\right\|^2+\sum_{i=1}^{n}\left\|\sum_{s+h=1}^{T}u_{i,s+h}f'_{s+h}\right\|^2\right]\right)\\
&=O_P\left(T^{-2}n^{-1/2}\left[O_P\left(T^{3/2}(n^{1/2}+T^{1/2})\right)+O_P(nT)\right]\right)\\
&=O_P\left(n^{-1/2}+T^{-1/2}+n^{1/2}T^{-1}\right)\overset{\text{(i)}}{=}o_P(1),
\end{aligned}$$

where (i) follows by $n/T^2=o(1)$.

Step 4: show that $n^{-1/2}\sum_{i=1}^{n}\zeta_{i,t+h,1}^2=o_P(1)$.

Let $q_{n,2}=(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda}\left(\sum_{s+h=1}^{T}\bar{u}_{s+h}\bar{e}'_{s+h}\right)\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}'_{s+h}\right)^{-1}\bar{e}_{t+h}$. Then $\zeta_{i,t+h}=-\lambda'_i q_{n,2}$. It follows that

$$n^{-1/2}\sum_{i=1}^{n}\|\zeta_{i,t+h}\|^2=n^{-1/2}\sum_{i=1}^{n}q'_{n,2}\lambda_i\lambda'_i q_{n,2}\leq O_P\left(n^{1/2}\|q_{n,2}\|^2\right). \tag{C.7}$$

By the previous argument, $\left(\sum_{s+h=1}^{T}\bar{e}_{s+h}\bar{e}'_{s+h}\right)^{-1}\bar{e}_{t+h}=O_P(T^{-1})$. Notice that

$$\sum_{s+h=1}^{T}\bar{u}_{s+h}\bar{e}'_{s+h}=\sum_{s+h=1}^{T}\bar{u}_{s+h}\bar{u}'_{s+h}+\sum_{s+h=1}^{T}\bar{u}_{s+h}f'_{s+h}\bar{\lambda}.$$

It is simple to show that $\sum_{s+h=1}^{T} \bar{u}_{s+h}\bar{u}'_{s+h} = O_P(T/n)$ and $\sum_{s+h=1}^{T} \bar{u}_{s+h}f'_{s+h}\bar{\lambda} = O_P(\sqrt{T/n})$. Therefore, $\|q_{n,2}\| = O_P\left(T/n + \sqrt{T/n}\right) \cdot O_P(T^{-1})$. By (C.7), we have

$$n^{-1/2}\sum_{i=1}^{n} \|\zeta_{i,t+h}\|^2 = O_P\left(n^{1/2}\|q_{n,2}\|^2\right) = O_P\left(n^{-3/2} + n^{-1/2}T^{-1}\right) = o_P(1).$$

This completes the proof. □

C.1.7 Theorem 3.3.3

Proof. By (3.12) and Lemma 3.3.1, we have that

$$\begin{aligned}
&\sqrt{n}\left[\overline{\Delta\hat{u}}^2_{t+h} - n^{-1}\sum_{i=1}^{n} E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F})\right] \\
&= n^{-1/2}\sum_{i=1}^{n}\Big[(u^2_{i,t+h,1} - u^2_{i,t+h,2}) - E(u^2_{i,t+h,1} - u^2_{i,t+h,2} \mid \mathcal{F}) \\
&\quad + 2(\lambda'_{i,1}f_{t+h}u_{i,t+h,1} - \lambda'_{i,2}f_{t+h}u_{i,t+h,2}) + u'_{i,t+h}D_{t+h}\Big] + o_P(1), \qquad \text{(C.8)}
\end{aligned}$$

where $D_{t+h} = \bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\left(n^{-1}\sum_{i=1}^{n}[\lambda_{i,1}\lambda'_{i,1} - \lambda_{i,2}\lambda'_{i,2}]\right) f_{t+h}$. Since $\hat{\lambda}'_i - \lambda'_i(\bar{\lambda}\bar{\lambda}')^{-1}\bar{\lambda} = o_P(1)$, $\bar{e}_{t+h} = \bar{\lambda}'f_{t+h} + o_P(1)$ and $(\bar{\lambda}\bar{\lambda}')^{-1}$ exists asymptotically, we have $\hat{D}_{t+h} = D_{t+h} + o_P(1)$. Since $\{u_{i,t+h,m}\}_{i=1}^{n}$ is independent across i, the result then follows by the classical CLT and a self-normalized CLT; see e.g., Theorem 4.1 of Chen, Shao and Wu (2016); Peña, Lai and Shao (2008). □

C.1.8 Theorem 3.3.4

Proof. By Lemma 3.3.1, we have that under the null hypothesis,

$$n^{-1/2}\sum_{i=1}^{n}\left[(\hat{\lambda}'_{i,1}\bar{e}_{t+h})^2 - (\hat{\lambda}'_{i,2}\bar{e}_{t+h})^2\right] = 2n^{-1/2}\bar{u}'_{t+h}\bar{\lambda}'(\bar{\lambda}\bar{\lambda}')^{-1}\sum_{i=1}^{n}\left(\lambda_{i,1}\lambda'_{i,1} - \lambda_{i,2}\lambda'_{i,2}\right) f_{t+h} + o_P(1).$$

Since $\{u_{i,t+h,m}\}_{i=1}^{n}$ is independent across i, the result then follows by the classical CLT and a self-normalized CLT; see e.g., Theorem 4.1 of Chen, Shao and Wu (2016); Peña, Lai and Shao (2008). □

C.1.9 Lemma 3.3.2

Proof. Under Assumptions A-F and Theorem 3 in Bai (2003), recall that the following result holds:

$$\hat{\lambda}_i'\hat{f}_{t+h} - \lambda_i' f_{t+h} = n^{-1}\lambda_i' \left(n^{-1}\sum_{j=1}^{n}\lambda_j\lambda_j' \right)^{-1} \sum_{j=1}^{n}\lambda_j u_{jt+h} + T^{-1} f_{t+h}' \left(T^{-1}\sum_{s+h=1}^{T} f_{s+h}f_{s+h}' \right)^{-1} \sum_{s+h=1}^{T} f_{s+h}u_{is+h} + O_P(1/\min\{n,T\}).$$

Using this result, we have $\hat{\lambda}_{i,m}' f_{t+h} = \lambda_{i,m}' f_{t+h} + \xi_{i,t+h,m}$ for $m \in \{1,2\}$, where

$$\xi_{i,t+h,m} = n^{-1}\lambda_i' \left(n^{-1}\sum_{j=1}^{n}\lambda_j\lambda_j' \right)^{-1} \sum_{j=1}^{n}\lambda_{j,m} u_{jt+h,m} + T^{-1} f_{t+h}' \left(T^{-1}\sum_{s+h=1}^{T} f_{s+h}f_{s+h}' \right)^{-1} \sum_{s+h=1}^{T} f_{s+h}u_{is+h,m} + O_P(1/\min\{n,T\}).$$

It follows that

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\left[(\hat{\lambda}_{i,1}'\hat{f}_{t+h})^2 - (\hat{\lambda}_{i,2}'\hat{f}_{t+h})^2 \right] &= n^{-1}\sum_{i=1}^{n}[(\lambda_{i,1}' f_{t+h})^2 - (\lambda_{i,2}' f_{t+h})^2] \\ &\quad + 2n^{-1}\sum_{i=1}^{n}[\lambda_{i,1}' f_{t+h}\xi_{i,t+h,1} - \lambda_{i,2}' f_{t+h}\xi_{i,t+h,2}] \\ &\quad + n^{-1}\sum_{i=1}^{n}[\xi_{i,t+h,1}^2 - \xi_{i,t+h,2}^2]. \end{aligned}$$

The last term is of order $1/\min\{n,T\}$, which is negligible if $\sqrt{n}/T = o(1)$. Under assumptions of weak (cross-sectional and serial) dependence in $u_{i,t+h,m}$ (e.g., Assumptions E and

F in Bai (2003)), we can show that

$$n^{-1}\sum_{i=1}^{n}\lambda'_{i,m}f_{t+h}\xi_{i,t+h,m} = n^{-1}\sum_{i=1}^{n}\lambda'_{i,m}f_{t+h}u_{i,t+h,m} + o_P(n^{-1/2}).$$

Using this, it follows that

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\left[(\hat{\lambda}'_{i,1}\hat{f}_{t+h})^2 - (\hat{\lambda}'_{i,2}\hat{f}_{t+h})^2\right] &= n^{-1}\sum_{i=1}^{n}[(\lambda'_{i,1}f_{t+h})^2 - (\lambda'_{i,2}f_{t+h})^2] \\ &\quad + 2n^{-1}\sum_{i=1}^{n}[\lambda'_{i,1}f_{t+h}u_{i,t+h,1} - \lambda'_{i,2}f_{t+h}u_{i,t+h,2}] + o_P(n^{-1/2}). \end{aligned}$$

The stated result follows from this. □

Table C.1: Coverage probabilities for 95% confidence intervals constructed to test the null of equal squared biases (5-cluster DGP)

	clustering						
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	88.1	86.6	84.2	83.8	81.8	78.5	77.1
25	97.3	96.0	95.0	94.2	94.2	91.6	90.6
50	98.2	97.5	96.1	95.9	95.0	94.3	93.9
100	98.2	97.4	96.8	95.6	95.0	94.5	94.6
200	98.4	96.5	95.8	95.3	94.9	94.6	95.1
1000	96.9	96.2	96.1	95.8	95.5	95.8	95.3

	CCE						
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	98.5	97.5	95.2	94.6	93.6	90.7	90.6
25	98.6	97.1	95.7	95.5	93.3	93.1	91.2
50	98.4	96.6	94.9	94.5	93.4	92.9	93.3
100	96.6	95.6	94.5	93.7	93.5	93.8	93.2
200	95.7	93.3	92.9	92.8	92.7	92.4	93.9
1000	94.3	93.0	93.5	93.2	93.1	93.7	93.4

	PCA						
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	99.7	99.4	98.7	99.5	99.7	99.8	99.9
25	99.7	99.2	99.9	99.9	99.8	100.0	100.0
50	99.2	99.7	100.0	100.0	99.9	100.0	99.9
100	98.8	99.7	99.9	100.0	100.0	100.0	100.0
200	99.7	99.8	100.0	100.0	100.0	100.0	100.0
1000	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Note: This table reports the coverage probability for a 95% confidence interval for the test of equal squared biases, using the Monte Carlo simulation setup described in Section 5.1 and 2,000 random samples. n refers to the number of cross-sectional units used in the pair-wise comparison of loss differences, while ρ^2 measures the predictive power of the underlying forecasts. We show coverage probabilities for the clustering, CCE, and PCA methods described in Section 3. The assumed time-series dimension is $T = 80$. The underlying data generating process assumes factor loadings that follow a cluster structure with 5 clusters and $n/5$ elements in each cluster.

Table C.2: Coverage probabilities for 95% confidence intervals constructed to test the null of equal idiosyncratic error variances (5-cluster DGP)

	clustering						
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	98.0	98.0	98.4	98.8	99.0	99.5	99.8
25	96.4	97.1	98.3	98.8	98.9	99.8	99.7
50	96.4	96.5	97.6	98.7	99.1	99.8	99.8
100	96.8	96.6	97.9	98.4	99.0	99.6	99.8
200	96.4	97.1	98.5	98.6	98.9	99.1	99.7
1000	95.7	96.7	97.9	97.9	98.7	99.4	99.5

	CCE						
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	95.8	95.4	96.6	96.6	97.0	98.4	98.9
25	95.4	96.8	97.4	97.5	97.7	98.8	99.3
50	95.8	95.9	96.6	97.6	98.3	99.3	99.4
100	97.0	96.6	97.5	98.2	98.1	99.0	99.6
200	95.8	96.2	97.5	98.5	98.4	98.4	99.2
1000	95.7	96.7	97.9	97.0	98.3	98.9	98.9

	PCA						
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	92.3	91.2	91.3	92.9	93.2	92.3	92.3
25	93.1	94.4	94.9	94.7	94.2	94.4	95.2
50	94.2	93.6	94.2	95.2	95.5	96.0	95.1
100	95.6	94.8	95.4	95.5	95.5	95.6	95.6
200	94.8	95.5	95.8	96.1	95.8	94.8	95.7
1000	95.1	94.8	95.8	94.1	96.4	95.3	95.7

Note: This table reports the coverage probabilities for 95% confidence intervals for the test of equal idiosyncratic variances using the Monte Carlo simulation setup described in Section 5.1 and 2,000 random samples. n refers to the number of cross-sectional units used in the pairwise comparison of loss differences, while ρ^2 measures the predictive power of the underlying forecasts. We show coverage probabilities for the clustering, CCE, and PCA methods described in Section 3. The assumed time-series dimension is $T = 80$. The underlying data generating process assumes that factor loadings follow a cluster structure with 5 clusters and $n/5$ elements in each cluster.

Table C.3: Coverage probability of 95% confidence intervals: 2 factors with heterogeneous loadings

CCE: squared bias							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	98.2	95.4	93.2	94.0	92.1	92.0	91.8
25	97.7	94.7	93.1	94.2	93.7	94.3	93.8
50	95.4	94.2	92.5	93.5	94.6	94.3	94.7
100	94.7	94.3	93.7	94.8	95.1	94.6	94.2
200	92.9	94.2	94.5	93.4	94.4	95.0	95.0
1000	93.7	94.2	94.8	93.9	94.0	95.4	94.1

PCA: squared bias							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	97.3	95.1	89.9	88.4	86.0	85.7	84.6
25	95.7	91.9	89.4	90.7	90.5	90.1	90.0
50	93.6	88.7	89.0	91.1	93.0	92.1	93.0
100	90.1	89.5	91.8	92.7	93.1	93.7	93.8
200	87.5	89.7	93.0	93.5	93.5	94.7	94.5
1000	87.3	91.5	93.0	94.7	92.9	94.6	94.7

CCE: variance							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	95.5	96.0	97.1	96.3	96.6	97.8	98.9
25	96.2	96.8	97.3	98.0	97.6	98.5	98.9
50	95.6	96.4	97.5	98.0	98.1	99.1	99.3
100	96.5	96.7	97.7	97.9	98.6	98.9	99.5
200	96.3	96.2	97.5	97.8	98.9	99.4	99.4
1000	95.7	96.5	97.7	98.2	98.3	99.3	99.4

PCA: variance							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	92.4	93.0	92.3	93.8	94.3	96.7	98.5
25	94.3	93.7	94.9	95.8	96.0	98.0	99.0
50	93.4	94.5	96.0	96.7	97.5	98.6	99.3
100	94.8	94.9	96.1	97.4	97.9	98.7	99.2
200	94.6	94.7	96.0	97.3	97.4	99.4	99.4
1000	94.1	95.4	96.5	97.2	97.9	98.9	99.2

Note: This table reports the coverage probability for a 95% confidence interval for the test of equal squared biases (top two panels) or equal idiosyncratic variances (panels 3 and 4), using the Monte Carlo simulation setup described in Section 5.1 and 2,000 random samples. n refers to the number of cross-sectional units used in the pair-wise comparison of loss differences, while ρ^2 measures the predictive power of the underlying forecasts. We show coverage probabilities for the CCE and PCA methods described in Section 3. The assumed time-series dimension is $T = 80$. The underlying data generating process assumes two factors.

Table C.4: Coverage probability of 95% confidence intervals: 3 factors with heterogeneous loadings

CCE method for difference in squared bias							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	96.8	95.2	91.9	92.7	91.1	89.1	88.9
25	96.5	94.4	92.8	90.3	91.7	88.9	85.9
50	95.2	94.0	90.6	90.9	88.4	84.6	77.6
100	92.8	92.3	88.4	88.1	84.6	75.1	66.1
200	93.0	89.4	84.8	81.1	75.5	63.9	53.8
1000	90.6	84.6	74.9	68.7	64.0	52.2	40.4

PCA method for difference in squared bias							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	98.1	96.9	92.7	91.3	88.7	88.3	88.0
25	98.2	95.1	93.2	93.0	93.3	92.4	93.9
50	96.0	94.3	93.1	92.2	92.9	95.0	94.4
100	94.3	91.9	94.2	94.8	94.8	95.1	94.6
200	92.1	93.3	94.0	95.6	94.7	95.0	95.4
1000	91.1	92.7	96.0	94.7	95.7	95.9	96.1

CCE method for difference in variance							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	94.5	97.1	96.3	97.5	97.5	98.3	99.0
25	96.3	97.5	97.7	98.0	97.4	98.5	98.8
50	96.5	96.4	98.4	97.5	98.1	98.3	97.6
100	97.2	96.3	97.8	97.8	96.7	94.6	90.5
200	95.4	97.1	97.4	95.7	94.9	86.5	77.6
1000	95.5	94.9	94.1	90.5	86.1	74.0	59.0

PCA method for difference in variance							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	92.1	92.8	91.3	93.6	93.8	96.1	97.5
25	93.7	94.4	94.6	95.9	95.9	98.0	98.8
50	94.4	93.9	96.9	96.0	97.0	98.4	98.7
100	95.1	94.2	96.7	97.1	96.7	99.1	98.6
200	93.5	95.2	96.7	97.6	98.3	98.9	98.9
1000	93.8	94.6	97.1	96.6	97.2	98.6	99.0

Note: This table reports the coverage probability for a 95% confidence interval for the test of equal squared biases (top two panels) or equal idiosyncratic variances (panels 3 and 4), using the Monte Carlo simulation setup described in Section 5.1 and 2,000 random samples. n refers to the number of cross-sectional units used in the pair-wise comparison of loss differences, while ρ^2 measures the predictive power of the underlying forecasts. We show coverage probabilities for the CCE and PCA methods described in Section 3. The assumed time-series dimension is $T = 80$. The underlying data generating process assumes three factors.

Table C.5: Coverage probability of 95% confidence intervals: Breaks in the number of factors with heterogeneous loadings

	CCE method							PCA method						
difference in squared bias (before the break)														
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	97.4	95.6	93.5	91.9	90.9	89.5	89.4	99.6	98.2	94.5	91.1	89.9	85.8	85.4
25	97.2	95.6	92.1	92.2	91.3	88.7	88.3	99.2	96.6	93.0	92.0	92.3	89.8	91.1
50	95.2	91.7	90.6	90.1	90.3	87.1	85.6	98.0	95.0	93.4	93.1	92.5	91.7	92.0
100	94.1	91.1	89.8	89.9	87.8	84.6	76.7	95.1	94.4	93.6	92.5	94.0	94.0	93.3
200	93.5	92.0	88.7	85.6	85.7	77.6	65.7	94.3	93.7	93.8	93.3	94.0	94.8	94.2
1000	90.9	91.0	83.7	80.2	77.2	62.6	47.4	92.5	93.1	93.8	94.2	95.3	95.1	94.5
diff in squared bias (after the break)														
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	96.8	95.9	93.2	91.9	90.7	90.4	90.2	99.5	97.3	93.0	90.7	88.6	88.8	87.3
25	96.9	93.9	93.0	92.5	91.8	90.6	86.9	98.6	95.6	94.5	93.1	93.6	93.4	92.4
50	94.5	93.8	91.7	90.9	90.6	84.4	79.8	97.4	93.7	93.9	94.0	95.2	93.9	94.1
100	94.2	92.2	89.9	87.0	86.4	76.0	67.4	96.0	95.3	94.6	94.8	95.8	95.4	95.5
200	92.7	90.9	85.2	81.9	76.9	65.0	54.3	94.1	94.6	94.5	94.9	95.7	96.0	96.2
1000	90.7	85.2	73.7	68.2	64.1	47.8	39.8	94.0	95.2	96.0	95.5	95.0	96.2	96.2
diff in variance (before the break)														
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	95.4	95.6	95.6	96.8	97.2	98.3	99.0	91.8	91.6	92.3	92.9	94.4	97.3	98.4
25	95.5	96.4	97.4	97.5	98.2	98.7	99.4	93.3	94.6	95.4	96.0	97.5	98.5	99.5
50	95.8	95.8	97.4	98.3	98.7	99.2	99.6	94.4	94.1	96.1	97.2	97.7	99.2	99.5
100	96.2	96.9	97.2	98.0	98.4	98.4	96.8	94.9	94.8	96.2	97.1	98.7	99.3	99.2
200	96.3	96.2	96.9	97.6	98.1	95.7	88.5	94.7	95.2	96.9	97.9	98.5	99.0	99.8
1000	96.0	96.7	95.4	95.6	93.9	83.6	69.2	94.6	95.9	96.3	96.9	98.4	99.6	99.2
diff in variance (after the break)														
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	95.3	96.0	96.7	97.2	97.6	98.9	99.2	91.8	91.7	92.3	93.6	94.4	96.9	98.0
25	96.4	96.2	98.3	98.3	98.7	99.1	98.7	93.1	92.8	96.0	95.5	96.7	98.6	98.2
50	96.1	97.7	98.0	98.0	98.2	97.9	98.2	93.8	94.4	95.7	96.6	97.2	98.5	99.3
100	97.2	97.0	97.7	97.4	96.7	94.6	89.9	94.8	95.2	96.7	97.2	97.4	98.6	99.0
200	95.9	97.2	96.3	96.4	94.2	87.9	74.8	94.4	95.4	96.5	97.4	97.6	99.0	99.0
1000	96.2	96.5	93.5	89.3	84.9	67.9	56.0	94.9	95.6	96.7	97.1	97.4	98.9	99.3

Note: This table reports the coverage probability for a 95% confidence interval for the test of equal squared biases (top two panels) or equal idiosyncratic variances (panels 3 and 4), using the Monte Carlo simulation setup described in Section 5.1 and 2,000 random samples. n refers to the number of cross-sectional units used in the pair-wise comparison of loss differences, while ρ^2 measures the predictive power of the underlying forecasts. We show coverage probabilities for the CCE and PCA methods described in Section 3. The assumed time-series dimension is $T = 80$. The underlying data generating process assumes that there are initially two factors but that this changes to three factors in the second half of the sample.

Table C.6: 95% Coverage probabilities for a 95% confidence interval for testing the null of equal conditionally expected loss under Linex loss

	Coverage probability (linex loss)						
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	89.5	90.5	89.2	88.9	89.0	88.1	87.8
25	92.6	93.0	92.9	92.5	93.1	92.1	90.5
50	95.4	94.5	94.3	94.2	93.8	93.0	92.3
100	95.1	94.2	94.3	94.1	94.2	94.0	93.6
200	95.0	94.9	94.5	95.0	95.1	94.9	94.1
1000	94.3	95.4	95.3	94.8	94.7	95.0	95.8

Note: This table reports the coverage probability for a 95% confidence interval for the test of equal expected loss, $E(\overline{\Delta L}_{t+h} \mid \mathcal{F}) = 0$, using the linex loss function. We use the Monte Carlo simulation setup described in Section 5.1 and 2,000 random samples. n refers to the number of cross-sectional units used in the pair-wise comparison of loss differences, while ρ^2 measures the predictive power of the underlying forecasts.

Table C.7: Coverage probabilities for a 95% confidence interval for the average difference in squared bias under conditionally heteroskedastic shocks

clustering							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	97.9	96.3	94.2	93.7	93.0	91.0	91.3
25	98.2	96.1	94.9	94.1	93.6	92.6	94.2
50	96.1	95.7	95.0	94.0	94.8	94.5	94.7
100	96.0	96.0	94.5	94.6	95.3	94.3	94.1
200	96.1	95.8	94.3	94.1	95.4	95.1	95.0
1000	95.3	94.8	95.1	94.5	94.5	94.5	95.7

CCE							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	97.8	95.9	94.0	93.3	92.0	90.2	91.2
25	98.3	96.0	94.6	94.1	93.2	92.2	93.9
50	96.1	95.2	94.8	93.5	94.6	94.0	94.3
100	95.7	95.8	94.3	94.4	95.1	94.2	93.8
200	95.8	95.2	94.1	93.8	94.9	94.6	94.8
1000	94.9	94.7	94.8	94.2	94.1	94.1	95.3

PCA							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	97.7	94.8	90.1	89.2	90.2	89.8	90.8
25	96.1	91.4	91.7	91.8	91.9	92.1	94.1
50	92.1	88.4	91.6	93.1	94.1	93.6	94.5
100	89.8	89.3	91.2	93.3	94.0	94.1	93.6
200	87.0	89.9	93.7	92.9	94.1	94.4	94.8
1000	86.7	91.0	93.9	93.4	94.0	93.8	95.2

Note: This table reports coverage probabilities for 95% confidence intervals for the test of equal squared biases using the Monte Carlo simulation setup described in Section 5.1 and 2,000 random samples. n refers to the number of cross-sectional units used in the pair-wise comparison of loss differences, while ρ^2 measures the predictive power of the underlying forecasts. We show coverage probabilities for the clustering, CCE, and PCA methods described in Section 3. The assumed time-series dimension is $T = 80$. The table replaces the assumption of i.i.d standard normal errors and factors with an assumption of ARCH dynamics.

Table C.8: Coverage probabilities for a 95% confidence interval for the average difference in variance under conditionally heteroskedastic shocks

clustering							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	94.3	94.4	95.7	95.8	96.5	97.9	98.3
25	95.1	96.4	96.8	96.4	98.1	99.2	99.5
50	95.4	96.5	97.5	98.0	98.0	98.7	99.6
100	95.6	96.7	97.3	96.9	98.1	99.0	99.7
200	95.5	96.2	96.4	97.4	97.7	98.9	99.6
1000	94.5	96.2	97.1	97.6	98.3	98.9	99.0

CCE							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	95.8	96.2	97.1	97.0	97.9	98.9	99.3
25	95.8	97.0	98.1	97.6	98.9	99.6	99.7
50	96.3	97.5	98.2	98.4	98.5	99.2	99.8
100	96.1	97.1	98.0	98.1	99.1	99.6	99.8
200	96.4	96.9	97.6	98.3	98.6	99.3	99.9
1000	95.1	96.9	98.1	98.4	98.7	99.5	99.6

PCA							
$n \setminus \rho^2$	0.05	0.1	0.2	0.25	0.3	0.4	0.45
10	93.4	92.7	92.1	93.1	95.1	97.4	98.1
25	93.1	93.4	94.9	95.3	97.7	99.3	99.2
50	94.0	94.4	96.9	97.2	97.7	98.6	99.6
100	94.1	95.0	96.4	96.6	97.8	99.1	99.7
200	94.4	94.5	95.7	97.2	97.8	98.9	99.6
1000	93.8	94.6	96.5	97.3	97.9	98.8	98.9

Note: This table reports the coverage probability for 95% confidence intervals for the test of equal idiosyncratic error variances, using the Monte Carlo simulation setup described in Section 5.1 and 2,000 random samples. n refers to the number of cross-sectional units used in the pair-wise comparison of loss differences, while ρ^2 measures the predictive power of the underlying forecasts. We show coverage probabilities for the clustering, CCE, and PCA methods described in Section 3. The assumed time-series dimension is $T = 80$. The table replaces the assumption of i.i.d standard normal errors and factors with an assumption of ARCH dynamics.

Table C.9: Expected length of 95% confidence intervals

	squared bias							variance						
	clustering													
$n \setminus \rho_e^2$	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	3.84	3.35	2.93	2.81	2.40	2.27	2.08	4.95	4.23	3.66	3.44	2.94	2.68	2.43
25	2.42	2.15	1.89	1.69	1.55	1.53	1.40	3.25	2.86	2.46	2.18	1.93	1.91	1.68
50	1.72	1.48	1.33	1.29	1.13	1.13	1.05	2.37	2.00	1.77	1.67	1.45	1.41	1.28
100	1.22	1.07	0.96	0.87	0.85	0.76	0.72	1.70	1.47	1.27	1.14	1.10	0.97	0.90
200	0.88	0.76	0.69	0.61	0.58	0.55	0.50	1.22	1.04	0.92	0.80	0.74	0.70	0.63
1000	0.55	0.47	0.42	0.40	0.37	0.35	0.33	0.77	0.65	0.57	0.53	0.48	0.44	0.41
	CCE													
$n \setminus \rho_e^2$	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	4.28	3.74	3.17	2.79	2.31	1.83	1.31	6.00	4.92	4.03	3.43	2.75	2.12	1.46
25	3.20	2.77	2.30	1.86	1.53	1.22	0.82	4.07	3.41	2.78	2.21	1.79	1.42	0.95
50	2.45	1.96	1.65	1.42	1.14	0.90	0.60	3.02	2.38	1.97	1.67	1.33	1.05	0.70
100	1.78	1.46	1.23	0.97	0.81	0.62	0.42	2.17	1.75	1.45	1.16	0.95	0.72	0.49
200	1.32	1.06	0.87	0.69	0.58	0.44	0.29	1.57	1.25	1.03	0.81	0.67	0.52	0.34
1000	0.83	0.66	0.54	0.45	0.36	0.28	0.18	0.99	0.79	0.64	0.53	0.43	0.33	0.22
	PCA													
$n \setminus \rho_e^2$	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	3.50	2.89	2.41	2.03	1.59	1.24	0.81	3.99	3.13	2.52	2.08	1.63	1.25	0.82
25	2.58	2.12	1.73	1.39	1.12	0.88	0.58	2.67	2.18	1.76	1.41	1.12	0.88	0.58
50	1.94	1.53	1.26	1.07	0.84	0.66	0.44	1.98	1.55	1.27	1.07	0.85	0.67	0.44
100	1.41	1.14	0.93	0.75	0.61	0.47	0.32	1.42	1.14	0.93	0.75	0.62	0.47	0.32
200	1.01	0.81	0.67	0.53	0.43	0.33	0.22	1.02	0.81	0.67	0.53	0.43	0.34	0.22
1000	0.65	0.51	0.42	0.35	0.28	0.22	0.14	0.65	0.52	0.42	0.35	0.28	0.22	0.14

Note: This table assumes a three-factor data generating process with random factor loadings. Parameters are set to match the average cross-sectional R^2 value observed in the empirical application.

Bibliography

Ahn, Seung C, and Alex R Horenstein. 2013. "Eigenvalue ratio test for the number of factors." *Econometrica*, 81(3): 1203–1227.

Ait-Sahalia, Yacine, Jonathan A. Parker, and Motohiro Yogo. 2004. "Luxury goods and the equity premium." *The Journal of Finance*, 59(6): 2959–3004.

Andrews, Donald WK, and Gustavo Soares. 2010. "Inference for parameters defined by moment inequalities using generalized moment selection." *Econometrica*, 78(1): 119–157.

Bai, Jushan. 2003. "Inferential theory for factor models of large dimensions." *Econometrica*, 135–171.

Bansal, Ravi, and Amir Yaron. 2004. "Risks for the long run: A potential resolution of asset pricing puzzles." *The journal of Finance*, 59(4): 1481–1509.

Bansal, Ravi, Dana Kiku, Ivan Shaliastovich, and Amir Yaron. 2014. "Volatility, the macroeconomy, and asset prices." *The Journal of Finance*, 69(6): 2471–2511.

Barro, Robert J. 2006. "Rare disasters and asset markets in the twentieth century." *The Quarterly Journal of Economics*, 121(3): 823–866.

Barro, Robert J., and José F. Ursúa. 2008. "Macroeconomic crises since 1870." National Bureau of Economic Research.

Barro, Robert J., and José F. Ursúa. 2012. "Rare macroeconomic disasters." *Annu. Rev. Econ.*, 4(1): 83–109.

Beeler, Jason, and John Y. Campbell. 2012. "The Long-Run Risks Model and Aggregate Asset Prices: An Empirical Assessment." *Critical Finance Review*, 1(1): 141–182.

Belo, Frederico, and Andres Donangelo. 2020. "Priceless Consumption." *Available at SSRN 3475267*.

Blake, David, Alberto G Rossi, Allan Timmermann, Ian Tonks, and Russ Wermers. 2013. "Decentralized investment management: Evidence from the pension fund industry." *The Journal of Finance*, 68(3): 1133–1178.

Bollerslev, Tim. 1986. "Generalized autoregressive conditional heteroskedasticity." *Journal of econometrics*, 31(3): 307–327.

Bollerslev, Tim, and Viktor Todorov. 2011. "Tails, fears, and risk premia." *The Journal of Finance*, 66(6): 2165–2211.

Breeden, Douglas T. 1979. "An intertemporal asset pricing model with stochastic consumption and investment opportunities." *Journal of Financial Economics*, 7(3): 265–296.

Campbell, John Y., and Robert J. Shiller. 1988. "The dividend-price ratio and expectations of future dividends and discount factors." *The Review of Financial Studies*, 1(3): 195–228.

Campbell, John Y., Stefano Giglio, Christopher Polk, and Robert Turley. 2018. "An intertemporal CAPM with stochastic volatility." *Journal of Financial Economics*, 128(2): 207–233.

Carter, Chris K., and Robert Kohn. 1994. "On Gibbs sampling for state space models." *Biometrika*, 81(3): 541–553.

Cecchetti, Stephen G., Pok-sang Lam, and Nelson C. Mark. 2000. "Asset pricing with distorted beliefs: are equity returns too good to be true?" *American Economic Review*, 90(4): 787–805.

Cheng, Xu, Zhipeng Liao, and Frank Schorfheide. 2016. "Shrinkage estimation of high-dimensional factor models with structural instabilities." *The Review of Economic Studies*, 83(4): 1511–1543.

Chen, Xiaohong, Qi-Man Shao, and Wei Biao Wu. 2016. "Supplement to "Self-normalized Cramér-Type Moderate Deviations under Dependence"." *The Annals of Statistics*.

Chen, Xiaohong, Qi-Man Shao, Wei Biao Wu, and Lihu Xu. 2016. "Self-normalized Cramér-type moderate deviations under dependence." *The Annals of Statistics*, 44(4): 1593–1617.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2013. "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors." *The Annals of Statistics*, 41(6): 2786–2819.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2015. "Comparison and anti-concentration bounds for maxima of Gaussian random vectors." *Probability Theory and Related Fields*, 162(1-2): 47–70.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2017. "Central limit theorems and bootstrap in high dimensions." *The Annals of Probability*, 45(4): 2309–2352.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2018. "Inference on causal and structural parameters using many moment inequalities." *Review of Economic Studies (Forthcoming), available at arXiv:1312.7614*.

Chib, Siddhartha. 1998. "Estimation and comparison of multiple change-point models." *Journal of econometrics*, 86(2): 221–241.

Chong, Yock Y, and David F Hendry. 1986. "Econometric evaluation of linear macro-economic models." *The Review of Economic Studies*, 53(4): 671–690.

Clark, Todd E, and Kenneth D West. 2007. "Approximately normal tests for equal predictive accuracy in nested models." *Journal of Econometrics*, 138(1): 291–311.

Clark, Todd E, and Michael W McCracken. 2001. "Tests of equal forecast accuracy and encompassing for nested models." *Journal of econometrics*, 105(1): 85–110.

Collin-Dufresne, Pierre, Michael Johannes, and Lars A. Lochstoer. 2016. "Parameter learning in general equilibrium: The asset pricing implications." *American Economic Review*, 106(3): 664–98.

Davies, Anthony, and Kajal Lahiri. 1995. "A new framework for analyzing survey forecasts using three-dimensional panel data." *Journal of Econometrics*, 68(1): 205–227.

Davies, Anthony, Kajal Lahiri, et al. 1995. "A new framework for analyzing survey forecasts using three-dimensional panel data." *Journal of Econometrics*, 68(1): 205–228.

Diebold, Francis X, and Roberto S Mariano. 1995. "Comparing Predictive Accuracy." *Journal of Business & Economic Statistics*, 253–263.

Elliott, Graham, Ivana Komunjer, and Allan Timmermann. 2005. "Estimation and testing of forecast rationality under flexible loss." *Review of Economic Studies*, 72(4): 1107–1125.

Fama, Eugene F., and Kenneth R. French. 1992. "The cross-section of expected stock returns." *the Journal of Finance*, 47(2): 427–465.

Gallant, Ronald A., Mohammad R. Jahan-Parvar, and Hening Liu. 2019. "Does Smooth Ambiguity Matter for Asset Pricing?" *The Review of Financial Studies*, 32(9): 3617–3666.

Geweke, John, and Yu Jiang. 2011. "Inference and prediction in a multiple-structural-break model." *Journal of Econometrics*, 163(2): 172–185.

Giacomini, Raffaella, and Halbert White. 2006. "Tests of conditional predictive ability." *Econometrica*, 74(6): 1545–1578.

Giacomini, Raffaella, Simon Lee, and Silvia Sarpietro. 2019. "Microforecasting with Individual Forecast Selection." *Unpublished working paper, UCL.*

Gordon, Robert J. 2007. *The American business cycle: Continuity and change.* Vol. 25, University of Chicago Press.

Gordon, Robert J. 2017. *The rise and fall of American growth: The US standard of living since*

the civil war. Vol. 70, Princeton University Press.

Granger, Clive William John. 1999. "Outline of forecast theory using generalized cost functions." *Spanish Economic Review*, 1(2): 161–173.

Green, Peter J. 1995. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82(4): 711–732.

Hansen, Lars Peter, John C Heaton, and Nan Li. 2008. "Consumption strikes back? Measuring long-run risk." *Journal of Political economy*, 116(2): 260–302.

Hansen, Peter R, Asger Lunde, and James M Nason. 2011. "The model confidence set." *Econometrica*, 79(2): 453–497.

Hansen, Peter Reinhard. 2005. "A test for superior predictive ability." *Journal of Business & Economic Statistics*, 23(4): 365–380.

Hansen, Peter Reinhard, and Allan Timmermann. 2015. "Equivalence Between Out-of-Sample Forecast Comparisons and Wald Statistics." *Econometrica*, 83(6): 2485–2505.

Jahan-Parvar, Mohammad R., and Hening Liu. 2014. "Ambiguity aversion and asset prices in production economies." *The Review of Financial Studies*, 27(10): 3060–3097.

Jegadeesh, Narasimhan, and Sheridan Titman. 1993. "Returns to buying winners and selling losers: Implications for stock market efficiency." *The Journal of finance*, 48(1): 65–91.

Johannes, Michael, Lars A. Lochstoer, and Yiqun Mou. 2016. "Learning about consumption dynamics." *The Journal of finance*, 71(2): 551–600.

Jorgenson, Dale W., Mun S. Ho, and Jon D. Samuels. 2014. "What will revive U.S. economic growth? Lessons from a prototype industry-level production account for the United States." *Journal of Policy Modeling*, 36(4): 674 – 691.

Ju, Nengjiu, and Jianjun Miao. 2012. "Ambiguity, learning, and asset returns." *Econometrica*, 80(2): 559–591.

Kacperczyk, Marcin, Stijn Van Nieuwerburgh, and Laura Veldkamp. 2014. "Time-varying fund manager skill." *The Journal of Finance*, 69(4): 1455–1484.

Keane, Michael P, and David E Runkle. 1990. "Testing the rationality of price forecasts: New evidence from panel data." *American Economic Review*, 80(4): 714–735.

Koop, Gary, and Simon M. Potter. 2007. "Estimation and forecasting in models with multiple breaks." *The Review of Economic Studies*, 74(3): 763–789.

LeRoy, Stephen F., and Richard D. Porter. 1981. "The present-value relation: Tests based on implied variance bounds." *Econometrica: Journal of the Econometric Society*, 555–574.

Lettau, Martin, Sydney C. Ludvigson, and Jessica A. Wachter. 2008. "The declining equity premium: What role does macroeconomic risk play?" *The Review of Financial Studies*, 21(4): 1653–1687.

Liu, Laura, Hyungsik Roger Moon, and Frank Schorfheide. 2018. "Forecasting with Dynamic Panel Data Models." *Unpublished working paper, University of Pennsylvania*.

Liu, Laura, Hyungsik Roger Moon, and Frank Schorfheide. 2019. "Forecasting with a Panel Tobit Model." *Unpublished working paper, University of Pennsylvania*.

Liu, Laura Xiaolei, and Lu Zhang. 2008. "Momentum profits, factor pricing, and macroeconomic risk." *The Review of Financial Studies*, 21(6): 2417–2448.

Mackowiak, Bartosz, and Mirko Wiederholt. 2009. "Optimal sticky prices under rational inattention." *American Economic Review*, 99(3): 769–803.

McCracken, Michael W. 2007. "Asymptotics for out of sample tests of Granger causality." *Journal of Econometrics*, 140(2): 719–752.

Mehra, Rajnish, and Edward C. Prescott. 1985. "The equity premium: A puzzle." *Journal of monetary Economics*, 15(2): 145–161.

Mehra, Rajnish, and Edward C. Prescott. 1988. "The equity risk premium: A solution?" *Journal of Monetary Economics*, 22(1): 133–136.

Onatski, Alexei. 2009. "Testing hypotheses about the number of factors in large factor models." *Econometrica*, 77(5): 1447–1479.

Patton, Andrew J, and Allan Timmermann. 2011. "Predictability of output growth and inflation: A multi-horizon survey approach." *Journal of Business & Economic Statistics*, 29(3): 397–410.

Patton, Andrew J, and Allan Timmermann. 2012. "Forecast rationality tests based on multi-horizon bounds." *Journal of Business & Economic Statistics*, 30(1): 1–40.

Peña, Victor H, Tze Leung Lai, and Qi-Man Shao. 2008. *Self-normalized processes: Limit theory and Statistical Applications.* Springer Science & Business Media.

Pesaran, M Hashem. 2004. "General diagnostic tests for cross section dependence in panels."

Pesaran, M Hashem. 2006. "Estimation and inference in large heterogeneous panels with a multifactor error structure." *Econometrica*, 74(4): 967–1012.

Piazzesi, Monika, Martin Schneider, and Selale Tuzel. 2007. "Housing, consumption and asset pricing." *Journal of Financial Economics*, 83(3): 531–569.

Qu, Ritong, Allan Timmermann, and Yinchu Zhu. 2019. "Do any economists have superior

forecasting skills?" *Working paper*.

Rietz, Thomas A. 1988. "The equity risk premium a solution." *Journal of monetary Economics*, 22(1): 117–131.

Romano, Joseph P, and Michael Wolf. 2005. "Stepwise multiple testing as formalized data snooping." *Econometrica*, 73(4): 1237–1282.

Romano, Joseph P, Azeem M Shaikh, and Michael Wolf. 2014. "A Practical Two-Step Method for Testing Moment Inequalities." *Econometrica*, 82(5): 1979–2002.

Schorfheide, Frank, Dongho Song, and Amir Yaron. 2018. "Identifying long-run risks: A Bayesian mixed-frequency approach." *Econometrica*, 86(2): 617–654.

Shiller, Robert J. 1981. "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?" *The American Economic Review*, 71(3): 421–436.

Smith, Simon C. 2017. "Noncommon Breaks." *Available at SSRN 3239963*.

Smith, Simon C., and Allan Timmermann. 2018. "Detecting Breaks in Real Time: A Panel Forecasting Approach." *Econometrics: Econometric & Statistical Methods - General eJournal*.

Smith, Simon C., and Allan Timmermann. 2020. "Break Risk." *The Review of Financial Studies*.

Stock, James H., and Mark W. Watson. 2002. "Has the business cycle changed and why?" *NBER macroeconomics annual*, 17: 159–218.

Timmermann, Allan. 2007. "An evaluation of the World Economic Outlook forecasts." *IMF Staff Papers*, 54(1): 1–33.

Wachter, Jessica A. 2013. "Can time-varying risk of rare disasters explain aggregate stock market volatility?" *The Journal of Finance*, 68(3): 987–1035.

West, Kenneth D. 1996. "Asymptotic inference about predictive ability." *Econometrica: Journal of the Econometric Society*, 1067–1084.

White, Halbert. 2000. "A reality check for data snooping." *Econometrica*, 68(5): 1097–1126.

Yogo, Motohiro. 2006. "A consumption-based explanation of expected stock returns." *The Journal of Finance*, 61(2): 539–580.

Zhu, Yinchu, and Jelena Bradic. 2018. "Significance testing in non-sparse high-dimensional linear models."