Essays on Econometric Forecasting

LARS E. SPRENG

Bayes Business School
Faculty of Finance
City, University of London

July 12, 2023

Lars E. Spreng: *Essays on Econometric Forecasting,* A Thesis submitted to Bayes Business School in partial fulfilment of the requirements for the degree of Doctor of Philosophy, © July 12, 2023

DECLARATION

I, Lars Spreng, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

London, July 12, 2023

Lars E. Spreng

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis was composed during my time as a PhD Candidate at Bayes Business School from 2019 to 2023. Financial support for this research was provided by a scholarship from City, University of London. Additionally, I was affiliated with the Centre for Econometric Analysis at Bayes, which supplied valuable resources for my work.

I wish to convey my gratitude to the following individuals who have supported me throughout my doctoral journey. First and foremost, I would like to thank my first supervisor, Giovanni Urga. Giovanni has time and again provided me with invaluable feedback and advice. Without his guidance, I would not have been able to complete this thesis to the same standard, and for that I hold his supervision in the highest possible regard. I also express my gratitude to my second supervisor, Ian Marsh. Ian's expertise in the field of foreign exchange rates has been crucial to my understanding of the market. I thank my co-authors, Eric Hillebrand and Jakob Mikkelsen, with whom I have collaborated on my first chapter. I would also like to extend my thanks to all the people who have provided me with their feedback on multiple occasions. In particular, Lynda Khalaf, Fa Wang, Roy Batchelor, Kate Phylaktis, Giuliano de Rossi, Jens Perch Nielsen, and participants at various seminars and talks. Furthermore, I wish to acknowledge several people involved in the publication process of two of my chapters. My first chapter has benefited greatly from the comments of the Editor of the Journal of Applied Econometrics, Barbara Rossi, as well as two anonymous referees. I also thank the Editor of the Journal of Business and Economic Statistics, Christian Hansen, an Associate Editor at the same journal, and two anonymous referees whose comments have led to an improved version of my second chapter.

Finally, I would like to express my appreciation to all my friends, family, and PhD colleagues who have accompanied me through my academic endeavours.

LIST OF PUBLICATIONS

The first two chapters of this thesis have appeared in the following publications

1. Hillebrand, E. Mikkelsen J. Spreng, L. and Urga, G., Exchange Rates and Macroeconomic Fundamentals: Evidence from Time-Varying Factor Loadings, *Journal of Applied Econometrics*, forthcoming, https://doi.org/10.1002/jae.2984

2. Spreng, L. and Urga G., Combining p-Values for Multivariate Predictive Ability Testing, *Journal of Business and Economic Statistics*, forthcoming, https://doi.org/10.1080/07350015.2022.2067545

INTRODUCTION

The ability to form accurate predictions of economic and financial variables is of paramount importance to government bodies, international organisations, and financial institutions alike. From forecasting macroeconomic variables that inform policy decisions to quantitative risk management, the practice of forecasting is widespread. Therefore, it constitutes a crucial area of econometrics, with a particular emphasis on (i) developing new forecasting methods and (ii) evaluating their performance. This thesis contributes to both these aspects of the forecasting literature.

The difficulty of identifying new forecasting models and selecting between them is exemplified by the challenging task of predicting foreign exchange rates. Many structural models derived from macroeconomic theory have been shown to be no more accurate than a random walk (Rossi, 2013). In other words, they have no predictive power. Numerous attempts have been made to identify the underlying reasons for this, with one possible explanation being that the parameters in foreign exchange rate models are time-varying [see, for example, Rossi (2006), Bekiros (2014), or Byrne et al. (2018)]. Indeed, survey evidence of UK and US based foreign exchange traders also indicates that their reliance on macroeconomic variables changes over time (Cheung and Chinn, 2001; Cheung et al., 2004). The first chapter of this thesis adds to this literature by analysing the unstable relationship between exchange rates and macroeconomic fundamentals through the lens of a factor model with time-varying loadings. Leveraging the findings of Mikkelsen et al. (2019), we estimate a theoretical model in which macroeconomic fundamentals are treated as latent factors. These factors are extracted as principal components from a novel real-time database that we curated for this chapter. The database encompasses 272 monthly datasets, each comprising over 100 variables from 15 countries. We have made this database publicly available as a contribution of this chapter. To gauge the significance of time-variation, we compare the out-of-sample performance of our model to a factor model with constant loadings and a random walk. Our

results demonstrate that the time-varying model consistently outperforms the constant loadings benchmark and even the random walk in multiple instances.

Building on the out-of-sample evaluation conducted in the first chapter, the second chapter introduces a new statistical test to evaluate forecasts. One of the earliest tests for this purpose is proposed by Diebold and Mariano (1995) and compares two primitive forecasts without considering the underlying models that generated them. However, when employing a forecast testing procedure for model selection, it is imperative to address potential concerns such as estimation errors, nested models, and forecast step sizes. In a seminal paper, Giacomini and White (2006) introduce the notion of Conditional Predictive Ability and a testing framework that enables the discrimination between various underlying forecasting methods. Since then, a number of alternative testing procedures have been introduced [see Clark and McCracken (2013) for a survey]. However, existing tests are mostly univariate and only evaluate two competing forecasts at a time. This is problematic because in many econometric applications dependence between variables and forecasting models is the norm. This, in turn, introduces dependencies between univariate test statistics and p-values, as shown in chapter two of this thesis. In consequence, such tests cannot be evaluated individually, which motivates the development of multivariate forecast tests that account for dependence (Qu et al., 2021). The novel approach we introduce in the second chapter combines univariate forecast accuracy tests, without making any assumptions regarding the joint distribution of the test statistics. Our approach builds upon recent advancements in the statistical literature on the combination of dependent p-values (Vovk and Wang, 2020). It allows for the implementation of whichever tests are most appropriate in a given scenario and evaluates whether predictive ability holds in the cross-section. We specify a global null hypothesis that is defined as the intersection of all individual null hypotheses, while also accounting for false discovery and dependence. We establish the statistical size properties of the test in finite samples as well as asymptotically for large cross-sections, and demonstrate its consistency in the asymptotic case. To examine the test further, we report extensive Monte-Carlo simulation results and conduct an empirical application using a large dataset of 84 daily exchange rates

Together, the first and second chapter have inspired the focus of the third chapter. Although it is widely recognised that individual models, such as predictive stock return regressions, exhibit predictive power only during certain periods (Timmermann, 2008), very few tests consider the possibility that the relative predictive ability of different models may also

vary over time. The first test for time-varying predictive ability is proposed by Giacomini and Rossi (2010) and can be viewed as a rolling t-test on the loss differential between forecasts. More recently, Odendahl et al. (2022) introduce a time-varying predictive ability test that accounts for state dependence. We add to this literature by proposing two novel forecast evaluation tests that consider the issue of time-variation in conjunction with dependencies between forecasts. The first test is a time-varying analogue to the Conditional Predictive Ability test proposed by Giacomini and White (2006), which evaluates the null hypothesis conditional on information up to the previous period. The second test, called Total Predictive Ability test, evaluates the null hypothesis conditional on the full-sample information set. Both tests can be applied in a univariate or multivariate framework, where dependencies between forecasts are explicitly modelled. To assess the performance of the tests, we conduct Monte-Carlo simulations and apply the tests in an evaluation of intraday volatility forecasts.

CHAPTER 1

EXCHANGE RATES AND MACROECONOMIC FUNDAMENTALS: EVIDENCE OF INSTABILITIES FROM TIME-VARYING FACTOR LOADINGS

1.1 INTRODUCTION

In this paper, we analyse the unstable relationship between exchange rates and macroeconomic fundamentals. To this end, we apply the two-step maximum likelihood approach proposed in Mikkelsen et al. (2019) that enables the estimation of time-varying loadings (TVL) in factor models.

Two-step estimation in large factor models was proposed in Doz et al. (2011) and Doz et al. (2012). These, together with the important result found in Bates et al. (2013) that principal components are consistent estimators of unobserved factors even in the presence of time-varying loadings, allowed Mikkelsen et al. (2019) to propose the consistent estimation of time-varying factor loadings in two-step maximum likelihood. Different approaches are taken in Su and Wang (2017), who estimate smoothly changing time-varying factor loadings using a local principal component estimator for latent factors, and Barigozzi et al. (2021), who introduce a generalised dynamic factor model in which factors are loaded with a time-varying filter.

Since Meese and Rogoff's (1983) key finding that structural exchange rate models perform no better than a random walk, arduous empirical work has been invested into this *disconnect puzzle* (Obstfeld and Rogoff, 2001); however, in many cases to no avail (Rossi, 2013). One possible solution for the disconnect puzzle is to model the relationship between macroeconomic fundamentals and exchange rates as time-varying. A theoretical explanation

for the presence of time-variation is provided by the scapegoat theory (Bacchetta and van Wincoop, 2004, 2013). An observed fundamental becomes a scapegoat if it is correlated with an unobserved shock and investors attribute exchange rate fluctuations to the fundamental instead of the actual unobserved shock. This leads investors to update their expectation about the effect fundamentals exert in a time-varying manner. Fratzscher et al. (2015) are the first to present empirical support for this theory by examining a survey of FX traders to obtain a measure of the scapegoat weights. In the same vein, Pozzi and Sadaba (2020) construct parameter expectations from survey data and use a Bayesian approach to determine the probability that variables are scapegoats. The disconnect puzzle may, however, also be a product of inaccurate model selection, as suggested by Sarno and Valente (2009): models would have to be altered frequently to optimally capture the information embedded in fundamentals and this implies a high degree of time-variation in their parameters. Kouwenberg et al. (2017) develop a dynamic model selection rule, which they find to produce better forecasts than several benchmark models. The reason behind this lies, again, in the rule's ability to incorporate time-variation. Further evidence for parameter instability in exchange rate regressions is provided by Rossi (2006), Bekiros (2014), and Byrne et al. (2018).

The present paper uses the results of Mikkelsen et al. (2019) to estimate a theoretical model in which the relationship between exchange rates and macroeconomic fundamentals is unstable. We treat macroeconomic fundamentals as latent factors, which are extracted in real-time from 272 newly compiled monthly vintage datasets. Specifically, we collate all real-time vintages of the McCracken and Ng (2016) FRED-MD database from 1999:08 to 2022:02. In addition, we compile large vintage datasets from the OECD statistical database for the same time periods, which we merge with the FRED-MD data. This yields a novel real-time database with each series starting in 1990:04 and ending one month prior to their release.[1] Therefore, rather than just relying on first releases, we include revisions made to past data with each new release. That is, we accurately replicate the information set available at each time step, which has been shown to improve forecasts of financial variables (Coroneo and Caruso, 2022). The information inherent in these series is extracted via principal components that serve as factor estimates. The factors can be interpreted as real economy, housing market, and interest rate factors. The model is tested for 14 different currencies *vis-à-vis* the US dollar. To examine whether accounting for time variation improves

[1] That is, the 1999:08 vintage dataset starts in 1990:04 and ends in 1999:07. The 1999:09 vintage starts in 1990:04 and ends in 1999:08, and so on.

exchange rate predictions, we compare the results to a factor model with constant loadings. As constant factor models have been found to deliver limited to no forecast improvements over a random walk (Engel et al., 2015; Rossi, 2013), we also analyse whether incorporating time-variation increases predictive ability relative to a random walk. The paper provides in-sample evidence that accounting for time-variation improves the model fit considerably as demonstrated by an R^2 ranging from 27% up to 88%. It correctly matches an appreciation and depreciation up to 89% of the time, whereas the constant loadings model can only explain a very small part of exchange rate variations, demonstrated by an R^2 between 0% and 6%. We show that taking the aforementioned instabilities into consideration improves the out-of-sample forecast accuracy of the model across benchmarks and evaluation procedures. First, the time-varying model displays better predictive ability than the constant loadings model across different forecasting horizons according to the conditional and unconditional Giacomini and White (2006) test. The constant loadings model never outperforms the time-varying model. Second, the time-varying model's forecasts display greater direction accuracy. Third, it performs better than the constant loadings model relative to a random walk. It achieves a lower Root Mean Squared Error (RMSE) than the random walk for up to 9 currencies and outperforms it statistically in 2 cases. In line with Engel et al. (2015), we find that the constant model can achieve a lower RMSE only at very long forecasting horizons. Fourth, when comparing predictive ability locally using the Giacomini and Rossi (2010) fluctuation test, we observe that the time-varying model improves forecasts during crises and, for several exchange rates, performs well against the random walk during periods in which the constant model displays poor performance.

The paper is structured as follows: Section 1.2 presents the theoretical model of structural instabilities between exchange rates and fundamentals. In Section 1.3, the model is mapped into state space form, and the econometric approach is described. Section 1.4 discusses the data, in Section 1.5 we report the in-sample, and in Section 1.6 the out-of-sample results. Section 1.7 concludes.

1.2 THEORETICAL MODEL

This section derives a model with instabilities in the relationship between exchange rates and macroeconomic fundamentals. The model belongs to the same class as the ones examined by Engel and West (2005). That is, it expresses the exchange rate as the discounted value of expected future

fundamentals and unobservable shocks (see equation (1) in Engel and West (2005)). Specifically, the relation is:

$$\Delta s_t = F_t + \sum_{j=1}^{\infty} \left(\frac{1}{\mu}\right)^j \mathbb{E}[F_{t+j} \mid I_t] - \sum_{j=0}^{\infty} \left(\frac{1}{\mu}\right)^{j+1} \mathbb{E}[\phi_{t+j}|I_t], \tag{1.1}$$

where s_t is the log of the exchange rate measured as the domestic price per unit of foreign currency, $\mathbb{E}[\cdot \mid I_t]$ is the expectation of the representative agent conditional on I_t, the information set available at time t, and $\mu \geq 0$. The value of the exchange rate is determined by the present and expected future macroeconomic fundamentals F_t. Finally, ϕ_t is the risk premium. The above equation results from two conditions. The first is an uncovered interest parity (UIP) condition:[2]

$$\mathbb{E}[s_{t+1}|I_t] - s_t = i_t - i_t^* + \phi_t, \tag{1.2}$$

where i_t is the nominal one-period interest rate. An asterisk denotes foreign variables, and deviations from UIP are accounted for by the risk premium ϕ_t. The expected change in the exchange rate is thus equal to the interest rate differential between the domestic and the foreign country plus a risk premium. The second condition relates the interest rate differential to macroeconomic fundamentals:

$$i_t - i_t^* = \mu \Delta s_t - \mu F_t. \tag{1.3}$$

Engel and West (2005) discuss a range of models that lead to Equation (1.1) and (1.3), for instance, a Taylor rule or monetary model, and it is also derived in Bacchetta and van Wincoop (2013, equation 3).[3] Combining equations (1.2) and (1.3) results in

$$\begin{aligned} \mathbb{E}[\Delta s_{t+1}|I_t] &= \mu \Delta s_t - \mu F_t + \phi_t \\ \Delta s_t &= \frac{1}{\mu} \{\mathbb{E}[\Delta s_{t+1}|I_t] - \phi_t\} + F_t. \end{aligned}$$

[2]See Engel (2014) for a survey of exchange rates and interest parity as well as the existence of the risk premium term.

[3]Note that Bacchetta and van Wincoop (2013) specify their model in levels and distinguish between observed fundamentals (F_t) and unobserved fundamentals (b_t). They obtain $i_t - i_t^* = \mu s_t - \mu(F_t + b_t)$ which they show to be consistent with several established exchange rate models, such as the monetary model. In an earlier version, they show that this relationship also holds in first differences (Bacchetta and van Wincoop, 2009).

Recursive substitution of Δs_t, assuming no bubbles, yields equation (1.1), which establishes the common result that the exchange rate equals the present value of expected future macroeconomic fundamentals and the foreign exchange risk premium. In their model, Bacchetta and van Wincoop (2013) specify $F_t = f_t'\beta$ as a linear combination of observable macroeconomic fundamentals, where $f_t = (f_{1,t}, ..., f_{n,t})'$ and $\beta = (\beta_1, ..., \beta_n)'$. We allow this combination to be time-varying:

$$F_t = f_t'\xi_t = f_t'(\beta + \kappa_t).$$

That is, consistent with the theory in Bacchetta and van Wincoop (2013), β describes a long-run equilibrium relationship between fundamentals and exchange rates, but we allow for transitory, zero-mean deviations in form of $\kappa_t = (\kappa_{1,t}, ..., \kappa_{n,t})'$. Consequently, the relative importance of fundamentals in determining the exchange rate is time-varying and affected by κ_t. This specification nests the constant coefficients case if $\kappa_t = 0$ for all t. To derive the effect of changes in observed fundamentals on the exchange rate, consider for simplicity the case of a single fundamental and assume that f_t, κ_t, and ϕ_t follow AR(1) processes:

$$\begin{aligned} f_t &= \rho_f f_{t-1} + v_t, \qquad v_t \sim i.i.d.(0, \sigma_f^2) \\ \kappa_t &= \rho_\kappa \kappa_{t-1} + u_t, \qquad u_t \sim i.i.d.(0, \sigma_\kappa^2) \\ \phi_t &= \rho_\phi \phi_{t-1} + w_t, \qquad w_t \sim i.i.d.(0, \sigma_\phi^2), \end{aligned} \tag{1.4}$$

where $|\rho_f|, |\rho_\kappa|, |\rho_\phi| < 1$. Clearly, $\mathbb{E}[f_{t+j}|I_t] = \rho_f^j f_t$ and $\mathbb{E}[\kappa_{t+j}|I_t] = \rho_\kappa^j \kappa_t$. Assuming f_t and κ_t are uncorrelated, equation (1.1) becomes:

$$\begin{aligned} \Delta s_t &= \sum_{j=0}^{\infty} \left(\frac{1}{\mu}\right)^j \rho_f^j f_t \beta + \sum_{j=0}^{\infty} \left(\frac{1}{\mu}\right)^j \rho_\kappa^j f_t \kappa_t - \frac{1}{\mu} \sum_{j=0}^{\infty} \left(\frac{1}{\mu}\right)^j \rho_\phi^j \phi_t \\ &= f_t \left(\frac{\mu}{\mu - \rho_f} \beta + \frac{\mu}{\mu - \rho_\kappa} \kappa_t \right) - \frac{1}{\mu - \rho_\phi} \phi_t. \end{aligned} \tag{1.5}$$

The derivative of the exchange rate with respect to the fundamentals is:

$$\frac{\partial \Delta s_t}{\partial f_t} = \left(\frac{\mu}{\mu - \rho_f} \beta + \frac{\mu}{\mu - \rho_\kappa} \kappa_t \right). \tag{1.6}$$

That is, the effect of variations in macroeconomic fundamentals on the exchange rate corresponds to a constant part, $\frac{\mu}{\mu - \rho_f}\beta$, and a time-varying

part, $\frac{\mu}{\mu-\rho_\kappa}\kappa_t$. Based on existing studies, one can expect that μ is close to one (Engel and West, 2005). In that case, if the transitory shocks are highly persistent relative to the fundamentals, the relationship between the latter and exchange rates is characterised by a greater degree of instability.

1.2.1 *Relation to the Scapegoat Theory*

The model relates to Bacchetta and van Wincoop's (2013) scapegoat theory in the sense that they also consider a model based on a single stochastic difference equation. In their framework, fundamentals, too, display temporary changes in their weights. In contrast to our model, however, this is the result of investors being unable to pin down the value of the structural parameters in β. If parameters were known, their model simply implies $\frac{\partial s_t}{\partial f_{i,t}} = \beta_i$. If parameters are unknown, on the other hand, agents form their expectations of β over time by updating their beliefs about the impact of fundamentals which takes the form $f_t'\beta + b_t$. Here, b_t are unobserved shocks that coincide with changes in fundamentals and thus introduce time-variation. Investors can observe a large value of the signal $f_t'\beta + b_t$ but are unable to distinguish whether this is due to β being greater than expected or a result of changes in the unobservables b_t. It becomes rational for agents to attribute at least some weight to a larger β, thereby raising their expectations of the structural parameters. Consequently, the relationship between fundamentals and the exchange rate becomes time-varying, in spite of the structural parameters being constant. This manifests itself in the derivative of the exchange rate with respect to fundamentals:

$$\frac{\partial s_t}{\partial f_{i,t}} = \theta\beta_i + (1-\theta)\mathbb{E}[\beta_i|I_t] + (1-\theta)f_t'\frac{\partial \mathbb{E}[\beta|I_t]}{\partial f_{i,t}},$$

where the first two terms on the right-hand side are a weighted average of the true structural parameters and their expectations. The time-varying last term reflects the gradual learning about β. In Bacchetta and van Wincoop (2013), unobserved fundamentals can i.a. reflect macroeconomic news.[4] However, while trying to explain the arising fluctuations rationally, heterogeneously informed investors attribute these shocks to an observable fundamental which temporarily receives an excessive weight as a result. Therefore, the Bacchetta and van Wincoop (2013) model leads to similar

[4]Newly released information often coincides with other events that can obfuscate the origin of a shock.

relationship between fundamentals and the exchange rate as the model in this paper.

1.3 MODELLING PARAMETER INSTABILITY

1.3.1 *State Space Formulation*

This subsection demonstrates how the theoretical model can be mapped into state space form. Focus on the single factor case for illustrative purposes and combine equation (1.5) with the autoregressive processes for the transitory shocks to obtain the system:

$$\begin{aligned} \kappa_t &= \rho_\kappa \kappa_{t-1} + u_t, \\ \Delta s_t &= f_t \left(\frac{\mu}{\mu - \rho_f} \beta + \frac{\mu}{\mu - \rho_\kappa} \kappa_t \right) - \frac{1}{\mu - \rho_\phi} \phi_t. \end{aligned} \tag{1.7}$$

To estimate the relation of exchange rates to fundamentals, $\frac{\mu}{\mu-\rho_f}\beta + \frac{\mu}{\mu-\rho_\kappa}\kappa_t$, write the system in the following state space representation:

$$\begin{aligned} \lambda_t - \bar{\lambda} &= b(\lambda_{t-1} - \bar{\lambda}) + \eta_t, & \eta_t &\sim i.i.d.(0, \sigma_\eta^2), \\ \Delta s_t &= f_t' \lambda_t + \epsilon_t, & \varepsilon_t &\sim i.i.d.(0, \sigma_\varepsilon^2), \end{aligned} \tag{1.8}$$

where the measurement error ε_t is an estimate of the risk premium,[5] and the state vector λ_t estimates the relation between macroeconomic fundamentals f_t and the exchange rate s_t. By comparing equations (1.7) and (1.8), it can be seen that $\lambda_t = \frac{\mu}{\mu-\rho_f}\beta + \frac{\mu}{\mu-\rho_\kappa}\kappa_t$; hence, the parameters of the state space representation (1.8) can be mapped to the parameters in equation (1.7):

$$\begin{aligned} \bar{\lambda} = \mathbb{E}[\lambda_t] &= \mathbb{E}\left[\frac{\mu}{\mu - \rho_f} \beta + \frac{\mu}{\mu - \rho_\kappa} \kappa_t \right] = \frac{\mu}{\mu - \rho_f} \beta, \\ \frac{\sigma_\eta^2}{1 - b^2} = \mathbb{V}[\lambda_t] &= \mathbb{V}\left[\frac{\mu}{\mu - \rho_f} \beta + \frac{\mu}{\mu - \rho_\kappa} \kappa_t \right] = \omega^2 \frac{\sigma_\kappa^2}{1 - \rho_\kappa^2}, \end{aligned}$$

where $\omega = \frac{\mu}{\mu-\rho_\kappa}$ and $\mathbb{V}$ denotes the variance. The autocorrelation parameter ρ_κ of κ_t corresponds to the autocorrelation parameter b of λ_t. Therefore, estimating state space system (1.8) will give estimates of the parameter vector $\frac{\mu}{\mu-\rho_f}\beta$ and estimates of the transitory shock process scaled by ω. The

[5] Specifically, we have $\varepsilon = -\frac{1}{\mu-\rho_\phi}\phi_t$ and $\sigma_\varepsilon^2 = \frac{\sigma_\phi^2}{(\mu-\rho_\phi)^2(1-\rho_\phi^2)}$.

state space representation (1.8) can easily be generalised to the multivariate case with r observed fundamentals $f_t = (f_{1t}, ..., f_{rt})'$ and state vector $\lambda_t = (\lambda_{1t}, ..., \lambda_{rt})'$:

$$\begin{aligned} B(L)(\lambda_t - \bar{\lambda}) &= \eta_t, \qquad \eta_t \sim i.i.d.(0, Q), \\ \Delta s_t &= f_t' \lambda_t + \varepsilon_t, \qquad \varepsilon_t \sim i.i.d.(0, \sigma_\varepsilon^2), \end{aligned} \tag{1.9}$$

where $B(L) = I - B_{t,1}^0 L - \cdots - B_{t,q}^0 L^q$ is a q^{th}-order lag polynomial with roots outside the unit circle. The covariance matrix of the state innovation, η_t, is $\mathbb{E}[\eta_t \eta_t'] = Q$.

1.3.2 *A Factor Model with Time-Varying Loadings*

To estimate the system (1.9) empirically, we specify a factor model with time-varying loadings. We use a large panel of macroeconomic data series $X_t = (X_{1t}, ..., X_{Nt})', t = 1, ..., T$, whereby we assume that X_{it} has a factor structure:

$$X_{it} = \alpha_{it}' f_t + \epsilon_{it},$$

where f_t is an $r \times 1$ vector of common factors, α_{it} are the corresponding time-varying factor loadings, and ϵ_{it} are idiosyncratic errors. In our application below, N and T are of the same order of magnitude, which renders one-step maximum likelihood estimation of the model infeasible due to the number of parameters to be estimated (Shumway and Stoffer, 1982; Bai, 2003).

Progress towards estimability of factor models by maximum likelihood has been made in Doz et al. (2011) and Doz et al. (2012), who established a two-step procedure that first uses principal components to estimate the factors. To identify the effect of fundamentals on the exchange rate in the presence of structural instabilities in this paper, we employ two important theoretical results: (i) the principal component estimator gives consistent factor estimates even in the presence of time-varying loadings (Bates et al., 2013). (ii) Maximising the likelihood of a factor model with principal components as estimators of the unobservable factors gives consistent estimates of stationary time-varying loadings (Mikkelsen et al., 2019). We refer to the latter paper for details on how the estimation error from the first step is controlled and consistency is established.

The factor model allows the idiosyncratic errors to have limited cross-sectional correlation. The number of factors, r, is considerably smaller than the number of series, N, such that the information in the large number of

macroeconomic variables is condensed into the r-dimensional factors. That is, by extracting the first r principal components of X_t, one can construct a set of macroeconomic factors that represents the information contained in the observable fundamentals. The principal components estimator treats the loadings as being constant over time, i.e. $\alpha_{it} \equiv \alpha_i$, and solves the minimisation problem:

$$(\tilde{f}, \tilde{\alpha}_i) = \min_{f, \alpha_i} (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it} - \alpha_i' f_t)^2,$$

where $\tilde{f}$ is a $T \times r$ matrix of common factors, and $\tilde{\alpha}$ is an $r \times 1$ vector of factor loadings. By concentrating out α_i and imposing the normalisation constraint $f'f/T = I_r$, the minimisation problem becomes equivalent to maximising $\text{tr}(f'(X'X)f)$, where X is the $T \times N$ matrix of observations. The resulting factor matrix is given by $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of the $T \times T$ matrix XX'. It follows from Bates et al.'s (2013) main result that the fundamentals in equation (1.9) can be represented through the r principal component estimates, $\tilde{f}_t$, in spite of the structural instability underlying the state vector λ_t.
Having obtained the principal component estimates, we estimate the parameters of the state space model (1.9) by forming the likelihood function:

$$\mathcal{L}_T(\Delta s | \tilde{f}; \theta) = -\frac{1}{2} \log(2\pi) - \frac{1}{2T} \log |\Sigma| - \frac{1}{2T} (\Delta s - \mathbb{E}[\Delta s])' \Sigma^{-1} (\Delta s - \mathbb{E}[\Delta s]),$$

where $\Delta s = (\Delta s_1, ..., \Delta s_T)'$ with mean $\mathbb{E}[\Delta s] = f\bar{\lambda}$ and variance matrix $\mathbb{V}[\Delta s] = \Sigma$. The parameter vector $\theta = (B(L), \bar{\lambda}, Q, \sigma_\varepsilon^2)$ is estimated as:

$$\tilde{\theta} = \arg\max_{\theta} \mathcal{L}_T(\Delta s | \tilde{f}; \theta).$$

The likelihood can be computed efficiently with the Kalman filter as (1.9) is a linear state space system. Mikkelsen et al. (2019) show that under standard assumptions and provided $T/N^2 \to 0$, the maximum likelihood estimator is consistent for the parameters of the time-varying factor loadings λ_t, i.e. $\tilde{\theta} \xrightarrow{p} \theta$. Once $\tilde{\theta}$ is obtained, the estimates of the factor loadings $\tilde{\lambda}_t$ for $t = 1..., T$ are computed with the state smoother. As emphasised in Mikkelsen et al. (2019), these estimates are consistent even under missing factors. In addition to the time-varying model, we estimate a constant loadings (CL) benchmark in order to assess the relative contribution of time-varying loadings in explaining exchange rate fluctuations. In that case, equation (1.7) reduces to $\Delta s_t = f_t \frac{\mu}{\mu - \rho_f} \beta - \frac{1}{\mu - \rho_\kappa} \phi_t$, i.e. a present value model

for exchange rates with constant parameters. Therefore, the reduced form relation between fundamentals and exchange rates, $\frac{\mu}{\mu-\rho_f}$, can simply be estimated via OLS by regressing Δs_t on the factors $\tilde{f}_t$. We denote $\frac{\mu}{\mu-\rho_f}\beta$ by $\tilde{\lambda}_{CL}$. Comparing the in- and out-of sample fit of the two models determines if, indeed, the TVL model fares better at explaining the relationship of exchange rates and fundamentals.

1.4 DATA

1.4.1 *Exchange Rate Data*

Table 1.1: Summary Statistics Exchange Rates

Currency	Mean	Std.Dev.	Min.	Max.	$\hat{\rho}(1)$	$\hat{\rho}(12)$	$\hat{\rho}(24)$	Q_{BP}
AUD	-0.000	0.026	-0.180	0.073	0.342	-0.084	-0.076	0.000
CAD	-0.000	0.017	-0.109	0.062	0.307	-0.091	-0.040	0.000
DKK	0.000	0.023	-0.078	0.062	0.309	-0.093	-0.005	0.000
JPY	0.001	0.025	-0.080	0.103	0.292	-0.035	-0.013	0.000
MXN	-0.005	0.033	-0.321	0.088	0.265	-0.085	0.005	0.000
NZD	0.001	0.026	-0.106	0.074	0.310	0.002	-0.029	0.000
NOK	-0.001	0.026	-0.131	0.057	0.371	-0.107	0.027	0.000
SEK	-0.001	0.026	-0.109	0.071	0.402	-0.104	-0.047	0.000
CHF	0.001	0.025	-0.112	0.081	0.214	-0.063	0.042	0.001
GBP	-0.000	0.022	-0.110	0.059	0.290	-0.036	-0.096	0.000
BRL	-0.033	0.085	-0.368	0.113	0.851	0.486	0.326	0.000
INR	-0.004	0.020	-0.194	0.061	0.200	-0.081	0.019	0.001
ZAR	-0.005	0.035	-0.190	0.152	0.290	-0.078	-0.053	0.000
EUR	-0.000	0.023	-0.079	0.062	0.319	-0.081	-0.021	0.000

Note: Sample Period: 1990:05 - 2021:09. $\hat{\rho}(m)$ denotes the autocorrelation at month m. Q_{BP} denotes the p-value of the Box-Pierce Q_{BP} test. The currency abbreviations stand for Australian Dollar (AUD), Brazilian Real (BRL), Canadian Dollar (CAD), Danish Krone (DKK), Indian Rupee (INR), Mexican Peso (MXN), New Zealand Dollar (NZD), Norwegian Krone (NOK), South African Rand (ZAR), Swedish Krona (SEK), Swiss Franc (CHF), British Pound (GBP), and Euro (EUR).

We use monthly averages of the US dollar exchange rate vis-à-vis 14 currencies between 1990:05 and 2021:09. The considered exchange rates are: the Australian Dollar (AUD), the Brazilian Real (BRL), the Canadian Dollar (CAD), the Danish Krone (DKK), the Indian Rupee (INR), the Mexican Peso (MXN), the New Zealand Dollar (NZD), the Norwegian Krone (NOK), the South African Rand (ZAR), the Swedish Krona (SEK), the Swiss Franc (CHF), the British Pound (GBP), and the Euro (EUR). The data are compiled from the

OECD database.[6] Table 1.1 reports summary statistics for the first difference of the 14 log exchange rates. Looking at the mean percentage changes, they are all either zero or very close to zero with standard deviations ranging from 1.7% to 8.5%. In terms of fluctuations, the Brazilian Real displays the largest downward movement with -36.8%, whereas the South African Rand appreciated the most over one month with 15.2%. All currencies are positively autocorrelated at one month and with two exceptions negatively correlated at 12 and 24 months. The Box-Pierce test implies that, across currencies, the first three autocorrelations are all statistically significant.

1.4.2 *Macroeconomic Data*

The factors are extracted from a large set of real-time macroeconomic fundamentals. To this end, we combine two different databases. First, we use McCracken and Ng's (2016) FRED-MD database which contains 128 monthly time series of the US economy, categorised into: (1) Output & Income, (2) Labour Market, (3) Housing, (4) Orders & Inventories, (5) Money & Credit, (6) Interest Rates, (7) Prices, and (8) Stock Markets. Following McCracken and Ng (2016), 5 time-series are removed to balance the panel; in addition, we remove the exchange rates in the dataset to exclude them from the factors. In order to ensure stationarity of all variables, we use the same transformations as McCracken and Ng (2016) and refer to their paper for a detailed description. We collate all real-time vintage releases of the FRED-MD database, starting in 1999:08 and ending in 2022:02. For each vintage of the database, observations are available up to the month that precedes its release date. We remove all time series whose latest observation lags the release date by two months or more. For instance, all time-series ending before 2019:12 are removed from the 2020:01 vintage. As the time lag with which each variable is released changes over time, the number of macro series in the dataset lies between 100 and 128 for each release.

Second, we compile large vintage datasets for the 14 remaining countries from the OECD statistical database. To increase the number of available variables, we collect data for the 3 largest Euro area economies, Germany, France, and Italy, instead of an aggregate measure. Specifically, for each country, we compile a maximum of 9 macro variables, 32 survey indicators, and 3- as well as 10-year yields on Government bonds. We collect real-

[6] Prior to 1999, the exchange rate for the ECU is used in place of the Euro, i.e. an weighted average of the Austrian Schilling, Belgian and Luxembourg Francs, Finnish Markka, French Franc, German Mark, Irish Pound, Italian Lira, Netherlands Guilder, Portuguese Escudo, and Spanish Peseta.

time vintages of this dataset from the OECD's Revision Analysis Database, starting with the first set of vintages released between 1999:02 and 2022:02. As with the FRED-MD data, we remove all variables whose latest observation predates the publication date by more than one month. In many countries, data are published less timely or consistently than in the US. For example, suppose industrial production data for Mexico are published with a time-lag of two months in 2000:08, i.e. the latest available observation for the respective vintage dataset is 2000:06. Then this particular time series will be dropped from the 2000:08 vintage for Mexico to ensure the panel is balanced over time. The first observation in each vintage is recorded in 1990:04. For instance, suppose India only started publishing retail sales data in 1995:05, then no vintage will include this particular variable for India. Therefore, upon removal, the number of macro series available in each vintage version of the dataset lies between 61 and 100, that is, the datasets are balanced across time but unbalanced across countries. We obtain a total of 272 real-time datasets for all countries, the first of which contains $T = 105$ observations per variable, and the last $T = 382$ observations per variable. In each vintage, we group the variables according to the categories in McCracken and Ng (2016). Consistent with McCracken and Ng (2016), the variables are transformed either by taking first log differences or first differences to ensure stationarity. Tables A.1 to A.6 summarise which variables are available for each country and how they are categorised in terms of McCracken and Ng's (2016) classifications.[7] We merge each real-time dataset with the corresponding vintage of the FRED-MD database, yielding 272 real-time vintages with a total number of variables between $N = 168$ and $N = 209$. We use all of these datasets in our out-of-sample estimation, as they replicate the information set available to investors during each month. Thereby, rather than simply using the first release of each observation, we take into account that investors have knowledge of revisions to published data prior to the current month, which could potentially impact their decisions. Coroneo and Caruso (2022) show that this improves forecasts of financial variables relative to just relying on first releases.

For our in-sample exercise, we use the 2022:02 vintage of the FRED-MD and OECD dataset and remove all variables with missing observations before 2021:09. Thus, the time dimension corresponds to $T = 380$. As several variables are released with a lag of more than 1 month, this increases the number of available macro series to $N = 224$. Not every series is available for all countries, hence the in-sample dataset is also balanced across time

[7] A complete description of all vintages in the dataset can be found in the Supplementary Material.

but unbalanced across countries.

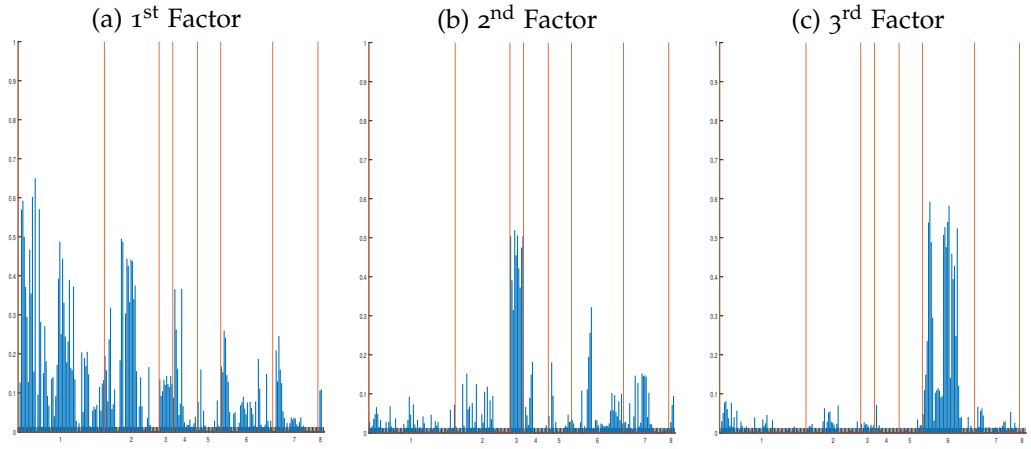


Figure 1.1: Marginal R^2 Between Factors and Macro Series

Note: R^2 from regression of each series on first, second, and third factor. Series categorised as (1) Output & Income, (2) Labour Market, (3) Housing, (4) Orders & Inventories, (5) Money & Credit, (6) Interest Rates, (7) Prices, and (8) Stock Market.

In practice, the optimal number of factors describing $X_{it} = \alpha_{it}' f_t + \epsilon_{it}$ is not apparent, and multiple factor selection criteria for models where N and T are large exist. Let $V(k) = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it} - \tilde{\alpha}_i^{k\prime} \tilde{f}_t^k)^2$. While setting the number of factors in the model, k, equal to N minimises $V(k)$, it does not imply N corresponds to the optimal number of factors r. Bai and Ng (2002) propose information criteria with a penalty function $g(N,T)$ such that: $r = \arg\min_{0 \leq k \leq k_{max}} \mathrm{IC}(k) = \arg\min_{0 \leq k \leq k_{max}} \log(V(k)) + kg(N,T)$, where $k_{max} \in \mathbb{N}$ is a maximum number of factors chosen by the researcher (here: $k_{max} = 9$). Due to the penalty term, $r \ll N$. Choi and Jeong (2019) compare the performance of different approaches and suggest using several criteria in combination. We follow their recommendation and first evaluate Bai and Ng's (2002) IC_{p2} and BIC_3 which both pick $r = 9$ factors. Subsequently, we consider several criteria with improved robustness to miss-specification that are found to perform well in Choi and Jeong (2019). Alessi et al. (2010) propose modifications of the penalty functions in Bai and Ng (2002) based around an arbitrary constant, c, as in Hallin and Liška (2007). We set $c \in (0, 10]$ which leads to the conclusion that the optimal number of factors is 1. Kapetanios (2010) suggests a criterion with improved robustness to cross-sectional dependence. When applying the Alessi et al. (2010) modification to this criterion, the optimal number of factors is again found to be 1. In accordance with the advice in Choi and Jeong (2019), we also considered the eigenvalue-based approaches in Ahn and Horen-

stein (2013). Both the ER and GR test imply $r = 5$, suggesting that a low number of factors is indeed a plausible choice. Therefore, 1 and 3 factors are deemed an appropriate choice for the empirical estimation in this paper.[8]

In Figure 1.1, we depict the squared correlation of the factors with each macro variable, categorised as described in the figure caption. The displayed correlations are based on the in-sample dataset between 1990:04 and 2021:09. The first factor exhibits strong correlations with measures of output, labour market indicators, and to a lesser extent with manufacturing orders and capacity utilisation. Therefore, we interpret the first factor as an indicator of real economic activity. The second factor correlates mainly with housing data, wherefore we interpret it as a housing factor. Regarding the third factor, it displays strong correlations with interest rates and we interpret it as an interest rate factor.[9]

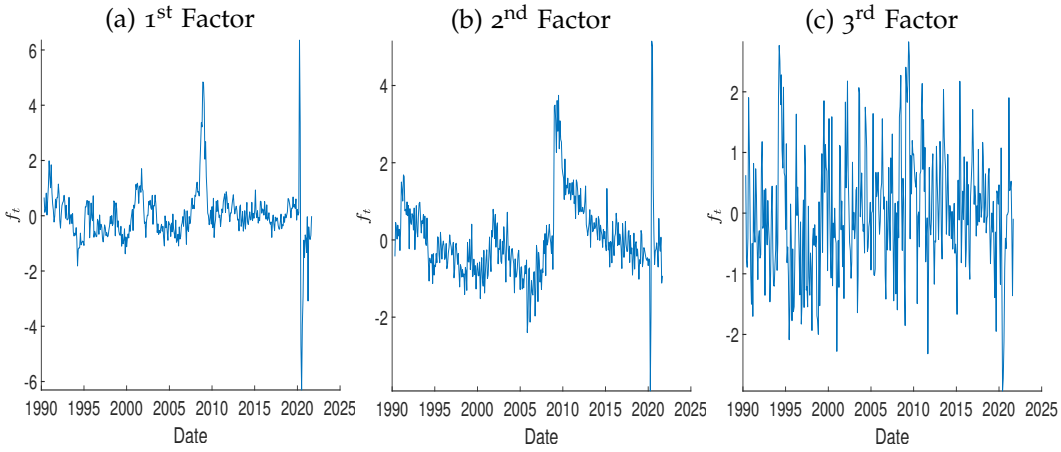


Figure 1.2: Principal Components

Note: The Figure displays the time-series of the first three principal components extracted from the macro-dataset.

In Figure 1.2, the individual factors are plotted. For the first factor, the drop in economic activity during the great recession is clearly visible and so is the COVID-19 crisis towards the end of the sample. The second factor exhibits a structural change after the subprime crisis and large short-dated fluctuations around the COVID-19 crisis. The factor loadings are not sign identified, hence, both the first and second factors are inversely related to the

[8] Appendix C also reports the results for 5 factors.

[9] We provide an animation of how the correlations change over time for the vintages used in our out-of-sample exercise in the Supplementary Material.

real economy and housing markets. Looking at the third factor, it displays less extreme fluctuations, with an overall minimum in 2020.

1.5 IN-SAMPLE RESULTS

This section presents the empirical results of estimating the state space model in equation (1.9) by comparing the constant and the time-varying loadings model in-sample. The discussion focuses on the GBP and the EUR while covering the remaining exchange rates more succinctly. To conduct a comparison of the two models, we use the squared correlation, R^2, between the exchange rate and the in-sample predictions. Furthermore, we report the hit rate (HR) of each model, i.e. the percentage of times the model matched the signs of the exchange rate changes. The hit rate indicates how often a model correctly predicts a depreciation or appreciation. The two criteria are shown in Table 1.2. In addition, the table reports the p-values of a likelihood-ratio test with a null hypothesis of no significant differences in the likelihoods of the two models.

Consider first the estimates for one factor, the real economy factor. The top panel of Figure 1.3 shows the results for the GBP and the EUR: the common component obtained from the TVL model (blue), the CL model (red), and the actual exchange rate changes (black). Particularly during the great recession, the time-varying model can capture the fluctuations in the exchange rate better. This is reflected in the R^2, according to which the model can explain 35% (23%) of the variation in the EUR (GBP). It assigns accurate directional changes in 73.7% of the cases for the EUR and 60.5% for the GBP. In contrast, the CL model only has an R^2 of 1% and 3%, respectively, i.e. it has almost no explanatory power. With 47.5%, the hit rate of the CL model for the EUR is as good as random, the same holds for the GBP with 52.3%. It should be noted that the CL results for the GBP are among the highest out of the 14 exchange rates. The lowest are the ones for CHF and JPY where 0% of the fluctuations are explained. Looking at the time-varying model, it has the highest explanatory power for the MXN with an R^2 of 54%, and the lowest R^2 for the INR (1%). In case of the latter, the TVL model and the CL model have the same R^2 which suggests the real economy factor offers no explanatory power for the INR. Nevertheless, the time-varying 1-factor model adds substantial explanatory power over the model with constant coefficients for all other currencies, as it consistently outperforms the CL model according to the two metrics. This conclusion is supported by the likelihood ratio tests. With three exceptions in case of the 1-factor model (INR, CHF, and JPY), the test always finds that the likelihood of the CL

Table 1.2: In-Sample Performance

Currency	1 Factor					3 Factors				
	R^2		Hit Rate		LR	R^2		Hit Rate		LR
	TVL	CL	TVL	CL	p-val.	TVL	CL	TVL	CL	p-val.
AUD	0.44	0.01	74.54	49.60	0.00	0.53	0.04	75.86	56.76	0.00
CAD	0.32	0.02	70.29	52.52	0.00	0.36	0.04	68.70	53.32	0.00
DKK	0.33	0.01	70.29	47.21	0.00	0.61	0.01	79.58	54.91	0.00
JPY	0.09	0.00	57.03	52.25	0.19	0.41	0.00	67.90	53.58	0.01
MXN	0.54	0.01	71.35	54.38	0.00	0.55	0.02	68.44	50.40	0.00
NZD	0.27	0.02	70.03	49.34	0.00	0.51	0.05	78.78	57.56	0.00
NOK	0.28	0.02	74.01	46.95	0.00	0.41	0.03	75.07	52.79	0.00
SEK	0.23	0.03	74.27	50.66	0.00	0.48	0.04	76.66	54.91	0.00
CHF	0.37	0.00	69.50	44.83	0.07	0.64	0.01	82.76	52.79	0.06
GBP	0.23	0.03	60.48	52.25	0.00	0.27	0.04	65.52	57.29	0.01
BRL	0.50	0.01	72.94	53.58	0.00	0.86	0.01	86.47	55.70	0.00
INR	0.01	0.01	50.13	50.13	1.00	0.88	0.01	89.39	52.25	0.00
ZAR	0.39	0.03	69.76	50.40	0.00	0.48	0.06	76.13	55.44	0.00
EUR	0.35	0.01	73.74	47.48	0.00	0.61	0.01	81.70	53.85	0.00
Mean	0.31	0.01	68.45	50.11		0.54	0.03	76.64	54.40	
Median	0.32	0.01	70.29	50.27		0.52	0.02	76.39	54.38	

Note: The table reports measures of in-sample fit to compare the TVL model in equation (1.9) and the CL model. Namely, both the squared correlations between changes in the exchange rate and the in-sample prediction of the TVL & CL model as well as the hit rate in %. The latter is the fraction of times the sign of the fitted values corresponded to the sign of the realised values. In addition, the table reports the p-values of a likelihood ratio test (LR) between the two models. The currency abbreviations stand for Australian Dollar (AUD), Brazilian Real (BRL), Canadian Dollar (CAD), Danish Krone (DKK), Indian Rupee (INR), Mexican Peso (MXN), New Zealand Dollar (NZD), Norwegian Krone (NOK), South African Rand (ZAR), Swedish Krona (SEK), Swiss Franc (CHF), British Pound (GBP), and Euro (EUR).

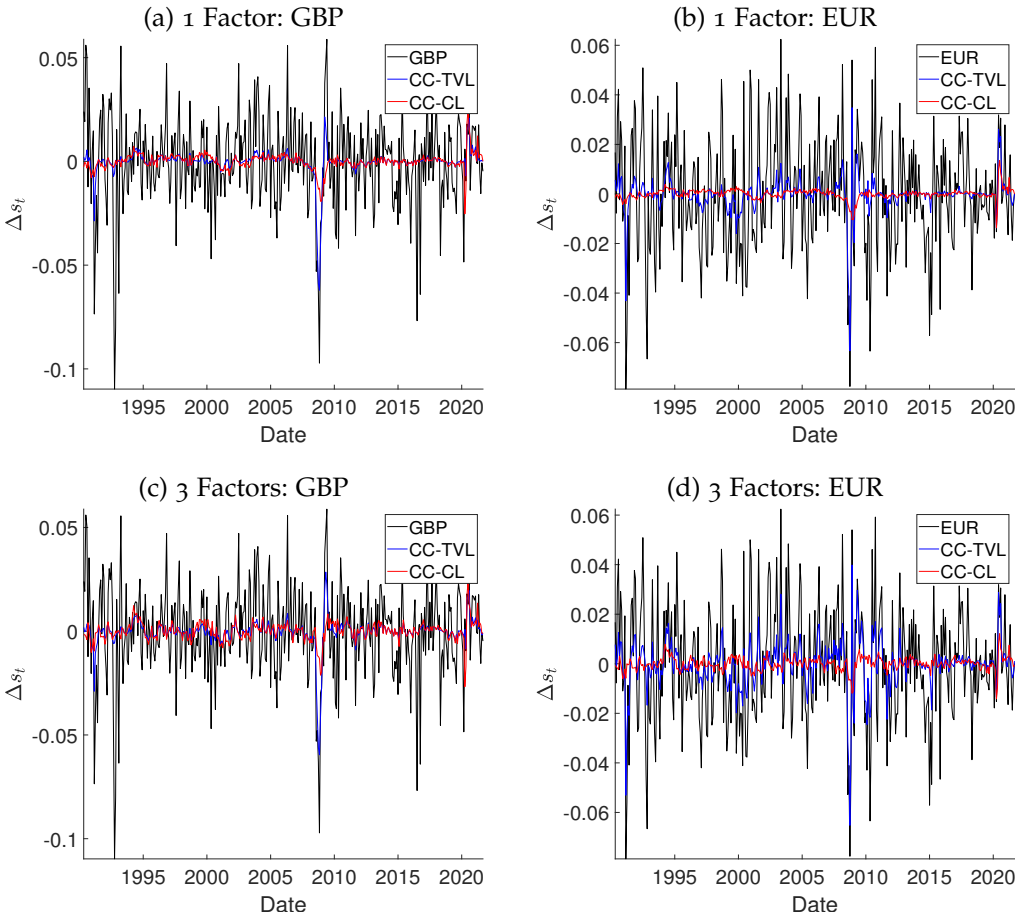


Figure 1.3: In-Sample Fit

Note: The figure displays the in-sample fit for the Euro (EUR) and the British Pound (GBP) of a model with 1 and 3 factors. The black line is the FX change, the blue line is the TVL model fit, and the red line is the CL model fit. CC-TVL stands for the Common Component of the State Space model, CC-CL for the Common Component of the CL model.

model is significantly smaller. In a next step, we also include the interest

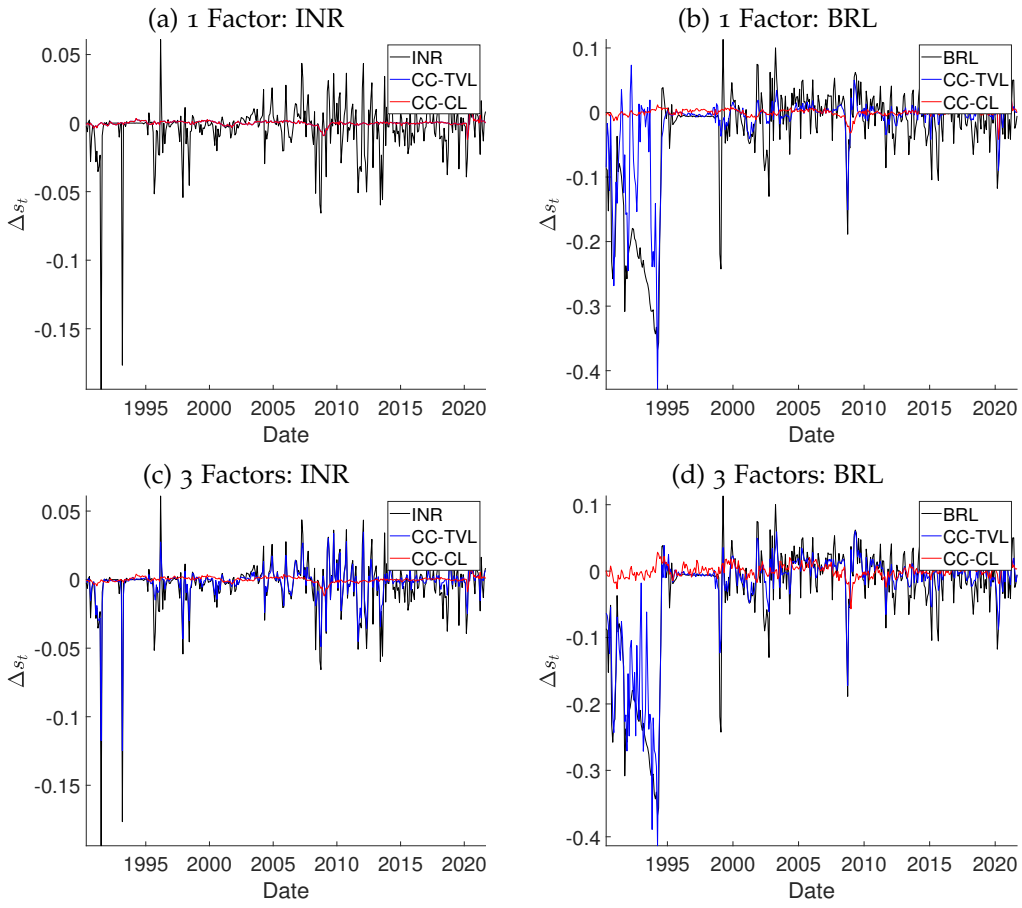


Figure 1.4: In-Sample Fit – BRL and INR

Note: The figure displays the results for the Brazilian Real (BRL) and the Indian Rupee (INR) of a model estimated with 1 and 3-factors. The black line is the FX change, the blue line is the TVL model fit, and the red line is the CL model fit. CC-TVL stands for the Common Component of the State Space model, CC-CL for the Common Component of the CL model.

rate and the housing factor into the model. The actual and fitted values for the GBP and the EUR are presented in the bottom panel of Figure 1.3. In particular for the EUR, the fit of the time-varying model improves visibly – the model tracks the depreciation during the Euro-crises in 2010, 2012, and 2015 remarkably well. The same holds true for the early 2000s. The fit for the GBP also appears to have improved, even though to a lesser extend. Note that, as before, the GBP exhibits the worst fit of all time-varying regressions with an R^2 of 27% followed by the CAD with 36%. Still, it manages to predict whether the exchange rate appreciates or depreciates in 65.5% of all cases. Regarding the EUR, the explanatory power of the time-varying model has

jumped to an R^2 of 61%, and the hit rate corresponds to 81%. On the other hand, the CL model still only manages to explain 1% of the variations in the EUR and 4% in the GBP. Overall, the time-varying model of the INR has now the highest R^2 and hit rate with 88% and 89.4%, respectively – having displayed the worst fit in the 1-factor model. This implies that only the second and third factors offer predictive ability for this currency. The INR is followed by BRL, CHF, EUR, and DKK for which the explanatory power always exceeds 60%. Generally, we see an improvement in the R^2 across exchange rates – the hit rate declines slightly for the MXN in the time-varying framework but is close to 80% in most cases; however, looking at the CL fit, it is still only slightly better than 50% and never above 58%. After the INR, the BRL exhibits the greatest improvement with the R^2 increasing to 86% from 50%. We graph the differences between the 1 and 3-factor model in Figure 1.4. Both INR and BRL were hit by a currency crisis at the beginning of the sample period. Regardless of whether 1- or 3-factors are used, the CL model can map neither fluctuations in INR nor BRL. In contrast, the TVL model with 3 factors is able to capture both currency crises as well as subsequent variations. For the remaining series, the 1-factor model is also able to capture exchange rate fluctuations better around the financial crisis (see Figure A.1). As is the case with the EUR, GBP, INR, and BRL, the 3-factor model substantiates the ability of the time-varying model to outperform the constant loadings framework. We further underpin the robustness of the in-sample findings by re-estimating the model using 5 factors and report the results in Appendix A.4.1. While the R^2 for the CL model improves somewhat to a maximum 7%, it remains considerably lower than the one of the TVL model across all currencies. The same holds true for the hit-rate which ranges between 66% and 90% for the time-varying and 50% to 68% for the CL model.

1.5.1 *Instabilities in Factor Loadings*

This subsection considers the role of parameter instability in greater detail. First, we revisit the GBP and the EUR and discuss the time-varying loadings on the real economy factor, depicted in the left panel of Figure 1.5.

The dashed red lines correspond to the confidence intervals of the constant loadings estimates. For both currencies, we observe a high degree of variation in the time-varying loadings, especially in 2008-9. In the GBP model, the loadings decline first and then rise sharply; therefore, they are considerably outside the CL confidence interval during the financial crisis. This is a

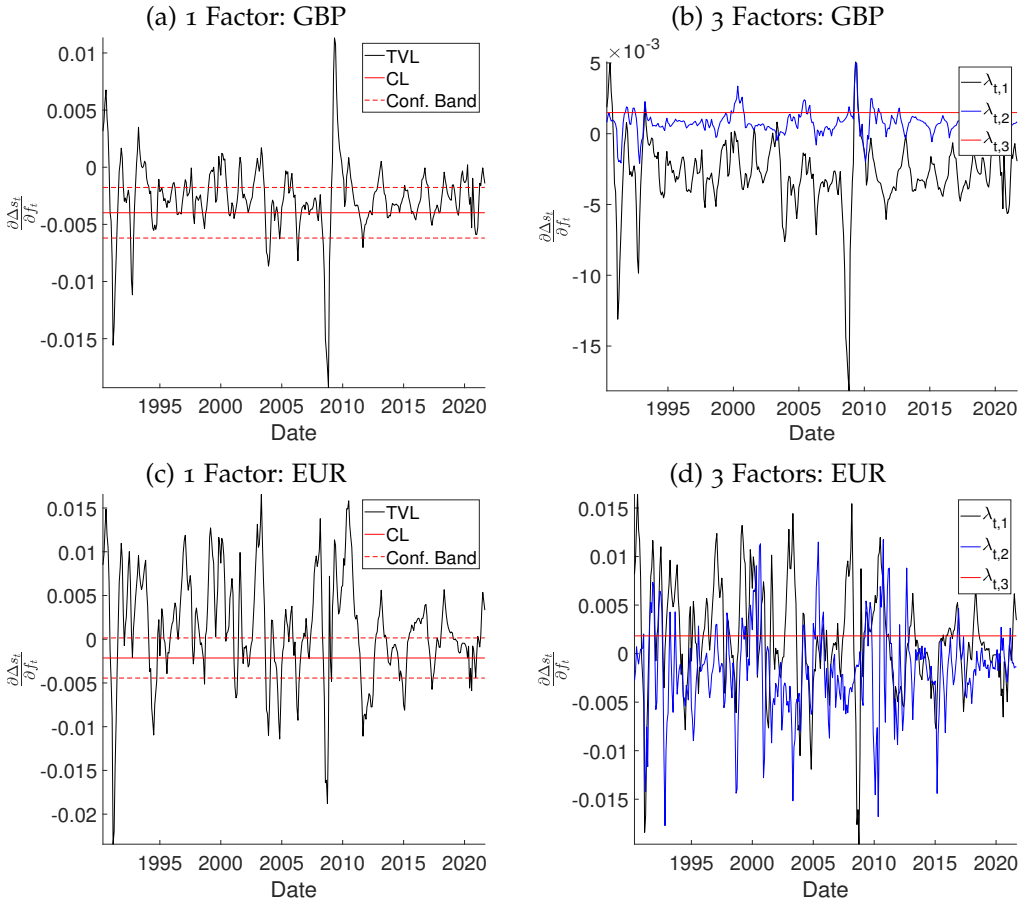


Figure 1.5: Loadings – GBP & EUR

Note: The left panel of the figure displays the loadings on the real economy factor. The black line is the time-varying loading (TVL), the red line the constant loadings (CL) estimate, and the dashed lines are the CL confidence intervals. The right panel displays the loadings of the 3-factor model. The black line is the loadings on the real economy factor, the blue line the loadings on the housing factor, and the red line the loadings on the interest rate factor.

consistent theme across exchange rates.[10] The loading in the EUR model crosses the CL confidence bands more often, displaying large fluctuations throughout the sample period. However, the EUR CL estimate itself is insignificant. Contrary to the GBP, the loadings exhibit large positive spikes in the early 2000s, consistent with the EUR depreciation *vis-à-vis* the dollar during that period. Although the loadings are not sign identified, Figure 1.5 shows that the loadings display frequent sign changes, in the sense that they fluctuate significantly above and below the OLS confidence bands. This points to significant instabilities in the factor loadings that the CL model cannot capture. The right panel of Figure 1.5 plots the loadings on the first three factors. It can be seen that the explanatory power of the second factor differs across the two currencies with the magnitude of the loadings being much greater for the EUR. What is more, as Mikkelsen et al.'s (2019) methodology is robust to missing factors, the first loadings in the 3-factor model are always identical to the loadings in the 1-factor model.

1.5.2 *Instabilities in Factor Structures*

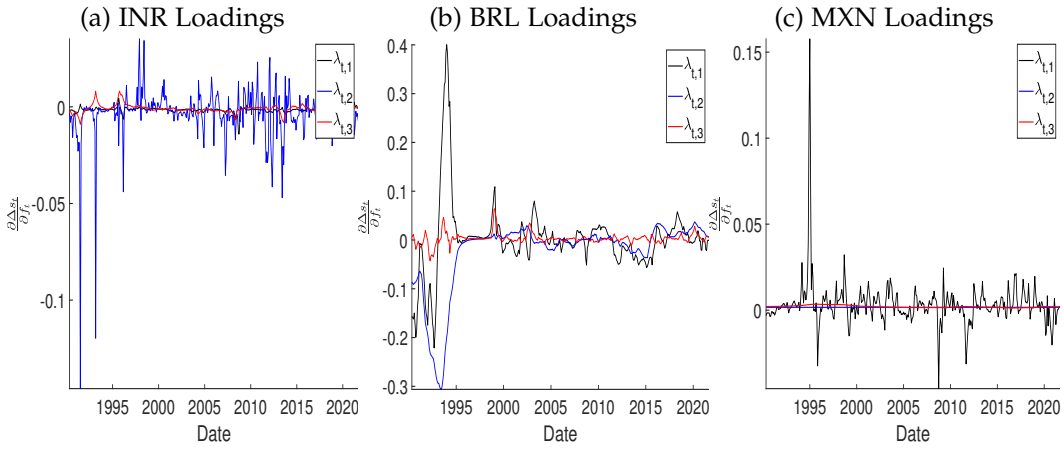


Figure 1.6: Loadings – Unstable Factor Structure

Note: The figure displays the loadings of the 3-factor model. The black line are the loadings on the real economy factor, the blue line the loadings on the housing factor, and the red line the loadings on the interest rate factor.

Changes in factor structures occur if either (i) the number of factors changes, (ii) related to this, some factors disappear and others appear, or (ii) the direction of the influence of a factor changes. Time-varying loadings

[10] The plots of the first factor loadings for the remaining 12 currencies are reported in Appendix B.

capture a reduction in the number of factors, since loadings can be (close to) zero for periods. They also capture changes in the direction of influence, since the sign of the loadings can change. The right panel of Figure 1.5 shows that in case of the GBP, the second loadings are always close to zero apart from few deviations. For both currencies, the third loadings are constant and only marginally above zero, meaning only the first and second factors affect these currencies. Next, we look at three emerging market currencies that were subject to crises at the beginning of the sample period. Therefore, they display greater instabilities in both loadings and factor structure. Figure 1.6 plots the loadings on the first three factors for INR, BRL, and MXN. Several insights emerge from the figure. First, the dominant loadings differ from currency to currency. While the first loadings are virtually zero for the INR, the second loadings exhibit large swings and the third loadings show only temporal deviations from zero. With respect to the latter, the same is true for the BRL. However, both the second and third factors load heavily onto the BRL during the early 1990s, while fluctuating less drastically in the 2000s where the loadings are close to zero in some months. This contrasts with the MXN, for which only the first loadings are markedly different from zero, with large spikes during the Peso crisis and the financial crisis.

Unlike the CL model, the TVL framework can account for such changes in the importance of certain factors. It is therefore able to model instabilities in both loadings and factor structures which explains the superior fit discussed above.

1.6 OUT-OF-SAMPLE RESULTS

This section compares the performance of the two models, CL and TVL out-of-sample using the 3-factor model. First, we elaborate on the chosen forecast evaluation methods; and second, assess the forecasting results. Specifically, we compare (i) the relative predictive ability of the two models, (ii) their direction accuracy, i.e. how well they forecast an appreciation or depreciation, and (ii) their relative performance over time. For the out-of-sample estimation, we use the large vintage datasets described above.

1.6.1 *Forecast Setup and Evaluation*

To forecast Δs_{t+h}, where h is the forecast horizon, we divide the sample into in-sample and out-of-sample portions. We denote the total number of observations by T, the number of in-sample observations by R, and the

number of out-of-sample predictions by P, so $T = R + P + h + 1$. We generate direct h-step ahead forecasts according to the following specification:[11]

$$\Delta s_{t+h|t} = f'_{t-1+h|t}\lambda_{t+h|t} + e_{t+h|t}$$

where $t = R, ..., T - h$ and $e_{t+h|t}$ is the forecast error. Note that although the factors, $f_{t-1+h|t}$, are extracted from data one date prior to the observed returns, the estimates are nonetheless conditional on the time-t information set. This is due to the fact that data are released with a lag, as, for example, inflation figures for March are published in April. Our setup ensures the model is truly real-time by accounting for the fact that exchange rates can react to data only upon release. As our dataset includes over 20 survey indicators, we also incorporate the impact of expectations that anticipate some of this response. We set $h = \{1, 6, 12\}$ and use a rolling window to compute P predictions. Forecasts for the loadings, $\hat{\lambda}_{t+h|t}$, are easily obtained as the one-step-ahead predictions of the Kalman filter. The Kalman filter generates optimal forecasts as it, by construction, minimises the Mean Squared Error (MSE) between predictions and observations. For the constant loadings benchmark, λ does not need to be forecast as it simply corresponds to the CL estimates at each iteration. Regarding the factors, we recompute the principal components for each of the 272 real-time datasets. To generate h-step ahead forecasts of the factors, we fit a VAR(1) to the real-time estimates at each of the P forecasting steps and use the coefficients to forecast $\hat{f}_{t-1+h|t}$. One then obtains the out-of-sample estimates as $\widehat{\Delta s}_{t+h|t} = \hat{f}'_{t-1+h|t}\hat{\lambda}_{t+h|t}$ and $\widetilde{\Delta s}_{t+h|t} = \hat{f}'_{t-1+h|t}\hat{\lambda}_{CL}$. Both forecasts are only based on the actual information set available in real-time at each period. To assess whether time-varying loadings also improve the out-of-sample fit of the factor model, we compare the TVL and CL forecasts using several tests. In addition we compare both models to a Random Walk (RW) for the level of the log exchange rate, i.e. $\mathbb{E}[s_{t+h|t} - s_t] = 0$.

Selecting adequate tests of predictive ability is of paramount importance in out-of-sample evaluation. The properties of tests for nested models are different because their forecast errors converge asymptotically (Clark and McCracken, 2001). Furthermore, window choice is an important determinant in forecast evaluation. A large P provides more forecast information, while a large R improves parameter accuracy. In fact, Mikkelsen et al. (2019) show through Monte-Carlo simulations that in order for the bias in the

[11] See Boivin and Ng (2005) for a comparison of different approaches to generating factor-based forecasts.

autoregressive parameters of the loadings to be below 10%, the estimation sample should be $R \geq 200$. However, a litmus test for every exchange rate forecast is the financial crisis. To put the model to this test, the forecasts need to be evaluated using a criterion that is robust to in-sample estimation errors, as one would have $R < 200$ for a prediction window starting prior to the crisis. Giacomini and White (2006) propose a test of Conditional Predictive Ability (CPA) that introduces estimation error under the null hypothesis. Define the forecast loss differential between TVL and CL as $\{\Delta\mathcal{L}_t(\hat{f}'_{t-1}\hat{\lambda}_t, \hat{f}'_{t-1}\hat{\lambda}_{CL})\}_{t=R+h}^{T} = \{\mathcal{L}(\Delta s_t, \hat{f}'_{t-1}\hat{\lambda}_t) - \mathcal{L}(\Delta s_t, \hat{f}'_{t-1}\hat{\lambda}_{CL})\}_{t=R+h}^{T}$, where $\mathcal{L}(\cdot)$ is a forecast loss function. The null hypothesis of the CPA test is

$$\mathcal{H}_0 : \mathbb{E}\left[\Delta\mathcal{L}_t(\hat{f}'_{t-1}\hat{\lambda}_t, \hat{f}'_{t-1}\hat{\lambda}_{CL}) \mid \mathcal{F}_t\right] = 0$$

and $\mathcal{F}_t$ is the time-t information set available to the forecaster. For the implementation of the test, we condition on lagged values of the loss differential. In addition, we also use Giacomini and White's (2006) Unconditional Predictive Ability (UPA) test.[12] As both tests are valid for nested as well as for non-nested models and their asymptotic properties are derived for $R < P \to \infty$, they are well-suited in this application. We choose $P = 238$ and $R = 142$ as the baseline configuration and report additional forecasts in Appendix C.

Leitch and Tanner (1991) argue that, while one model may produce a smaller forecasting error than another model, it can still perform worse when it comes to predicting sign changes. In the case of exchange rates, a desirable feature of a model is its ability to forecast an appreciation or depreciation. To assess this statistically, we use Pesaran and Timmermann's (1992) nonparametric test of predictive performance. The test compares the signs of the predicted and realised values and, in doing so, uses no additional information. Thus, it does not require knowledge of the underlying probability distribution of the forecast. Although the test does not put two models in relation to one another, it indicates which model is able to identify a higher number of predictable relationships.

Given the structural instabilities in the exchange rate regression, it may well be the case that the relative forecasting performance of the models is itself unstable. Indeed, Rossi (2013) finds that the forecasting power of many exchange rate models breaks down over time. Notably, however, parameter instability itself does not necessarily engender unstable relative forecast

[12] The UPA test is identical to the popular Diebold and Mariano (1995) test, but derived under different assumptions that render it valid for our purposes.

performance.[13] While the Giacomini and White (2006) test selects the best global model, Giacomini and Rossi (2010) propose a fluctuation test that compares the performance of two competing models at each point in time and allows for nested models by adopting the same asymptotic framework as Giacomini and White (2006). The test statistic is computed over a rolling window and equal to:

$$GR_{t,m} = \hat{\sigma}^{-1} m^{-1/2} \sum_{j=t-m/2}^{t+m/2-1} \Delta\mathcal{L}_j(\hat{f}'_{j-1,R}\hat{\lambda}_{j,R}, \hat{f}'_{j-1,R}\hat{\lambda}_{CL})$$

where $t = R + h + m/2, ..., T - m/2 + 1$, $\hat{\sigma}$ is the HAC estimator of the variance of the loss differential, and m is the size of the rolling window over which it is computed.

1.6.2 *Forecast Results*

Focusing foremost on GBP and EUR, we now discuss the forecasting results. Figure 1.7 depicts the TVL and CL forecasts for the two currencies as well as the realised values. At first glance, it appears the time-varying model performs slightly better for the GBP during the financial crisis – and for the EUR also during the subsequent years. Figure 1.8 plots the forecasts of the remaining series which paint a similar picture; in particular, the INR (Figure 1.8 (f)) is forecast remarkably well.

We report statistical forecast accuracy tests in Table 1.3. Panel A presents the results for one-step-ahead predictions. Columns 2 and 3 of the table report the Root Mean Square Error (RMSE) of the forecasts relative to the RMSE of a random walk forecast. That is, if the value in columns 2 and 3 is smaller than one, the respective model has a lower RMSE smaller than a random walk. The TVL model achieves an RMSE that is between 4.4% and 0.3% smaller than a random walk for 8 exchange rates, and between 0.2% and 2.7% greater for 4 exchange rates. In contrast, the RMSE of the CL model is always between 0.2% and 2.9% larger. We proceed by comparing the TVL and CL models directly. Columns 4 and 5 contain the p-values of the CPA and UPA test, computed for a quadratic loss differential. At the 5% level, the former rejects 4 times in favour of the TVL model, and the latter 6 times. We conduct the same tests using an absolute loss differential (Columns 6 and 7) for which they reject 4 and 3 times, respectively, always in favour of the TVL model. In Columns 8 and 9, we report the p-values of the direction accuracy test, which rejects the null hypothesis of no directional

[13] For a detailed discussion of forecasting under instabilities, see Rossi (2021).

Table 1.3: Forecast Statistics

	RMSE		TVL vs. CL						TVL vs. RW				CL vs. RW			
			Quad.		Abs.		DA		Quad.		Abs.		Quad.		Abs.	
	TVL	CL	CPA	UPA	CPA	UPA	TVL	CL	CPA	UPA	CPA	UPA	CPA	UPA	CPA	UPA
Panel A: 3 Factors, $h = 1$																
AUD	0.972	1.012	0.075	0.024	0.117	0.086	0.029	0.717	0.252	0.119	0.237	0.104	0.502	0.281	0.717	0.931
CAD	0.985	1.004	0.177	0.065	0.329	0.136	0.017	0.187	0.133	0.225	0.340	0.226	0.545	0.752	0.991	0.907
DKK	1.002	1.015	0.445	0.431	0.238	0.110	0.020	0.685	0.698	0.940	0.380	0.566	0.148	0.114	0.240	0.176
JPY	1.027	1.029	0.507	0.908	0.849	0.819	0.323	0.712	0.168	0.088	0.374	0.151	0.042	0.011	0.262	0.096
MXN	1.004	1.019	0.411	0.398	0.053	0.051	0.006	0.429	0.169	0.775	0.203	0.285	0.413	0.202	0.478	0.290
NZD	0.966	1.025	0.029	0.011	0.049	0.055	0.036	0.378	0.217	0.123	0.278	0.173	0.234	0.108	0.369	0.494
NOK	0.976	1.009	0.047	0.033	0.037	0.029	0.003	0.680	0.266	0.166	0.545	0.343	0.090	0.220	0.146	0.074
SEK	0.956	1.014	0.005	0.009	0.003	0.010	0.001	0.099	0.008	0.034	0.023	0.082	0.262	0.283	0.293	0.252
CHF	1.004	1.009	0.673	0.691	0.746	0.949	0.914	0.830	0.938	0.775	0.676	0.409	0.022	0.276	0.180	0.209
GBP	0.985	1.023	0.258	0.136	0.538	0.511	0.178	0.454	0.722	0.524	0.358	0.597	0.149	0.232	0.262	0.247
BRL	0.972	1.090	0.004	0.012	0.004	0.006	0.123	0.509	0.408	0.331	0.212	0.373	0.047	0.021	0.092	0.041
INR	0.993	1.002	0.447	0.766	0.919	0.811	0.211	0.039	0.429	0.823	0.960	0.950	0.983	0.851	0.878	0.663
ZAR	0.997	1.023	0.099	0.039	0.202	0.078	0.113	0.493	0.615	0.885	0.862	0.581	0.196	0.098	0.171	0.128
EUR	1.000	1.015	0.349	0.420	0.252	0.116	0.020	0.751	0.557	0.994	0.353	0.567	0.186	0.131	0.265	0.176
$\sum$	9	0	4	6	4	3	8	1	1	1	1	0	3	2	0	1
Panel B: 3 Factors, $h = 6$																
AUD	0.996	1.013	0.254	0.209	0.611	0.328	0.102	0.901	0.027	0.195	0.087	0.035	0.084	0.102	0.073	0.631
CAD	1.017	1.020	0.179	0.729	0.422	0.736	0.076	0.565	0.101	0.195	0.163	0.324	0.097	0.098	0.146	0.105
DKK	1.002	1.007	0.299	0.503	0.732	0.513	0.399	0.487	0.057	0.677	0.225	0.616	0.030	0.178	0.142	0.090
JPY	1.015	1.012	0.713	0.482	0.967	0.801	0.444	0.668	0.040	0.015	0.425	0.215	0.059	0.032	0.243	0.111
MXN	1.003	1.011	0.581	0.317	0.413	0.411	0.712	0.994	0.784	0.563	0.105	0.286	0.103	0.152	0.034	0.089
NZD	1.004	1.019	0.435	0.342	0.730	0.468	0.980	0.972	0.293	0.517	0.771	0.844	0.127	0.066	0.129	0.219
NOK	0.996	1.007	0.276	0.129	0.270	0.114	0.120	0.596	0.015	0.311	0.151	0.160	0.095	0.247	0.122	0.271
SEK	1.003	1.025	0.485	0.261	0.055	0.115	0.071	0.758	0.062	0.530	0.190	0.690	0.262	0.140	0.092	0.035
CHF	1.009	1.000	0.094	0.056	0.114	0.107	0.448	0.382	0.211	0.124	0.448	0.355	0.299	0.972	0.637	0.905
GBP	1.008	1.022	0.321	0.132	0.533	0.304	0.551	0.680	0.373	0.364	0.382	0.192	0.195	0.078	0.257	0.105
BRL	1.006	1.035	0.359	0.092	0.540	0.282	0.000	0.049	0.850	0.669	0.276	0.562	0.112	0.116	0.075	0.433
INR	0.993	1.003	0.395	0.189	0.792	0.727	0.273	0.026	0.215	0.239	0.557	0.503	0.375	0.452	0.892	0.881
ZAR	1.002	1.020	0.342	0.127	0.196	0.060	0.487	0.980	0.732	0.647	0.966	0.946	0.031	0.011	0.029	0.008
EUR	1.003	1.006	0.359	0.611	0.595	0.572	0.598	0.597	0.059	0.621	0.166	0.616	0.034	0.221	0.158	0.124
$\sum$	3	0	0	0	0	0	1	2	3	1	0	1	3	2	2	2
Panel C: 3 Factors, $h = 12$																
AUD	0.999	1.004	0.274	0.172	0.388	0.428	0.124	0.992	0.044	0.705	0.338	0.145	0.655	0.380	0.445	0.894
CAD	1.035	1.005	0.333	0.188	0.218	0.113	0.142	0.425	0.181	0.168	0.313	0.131	0.532	0.416	0.728	0.699
DKK	1.009	0.997	0.215	0.209	0.507	0.189	0.362	0.333	0.052	0.141	0.058	0.237	0.117	0.538	0.336	0.458
JPY	1.006	1.003	0.464	0.170	0.155	0.150	0.916	0.333	0.387	0.206	0.054	0.051	0.598	0.552	0.336	0.157
MXN	1.022	1.004	0.527	0.322	0.478	0.367	0.214	0.998	0.140	0.179	0.077	0.031	0.159	0.215	0.035	0.009
NZD	1.004	1.001	0.407	0.441	0.588	0.841	0.987	0.978	0.079	0.246	0.948	0.745	0.951	0.923	0.556	0.736
NOK	0.998	0.999	0.046	0.840	0.269	0.549	0.099	0.525	0.009	0.716	0.052	0.696	0.140	0.814	0.562	0.962
SEK	1.002	1.002	0.971	0.953	0.258	0.120	0.075	0.236	0.024	0.614	0.019	0.687	0.148	0.684	0.423	0.399
CHF	1.020	0.998	0.215	0.243	0.258	0.135	0.555	0.367	0.417	0.216	0.456	0.206	0.403	0.660	0.436	0.416
GBP	1.002	0.996	0.127	0.119	0.092	0.147	0.710	0.784	0.435	0.812	0.642	0.387	0.878	0.671	0.886	0.994
BRL	0.993	0.997	0.168	0.625	0.687	0.978	0.006	0.154	0.651	0.422	0.271	0.269	0.563	0.775	0.489	0.371
INR	0.995	1.000	0.426	0.317	0.829	0.531	0.915	0.348	0.343	0.415	0.884	0.977	0.948	0.940	0.501	0.523
ZAR	0.998	1.007	0.065	0.012	0.467	0.155	0.920	1.000	0.241	0.582	0.773	0.809	0.062	0.018	0.080	0.192
EUR	1.009	0.996	0.314	0.188	0.385	0.134	0.621	0.433	0.045	0.134	0.061	0.188	0.106	0.467	0.302	0.397
$\sum$	5	6	1	1	0	0	1	0	4	0	1	1	0	1	1	1

Note: Columns 2 and 3 report the Root Mean square Error (RMSE) of TVL and CL forecasts divided by the RMSE of forecasts by a Random Walk (RW). A value < 1 implies the respective model has a smaller RMSE than the RW. Columns 4 to 9 compare the TVL and CL models and report the p-values of the Conditional Predictive Ability (CPA) and Unconditional Predictive Ability (UPA) test of Giacomini and White (2006) using a quadratic (Quad.) and an absolute (Abs.) loss function. Further, they report the p-values of the Pesaran and Timmermann (1992) nonparametric Direction Accuracy (DA) test for TVL and CL. Columns 10 to 13 compare TVL and RW, reporting p-values of the CPA and UPA test for quadratic and absolute loss differentials. Columns 14 to 17 compare CL and RW, reporting p-values of the same tests. The rows denoted by $\sum$ report the total number of p-values $\leq 5\%$ and in the case of the second and third columns, the number of RMSE smaller than those of a random walk. Results are shown for forecast horizons of $h = 1, 6, 12$. The currency abbreviations stand for Australian Dollar (AUD), Brazilian Real (BRL), Canadian Dollar (CAD), Danish Krone (DKK), Indian Rupee (INR), Mexican Peso (MXN), New Zealand Dollar (NZD), Norwegian Krone (NOK), South African Rand (ZAR), Swedish Krona (SEK), Swiss Franc (CHF), British Pound (GBP), and Euro (EUR).

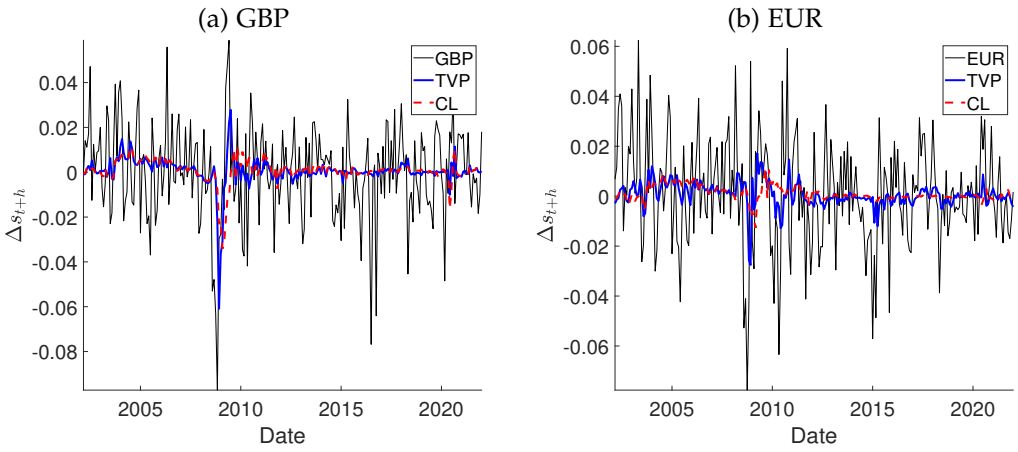


Figure 1.7: Rolling Window Forecast, $h = 1$

Note: This figure plots the out-of-sample, one-step-ahead, rolling window forecasting results of the time-varying loadings (TVL, blue line) and the constant loadings (CL, red line) model. The black line corresponds to the actual exchange rate.

accuracy 8 times for the TVL model and only once for the CL model. All three tests provide consistent evidence of improved predictive ability when time variation is accounted for. Next, we compare the TVL model to random walk forecasts using the CPA and UPA tests (p-values reported in Columns 10 to 13). We do not report the direction accuracy test, as the random walk forecast, by definition, has zero directional accuracy. With a quadratic loss function, both CPA and UPA reject once in favour of the TVL model and never in favour of the random walk. For an absolute loss function, the CPA test rejects once and again in favour of the TVL model. Finally, we compare the CL model with the random walk (p-values reported in Columns 14 to 17). Now, the CPA test rejects 3 times, and the UPA test twice, in favour of the random walk. Using an absolute loss function, the UPA test rejects once. The results imply that the TVL model does not only have better predictive ability than the CL model but also that it produces better forecasts than the CL model when comparing both to a random walk. We repeat this exercise for a forecast horizon of $h = 6$ in Panel B of Table 1.3 and for $h = 12$ in Panel C. While TVL and CL have equal predictive ability for $h = 6$ according to the CPA and UPA test, the CPA test based on a quadratic loss function now rejects the null hypothesis that the TVL model and a random walk have equal predictive ability three times: twice in favour of the TVL model, for AUD and NOK, and in case of the JPY in favour of the random walk. The UPA test also rejects for the JPY, however using an absolute loss function, it rejects for the AUD (i.e. in favour of the TVL model). On the other hand, the

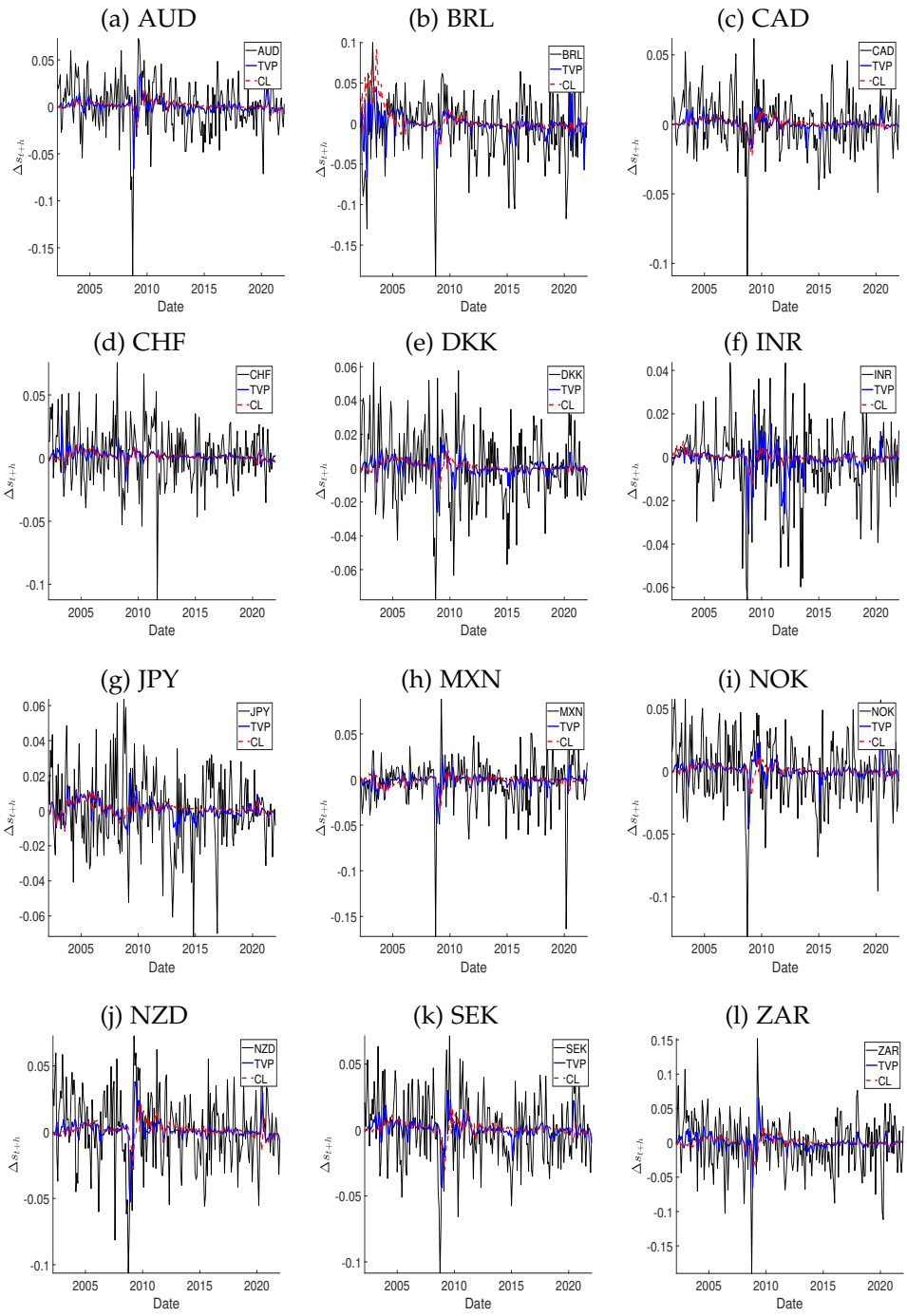


Figure 1.8: Rolling Window Forecast, $h = 1$

Note: This figure plots the out-of-sample, one-step-ahead, rolling window forecasting results of the time-varying loadings (TVL, blue line) and the constant loadings (CL, red line) model. The black line corresponds to the actual exchange rate.

tests reject in favour of the random walk throughout when comparing it to the CL model. All in all, the results demonstrate that time-varying loadings improve the fit of the model, especially for one-step-ahead forecasts. In a companion working paper (Hillebrand et al., 2020), we conducted the out-of-sample estimation using the in-sample dataset, i.e. without real-time data. The TVL model also consistently outperformed the CL model, albeit with fewer rejections overall. We conjecture that the Kalman filter is able to filter out noise in the real-time data, which led to an even better performance of the TVL relative to the CL model.

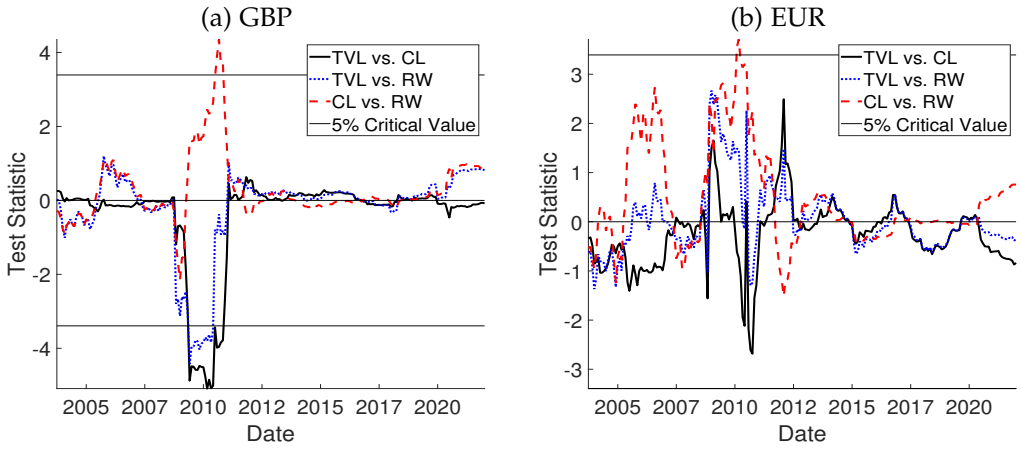


Figure 1.9: Fluctuation Test, $h = 1$

Note: This figure plots the results of the Giacomini and Rossi (2010) fluctuation test. The solid blue line is the test statistic, the dotted red lines are the 5% critical values. The results are based on a rolling window of $m = 20$ and a quadratic loss function.

In addition to selecting a model based on relative global predictive ability, it is interesting to examine how the relative predictive ability of two models changes over time. Figure 1.9 plots the results of the Giacomini and Rossi (2010) fluctuation test for GBP and EUR at each point in time for $h = 1$.[14] The solid black line represents the test statistic for the TVL compared to the CL model, the dotted blue line the statistic for the TVL model against a random walk, and the dashed red line corresponds to the test statistic for the CL model against the random walk. The horizontal lines are the 5% critical values. The sign of the test statistic corresponds to the sign of the MSE. When the test statistic falls below the negative critical value, the test rejects the null hypothesis of equal predictive ability in favour of the TVL model (or the CL model against the random walk). The results are obtained using a rolling window of $m = 20$ and a quadratic loss function.

[14] Plots for $h = 6, 12$ are reported in Appendix B.

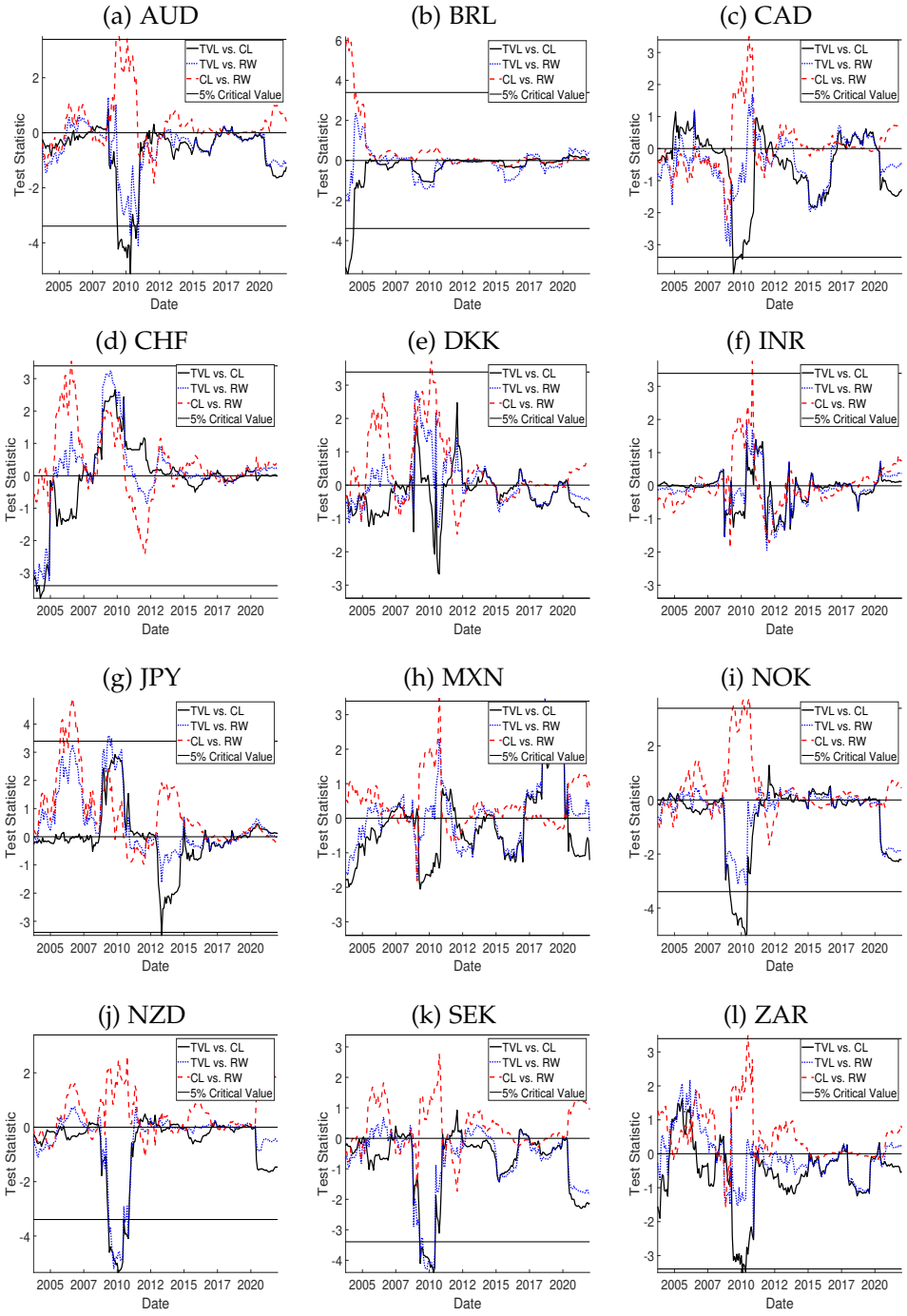


Figure 1.10: Fluctuation Test, $h = 1$

Note: This figure plots the results of the Giacomini and Rossi (2010) fluctuation test. The solid blue line is the test statistic, the dotted red lines are the 5% critical values. The results are based on a rolling window of $m = 20$ and a quadratic loss function.

Although the Giacomini and White (2006) test does not provide evidence that the time-varying loadings model is globally more accurate in case of the GBP (see Table 1.3), the fluctuation test rejects in favour of the TVL model both against the CL model and the random walk during the financial crisis. Figure 1.10 shows the results for the remaining currencies. The fluctuation test rejects the null hypothesis that TVL and CL have equal predictive ability for 10 currencies (AUD, BRL, CAD, CHF, GBP, JPY, NOK, NZD, SEK, and ZAR). The Giacomini and White (2006) test only rejects the null hypothesis of equal predictive ability for 4 currencies (BRL, NOK, NZD, and SEK, Table 1.3). That is, for an additional 6 currencies, we find evidence that accounting for instabilities in factor loadings leads to improved local predictive ability. In most cases, these pockets of predictability occur during the financial crisis, which indicates that the TVL model performed particularly well in that period. The fluctuation test statistic never attains the positive critical value, meaning the null hypothesis is never rejected in favour of the constant loadings model. The test further indicates that the random walk has superior predictive ability when compared against the CL model in all except 2 cases (NZD and SEK). In contrast, the fluctuation test only rejects twice in favour of the random walk against the TVL model (JPY and MXN) and 5 times in favour of the TVL model (AUD, CHF, GBP, NZD, and SEK). On closer examination, the CL model performs particularly poorly against the random walk during periods in which the TVL model does well, especially for AUD, BRL, CAD, and GBP. For these 4 currencies, the random walk significantly outperforms the CL model during periods in which the TVL model beats the random walk.

To underscore the robustness of our out-of-sample results, we use the 1-factor model to generate forecasts over the same horizon as above and report the statistical evaluation (see Appendix A.4.2). The results are equally – if not more – affirmative. Neither the Giacomini and White (2006) nor the Pesaran and Shin (1998) test find any predictive ability on the part of the constant loadings model. In addition, we include the forecast evaluation for the 3-factor model over different prediction horizons ($P = 260$ and $P = 200$). In all cases, the evidence stands clearly in favour of the time-varying loadings model. The computational complexity of the algorithm does not allow for the use of forecast evaluation criteria as in Rossi and Inoue (2012) that

are robust to estimation window size.[15] However, the robustness checks in conjunction with the three different forecast evaluation criteria should eliminate concerns that the out-of-sample results are driven by window size.

1.7 CONCLUSIONS

In this paper, we studied the unstable relationship between exchange rates and macroeconomic factors. Using a novel econometric approach, proposed in Mikkelsen et al. (2019), we showed that allowing for time-varying factor loadings increases the percentage of explained variation in exchanges rates by an order of magnitude. In addition, taking the aforementioned instabilities into consideration improves the relative out-of-sample predictive ability of the model globally, and yields better forecast of sign changes in exchange rates.

We extracted macroeconomic fundamentals as principal components from a new dataset that combines all 272 vintage releases of the FRED-MD database and a large number of variables sourced from the OECD. The unobserved time-varying loadings are estimated using a recently proposed two-step maximum likelihood estimator for high-dimensional factor models (Mikkelsen et al., 2019). The model is applied to 14 currencies *vis-à-vis* the US Dollar and the results show that failure to account for the instabilities between exchange rates and fundamentals is by no means innocuous. In-sample, the time-varying loadings model achieves a median R^2 of 52% percentage points compared to 2% for the constant loadings benchmark. We showed that out-of-sample, the time-varying loadings model exhibits significantly better forecast accuracy. This holds true both when comparing their predictive ability directly, and in terms of the improvements the time-varying model generates relative to a random walk. When evaluating the forecasts individually, the time-varying loadings model outperforms the constant loadings model at predicting directional exchange rate changes. To consider potentially unstable forecasting performance, we evaluated the relative predictive accuracy of the forecasts using the Giacomini and Rossi (2010) fluctuation test. In addition to higher global forecast accuracy, time-varying

[15] To eliminate the effects of in- and out-of-sample window choices, Rossi and Inoue (2012) propose a forecast test which is robust to window size. They show that for a test-statistic $S_T(R)$,

$$\frac{1}{[\bar{\mu}T] - [\underline{\mu}T + 1]} \sum_{R=[\underline{\mu}T]}^{[\bar{\mu}T]} S_T(R) \xrightarrow{d} \int_{\underline{\mu}}^{\bar{\mu}} S(\mu) d\mu$$

However, this procedure is computationally infeasible in our setting.

loadings improved forecasts locally around the financial crisis. This paper provides strong evidence that the relationship between macroeconomic fundamentals and exchange rates is highly unstable.

CHAPTER 2

COMBINING *P*-VALUES FOR MULTIVARIATE PREDICTIVE ABILITY TESTING

2.1 INTRODUCTION

In this paper, we propose a computationally efficient test for multivariate predictive ability that is valid for any number of univariate forecast accuracy tests and arbitrary dependence structures, without specifying the underlying multivariate distribution.

One of the main goals of econometric analysis is to make accurate predictions of a large range of variables such as inflation, exchange rates, stock returns, or volatility to name only a few. To this purpose, there exist a large number of candidate models and the main challenge is to select the one with the best predictive ability. Thus far, the literature has proposed a variety of testing procedures, most of which are univariate and evaluate two competing forecasts of a single variable. Important examples include Diebold and Mariano (1995, DM) and Giacomini and White (2006, GW). [See Clark and McCracken (2013) for a comprehensive overview of existing testing procedures.] However, researchers are often interested in evaluating more than two forecasts, either because multivariate models are used (e.g. Laurent et al., 2013) or multiple variables are forecast (e.g. Carriero et al., 2019). In such cases, independence between variables and forecasting models seldom holds true. Consequently, univariate test statistics and their p-values can also exhibit dependencies, which means they cannot necessarily be evaluated individually. This motivates the development of multivariate forecast tests that account for dependence. Notably, the joint distribution of dependent

variables is difficult or impossible to obtain analytically without several assumptions. The existing literature approaches this issue by evaluating a set of forecasts directly with a single test that adequately captures dependencies. For instance, Qu et al. (2021) condition on a common factor within forecast errors to capture any common components. Mariano and Preve (2012) extend the DM test into a multivariate setting without directly addressing the dependence structure. However, approaches that evaluate forecasts jointly suffer from drawbacks that may render them infeasible in certain situations. First, their limiting framework is valid only under *either* a large or a small cross-section, restricting their applicability. That is, one cannot use the same test with datasets of considerably different dimensions. Second, the design of the forecasting scenario can require the use of multiple different tests, e.g. because one compares nested and non-nested models or uses different estimation windows that affect the asymptotic properties of tests. Third, it is not always obvious when a test rejects in a multivariate setting in the sense that it is undefined how many individual forecasts must be equally accurate for the null to be sustained.

The testing framework we propose combines univariate tests, taking advantage of both recent advances in the statistical literature on combining dependent p-values as well as in the econometric literature on multivariate forecast evaluation. The resulting test allows researchers to estimate any number of univariate forecast accuracy tests – provided they fulfill some nonrestrictive assumptions – and corrects for dependence in a subsequent step that combines their p-values. Thus, one can implement tests that are most appropriate in a given scenario and examine whether predictive ability holds in the cross-section. We specify a global null hypothesis that is clearly defined as the intersection of all individual null hypotheses, accounting for false discovery and dependence. Furthermore, our method can be applied to p-values from different tests, meaning that when faced with mixed or inconclusive evidence, one can obtain a more conclusive result. Hence, our test can be used in a plethora of, if not all, forecasting scenarios. To the best of our knowledge, we are the first to propose such a test.

Specifically, we propose an intersection-union (IU) test by applying the theoretical results in Vovk and Wang (2020). We show that one can construct a global hypothesis test, based on a single or several of the existing univariate tests for forecast accuracy, that is level-α under any form of dependence. Crucially, our global test does not require knowledge of the joint distribution of the p-values of the individual test statistics which cannot be derived analytically. In addition, we study the power properties of the test and show under what conditions Type I and II errors vanish. We demonstrate the

good size and power properties of the test through a battery of Monte-Carlo simulations. For this purpose, we use three benchmark tests: DM, GW, and Clark and West (2007, CW). The three tests are widely used and suitable in different scenarios. However, we emphasize that our method is not restricted to these tests, meaning it can be deployed in a range of forecasting situations. The simulations evaluate the performance of our test in low-dimensional as well as high-dimensional scenarios. We compare our method to the Mariano and Preve (2012) test, a multivariate GW test, and methods that do not account for dependence between p-values. The simulations show that these procedures display considerable size distortions, effectively rendering them inapplicable in practical scenarios. When investigating their rejection accuracy, we find that contrary to our test, the other multivariate procedures, which are both of Wald-type, exhibit distinctly different rejection rates when dimensions change. This highlights how the interpretation of test results can benefit from our intersection null in practice. We illustrate the empirical validity of our testing procedure via an application involving a large dataset of 84 daily exchange rates, running from 1 January 2011 to 1 April 2021, quoted against the US-Dollar, the British Pound, and the Euro. The empirical illustration highlights the wide-ranging applicability of our test both in small and high dimensional cases. Moreover, it exemplifies how our test addresses inconclusive results that arise often in practice.
The remainder of the paper is structured as follows: Section 2.2 describes the forecasting setup, introduces the IU testing framework, and also reports a way to apply the GW test in a multivariate setting. Section 2.3 analyzes the size and power properties of the test in various simulations and Section 2.4 provides an empirical illustration of the test. Section 2.5 concludes.

2.2 THEORY

This section presents our theoretical contribution. First, we outline a general framework for univariate forecast evaluation that is consistent with our test. Second, we propose treating univariate tests as sub-tests of a global null hypothesis and discuss the assumptions imposed upon them. Third, we introduce our methodology for multivariate forecast comparison which is based on the intersection of the sub-tests. Finally, we show how the Wald-type GW test can be applied in a multivariate setting and serve as a benchmark for our IU test.

2.2.1 *Forecasting Setup*

Suppose we observe the vector $\mathbf{V}_t \equiv (Y_t', X_t')'$, where $Y = \{Y_t : \mathcal{Y} \mapsto \mathbb{R}^n, n \in \mathbb{N}\}$ are the variables one wishes to forecast and $X = \{X_t : \mathcal{X} \mapsto \mathbb{R}^s, s \in \mathbb{N}\}$ are predictor variables. Define $\mathscr{F}_t$ as the σ-field generated by the infinite history of $\mathbf{V} \equiv \{\mathbf{V}_t : \Omega \mapsto \mathbb{R}^{s+n}\}$ and $\mathcal{Y} \cup \mathcal{X} = \Omega$ such that $\mathbf{V}$ is defined on the complete probability space $(\Omega, \mathscr{F}, P)$. The forecasting equation for Y_t takes the form of an $\mathscr{F}_t$-measurable function $\psi : \Omega \mapsto \mathcal{F} \subset \mathbb{R}$. The function can include lagged values of Y_t as well as X_t and can be parametric or non-parametric. It produces τ-step-ahead forecasts $\hat{Y} = \{\hat{Y}_{t+\tau} : \mathcal{F} \mapsto \mathbb{R}^n\}$ of Y based on the information set $\mathscr{F}_t$. Notation-wise, we define the estimation window of the parameters as R and place no restrictions on whether R is a fixed, rolling, or expanding estimation window. Further, we define the out-of-sample forecasting window as p such that $T = R + p + \tau$ is the total sample the forecaster observes. Multiple procedures to evaluate the resulting forecasts have been introduced, most of which rely on a forecast loss function. The forecast loss function is defined as $L(\hat{Y}, Y, X) : \mathcal{F} \times \Omega \mapsto \Lambda$. In many cases, the loss function is defined such that $\Lambda \subset \mathbb{R}_+$, with the most common type being the quadratic loss:

$$L_{i,R,t+\tau} = \left(Y_{i,t+\tau} - \hat{Y}_{i,t+\tau}\right)^2, \quad i \in \{1, ..., n\}.$$

The function can take many other forms and gives a vector $\{L_{R,t+\tau}\}_{t=R}^{T}$. One can assess forecasts based on a single loss function by testing whether it is statistically different from zero. For two forecasts, $\hat{Y}^{(1)}$ and $\hat{Y}^{(2)}$, one can define a loss differential, $\Delta L \equiv L(\hat{Y}^{(1)}, Y, X) - L(\hat{Y}^{(2)}, Y, X)$. That is, in the quadratic case we have

$$\Delta L_{i,R,t+\tau} = \left(Y_{i,t+\tau} - \hat{Y}^{(1)}_{i,t+\tau}\right)^2 - \left(Y_{i,t+\tau} - \hat{Y}^{(2)}_{i,t+\tau}\right)^2, \quad i \in \{1, ..., n\}.$$

The vector $\Delta \mathbf{L}_{R,t+1}$ then stacks the n univariate loss differentials. Note that one can use several different loss functions to evaluate the same forecasts. However, the majority of existing forecast accuracy tests evaluate forecasts based on one single loss differential. Most commonly, by either formulating the unconditional null hypothesis $\mathcal{H}_{i,0} : \mathbb{E}[\Delta L_{i,R,t+\tau}] = \mu_{i,0}$ or the conditional null $\mathcal{H}_{i,0} : \mathbb{E}[\Delta L_{i,R,t+\tau} \mid \mathscr{F}_t] = \mu_{i,0}$. The parameter $\mu_{i,0}$ is known and set to be zero when testing for equal predictive ability. The second type of tests is then called conditional equal predictive ability test and can ascertain if a particular model has superior forecasting abilities. This is a notable difference to unconditional tests that only assess if there are statistically

significant differences between two forecasts. We use the notation of GW, $\mathbb{E}[\Delta L_{i,R,t+\tau} \mid \mathscr{G}_t] = \mu_i$, where $\mathscr{G}_t$ corresponds to either the natural filtration $\mathscr{F}_t$ or the trivial σ-field $\{\emptyset, \Omega\}$, thereby referring to either of the two test types. The parameter μ_i characterizes either the unconditional or the conditional mean of the loss differential of the two forecasts. In this paper we consider the global null hypothesis:

$$\mathcal{H}_0 : \mathbb{E}[\Delta \mathbf{L}_{R,t+\tau} \mid \mathscr{G}_t] = \boldsymbol{\mu}_0.$$

Constructing a test for this hypothesis is less straightforward. One approach is to construct a single test, based on the entire sample space Λ that jointly evaluates all elements in $\Delta \mathbf{L}_{R,t+\tau}$. Dependencies between these elements pose an analytical and computational obstacle in the construction of a valid test. Therefore, we propose to test the intersection of the local hypotheses $\mathcal{H}_{i,0}$. Our methodology is compatible with all univariate test types that fulfill the assumptions outlined in the next section.

2.2.2 *Univariate Sub-Tests*

Consider a series of n forecast accuracy tests, each of which is based on the random variables $Y_{i,t}$ and X_t. The tests all examine the local hypothesis $\mathcal{H}_{i,0}$, i.e. compare the accuracy of two forecasts of the variable $Y_{i,t}$. We treat these tests as sub-test for the global null hypothesis that all forecasts exhibit equal accuracy. Each univariate sub-test evaluates the sub-hypothesis:

$$\mathcal{H}_{i,0} : \mu_i \in \mathsf{M}_{i,0}, \qquad \text{for } i \in \{1, ..., n\},$$

against the alternative

$$\mathcal{H}_{i,A} : \mu_i \in \mathsf{M}_{i,A}, \qquad \text{for } i \in \{1, ..., n\},$$

where $\mathsf{M}_{i,0}$ is the set of admissible values for μ_i under the sub-hypothesis, $\mathsf{M}_{i,A}$ is the set of admissible values for μ_i under the alternative hypothesis. Suppose the tests construct a test statistic $S_i = s_i(\mathbf{V})$ with realization $\hat{s}_i$ which, under $\mathcal{H}_{i,0}$, has a density $f_i(x)$. Then, the p-value of the test statistic corresponds to:

$$p_i \equiv \mathbb{P}_{\mu_i}[\{\mathbf{V} \in \Omega : s_i(\mathbf{V}) > \hat{s}_i\}] = 1 - F_i(\hat{s}_i), \quad \text{for all } \mu_i \in \mathsf{M}_{0,i},$$

for the cumulative distribution of $f_i(\cdot)$, $F_i(\cdot)$. One rejects the sub-hypothesis if $\mathbb{P}_{\mu_i}[\hat{s}_i \geq c_i] \leq \alpha$, i.e. $p_i \leq \alpha$, where α is the significance level chosen by the

researcher and $c_i = F_i^{-1}(1-\alpha)$. To be consistent with our methodology, the sub-tests must satisfy the following assumptions:

Assumption 2.1. (i) Under each sub-hypothesis, the choice for $\mathsf{M}_{i,0}$ is $\mathsf{M}_{i,0} = \{\mu_{i,0}\}$ for some known parameter value $\mu_{i,0} \in \mathbb{R}$ and for all $i \in \{1, ..., n\}$. (ii) Under each alternative, the choice for $\mathsf{M}_{i,A}$ is $\mathsf{M}_{i,A} = \{\mu_i; \mu_i \notin \mathsf{M}^*_{i,A} \cup \mathsf{M}_{i,0}\}$ where $\mathsf{M}^*_{i,A}$ represents the set of all parameters that are local alternatives to the sub-hypothesis $\mathcal{H}_{i,0}$.

Assumption 2.2. $f_i : \mathbb{R} \to [0, \infty]$, and $\mathbb{P}_{\mu_i}[a \leq X \leq b] = \int_a^b f_i(x)dx$ for all $i \in \{1, .., n\}$.

Assumption 2.3. (i) Under each sub-hypothesis $\mathcal{H}_{i,0}$, $\sup \mathbb{P}_{\mu_i}[\hat{s}_i \geq c_i] \leq \alpha$. (ii) Under each alternative $\mathcal{H}_{i,A}$, $\mathbb{P}_{\mu_i}[\hat{s}_i \geq c_i] \to 1$ for all $i \in \{1, ..., n\}$.

Assumption 2.1 (i) imposes that the parameter values for the null hypothesis of each sub-test consist of a single value, i.e. they are not composite, while (ii) ensures the sub-hypotheses are not tested against local alternatives that are too close to the null to be detected [see van der Vaart (1998, Ch. 7)]. Assumption 2.2 imposes that the density of the test statistic is absolutely continuous, as is the case for most econometric tests. It ensures that for any $\mu_i \in \mathsf{M}_{i,0}$, $\mathbb{P}_{\mu_i}[\{\mathbf{V} \in \Omega : s_i(\mathbf{V}) > \hat{s}_i\}]$ is known for all $i \in \{1, ..., n\}$. It is easy to see that both assumptions together imply that the p-values be uniform over $[0, 1]$: $p_i \sim \mathsf{Un}[0, 1]$ under $\mathcal{H}_{i,0}$. In this context, Assumption 2.1 is crucial as Robins et al. (2000, p.1144) show that p-values are not necessarily uniform if the null hypothesis of a test is composite. This is important as it implies our method is not applicable for tests with a null of *superior* predictive ability. Assumption 2.3 (i) ensures the univariate sub-tests are of level-α, while (ii) stipulates that their asymptotic power approaches one.
If Assumptions 2.1 and 2.2 hold true and we observe n independent p-values $p_1, ..., p_n \in [0, 1]$, then the variable $\mathbf{P} = (p_1, ..., p_n) \in [0, 1]^n$ is uniform on the hypercube $[0, 1]^n$. Indeed, under independence, it is easy to derive the distribution of various possible combinations of the n p-values. One of the most well known of such methods dates back to Fisher (1934) who shows that $S_F = -2\sum_i \log(p_i) \sim \chi^2_{2n}$ under $\mathcal{H}_0$ [see Heard and Rubin-Delanchy (2018) for a detailed review of different methods]. Under dependence, however, the distribution does not admit an analytical solution (Liu and Xie, 2019; Kost and McDermott, 2002). Notably, independence rarely holds in practice, particularly when forecasting multiple variables or comparing related models. To prevent size distortions, multivariate forecast evaluation methods must take dependence structures into account. As the latter are unknown in most scenarios, an essential requirement for a multivariate forecast accuracy test

is that it exhibits good size and power properties under arbitrary forms of dependence.
In the next section, we develop a testing framework to address these important features encountered in most economic and financial applications.

2.2.3 *An Intersection-Union Test of Multivariate Forecast Accuracy*

Consider a scenario where one has conducted a total of n sub-tests. Each test $i \in \{1, ..., n\}$ constructs a statistic $\hat{s}_i$, yielding a p-value p_i, both stacked in the vectors $\mathbf{S} = (\hat{s}_1, ..., \hat{s}_n)$ and $\mathbf{P} = (p_1, ..., p_n) \in \mathcal{P}^n$, where $\mathcal{P}$ is the set of all p-values. We do not assume that $\hat{s}_i$ and $\hat{s}_j$ are independent for $i \neq j \in \{1, ..., n\}$. The previous section discussed the properties of the sub-tests in detail. Now suppose we are interested in the global null hypothesis $\mathcal{H}_0 : \mathbb{E}[\Delta \mathbf{L}_{R,t+1} \mid \mathscr{G}_t] = \boldsymbol{\mu}_0$. Rather than developing a statistic that tests $\mathcal{H}_0$ directly, we formulate the global null as the intersection of the sub-hypotheses $\mathcal{H}_{i,0}$. That is, if we define the set $R = \{i \in \{1, ..., n\} : \mu_i \in \mathsf{M}_{i,A}\}$ with cardinality R_0, we wish to test if $R = \emptyset$. Formally, the intersection null hypothesis can be defined as

$$\mathcal{H}_0 = \bigcap_{i \in \mathsf{N}} \mathcal{H}_{i,0} : \boldsymbol{\mu} \in \bigcap_{i \in \mathsf{N}} \mathsf{M}_{i,0}, \tag{2.1}$$

with the index set $\mathsf{N} = \{1, ..., n\}$. It is tested against the alternative

$$\mathcal{H}_A = \bigcup_{i \in \mathsf{N}} \mathcal{H}_{i,A} : \boldsymbol{\mu} \in \bigcup_{i \in \mathsf{N}} \mathsf{M}_{i,A}.$$

We can write $\mathsf{M}_0 = \cap_{i \in \mathsf{N}} \mathsf{M}_{i,0}$ and $\mathsf{M}_A = \cup_{i \in \mathsf{N}} \mathsf{M}_{i,A}$. To test for equal predictive ability, we can set $\mu_{i,0} = 0$ for all $i \in \mathsf{N}$. In that case, the global null hypothesis $\mathcal{H}_0$ is that equal predictive ability holds for each of the n pairs of forecasts and it is rejected if any of the n sub-hypotheses is false. One can select whichever sub-test is most appropriate to examine each sub-hypothesis $\mathcal{H}_{i,0}$. This is a decisive advantage if one analyzes characteristically different datasets or models. The clearly defined rejection set of our IU test stands in contrast to other Wald-type tests of multivariate predictive ability whose rejection set is undefined. We demonstrate this in our Monte-Carlo simulations. When the test statistics are not independent, the global Type I error of the tests depends on the *joint* distribution of $\mathbf{S}$ which is unknown. The obvious implication is that one cannot simply consider the p-values individually to test $\mathcal{H}_0$. If one consults a statistic that assumes p-values are independent, one is, in fact, not testing $\mathcal{H}_0$ but rather a composite of $\mathcal{H}_0$ and

$\mathcal{A}_0 := \{p_i \perp\!\!\!\perp p_j \text{ for all } i \neq j \in \mathsf{N}\}$, where $\perp\!\!\!\perp$ denotes independence between variables. This can lead to considerable size distortions, as a rejection may simply be due to a false independence assumption – a point illustrated in our Monte-Carlo simulations. The main complication to testing $\mathcal{H}_0$ is finding a test statistic $s(\mathbf{P})$ for which it can be shown that

$$\mathbb{P}_{\kappa(\cdot)}\left[s(\mathbf{P}) \in C_{\kappa(\cdot)} \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1]\right] \leq \alpha.$$

Here, $\kappa(\mathbf{p})$ denotes an unknown reference density for the joint distribution of the p-values under the intersection null hypothesis and $C_{\kappa(\cdot)}$ is a critical region for the significance level α. Note that we do not condition on $\mathcal{A}_0$. In what follows, we propose a simple, widely applicable, and computationally convenient methodology to circumvent the problem of defining $\kappa(\cdot)$.
Based on recent results on the precision of merging functions for p-values of Vovk and Wang (2020), we define the following test statistic:

$$P_{r,n} = n^{-1}\left(\sum_{i=1}^{n} p_i^{-r}\right)^{1/r}, \qquad \text{for any } r \in (1,\infty). \tag{2.2}$$

Unlike methods of minimum p-values like the Bonferroni correction, the statistic above incorporates p-values of all sub-tests. The negative exponent ensures small p-values increase the statistic by more relative to large values. It can be seen that the statistic is permutation invariant, i.e. the order in which the individual tests are conducted does not change the outcome of the IU test. We apply Proposition 5 in Vovk and Wang (2020) to study the properties of the test both in a finite sample environment and asymptotically. The results are formulated in the following theorem:

Theorem 2.1. *Suppose Assumptions 2.1 (i), 2.2, and 2.3 (i) hold. Let $\{s_1, ..., s_n\}$ be test statistics from level-α tests of $\{\mathcal{H}_{1,0}, ..., \mathcal{H}_{n,0}\}$ with unknown dependence structure and p-values $\{p_1, ..., p_n\}$. Under the intersection null hypothesis $\mathcal{H}_0$, for the test statistic $P_{r,n}$ in (2.2) we obtain the finite sample result:*

$$\mathbb{P}\left[P_{r,n} \in C_{r,n} \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1]\right] \leq \alpha, \tag{2.3}$$

and the asymptotic result:

$$\limsup_{n\to\infty} \mathbb{P}\left[P_{r,n} \in C_{r,n} \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1]\right] = \alpha, \qquad \forall\ (p_1, ..., p_n) \in \mathcal{P}^n, \tag{2.4}$$

for all $\alpha \in (0,1)$*, any* $r \in (1,\infty)$*, and the critical region* $C_{r,n} = \left\{c_{r,n} \geq \frac{r}{\alpha(r-1)}\right\}$.

Theorem 2.1 shows that the statistic is level-α for finite n and size-α for $n \to \infty$. Note that the asymptotic result holds irrespective of whether the sub-tests are size-α or level-α. The size properties follow as it can be shown the test statistic (2.2) falls into the category of increasing Borel functions $F : [0,1]^n \to [0,\infty)$ for which Vovk and Wang (2020) show $\sup[\mathbb{P}\{F(\mathbf{U}) \leq \epsilon\} \mid \mathbf{U} \in [0,1]^n] \leq \epsilon$ for any $\epsilon \in (0,1)$, regardless of the joint distribution of $\mathbf{U}$. To construct and analyze the statistic, we do not need to impose any assumptions about the degree of dependence on the individual p-values. Nor does the computation require knowledge of the joint distribution of $\mathbf{P}$, as long as the uniformity of the p-values under their individual null hypotheses is satisfied. Indeed, this is a decisive advantage of our approach as it allows researchers to compare the accuracy of dependent forecasts without making restrictive assumptions about the joint distribution of their tests statistics and p-values, respectively. Importantly, if one decides to compare multiple forecasts through individual tests *without* our procedure (or a comparable method) one is implicitly assuming independence. Thereby, one is also testing the independence assumption which can increase the Type I error. Furthermore, our methodology controls the False Discovery Rate (FDR) which is defined as: $\text{FDR} = \mathbb{E}[FP/(FP+TP)\mathbf{1}_{\{FP+TP\geq 1\}}]$, where FP are false positives and TP are true positives. It can be seen that Theorem 2.1 keeps the FDR lower or equal to α. Notice also that we do not impose any restrictions on n relative to T. One implication of Theorem 2.1 is that, in finite samples, under $\mathcal{H}_0$ the statistic has size smaller or equal to the minimum size of any of the individual tests under the global null hypothesis:

$$\pi(\boldsymbol{\mu}) \leq \bigwedge_{i \in \mathsf{N}} \pi_i(\mu), \quad \boldsymbol{\mu} \in \mathsf{M}_0, \tag{2.5}$$

where $\pi_i(\mu) = \mathbb{P}_i[S_i \in C_i \mid \mu_i]$ is the power function for each sub-test, $\wedge_{i\in\mathsf{N}}$ denotes the minimum over all $\pi_i(\mu)$ for all $i \in \mathsf{N}$, and $\pi(\boldsymbol{\mu}) = \mathbb{P}[P_{r,n} \in C_{r,n} \mid \boldsymbol{\mu}]$ is the power function for the global test. In the next theorem, we turn to the behavior of the test statistic under the alternative hypothesis $\mathcal{H}_A$.

Theorem 2.2. *Suppose Assumptions 2.1-2.3 hold. Let* $\{s_1, ..., s_n\}$ *be a sequence of test statistics from level-*α *tests of* $\{\mathcal{H}_{1,0}, ..., \mathcal{H}_{n,0}\}$ *with unknown dependence structure and p-values* $\{p_1, ..., p_n\}$*. Then under the alternative* $\mathcal{H}_A$*:*

$$\mathbb{P}\left[P_{r,n} \in C_{r,n} \mid \boldsymbol{\mu} \in \mathsf{M}_A\right] \to 1,$$

for all $\alpha \in (0,1)$*, any* $r \in (1,\infty)$*, and the critical region* $C_{r,n} = \left\{c_{r,n} \geq \frac{r}{\alpha(r-1)}\right\}$.

Theorem 2.2 presents a general finite sample case that shows the test rejects with probability approaching 1 if the intersection null hypothesis is false. It is not necessary, and often impossible, to impose a particular distribution on the p-values under the alternative and analyze the finite sample power of the test. However, we can derive a specific asymptotic result if the test statistics are jointly normally distributed and the global null hypothesis is tested against sparse alternatives. This necessitates the following assumption:

Assumption 2.4. (i) $\mathbf{S} = (s_1, ..., s_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has off diagonal elements $\sigma_{i,j} = 0$ for any $|i - j| > 1$, (ii) $R_0 = n^\gamma$ for $\gamma \in [0, 0.5]$, (iii) $\mathsf{M}_{i,0} = \{i \in \mathsf{N} : \mu_{i,0} = 0\}$ and $\mathsf{M}_{i,A} = \{i \in \mathsf{N} : \mu_i = \sqrt{2\delta \log n}\}$ for all $\delta > -2\sqrt{\gamma(2r-1)/r} + \gamma - 1/r + 2$.

Assumption 2.4 (i) imposes the vector of test statistics be multivariate normal with banded correlation matrix, (ii) ensures only a relatively small number of tests rejects, by restricting the number of rejected sub-tests to be a function n^γ of the total number of sub-tests, while (iii) replaces the conditions imposed on local alternatives in Assumption 2.1. Since the parameter δ, which controls the magnitude of μ_i under the alternatives, depends negatively on γ, the magnitude of μ_i for which the test rejects decreases in the relative number of rejected sub-tests. The choice of $\mu_{i,0} = 0$ is standard in most forecast accuracy tests. This setup follows Liu and Xie (2020) and Donoho and Jin (2004) and embeds our test in the existing literature on combining p-values. In addition, we can define a specific range for δ that maximizes the power of our test. Under Assumption 2.4, Liu and Xie (2020, Theorem 3) show that the power of the statistic $\sum_{i=1}^n \omega_i \tan\{(0.5 - p_i)\pi\}$ converges to 1 for non-negative ω_i, with $\sum_i \omega_i = 1$. The following proposition extends their result to our IU test:

Proposition 2.1. *Suppose Assumptions 2.3 (i), 2.4 (i)-(iii) hold and we observe* $\mathbf{S} = (s_1, ..., s_n)$ *as well as* $\mathbf{P} = (p_1, ..., p_n)$*. Then under the alternative* $\mathcal{H}_A$*:*

$$\lim_{n\to\infty} \mathbb{P}\left[P_{r,n} \in C_{r,n} \mid \boldsymbol{\mu} \in \mathsf{M}_A\right] = 1, \tag{2.6}$$

for all $\alpha \in (0, 1)$*, any* $r \in (1, \infty)$*, and the critical region* $C_{r,n} = \left\{c_{r,n} \geq \frac{r}{\alpha(r-1)}\right\}$.

Proposition 2.1 represents a special case of our test for normally distributed sub-tests. If Assumption 2.4 holds and the magnitude of μ_i under the alternative is known, we can derive a more explicit lower bound for r that ensures the power of the test equals 1 asymptotically:

Corollary 2.1. *Suppose Assumptions 2.3 (i) and 2.4 (i)-(iii) hold. If* $\delta \in (\gamma - 2\sqrt{\gamma} + 1, \gamma - 2\sqrt{2}\sqrt{\gamma} + 2)$*, then the sum of Type I and II errors vanishes asymptotically for any* $r \in (1, 2 - 2\sqrt{\gamma\delta} - \gamma - \delta)^{-1})$.

Corollary 2.1 places an upper bound on r which depends positively on γ and δ. The result implies that the larger γ, the larger the range of admissible values for r. In the general case, we cannot specify a particular joint distribution. Therefore,we cannot simply obtain a p-value for the statistic as $1 - F(P_{r,n})$, where $F(\cdot)$ is any CDF. It is, however, possible to compute a p-value according to the proposition below:

Proposition 2.2. *Suppose Assumptions 2.1-2.3 hold and we observe* $\mathbf{S} = (s_1, ..., s_n)$ *as well as* $\mathbf{P} = (p_1, ..., p_n)$*. Then a p-value for the test statistic in (2.2) can be computed as*

$$h(P_{r,n}) = \frac{r}{r-1} \frac{1}{P_{r,n}} \wedge 1. \tag{2.7}$$

Proposition 2.2 defines a variable $h(P_{r,n}) \in [0, 1]$ that can be interpreted as a p-value to the test statistic. Notably, we do not necessarily obtain $h(P_{r,n}) \sim \mathsf{Un}[0, 1]$, nor are we able to analyze the distribution of $h(P_{r,n})$.

Theorems 2.1 and 2.2 hold regardless of whether we merge n different or identical tests, as long as they satisfy Assumptions 2.1-2.3. This is an important feature, as data availability and alternative model specifications, respectively, may render some univariate tests impractical or change the degrees of freedom of their test statistics. To the best of our knowledge, this is the first paper to propose an IU framework to test for forecast accuracy. Regarding the performance of our test, we are interested in (i) how the IU test compares to other methods of combining p-values, and (ii) how it compares to a test that jointly evaluates all individual forecasts in a single step. However, there are not many suitable benchmarks to assess the second point. A requirement in this regard is that the multivariate benchmark displays similar properties as the univariate sub-tests. One example for such a test is the multivariate DM extension of Mariano and Preve (2012, MP henceforth) which is comparable to the IU test based on DM sub-tests. Other multivariate tests, such as Qu et al. (2021), are not extensions of a univariate test. Thus, they are less suitable as a benchmark: it is possible that the IU test, based on, say, univariate GW p-values, has high (low) power relative to a test in Qu et al. (2021), while having relatively low (high) power when combining p-values from, say, CW tests. In the next subsection, we suggest an additional competitor for our test in the form of a multivariate GW test.

2.2.4 *A Wald Test for Multivariate Forecast Accuracy*

The framework of GW presents a natural point of comparison for our IU test. In what follows, we discuss how the GW test can be applied in multivariate settings and report details of the derivation in Appendix B.1. Like the IU test introduced above, the multivariate GW test evaluates the global null hypothesis $\mathcal{H}_0 : \mathbb{E}[\Delta \mathbf{L}_{R,t+1} \mid \mathscr{G}_t] = \boldsymbol{\mu}_0$ and takes into account cross-dependencies between forecasts. We seek to investigate if models have equal predictive ability relative to a benchmark across different variables, based on the information set $\mathscr{F}_t$. If this is the case, $\Delta \mathbf{L}_{R,t+1}$ is a martingale difference sequence under the null hypothesis. Adopting the notation of GW, the global null can be written as a moment condition, $\mathcal{H}_0 : \mathbb{E}[\tilde{\mathbf{h}}_t \otimes \Delta \mathbf{L}_{R,t+1}] = \mathbf{0}$ based on a $q \times 1$ dimensional, $\mathscr{F}_t$-measurable vector $\tilde{\mathbf{h}}_t$. Define $\underset{qn \times 1}{\mathbf{Z}_{R,t+1}} = \tilde{\mathbf{h}}_t \otimes \Delta \mathbf{L}_{R,t+1}$, $\bar{\mathbf{Z}}_{R,n} = p^{-1} \sum_{t=R}^{T-1} \mathbf{Z}_{R,t+1}$, and $\underset{qn \times qn}{\hat{\boldsymbol{\Omega}}_n} = p^{-1} \sum_{t=R}^{T-1} \mathbf{Z}_{R,t+1} \mathbf{Z}'_{R,t+1}$. The multivariate version of GW is a Wald test:

$$T^h_{R,n} = p \bar{\mathbf{Z}}'_{R,n} \hat{\boldsymbol{\Omega}}_n^{-1} \bar{\mathbf{Z}}_{R,n} \xrightarrow{p} \chi^2_{qn}, \qquad \text{as } p \to \infty. \tag{2.8}$$

The crucial difference compared to the univariate version lies in the dimension of the matrices: both $\bar{\mathbf{Z}}'_{R,n}$ and $\hat{\boldsymbol{\Omega}}_n$ are a multiple n of the dimension of $\tilde{\mathbf{h}}_t$. Therefore, the degrees of freedom of the test distributions differ: the univariate test converges to a χ^2_q, rather than a χ^2_{qn} distribution. The test is still consistent against the alternatives in GW, meaning it is straightforward to implement and its properties are readily available. As the matrix $\hat{\boldsymbol{\Omega}}_n^{-1}$ includes the covariance between loss differentials, the multivariate test also evaluates dependence in the cross-section of forecasts, whereas the univariate test only accounts for serial correlation. However, similarly to the MP test, it quickly encounters inconsistency problems as the number of variables increases. If one follows the suggestion of GW and uses lagged values of $\Delta \mathbf{L}_{R,t+1}$ as $\tilde{\mathbf{h}}_t$, $\hat{\boldsymbol{\Omega}}_n$ is consistent and invertible when n is small. Vice-versa, tests that rely on $n \to \infty$ in the presence of cross-sectional dependence are inconsistent in a small n environment. In the next section, we compare our IU test to the multivariate GW and MP Wald tests in high- and low-dimensional settings.

2.3 MONTE-CARLO SIMULATIONS

In this section, we report the results of an extensive set of Monte Carlo simulations to evaluate size and power properties of the test. Most univariate

forecast accuracy tests are only valid asymptotically and we analyze how this affects our IU test. We are interested in the question how our test compares to (i) tests that directly evaluate the global null hypothesis of equal predictive ability across n forecasts and (ii) other methods of combining p-values. To this end, we construct different simulation designs. We construct actual forecasting scenarios and illustrate the properties of our method using three different sub-test: GW, DM, and CW. All three tests are widely used which underlines the relevance of our simulations for practitioners. Furthermore, they allow us to test for both conditional as well as unconditional predictive ability. For our small-sample power simulations, we only use GW and DM as sub-test which allows us to compare the results of the IU test directly with the multivariate GW test and the MP test. As GW and DM have different null hypotheses, our simulations also study scenarios where one sub-test will plausible reject its null while the other will sustain it. In our final, high-dimensional, power-simulation, we focus on nested models and present results using GW and CW as sub-tests for our global null. We cannot present results for MP and the multivariate GW test due to the large number of forecasts we evaluate. In all simulations, we present results from Fisher's method of combining p-values for further comparison. All results are generated through 5000 Monte-Carlo iterations.

2.3.1 *The Choice of r*

This subsection provides guidance on the choice of r. Theorem 2.1 shows that any $r \in (1, \infty)$ controls the asymptotic size of the test, and suffices to control the level in finite samples. In practice, one seeks to minimize the difference between empirical and nominal size.

Corollary 1 provides an upper bound for r based on values of δ and γ. The Corollary says $r \in (1, (2 - 2\sqrt{\gamma\delta} - \gamma - \delta)^{-1})$ for $\delta \in (\gamma - 2\sqrt{\gamma} + 1, \gamma - 2\sqrt{2}\sqrt{\gamma} + 2)$ and $\gamma \in [0, 0.5]$. In Figure 2.1, we plot implied upper bound for r based on values of δ slightly below $\gamma - 2\sqrt{2}\sqrt{\gamma} + 2$. We denote the difference by $\tilde{\delta} = (0.01, 0.011, 0.012, ..., 0.05)$. The figure shows r on the y-axis, δ on the x-axis, and γ on the z-axis The upper bound lies between 8.5 and 50, depending on γ. Generally, the larger δ, i.e. the smaller $\tilde{\delta}$, the larger r can be be. Unreported results show that for $\tilde{\delta}$ very small, the bound on r increases exponentially.

We now conduct an extensive simulation in which we skip the sub-testing step and simulate p-values directly to analyze the performance of the test under various dependence structures, and different values of r and n. We analyze different degrees of dependence and examine if and

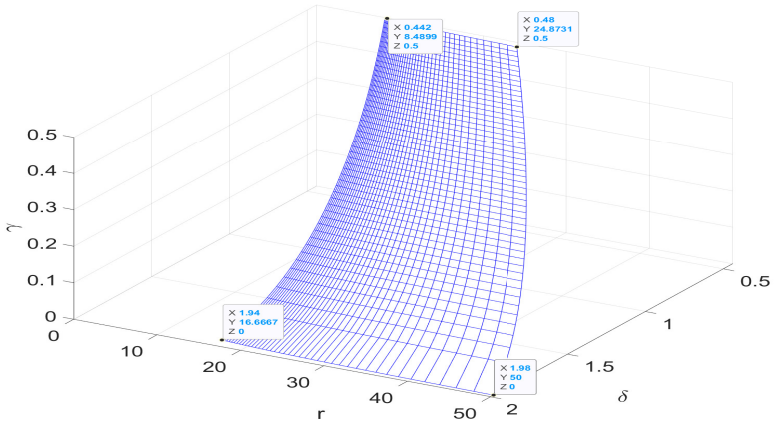


Figure 2.1: Upper Bound on r Implied by Corollary 1

Note: The figure plots the upper bound placed on r according to Corollary 1. The figure shows r on the y-axis, δ on the x-axis, and γ on the z-axis.

how the size changes in r. Further, we are interested in high- and low-dimensional settings and compare our test to conventional methods of combining p-values, namely $S_F = -2\sum_{i=1}^{n}\log(p_i) \sim \chi^2_{2n}$ (Fisher, 1934), $S_P = -2\sum_{i=1}^{n}\log(1-p_i) \sim \chi^2_{2n}$ (Pearson, 1933), and $S_T = \min_{1\leq i\leq n} p_i \sim \text{Beta}(1,n)$ (Tippett, 1931). To illustrate the behavior of the combined p-values for different values of r, we generate a vector of dependent random variables u that are uniform on [0,1]. We proceed as follows: First, we simulate a vector of multivariate normal random variables, $z \sim \mathcal{N}(0,\mathbf{\Sigma})$, with covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{n1} & \cdots & \cdots & 1 \end{bmatrix}$$

whose off-diagonal elements σ_{ij} are themselves random. However, in high-dimensional settings, it is not possible to simply draw each σ_{ij} from some distribution as the resulting $\mathbf{\Sigma}$ will not be symmetric positive definite. To ensure $\mathbf{\Sigma}$ is a covariance matrix, we first generate a random matrix $\mathbf{A} \sim \mathcal{N}(0,\mathbf{I}_n)$, drawn from a standard normal distribution. We then transform each row of $\mathbf{A}$ such that $\mathbf{A}^*_i = \mathbf{A}_i + (\xi_i - 1/2) * \sigma_a$, where the random variable ξ_i is uniform on the interval [0,1] and different for each row $i = 1,...,n$ of $\mathbf{A}^*$. The parameter σ_a characterizes the degree of dependence implied by $\mathbf{\Sigma}$. To be precise, σ_a controls the standard deviation of the distribution of

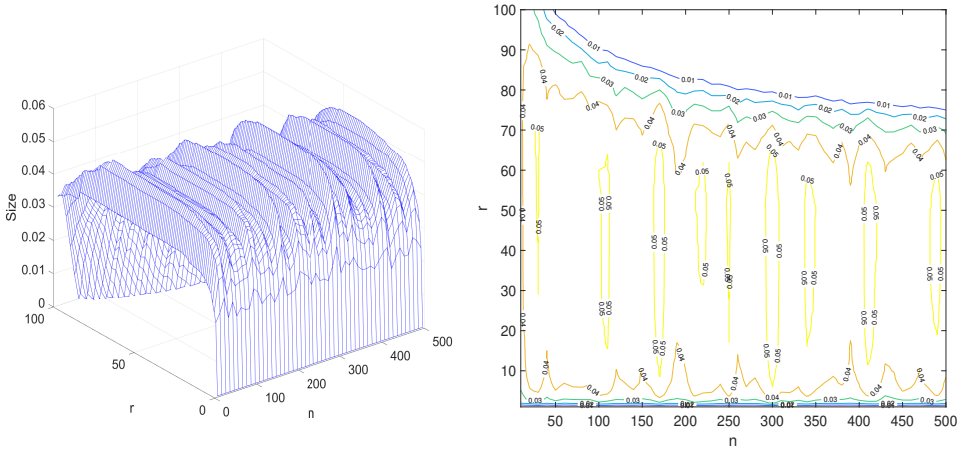
the off-diagonal elements σ_{ij}, i.e. the higher σ_a, the higher the probability of large covariances. That is, by setting σ_a, we are able to control the dependence of the data and test statistics. We transform the matrix $\mathbf{A}^*$ again such that $\tilde{\mathbf{A}}^* = (n-1)^{-1} * (\mathbf{A}^{*\prime}\mathbf{A}^*)$. Let $\tilde{\mathbf{a}}^*$ be a vector containing the diagonal elements of $\tilde{\mathbf{A}}^*$. The random positive definite covariance matrix $\mathbf{\Sigma}$ is then given as: $\mathbf{\Sigma} = (\tilde{\mathbf{a}}^*)^{-1/2}\tilde{\mathbf{A}}^*(\tilde{\mathbf{a}}^*)^{-1/2}$. It can easily be checked that $\mathbf{\Sigma}$ is indeed positive definite. We transform the dependent normal variables in z through $F(z)$, where F is the Gaussian CDF, to obtain a vector of dependent uniform random variables u. Repeating this process 5000 times allows us to compute the size for each $(1.1, 2.1, 3.1, ..., 100.1)$ and $n = (10, 20, ..., 500)$ at a nominal level of $\alpha = 5\%$. The results are shown as surface and contour plots in Figure 2.2 for $\sigma_a = (0, 2, 5)$. When the correlation between p-values is small ($\sigma_a = 0$), the size converges and stabilizes quickly at the nominal level. The test is visibly undersized for values of $r < 2$; we do not observe any differences across values of n. Interestingly, for large $r > 70$, the empirical size of the test declines again, with the decline being more pronounced for large n. The results are similar for $\sigma_a = 2$, although the empirical size stabilizes at a lower level, and starts to decline earlier. For very strong forms of dependence, the test becomes more conservative: when σ_a is increased to 5, the size falls below the nominal level between 3-2%. It now starts to decline for $r > 50$.

This stands in stark contrast to the three other methods of combining p-values we present as a comparison (Figure 2.3). When dependence is low ($\sigma_a = 0$), all methods are slightly oversized; interestingly, more so for small values of n. Conversely, as we increase the degree of dependence, the size becomes highly unstable and more distorted the greater n is. When dependence is high ($\sigma_a = 5$), the Type I error of the tests is large – above 30% for the Tippett (1931) method. The reason for this lies in the fact that these test simultaneously examine the independence assumption. This illustrates the importance of using our test in the presence of dependence to avoid size distortions, and highlights the fact that one cannot simply choose the minimum p-value as this will leave researchers highly prone to Type I errors.

The simulations confirm that any choice of $r \in (1, \infty)$ ensures the test is level-α. However, values of $r < 5$ result in the test being undersized, while the empirical size quickly approaches the nominal size for $r > 5$ and declines again for large values of $r > 50$, depending on the degree of dependence. In general, a similar analysis for the power of the test can only be conducted on a case-by-case basis. However, under Assumption 2.4, Corollary 2.1 narrows down the optimal range for r that maximizes the power of the test by providing an upper bound for r based on admissible values for δ and γ. As a rule of thumb, it implies that for a sparse number

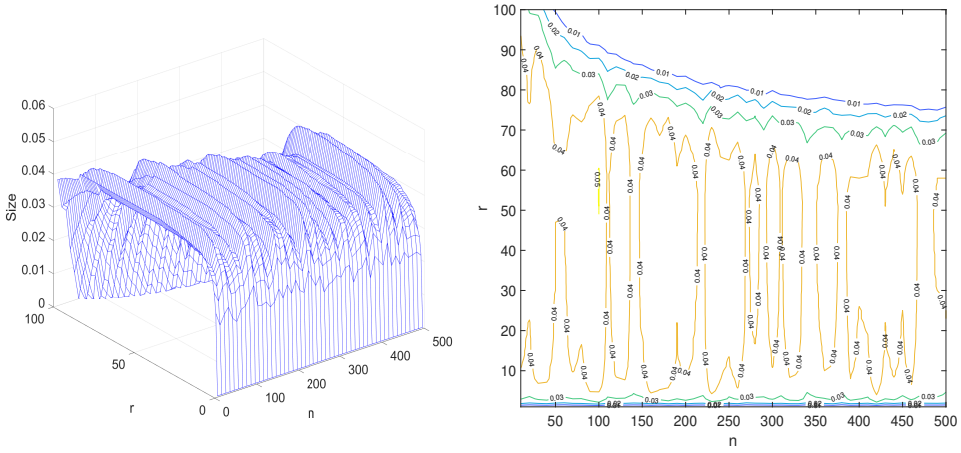
(a) $\sigma_a = 0$, Surface

(b) $\sigma_a = 0$, Contour



(c) $\sigma_a = 2$, Surface

(d) $\sigma_a = 2$, Contour



(e) $\sigma_a = 5$, Surface

(f) $\sigma_a = 5$, Contour

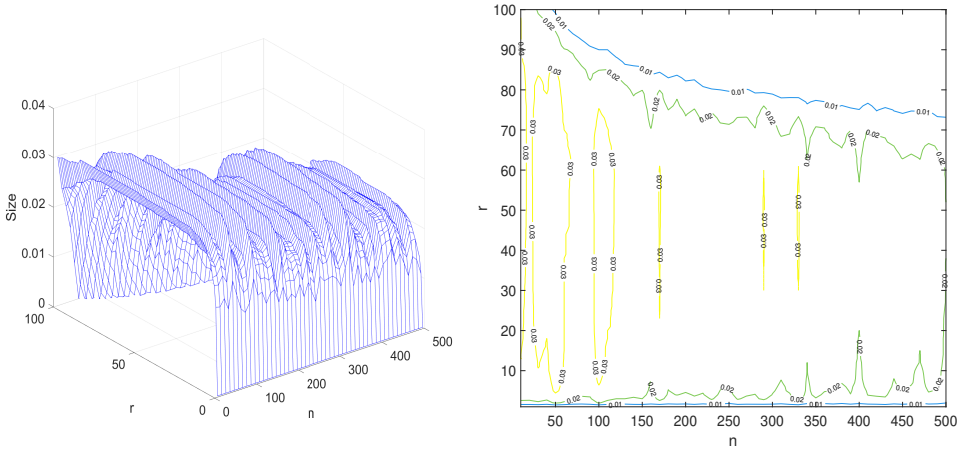


Figure 2.2: Simulated Size

Note: The figure reports the size of test statistic for $r = (1.1, 2.1, 3.1, ..., 100.1)$ and $n = (10, 20, ..., 500)$.

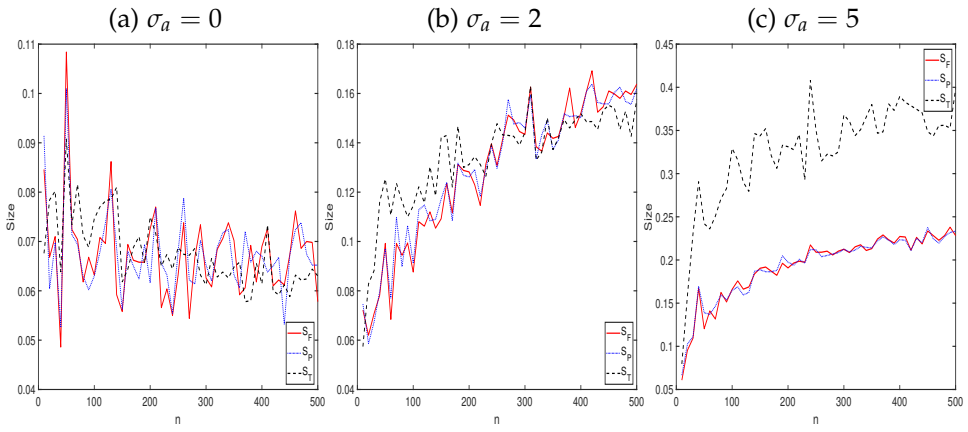


Figure 2.3: Size of Other Methods

Note: In the figure, the red line represents size of Fisher statistic (S_F), dotted blue line represents size of Pearson statistic (S_P), and dashed black line size of Tippett statistic (S_T).

of rejected sub-hypotheses, $r \in (1, 50)$ maximizes the power of the test. On that basis, we suggest to implement our test with any $r \in (5, 50)$. In the following sections, we set $r = 20$ and study the size and power of our test in detail.

2.3.2 *Size Properties*

This section analyzes the size of the test. We conduct univariate sub-tests using GW, DM, and CW tests by simulating different forecasting scenarios. We show results for low and high dimensions and compare them to other p-value combination methods. We also include results from the multivariate GW test as a reference but emphasize that the high-dimensional scenarios we consider are expected to result in size distortions.

We are interested in analyzing the size of our test in small n and large n settings. To this purpose, we proceed as follows: First, we generate an $n \times T$ matrix $\mathbf{Z}$ of cross-sectionally dependent Gaussian random variables. Each column of $\mathbf{Z}$ is drawn from a multivariate normal distribution with covariance matrix computed as described in the previous subsection. The parameter σ_a describes the degree of dependence between variables, i.e. the higher σ_a the more dependent are the forecasts. We use the i-th column of $\mathbf{Z}$ to construct n variables $Y_{i,t} = \phi_1 Y_{i,t-1} + Z_{i,t}$. We set $\phi = 0.5$ and generate two

Table 2.1: Test Size

		$\alpha = 1\%$				$\alpha = 5\%$				$\alpha = 10\%$			
	n:	10	50	100	200	10	50	100	200	10	50	100	200
Panel A: $\sigma_a = 0$													
$P_{r,n}$	*CW*	0.002	0.004	0.004	0.005	0.023	0.021	0.013	0.019	0.035	0.030	0.049	0.050
	DM	0.009	0.015	0.008	0.011	0.043	0.049	0.050	0.044	0.077	0.081	0.102	0.112
	GW	0.012	0.009	0.003	0.009	0.058	0.047	0.045	0.043	0.103	0.096	0.095	0.106
S_F	*CW*	0.008	0.004	0.002	0.000	0.044	0.016	0.002	0.000	0.059	0.023	0.014	0.001
	DM	0.300	0.975	1.000	1.000	0.546	0.996	1.000	1.000	0.720	0.999	1.000	1.000
	GW	0.075	0.332	0.563	0.875	0.215	0.546	0.777	0.962	0.311	0.671	0.865	0.979
	MGW	0.116	0.747	0.991	1.000	0.357	0.912	0.998	1.000	0.551	0.972	1.000	1.000
Panel B: $\sigma_a = 0.5$													
$P_{r,n}$	*CW*	0.002	0.004	0.004	0.005	0.023	0.021	0.013	0.019	0.035	0.030	0.049	0.050
	DM	0.009	0.015	0.008	0.011	0.043	0.049	0.050	0.044	0.077	0.081	0.102	0.112
	GW	0.012	0.009	0.003	0.009	0.058	0.047	0.045	0.043	0.103	0.096	0.095	0.106
S_F	*CW*	0.008	0.004	0.002	0.000	0.044	0.016	0.002	0.000	0.059	0.023	0.014	0.001
	DM	0.300	0.975	1.000	1.000	0.546	0.996	1.000	1.000	0.720	0.999	1.000	1.000
	GW	0.075	0.332	0.563	0.875	0.215	0.546	0.777	0.962	0.311	0.671	0.865	0.979
	MGW	0.116	0.747	0.991	1.000	0.357	0.912	0.998	1.000	0.551	0.972	1.000	1.000
Panel C: $\sigma_a = 2$													
$P_{r,n}$	*CW*	0.005	0.003	0.006	0.003	0.029	0.022	0.018	0.019	0.051	0.031	0.045	0.035
	DM	0.010	0.014	0.004	0.012	0.049	0.053	0.049	0.033	0.078	0.083	0.071	0.092
	GW	0.005	0.002	0.008	0.004	0.059	0.040	0.044	0.034	0.129	0.091	0.069	0.102
S_F	*CW*	0.019	0.029	0.038	0.063	0.041	0.051	0.062	0.077	0.077	0.067	0.072	0.092
	DM	0.290	0.940	0.999	1.000	0.540	0.990	1.000	1.000	0.701	0.998	1.000	1.000
	GW	0.083	0.346	0.487	0.668	0.223	0.496	0.675	0.806	0.327	0.625	0.730	0.834
	MGW	0.190	0.891	0.998	1.000	0.474	0.973	1.000	1.000	0.636	0.994	1.000	1.000
Panel D: $\sigma_a = 5$													
$P_{r,n}$	*CW*	0.005	0.004	0.007	0.002	0.019	0.018	0.019	0.019	0.034	0.035	0.031	0.014
	DM	0.009	0.012	0.009	0.005	0.027	0.033	0.030	0.023	0.052	0.064	0.051	0.043
	GW	0.009	0.006	0.010	0.002	0.058	0.038	0.040	0.031	0.093	0.090	0.067	0.054
S_F	*CW*	0.051	0.122	0.130	0.154	0.084	0.138	0.158	0.178	0.121	0.155	0.189	0.182
	DM	0.285	0.710	0.919	0.997	0.489	0.885	0.976	1.000	0.581	0.952	1.000	1.000
	GW	0.125	0.365	0.448	0.507	0.261	0.471	0.511	0.581	0.326	0.508	0.556	0.609
	MGW	0.665	0.992	1.000	1.000	0.844	1.000	1.000	1.000	0.925	1.000	1.000	1.000

Notes: The table reports the size of intersection-union test ($P_{r,n}$) for different sub-tests (CW, DM, GW) and different degrees of dependence (σ_a). The Fisher statistic (S_F) is provided as a comparison. Results obtained through 5000 Monte Carlo iterations. CW stands for Clark and West (2007), DM for Diebold and Mariano (1995), GW for Giacomini and White (2006), and MGW for the multivariate GW test. To compute the latter, loss differentials were averaged in higher dimensions, as described in the supplementary material, to make the covariance matrix invertible.

one-step-ahead rolling window forecasts of each $\mathbf{Y}_i = (Y_1, ..., Y_n)$ according to

$$\hat{Y}^{(1)}_{i,t+1} = \hat{\beta}_{i,1} Y_{i,t},$$
$$\hat{Y}^{(2)}_{i,t+1} = \hat{\mu} + \hat{\beta}_{i,2} Y_{i,t}.$$

The estimation window is set to be of length $h = 100$ and the out-of-sample window $p = 300$. The loss differential is specified to be $\Delta L_{i,t+1} = \{(Y_{i,t+1} - \hat{Y}^{(1)}_{i,t+1})^2 - (Y_{i,t+1} - \hat{Y}^{(2)}_{i,t+1})^2\}^n_{i=1}$. For the DM test, we set $\Delta L_{i,t+1} = \{\mathbf{Z}_{i,t}\}^n_{i=1}$ as its denominator is limiting to zero under the null for nested models. We compute the size at significance levels $\alpha = (0.01, 0.05, 0.1)$ for $n = (10, 50, 100, 200)$ and for different degrees of dependence $\sigma_a = (0, 0.5, 2, 5)$. We compare the results to the Fisher statistic, denoted by S_F. The results are reported in Table 2.1. The size properties of our IU test are good across the different sub-tests, albeit slightly undersized for CW's test. The size appears stable across different values of n with the global tests based on GW sub-tests being slightly oversized for small n at a nominal level of 10%. As we increase dependence up to $\sigma_a = 2$, there are no notable differences in the size of our statistic, although the size increases by some tenth of a percentage point. When dependence is increased further to very high levels ($\sigma_a = 5$), the test is noticeably more undersized. This indicates the test is conservative in the presence of strong dependence. In contrast, the Fisher statistic displays a high degree of variation in size paired with high distortions. What is more, there are stark differences across the underlying tests and for different values of n. For CW sub-test, the Fisher statistic is undersized for some n and oversized for others. As dependence increases, the size distortions of the Fisher statistic become greater; for DM and GW sub-tests the size reaches 1 for large n. Clearly, dependence renders the Fisher test impractical. The simulations also highlight the size distortions of the multivariate GW test, mirrored by other Wald-type tests whose results are corrupted by inconsistent high-dimensional covariance matrices. In contrast, these simulations demonstrate that our test has good size properties in small and high-dimensional settings and for different degrees of dependence.

2.3.3 *Power Properties*

We investigate the power of our test in three different settings, each designed with a specific purpose. The first scenario entertains a low dimensional, small n environment with changing cross-dependence. The second analyzes the rejection accuracy of the test by simulating combinations of true and false

null models. The third scenario considers the performance of the test in a high dimensional, large n case. Throughout, we evaluate the global null hypothesis in (2.1) against the alternative $\mathcal{H}_A = \cup_{i \in \mathbb{N}} \mathcal{H}_{i,A}$.

2.3.3.1 *Low Dimensions*

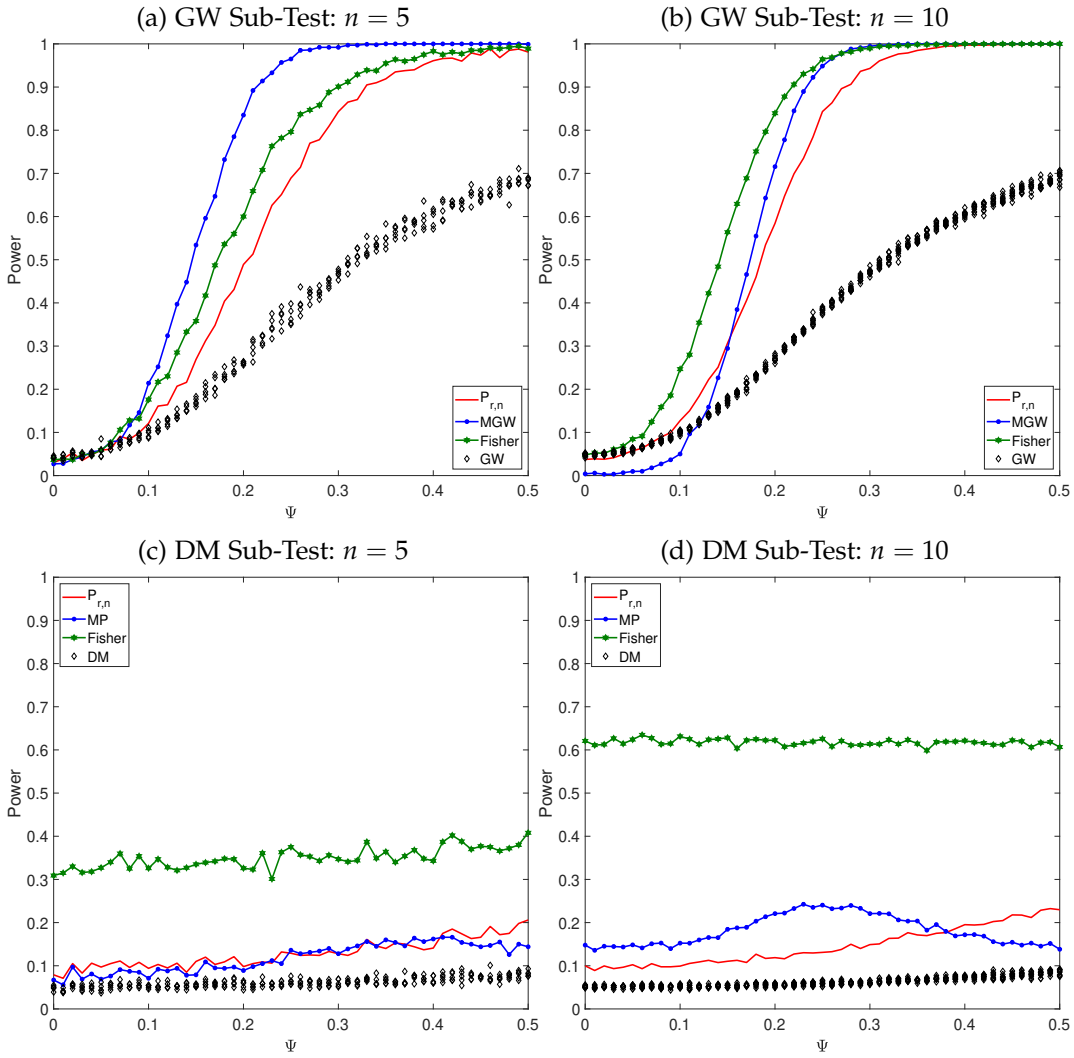


Figure 2.4: Power Functions IU Test Low Dimensions

Note: The figure reports the conditional predictive ability test of Giacomini and White (2006), GW, the multivariate GW test in Section **??**, MGW, the Diebold and Mariano (1995) test, DM, the Mariano and Preve (2012) test, MP, and the Fisher (1934) test. The x-axis reports the absolute values of the boundaries imposed on the distribution of $\mathbf{\Psi}$.

In the first Monte-Carlo study, we simulate the loss differentials directly as a VAR(1) process to ensure both cross- and serial correlation:

$$\Delta \mathbf{L}_{t+1} = \mathbf{\Psi} \Delta \mathbf{L}_t + \boldsymbol{\varepsilon}_{t+1}, \qquad \varepsilon \sim \mathcal{N}(0,1),$$

where $\Delta \mathbf{L}_{t+1} = (\Delta L_{1,t+1}, ..., \Delta L_{n,t+1})'$ and $t = 1, ..., p$. The coefficients in the $n \times n$ matrix $\mathbf{\Psi}$, $\psi_{i,j} \in [\psi_l, \psi_u]$, are drawn randomly from a truncated standard normal distribution with upper and lower bounds ψ_u and ψ_l. We ensure the roots of $\mathbf{\Psi}$ lie inside the unit circle by conditioning that its eigenvalues are smaller than one in absolute value. Additionally, to generate differences in test statistics, we impose $\mathbf{\Psi}$ be lower triangular. The number of forecasts is set to $n = (5, 10)$ with $p = 200$ periods each. The bounds are parameterized as $\psi_l = (0, -0.05, ..., -1)$ and $\psi_u = (0, 0.05, ..., 1)$. We then conduct GW and DM sub-tests for each of the n loss differentials. Notably, as soon as $\psi_l < 0$ and $\psi_u > 0$, the sub-hypothesis of all GW tests is no longer true. In contrast, the sub-hypothesis of the DM test, $\mathbb{E}[\Delta L_{i,t+1}] = 0$, remains true on average. Based on the p-values of the individual tests, we compute the global test statistic $P_{r,n}$ and its power function. In addition, we conduct the multivariate GW test as well as the adjusted MP test jointly for all n loss differentials. For further comparison, we also report the Fisher statistic. The results are shown in Figure 2.4. The solid red line is the power function of the intersection union test. The black diamonds represent the power with which each sub-test rejects its sub-hypothesis, the blue line corresponds to the multivariate GW and DM test, and the green line plots the Fisher test. We first consider the case of GW sub-tests. The power of our combined p-value statistic, is high regardless of n and higher than the power of individual GW tests. Its size also corresponds to the nominal level. The multivariate GW test performs well when n is small. However, it becomes increasingly undersized as n increases. Unreported simulations show that for $n > 10$, the covariance matrix of the multivariate GW test will be close to singular when $p \leq 200$, meaning it is no longer consistent. The Fisher statistic exhibits slightly greater power than our test, and also greater power than the multivariate GW test for $n = 10$. Moving to the DM sub-test, however, the Fisher statistic is extremely oversized. The MP and our test have roughly equal size for $n = 5$, but the former shows greater size distortions for $n = 10$. This simulation illustrates that, overall, our test has the best performance in a small dimensional scenario, regardless of the test type.

2.3.3.2 *Rejection Accuracy*

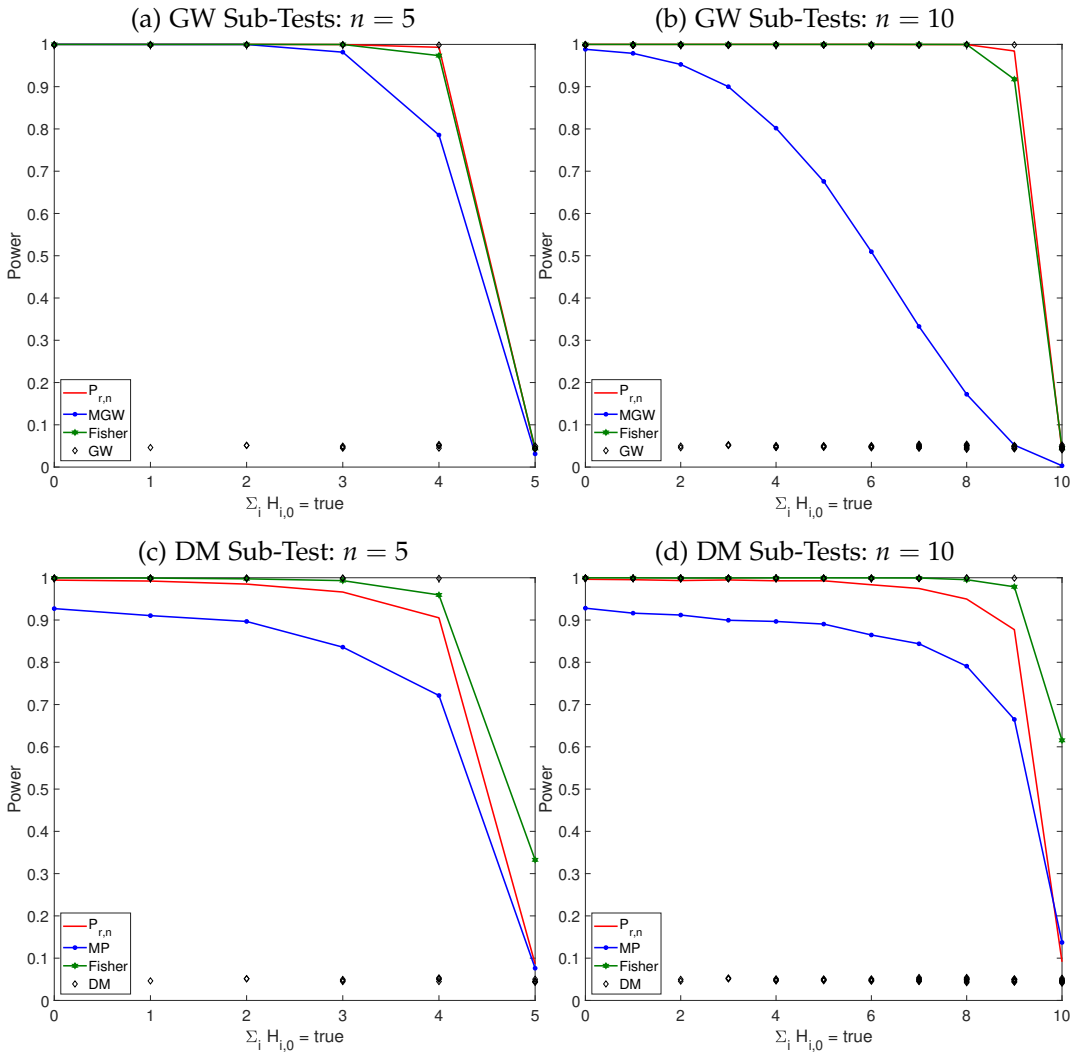


Figure 2.5: Power Functions IU Test Rejection Accuracy

Note: The figure reports the conditional predictive ability test of Giacomini and White (2006), GW, the multivariate GW test in Section 2.2.4, MGW, the Diebold and Mariano (1995) test, DM, the Mariano and Preve (2012) test, MP, and the Fisher (1934) test. The x-axis reports the absolute number of the true null hypotheses.

The second scenario we consider is one where $I = (0, 1, .., n)$ sub-hypotheses $\mathcal{H}_{i,0}$ are true, while $n - I$ sub-hypotheses are false. More precisely, we increase the number of true sub-hypotheses from 0 to n and are interested in the question how accurately our test rejects in each case and in comparison

to the multivariate GW test and the MP test. We use the same benchmarks as in the previous subsection. Specifically, we simulate

$$\begin{aligned} \Delta L_{i,t+1} &= \varepsilon_{t+1}, && \text{for} \quad i = 0, ..., I, \\ \Delta L_{j,t+1} &= F_{t+1} + \psi \Delta L_{j,t} + \nu_{t+1}, && \text{for} \quad j = I+1, ..., n, \\ F_{t+1} &= \mu + \psi F_t + \eta_{t+1} \end{aligned}$$

Here, $\varepsilon_t, \nu_t, \eta_t \sim \mathcal{N}(0,1)$. The coefficient ψ is fixed at 0.3 and μ is set to be 0.5. We ran unreported simulations with different values for the coefficients which did not change the overall picture of the results. F_t, the common factor across loss differentials, is the source of dependence. We set $n = (5, 10)$, noting that the ratio of n and p will impact the power of the multivariate GW and MP tests. The results are reported in Figure 2.5. The top panel shows the power functions based on GW sub-tests. Our test has consistently greater power than the multivariate GW test as well as the Fisher test. Indeed, for $n = 10$, the rejection pattern of the multivariate GW test looks remarkably different. The bottom two figures consider DM sub-tests. The MP test has lower power than our test and appears oversized when all null hypotheses are true for $n = 10$. Its power curve has steepened slightly, albeit less than the multivariate GW test. Although the Fisher test has slightly higher power than our test, it has a high likelihood of incorrectly rejecting the global null hypothesis when it is, in fact, true. Overall, the IU test exhibits the highest rejection accuracy. This simulation highlights that it is not obvious when the multivariate DM and GW tests reject their null hypotheses, obscuring the interpretability of their results.

2.3.3.3 *High Dimensions*

In the third scenario, we consider a large $n \times T$ framework and generate artificial rolling-window one-step-ahead forecasts instead of simulating the loss differential directly. We simulating nested models and therefore report results for GW and the CW sub-tests. Define the estimation window as h and the out-of-sample window as p such that $T = h + p + 1$ equals the total number of observations. First, we generate a random matrix $\mathbf{U}$ whose elements are uniform on $[-0.5, 0.5]$ which can easily be transformed into a symmetric positive definite matrix $\mathbf{\Sigma} = \mathbf{U}\mathbf{U}'$. This, in turn, can be used to generate a random $n \times T$ matrix with dependent rows, $\mathbf{Z}_i = \mathcal{N}(0, \mathbf{\Sigma})$, such that $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_n)$, where $\mathbf{Z}_i$ are $T \times 1$. The $\mathbf{Z}$ matrices are used to generate $n \times 1$ vectors $\mathbf{X}_t = \phi \mathbf{X}_{t-1} + \mathbf{Z}_t$, with $\phi = 0.3$. We summarize the information in $\mathbf{X}$ in form of a common factor using the principal components

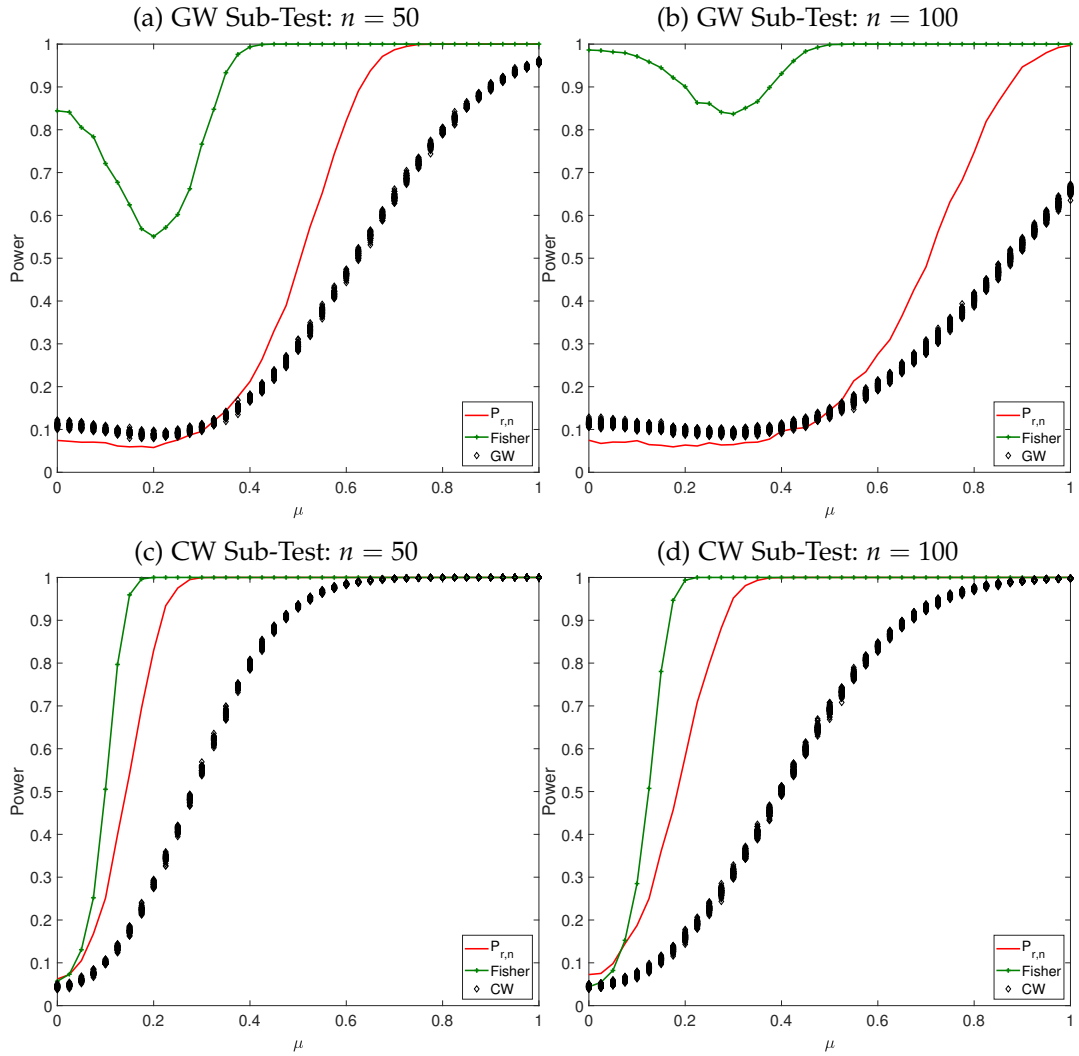


Figure 2.6: Power Functions IU Test High Dimensions

Note: The figure reports the conditional predictive ability test of Giacomini and White (2006), GW the Diebold and Mariano (1995) test, DM, and the Fisher (1934) test. The x-axis reports the mean of the first DGP.

estimator laid out in Bai (2003) and specify the process $\mathbf{Y_t} = \boldsymbol{\mu} + \mathbf{X}_t$, for $t = 1, ..., T$, and generate two different forecasts for each of the n variables in $\mathbf{Y}_t$:

$$\hat{Y}^{(1)}_{i,t+1} = \hat{\mu} + \hat{\beta}_{1,i}\tilde{F}_{1,t+1},$$
$$\hat{Y}^{(2)}_{i,t+1} = \hat{\beta}_{2,i}\tilde{F}_{1,t+1},$$

for $i = 1, ..., n$. The coefficients are estimated using OLS over a rolling window of size h. The total number of forecasts for each model is p. $\tilde{F}_{1,t}$ is the first principal component estimate of $\mathbf{X}$ and for simplicity its value in $t+1$ is assumed known. The forecast loss is specified to be quadratic, i.e. $L^{(1)}_{i,t+1} = (Y_{i,t+1} - \hat{Y}^{(1)}_{i,t+1})^2$. We parameterize $\boldsymbol{\mu}$ such that each $\mu_i = (0, 0.1, ..., 1)$. That is, for $\mu_i = 0$, the sub-hypothesis is true for both GW and the CW sub-test. The number of variables is set to $n = (50, 100)$, the estimation window to $h = 100$ and $p = 200$ such that $T = 301$. In a large n framework, the consistent computation of the multivariate GW and MP tests is no longer feasible. Therefore, we only present the results of the Fisher test. The power functions are depicted in Figure 2.6. Considering first the GW sub-tests, the top figures bear out the fact that Fisher's statistic exhibits considerable size distortions – exemplifying the need for appropriate corrections when considering dependent p-values. On the other hand, the IU test has low power for smaller values of μ, in line with the individual GW tests. As μ increases, however, the power of our test surpasses that of GW sub-tests. In contrast, for the CW test, it is slightly oversized and the divergence relative to the power of the individual tests is even more visible. Fisher's method displays moderately greater power.
The simulations highlight that the IU test we propose in this paper is a very reliable test for forecast accuracy in the presence of dependence. Whilst other tests may have higher power in some scenarios, they exhibit large Type I errors in others. We are able to confirm that our test most accurately rejects (sustains) the false (true) global null hypothesis of equal predictive ability across forecasts. The simulations substantiate that the intersection union test has good size and power properties regardless of dependence structures or dimensions.

2.4 EMPIRICAL ILLUSTRATION

This section provides an empirical illustration of the test. As interdependencies are ubiquitous in financial data, exchange rate forecasts are well suited

to apply our test. If a currency appreciates against the US-Dollar (USD) following the release of positive macroeconomic data, we would expect to see similar movements in its exchange rate against, say, the Euro (EUR). Dependencies are also reflected in common factors that explain currency variations, for example carry, momentum, or value factors. As such factors affect multiple currencies simultaneously, one can expect a model that is able to predict these elements for one FX rate to have an elevated likelihood of predicting them for others. Likewise, if a test only indicates predictive ability in a single instance out of many, this may well be a false positive (Harvey et al., 2016). Altogether, this strengthens the argument that one cannot disregard dependencies in the evaluation of exchange rate forecasts. We compile a large daily dataset of 84 exchange rates, consisting of 39 currencies vis-a-vis the USD, 23 currencies against the EUR, and 22 currencies against the Great British Pound (GBP). The three exchange rates USD-GBP, USD-EUR, and GBP-EUR are only included once. USD currency pairs are obtained from the Bank for International Settlements (BIS), EUR currency pairs from the European Central Bank (ECB), and GBP currency pairs from the Bank of England (BoE). The dataset spans from January 4, 2011 to April 1, 2021, a total number of 2558 observations for the USD, 2590 for the GBP, and 2622 for the EUR. We use the dataset to generate out-of-sample forecasts for each exchange rate and compare the performance of different models across currencies and tests. The aim is to show the main characteristics of the test in a scenario where some, but not all, individual tests reject for some models. Thereby, we illustrate how our test addresses mixed evidence problems. Moreover, we demonstrate how the test can be applied to different combinations of exchange rates, models, and individual tests. To this end, we estimate three models for each of the 84 exchange rates in our sample, a constant coefficients (CC) AR(1) model, and AR(2) model as well as a time-varying parameter (TVP) AR(1) model, estimated via maximum likelihood:

$$
\begin{aligned}
\Delta e_{i,t} &= \beta_{i,1}\Delta e_{i,t-1} + \nu_{i,t}, \\
\Delta e_{i,t} &= \beta_{i,2}\Delta e_{i,t-1} + \beta_{i,3}\Delta e_{i,t-2} + \eta_{i,t}, \\
\Delta e_{i,t} &= \gamma_{i,t}\Delta e_{i,t-1} + \varepsilon_{i,t}, \\
\gamma_{i,t} &= \rho_i \gamma_{i,t-1} + \epsilon_{i,t}.
\end{aligned}
$$

Here, $i = 1, 2, ..., 84$ and $\Delta e_{i,t}$ is the first difference of the log-FX rate. For each model, we generate one-step-ahead rolling window forecasts with an estimation window R of 750 for all exchange rates. We rely on the simulations of CW that show their MSPE-adjusted statistic performs well if p/R

converges to a finite constant. The GW test requires $p \to \infty$ whilst R remains fixed. However, their simulations show the test statistic exhibits excellent properties for the p/R ratios used here. This yields a total of 3×84 forecasts. We compare both the TVP-AR(1) and CC-AR(2) forecasts with the CC-AR(1) forecasts using GW, CW, and DM sub-test. The latter is not designed for nested models (Diebold, 2015); however, as it remains one of the most widely used tests, we include it nonetheless for illustrative purposes, emphasizing that its results should be taken with a grain of salt. For both GW and DM sub-test we use the loss differentials $\Delta L^{(1)}_{i,t+1} = \{L^{AR(1)}_{i,t+1} - L^{AR(2)}_{i,t+1}\}$ and $\Delta L^{(2)}_{i,t+1} = \{L^{AR(1)}_{i,t+1} - L^{TVP}_{i,t+1}\}$, where $L^{(m)}_{i,t+1}$ is the quadratic loss function of model m. This results in 3 different forecast accuracy tests being applied to compare the predictive ability of 2 models relative to a CC-AR(1) process for 84 exchange rates (USD + EUR + GBP), i.e. $3 \times 2 \times 84 = 504$ test statistics and p-values, respectively. Table 2.2 fleshes out the absolute number of

Table 2.2: Rejections for Individual Tests

	USD				EUR				GBP			
	AR(2)	(%)	TVP	(%)	AR(2)	(%)	TVP	(%)	AR(2)	(%)	TVP	(%)
CW	1	2.6	16	41.0	0	0	5	21.7	1	4.5	6	27.3
DM	1	2.6	3	7.7	4	17.4	1	4.3	6	27.3	1	4.5
GW	2	5.1	4	10.3	0	0	1	4.3	4	18.2	0	0

Note: The table contains the total number of rejections for each test as well as the number of rejections in percent of the total number of forecasts. AR(2) refers to CC-AR(2) compared to CC-AR(1) forecasts, while TVP refers to TVP-AR(1) forecasts compared to CC-AR(1) forecasts.

rejections of each sub-hypothesis at the 5%-level per model and currency. In addition, the table reports the number of rejections relative to the total number of tests conducted in each category. For instance, the GW sub-test rejects the null hypothesis that the CC-AR(2) and CC-AR(1) model display equal predictive ability twice for USD currency pairs. We have performed this test for each USD exchange rate in the dataset, i.e. 39 times, and only rejected in 5.1% of all cases. In several cases, the rejection rate is below 5%, i.e. in a range one would expect given a false discovery rate equal to the nominal size of the tests. The null hypothesis of equal predictive accuracy between TVP-AR(1) and CC-AR(1) is rejected more frequently, especially by the CW sub-test.

Figure 2.7 (a) - (b) show the p-values grouped by sub-tests. Subfigure (a) plots the results for the CC-AR(2) forecasts and Subfigure (b) the results for the TVP-AR(1). The dotted red values correspond to the p-values of the

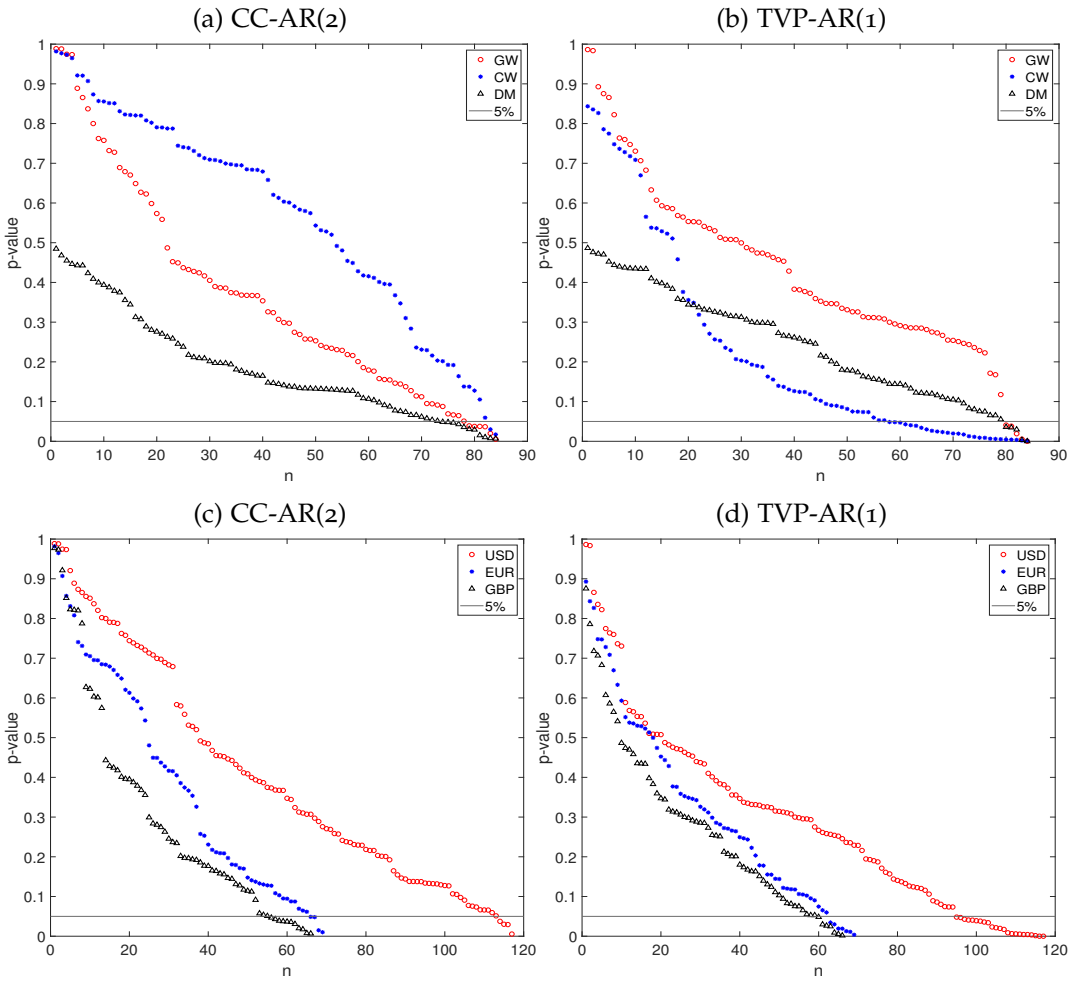


Figure 2.7: Sorted Individual p-Values

Note: The figure reprots the individual p-values for each of the n forecasts, grouped by test type (upper panel) and currency (lower panel). The solid line drawn at significance level $\alpha = 0.05$.

GW sub-test for USD, EUR, and GBP. The blue stars represent the p-values for the CW sub-test, while the black triangles correspond to the p-values of the DM sub-test. In all cases, with the exception of the CW test for the TVP-AR(1), only very few tests reject at the 5%-level across currencies for either model. Rather than assessing the results by test, one can also assess the findings per currency. Therefore, we group the p-values by currency and display the results for all tests in Figure 2.7 (c) - (d). The red dots show the p-values of all tests for exchange rates against the USD, the blue stars against the EUR, and the black triangles against the GBP. Again, with the exception of the TVP-AR(1) forecasts of USD exchange rates, there are only few rejections per currency. The figure visualises that we observe considerable differences across currencies, and the individual tests do not provide a conclusive answer to the question which model has the best predictive accuracy for each or all currencies.

We proceed by demonstrating that combining the individual p-values through our test is a convenient way to compute a global test statistic both when looking at USD, EUR, or GBP in isolation (small n) and for combinations of the three currencies or tests (large n). Panel A in Table 2.3 reports the test statistic of the IU test for USD, EUR, and GBP. The first three rows display the IU test statistic combining the p-values from each of the sub-tests: GW, CW, and DM. The first two columns CC-AR(2) and TVP-AR(1) contain the results combining the p-values of each sub-test, comparing the CC-AR(2) and TVP-AR(1) forecasts, respectively, with the CC-AR(1) forecasts for exchange rates against the USD. The statistics in these columns indicate whether there is evidence against the global null hypothesis of equal predictive ability between the two forecasts and a CC-AR(1) across all exchange rates in the sample that are quoted against the USD. The third column, Combined (Comb.), combines the p-values of the two preceding columns. The global null hypothesis is now that there is equal predictive ability between either of the two forecasts and a CC-AR(1), put differently, no available model produced better or worse forecasts than an CC-AR(1). In the fifth row, the p-values of the three sub-tests are combined together to ascertain whether there is evidence for predictive accuracy across different test types. The results for exchange rates quoted against EUR and GBP are reported analogously in the subsequent columns. It is common to use different individual tests to assess forecasting performance, hence we view this as an important application for our methodology, as the sub-tests may yield conflicting results. Our methodology allows researchers to formulate and evaluate a global null hypothesis of equal predictive accuracy

across different types of sub-tests.

Panel A presents scenarios with small to medium n which are reported in

Table 2.3: Test Statistic for Combined p-Values

Panel A: Individual Currencies

	USD			EUR			GBP		
	AR(2)	TVP	Comb.	AR(2)	TVP	Comb.	AR(2)	TVP	Comb.
CW	0.9	17.9*	8.9	0.1	12.2*	6.1	2.6	26.9**	13.4*
DM	0.9	2923.2***	1461.6***	4.3	1.3	2.2	6.9	1.5	3.4
GW	4.7	374.2***	187.1***	0.7	2.3	1.1	2.3	0.3	1.1
n	39	39	78	23	23	46	22	22	44
All	1.6	974.4***	487.2***	1.4	4.1	2.0	2.3	9.0	4.5
n	117	117	234	69	69	138	66	66	132

Panel B: Combined Currencies

	USD-EUR			USD-GBP			EUR-GBP			All		
	AR(2)	TVP	Comb.	AR(2)	TVP	Comb.	AR(2)	TVP	Comb.	AR(1)	TVP	Comb.
CW	1.0	11.4*	5.7	0.5	11.2*	5.6	1.3	13.1*	6.6	0.7	8.3	4.2
DM	2.5	1869***	934.5***	1.6	1838.8***	919.4***	3.4	0.8	1.7	1.8	1357.2***	678.6***
GW	3.0	239.3***	119.6***	2.9	235.4***	117.7***	1.1	1.2	0.6	2.2	173.7***	86.9**
n	62	62	124	61	61	122	45	45	90	84	84	168
All	1.0	612.9***	306.5***	1.0	623.0***	311.5***	1.1	4.4	2.2	0.7	452.4***	226.2***
n	186	186	372	183	183	366	135	135	270	252	252	504

Note: AR(2) refers to CC-AR(2) compared to CC-AR(1) forecasts, while TVP refers to TVP-AR(1) forecasts compared to CC-AR(1) forecasts. GW stands for Giacomini-White test, CW for Clark-West test, and DM for Diebold-Mariano test. Panel A: First two columns and first three rows contain test statistic of combined p-values from each tests for USD for each model. Third column contains test statistic for both models. Forth row contains test statistic for combined p-values from all tests for each model and for both models. Panel B contains test statistic for combined p-values from currency combinations for each test for each model and for both models as well as p-values from all tests combined for each model and for both models. Test rejects if statistic exceeds the critical values, obtained as $r/(\alpha(r-1))$: $\alpha = 1\%$: 105.263, denoted by ***, $\alpha = 5\%$: 21.053, denoted by **, $\alpha = 10\%$: 10.526, denoted by *.

the respective rows. Stars indicate the significance level at which the test rejects. Starting with USD exchange rates, the global null hypothesis that the CC-AR(2) and the CC-AR(1) forecasts exhibit equal predictive ability across exchange rates is sustained for all test types for all three currencies. In contrast, the same global null hypothesis for the TVP-AR(1) forecasts is rejected when combining the individual p-values of both GW and DM test (the CW test only rejects at the 10% level). When combining the p-values from all three tests, the global null hypothesis of equal predictive ability across exchange rates against the USD is also rejected. Likewise, the global null that neither model under- or outperforms a CC-AR(1) is rejected using the p-values of GW and DM tests as a basis as well as for all three tests combined.

On the contrary, the same global null hypotheses can only be rejected at the 10%-level for CW tests considering GBP exchange rates. Barring a 5%-level rejection of the CW test for the TVP-AR(1) model in case of the GBP, and a 10%-level rejection for the EUR, no other global null is rejected. Suppose one faces the decision whether to use the TVP model or not. The findings give rise to the question whether evidence against equal predictive ability exists only for the USD or also for combinations of the three currencies. Our test can analyze the null hypothesis of equal predictive ability across such combinations, and thereby provide an indication on whether a model can be deemed suitable for USD modeling or also in more general scenarios. Panel B presents the results for currency combinations, with n taking medium to large values. The first three columns combine the p-values of USD and EUR with the three corresponding sub-columns defined as in Panel A. That is, the values in the first column reflect the global null that there is equal predictive ability between CC-AR(2) and CC-AR(1) across exchange rates vis-a-vis USD and EUR. The second column, which reports the results for an analogous null hypothesis for the TVP-AR(1), shows that the latter is rejected at the 1% level according to both GW and DM test (and the CW test at the 10% level) as well as for all three tests combined. The same holds true for the global null that no model under- or outperforms a CC-AR(1) for USD and EUR combined. The results are the identical for USD and GBP combined. On the contrary, there is no evidence that any or both models perform differently than a CC-AR(1) when only considering p-values of EUR and GBP. The results should be interpreted bearing in mind that there are several rejections by the underlying individual tests, as reported in Table 2.2. However, this does not translate into an automatic rejection of the global null hypothesis, as the IU test accounts for dependence and false discovery. Finally, in the last column, we combine the p-values of all three currencies. The global null of equal predictive ability of TVP- and CC-AR(1) across USD, EUR, and GBP is rejected when combining the p-values of GW and DM test. The global null that no forecast is better or worse than a CC-AR(1) forecast is rejected by both GW and DM test at the 1% and 5% level. Next, we combine the p-values of all three tests, leading to the global null hypothesis that there is equal predictive ability across all currencies regardless of the underlying test. When tested for the TVP forecasts (penultimate column, fifth row) and for all forecasts (final column, fifth row), the global null is rejected at the 1% level. That is, there is evidence that, across all currencies, and all tests the TVP-AR(1) forecasts differ significantly from the CC-AR(1) forecasts. What is more, there is evidence that across currencies, and tests, forecasts of either model differ significantly from CC-AR(1) forecasts.

To summarize, the set of univariate tests conducted presents mixed evidence with generally few rejections. Through our IU test, we are able to formulate a range of global null hypotheses, based on different combinations of univariate tests. Thereby, we can present statistically significant evidence that a time-varying AR(1) model is able to produce superior forecasts compared to a constant coefficients model for currencies quoted against the USD. These results continue to hold when considering USD together with EUR or GBP exchange rates.

2.5 CONCLUSIONS

In this paper, we proposed an intersection-union multivariate forecasting accuracy test. The test is constructed using p-values of existing univariate tests that are treated as sub-tests and combined to evaluate a global null hypothesis of equal predictive ability across forecasts. Our test does not require any assumptions on the dependence structure between tests, and has a clearly defined rejection set. This is an important feature, as independence rarely holds in most forecasting exercises, and assuming independence may lead to considerable size distortions. In contrast, we proved that our test is level-α and consistent under the alternative. An extensive Monte Carlo simulation showed very good size properties of our test compared to conventional procedures of combining p-values, by conducting the intersection-union test using three popular univariate sub-tests of predictive ability. We showed that the properties of our multivariate procedure are unaffected by the number of sub-tests. To examine the power of the test, we simulated three different forecasting scenarios: a low dimensional scenario with changing cross-dependence, one that illustrated the rejection accuracy of our test, and, finally, a high-dimensional scenario. Our test showed high power in all cases. We compared our test to alternative benchmark procedures, each of which exhibited considerable limitations. An empirical illustration underpinned the wide applicability of our test. We compiled a large dataset of 84 daily exchange rates, quoted against USD, GBP, and EUR, to examine whether a time-varying AR(1) or a constant coefficients AR(2) model delivered different forecasts with respect to a constant coefficients AR(1) model. We also analyzed this across various combinations of currency pairs. While the results of the sub-tests themselves were mixed, our intersection-union test provided statistically significant evidence that the time-varying parameter model outperformed the constant coefficient AR(1) across combinations of currencies.

CHAPTER 3

TESTING FOR POINTWISE PREDICTIVE ABILITY WITH AN APPLICATION TO INTRADAY VOLATILITY FORECASTING

3.1 INTRODUCTION

In this paper, we propose new tests for predictive ability that compare two forecasts at each point in time. The tests are pointwise consistent and allow for the identification of breakpoints in forecasting performance.
Evaluating the out-of-sample performance of econometric models is crucial to determine which forecasting method to use under what circumstances. In particular for predictive stock return regressions, empirical evidence suggests that dependent variables are only able to predict returns in certain *pockets of predictability* (Timmermann, 2008). To address such findings, Georgiev et al. (2018) propose a test for structural breaks in predictive regressions and Demetrescu et al. (2022) introduce a method to evaluate the presence of episodic predictive ability in models with highly persistent endogenous predictors. Their empirical applications support the notion of temporal predictability in such regressions. Evidence like this suggests the relative predictive ability of different forecasting methods can also be time-varying, which gives rise to the need for appropriate evaluation procedures. Although numerous tests have been proposed that formulate a global test statistic for the joint evaluation of models throughout an entire forecasting period,[1] such global tests conceal time-variations in forecast accuracy. For

[1] See, for example West (1996), Diebold and Mariano (1995), Giacomini and White (2006), Clark and West (2006, 2007), Clark and McCracken (2001, 2015), or Li et al. (2022).

instance, we would expect a global test to reject a null hypothesis of equal predictive ability, even though such differences are only present during the first half of the evaluation period. Rossi (2021) provides a review of practical issues surrounding forecasting in unstable environments. The first formal time-varying method for the comparison of forecasts was introduced by Giacomini and Rossi (2010). They propose what is essentially a t-test on the mean of the difference in forecast loss between two models, computed over a rolling window. Thereby, the test is evaluating the Unconditional Predictive Ability (UPA) hypothesis of Giacomini and White (2006).[2] More recently, Odendahl et al. (2022) propose a tests for absolute and relative out-of-sample predictive ability of two models under state dependence.

Our paper adds to the literature on time-varying predictive ability tests in several ways. First, we formalise a test that evaluates the null hypothesis of equal Conditional Predictive Ability (CPA) at each point in time. We dub the test *pointwise* CPA test and it can be viewed as a time-varying analogue to the CPA test of Giacomini and White (2006). Second, we introduce a novel null hypothesis that examines whether, at each point in time, one model could have outperformed the other, conditional on all past and future information in the sample. We refer to this as pointwise Total Predictive Ability (TPA) test. Third, we are the first to propose a multivariate framework for the evaluation of time-varying predictive ability. We demonstrate that, at each point in time, both the pointwise CPA and TPA tests are consistent when jointly evaluating a cross-section of forecasts. Notably, our tests further distinguish themselves from the existing literature through their pointwise consistency. That is, they accurately reject at each time period, whereas existing tests reject accurately across time periods. This is achieved by evaluating the null hypothesis of equal relative predictive ability through a Kalman filter smoother algorithm.

We conduct extensive Monte-Carlo simulations to demonstrate the finite sample properties of our tests. The simulations show that both the TPA and CPA test have very good size and power in different scenarios. We apply the tests to compare intraday volatility forecasts of a GARCH model and the Markov-Switching Multifractal (MSM) model of Calvet and Fisher (2001, 2004). GARCH models are a natural benchmark in volatility forecasting due to their computational efficiency. MSM models are highly non-linear and, unlike a classical GARCH model, capture different states of volatility processes. Therefore, one might expect them to produce more accurate forecasts. Indeed, at lower frequencies, the evidence tilts in favour of the MSM model (e.g. Calvet and Fisher, 2004; Lux, 2022). When combined with a Support Vector Regression, Khashanah and Shao (2022) find that the MSM

[2]Their test can also be applied to other forecast comparison tests.

model also outperforms different GARCH models at one-minute frequency. However, intraday volatility exhibits strong diurnality (Andersen et al., 2019), which suggests that differences in predictive ability are periodic. We employ our test to investigate the relative forecast accuracy of MSM and GARCH models using one-minute NASDAQ index values. The results illustrate several points. First, the predictive ability of the two volatility models is highly time-varying. Second, across trading days, differences in predictive ability occur in clusters, particularly during market opening hours. Third, a simple GARCH model performs well in an estimation without overnight returns. Fourth, the MSM model has superior predictive ability when it comes to forecasting overnight returns.

The remainder of the paper is structured as follows: Section 3.2 sets out the theoretical framework behind the pointwise CPA and TPA tests. Section 3.3 presents the Monte-Carlo simulations and Section 3.4 discusses the empirical application. Section 3.5 concludes.

3.2 THEORY

3.2.1 *Forecasting Setup*

We consider both a univariate and a multivariate forecasting setup in which $2n$ different forecasts are being compared. The forecasts are based on the variables $\mathbf{v}_t \equiv (\mathbf{y}_t, \mathbf{x}_t)'$, with σ-field $\mathcal{G}_t := \sigma(\mathbf{v}_1', ..., \mathbf{v}_t')$ for $t = 1, ..., T$. Here, $\mathbf{y}_t$ is a $n \times 1$ dimensional vector of forecast variables, and $\mathbf{x}_t$ is an $s \times 1$ vector of predictors. We let $n \geq 1$ and $s \geq 0$, $t = 1, ..., T$, and define $\mathsf{N} := \{1, ..., n\}$. Forecasts of the n variables are generated using two $\mathcal{G}_t$-measurable functions $\hat{f}^{(j)}_{i,t+\tau}(\mathbf{v}_t; \hat{\boldsymbol{\beta}}_{i,t})$, for $j \in \{1, 2\}$ and $i \in \mathsf{N}$. The functions yield $\tau \geq 1$ step ahead forecasts and are dependent on a, possibly empty, set of parameters, $\hat{\boldsymbol{\beta}}_t$, which are estimated over a horizon r such that one obtains $p = T - r - \tau$ forecasts for each of the two forecasting methods. Let $\mathsf{T} := \{r, ..., T - \tau\}$. The estimation window $r < \infty$ can be different for each function. All $2n$ forecasts are evaluated using the same loss function,

$$L^{(j)}_{i,t+\tau} := L^{(j)}_{i,t+\tau}\left(Y_{i,t+\tau}, \hat{f}^{(j)}_{i,t+\tau}(\mathbf{v}_t; \hat{\boldsymbol{\beta}}_{i,t})\right) \in \mathbb{R}^+_0, \qquad \text{for } i \in \mathsf{N} \text{ and } t \in \mathsf{T}.$$

Since the loss function is restricted to be positive, the loss-differential $\Delta L_{i,t+\tau} := L^{(1)}_{i,t+\tau} - L^{(2)}_{i,t+\tau}$ indicates which forecasting method produced the smaller forecast loss in period t. The tests proposed in this paper provide

the statistical means to assess the hypothesis that both methods have equal predictive ability for each $t \in \mathsf{T}$.

3.2.2 *Null Hypotheses*

3.2.2.1 *Conceptual Remarks*

This paper develops time-varying predictive ability tests for both a single loss differential as well as multiple, possibly dependent, loss differentials. Our testing framework examines loss differentials by conditioning on two different σ-fields. The first way of conditioning builds on the existing literature, and the second one is a novel contribution of this paper. The concept of Conditional Predictive Ability (CPA) was introduced by Giacomini and White (2006). Their approach is based on evaluating the null hypothesis $\mathcal{H}_0 : \mathbb{E}[\Delta L_{t+\tau} \mid \mathcal{G}_t] = 0$, i.e. the expected value of the loss differential conditional on the σ-field $\mathcal{G}_t$. Thus, it accounts for the uncertainty and bias that can arise from miss-specification of the forecasting model and from changes in the information set over time. Giacomini and White (2006) exploit the fact that under the conditional null hypothesis, the loss differential is a martingale difference sequence and write: $\mathcal{H}_0 : \mathbb{E}[\tilde{h}_t \Delta L_{t+1}] = 0$, for all $\mathcal{G}_t$-measurable functions $\tilde{h}_t$. They restrict themselves to a subset of such functions, denoted by h_t. Conventionally, their test is implemented using lagged values of ΔL_{t+1} as a test function in which case one essentially tests for the presence of serial correlation in the loss differential. Hence, under homoskedasticity, the null hypothesis becomes:

$$\Delta L_{i,t+1} = \alpha + \beta \Delta L_t + \epsilon_{t+1}, \qquad \mathcal{H}_0 : \alpha = \beta = 0.$$

Their test is examined against the alternative that $\mathbb{E}[\bar{Z}_r]\mathbb{E}[\bar{Z}_r] \geq \delta > 0$ for $\bar{Z}_r = \frac{1}{p} \sum_{t=r}^{T} h_t \Delta L_{i,t+1}$. The choice of h_t is therefore crucial for the power of the test statistic. More recently, the literature has emphasised the issue of forecasting under instabilities (see Rossi (2021) for a survey). Giacomini and Rossi (2010) stress that the forecasting performance of models may change and break down locally. They propose a test which can be understood as a sequence of t-tests on:

$$\Delta \hat{L}_{i,m,t} = \alpha + \epsilon_t, \qquad \mathcal{H}_0 : \mu = 0,$$

where $\Delta \hat{L}_{i,m,t}$ is computed over a rolling window of size m yielding a sequence of $p - m$ test statistics. However, a rolling window will inevitably

smooth over breakpoints and mask the exact time in which their relative performance changes.

3.2.2.2 *Univariate Null Hypotheses*

We begin by formulating null hypotheses for the evaluation of two competing models based on a single loss differential. Tests of such hypotheses are henceforth referred to as univariate tests. Let the expectation of the i-th loss differential conditional on a period-t σ-field be:

$$\mathbb{E}[\Delta L_{i,t+\tau} \mid \mathcal{F}_t] = \mu_{i,t+\tau|t} \qquad \text{for each } t \in \mathsf{T} \text{ and any } i \in \mathsf{N}. \tag{3.1}$$

Here, $\mathcal{F}_t = \sigma(\Delta L_{i,r+\tau}, ..., \Delta L_{i,t})$ is the σ-field generated by the history of the loss differentials and clearly $\mathcal{F}_t \subseteq \mathcal{G}_t$. By analysing this expectation, we provide a time-varying analogue to the CPA test of Giacomini and White (2006). It should, however, be noted that the power of Giacomini and White's (2006) test depends on the choice of test function h_t, which may generate a σ-algebra that is a smaller subset of $\mathcal{G}_t$ than $\mathcal{F}_t$. Second, their test is a global test that forms a single test statistic for all t. In contrast, we form a null hypothesis of pointwise CPA, which says that at time t one cannot predict that one model outperforms the other

$$\mathcal{H}_{i,t,0} : \mu_{i,t+\tau|t} = 0, \qquad \text{for } t \in \mathsf{T} \text{ and } \; i \in \mathsf{N}. \tag{3.2}$$

The null hypothesis is tested against the local alternative

$$\mathcal{H}_{i,t,A} : \mu_{i,t+\tau|t} \in \mathsf{M}_{i,t,A}, \qquad \text{for } t \in \mathsf{T} \text{ and } \; i \in \mathsf{N}, \tag{3.3}$$

where $\mathsf{M}_{i,t,A}$ is a set of admissible values under the alternative.

In addition, we introduce a new hypothesis to the literature, which we call pointwise Total Predictive Ability (TPA) hypothesis. It evaluates the following conditional expectation

$$\mathbb{E}[\Delta L_{i,t+\tau} \mid \mathcal{F}_T] = \mu_{i,t+\tau|T} \qquad \text{for each } t \in \mathsf{T} \text{ and any } i \in \mathsf{N}, \tag{3.4}$$

where $\mathcal{F}_T = \sigma(\Delta L_{i,r+\tau}, ..., \Delta L_{i,T})$. That is, we examine the expectation of the loss differential at time t conditional on all future and past information embedded in $\mathcal{F}_T$:

$$\mathcal{H}^*_{i,t,0} : \mu_{i,t+\tau|T} = 0, \qquad \text{for } t \in \mathsf{T} \text{ and } \; i \in \mathsf{N}, \tag{3.5}$$

This hypothesis is tested against the alternative

$$\mathcal{H}^*_{i,t,A} : \mu_{i,t+\tau|T} \in \mathsf{M}_{i,t,A}, \qquad \text{for } t \in \mathsf{T} \text{ and } \ i \in \mathsf{N}, \tag{3.6}$$

The null hypotheses of CPA and TPA differ fundamentally. Under the CPA null hypothesis in (3.2) one cannot predict whether one forecasting method will outperform the other at time $t+\tau$ based on information up to time t. On the other hand, the TPA hypothesis in (3.5) tests if one method could have outperformed the other at time t based on all information in T. It is possible to write the alternative hypothesis to the null of CPA as,

$$\mathcal{H}_{t,A} : \mathbb{E}[\Delta L_{t+\tau} \mid \mathcal{F}_t]\mathbb{E}[\Delta L_{t+\tau} \mid \mathcal{F}_t] \geq \delta > 0, \qquad \text{for all } t \in \mathsf{T} \text{ and } i \in \mathsf{N}, \tag{3.7}$$

which encapsulates the alternative in Giacomini and White (2006).[3] However, it has potentially greater power in a range of scenarios, as the properties of the test do not depend on a test function. We permit $\mathsf{M}_{i,t,A} \neq \mathsf{M}_{i,j,A}$ for any $t \neq j$ and $t, j \in \mathsf{T}$. By allowing $\mathsf{M}_{i,t,A}$ to be time-varying, we can examine the null hypothesis against a range of changing alternatives. For instance, by specifying $\mathsf{M}_{i,t,A} = \{x \in \mathbb{R} : x < 0\}$ for $t = R, ..., R + p/2$ and $\mathsf{M}_{i,t,A} = \{x \in \mathbb{R} : x > 0\}$ for $t = r + p/2 + 1, ..., T - \tau$, we can test the null hypothesis against the specific alternatives that the first model performs better during the first half of the sample, while the second model is superior in the second half.

3.2.2.3 *Multivariate Null Hypotheses*

Whilst the null hypotheses above focus on univariate forecast evaluation, considering multiple forecasts jointly is another leading case of interest. In this context, it is crucial to consider potential cross-dependencies between forecasts, as failing to do so can lead to considerable size distortions (e.g. Qu et al., 2021; Spreng and Urga, 2022). Our paper is the first to examine pointwise multivariate hypotheses. Let $\Delta\mathbf{L}_t = (\Delta L_{1,t}, ..., \Delta L_{n,t})'$. The multivariate version of the CPA hypothesis examines the conditional expectation

$$\mathbb{E}[\Delta\mathbf{L}_{t+\tau} \mid \boldsymbol{\mathcal{F}}_t] = \boldsymbol{\mu}_{t+\tau|t} \qquad \text{for all } t \in \mathsf{T} \text{ and any } i \in \mathsf{N}, \tag{3.8}$$

where $\boldsymbol{\mu}_{t+\tau|t} := (\mu_{1,t+\tau|t}, ..., \mu_{n,t+\tau|t})'$ and $\boldsymbol{\mathcal{F}}_t := \sigma(\Delta\mathbf{L}'_{r+\tau}, ..., \Delta\mathbf{L}'_t)$. Here, the σ-field contains cross-sectional as well as time-series information. Thus, we now condition on past information about variables in conjunction with the

[3] Clearly, a similar formulation can be obtained for the alternative TPA hypothesis.

covariation between them. The multivariate null hypothesis implies that pointwise CPA holds in the cross section at each point in time:

$$\mathcal{H}_{t,0} = \bigcap_{i \in \mathsf{N}} \mathcal{H}_{i,t,0} : \boldsymbol{\mu}_{t+\tau|t} = \mathbf{0}, \qquad \text{for } t \in \mathsf{T}, \tag{3.9}$$

The null hypothesis is tested against the local alternative

$$\mathcal{H}_{t,A} : \bigcup_{i \in \mathsf{N}} \mathcal{H}_{i,t,A} : \boldsymbol{\mu}_{t+\tau|t} \in \bigcup_{i \in \mathsf{N}} \mathsf{M}_{i,t,A}, \qquad \text{for } t \in \mathsf{T}. \tag{3.10}$$

The alternative hypothesis is that, considering previous information about variables and their dependencies, we can predict that for the i-th forecasting variable, one method outperforms the other. Define $\boldsymbol{\mu}_{t+\tau|T} := (\mu_{1,t|T}, ..., \mu_{n,t|T})'$, such that the multivariate TPA analogue is:

$$\mathbb{E}[\Delta \mathbf{L}_t \mid \boldsymbol{\mathcal{F}}_T] = \boldsymbol{\mu}_{t+\tau|T} \qquad \text{for all } t \in \mathsf{T} \text{ and any } i \in \mathsf{N}, \tag{3.11}$$

with $\boldsymbol{\mathcal{F}}_T := \sigma(\Delta \mathbf{L}'_{r+\tau}, ..., \Delta \mathbf{L}'_T)$. Importantly, the TPA test now conditions on all past and future covariations between the loss differentials. The local TPA null reads

$$\mathcal{H}^*_{t,0} = \bigcap_{i \in \mathsf{N}} \mathcal{H}^*_{i,t,0} : \boldsymbol{\mu}_{t+\tau|T} = \mathbf{0}, \qquad \text{for } t \in \mathsf{T}, \tag{3.12}$$

and is tested against the local alternative

$$\mathcal{H}^*_{t,A} : \bigcup_{i \in \mathsf{N}} \mathcal{H}^*_{i,t,A} : \boldsymbol{\mu}_{t+\tau|T} \in \bigcup_{i \in \mathsf{N}} \mathsf{M}_{i,t,A}, \qquad \text{for } t \in \mathsf{T}. \tag{3.13}$$

The testing framework is flexible in the sense that it allows for $\mathsf{M}_{i,t,A} \neq \mathsf{M}_{j,t,A}$, for $i, j \in \mathsf{N}$ and $i \neq j$. Thereby, it enables the testing of different null hypotheses, whilst still considering cross-dependencies.

3.2.3 *Assumptions*

Assumption 3.1. For all $t = r, ..., T - \tau$, the joint process for all n loss differentials is $\Delta \mathbf{L}_{t+\tau} = \boldsymbol{\alpha}_{t+\tau} + \boldsymbol{\varepsilon}_{t+\tau}$, where $\boldsymbol{\alpha}_{t+\tau} = \boldsymbol{\alpha}_{t+\tau-1} + \boldsymbol{\eta}_{t+1}$, such that $\Delta L_{i,t+\tau} = \alpha_{i,t+\tau} + \varepsilon_{i,t+\tau}$, with $\alpha_{i,t+\tau} = \alpha_{i,t+\tau-1} + \eta_{i,t+\tau}$, for all $i \in \mathsf{N}$.

Assumption 3.2. For all $t \in \mathsf{T}$:

A. $\{\boldsymbol{\varepsilon}_t\}$ and $\{\boldsymbol{\eta}_t\}$ are *i.i.d.* Gaussian processes,

B. $\mathbb{E}[\boldsymbol{\varepsilon}_t] = \mathbf{0}$, $\mathbb{E}[\boldsymbol{\eta}_t] = \mathbf{0}$, $\mathbb{E}\|\boldsymbol{\eta}_t\|^\delta < \infty$ and $\mathbb{E}\|\boldsymbol{\varepsilon}_t\|^\delta < \infty$, for $0 < \delta \leq 4$,

C. $\mathbb{E}\|\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_{t+k}\| = \boldsymbol{\Sigma}_\varepsilon \mathbf{1}_{\{k=0\}} < \infty$ and $\mathbb{E}\|\boldsymbol{\eta}_t\boldsymbol{\eta}'_{t+k}\| = \boldsymbol{\Sigma}_\eta \mathbf{1}_{\{k=0\}} < \infty$,

D. $\mathbf{w}'\boldsymbol{\Sigma}_\eta\mathbf{w} > 0$ and $\mathbf{w}'\boldsymbol{\Sigma}_\varepsilon\mathbf{w} > 0$ for all $\mathbf{w} \in \mathbb{R}^n$,

E. $\mathbb{E}\|\boldsymbol{\varepsilon}_{t+k}\boldsymbol{\eta}'_{t+j}\| = \mathbf{0}$ for all k and j.

Assumption 3.3. $\boldsymbol{\alpha}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{V}_0)$.

Assumption 3.1 stipulates that the loss differential is the sum of a local level component and a noise component. Notably, we do not require the loss differential to be stationary. Assumption 3.2 imposes several conditions on the noise components, ensuring they are zero-mean, white, and *i.i.d.* Gaussian. In Appendix C.2, we discuss the effects of removing the Gaussian assumption. Further, the moments of the disturbance terms exist up to order four, and their covariance matrix is assumed to be positive definite. Assumption 3.3 is a standard assumption in the literature on state-space models (Durbin and Koopman, 2012), and ensures the density of the initial first state is Gaussian with mean $\boldsymbol{\mu}_0$ and variance $\mathbf{V}_0$. Assumption 3.1, 3.2-A to 3.2-E and 3.3 together imply the joint density of the loss differentials corresponds to:

$$f_{\boldsymbol{\theta}} \equiv f(\Delta\boldsymbol{L}_{r+\tau}, ..., \Delta\boldsymbol{L}_T; \boldsymbol{\theta}) = \prod_{t=r+\tau}^{T} p(\Delta\mathbf{L}_t \mid \boldsymbol{\mathcal{F}}_{t-1}), \tag{3.14}$$

for the parameter vector $\boldsymbol{\theta} := \left(vec(\boldsymbol{\Sigma}_\eta)', vec(\boldsymbol{\Sigma}_\varepsilon)'\right)'$.

Assumption 3.4. **A**. $\boldsymbol{\theta}$ is an interior point in the parameter space $\boldsymbol{\Theta} \subset \mathbb{R}$,

B. If $f_{\boldsymbol{\theta}_1} = f_{\boldsymbol{\theta}_2}$ then $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$,

C. $\frac{\partial^3 \log f}{\partial \boldsymbol{\theta}^3}$ exists and is continuous in the neighbourhood of $\boldsymbol{\theta}$.

Assumption 3.4 imposes standard regularity conditions that, in conjunction with Assumption 3.2-B and 3.2-D, ensure the parameter vector $\boldsymbol{\theta}$ can be estimated.

Assumption 3.5. $\mathbb{E}[\varepsilon_{i,t}\varepsilon_{j,t}] = 0$ and $\mathbb{E}[\eta_{i,t}\eta_{j,t}] = 0$ for all $i \neq j$ and $i, j \in \mathsf{N}$.

Assumption 3.5 is necessary to ensure the consistency of the univariate tests by restricting the covariance matrices to be diagonal. If Assumption 3.5 holds in addition to Assumptions 3.1-3.3, Equation (3.14) reduces to

$$f_{\theta_i} \equiv \prod_{t=r+\tau}^{T} p(\Delta L_{i,t} \mid \mathcal{F}_{t-1}), \qquad \text{for all } i \in \mathsf{N},$$

with the parameter vector $\theta_i := (\sigma_{i,\eta}, \sigma_{i,\varepsilon})'$. This serves as a useful illustration of the differences between the univariate and the multivariate framework, which applies to all existing forecast accuracy tests. By evaluating a single loss differential, one implicitly imposes Assumption 3.5, altering the information set that is used to conduct the test.[4] The σ-field in the univariate case does not include information about past cross-correlations.

Assumption 3.6. **A**. $n/p \to 0$,

B. $\boldsymbol{\Sigma}_\varepsilon = q\boldsymbol{\Sigma}_\eta$ for any scalar $q \in \mathbb{R}_0^+$.

Assumption 3.6-A rules out high-dimensional cases in which the number of forecast variables is larger than the number of out-of-sample periods and 3.6-B imposes that the covariance matrix of the n level components is proportional to the covariance matrix of the loss differentials.

3.2.4 *Pointwise Conditional Predictive Ability Test*

3.2.4.1 *Univariate Test*

In this section, we introduce the pointwise CPA test. For simplicity, we first focus on the univariate case where either $n = 1$, or $n > 1$ and Assumption 3.5 holds true. It follows from Assumption 3.1-3.3 that the expectation of the loss differential conditional on $\mathcal{F}_t$ is equal to

$$\mathbb{E}[\Delta L_{i,t+\tau} \mid \mathcal{F}_t] = \mathbb{E}[\alpha_{t+\tau} \mid \mathcal{F}_t] = \mu_{i,t+\tau|t}, \tag{3.15}$$

That is, under the null hypothesis, the expected value of the level component and the loss differential both equal zero. It can easily be shown that under Assumptions 3.1-3.3, the time-t conditional expectation and variance of the loss differential is obtained recursively as:

$$\begin{aligned} \mu_{i,t+\tau|t} &= \mu_{i,t|t-1} + \frac{p_{i,t|t-1}}{p_{i,t|t-1}+1}(\Delta L_{i,t} - \mu_{i,t|t-1}), \\ p_{i,t+\tau|t} &= \frac{p_{i,t|t-1}}{p_{i,t|t-1}+1} + \tau q_i, \\ v_{i,t+\tau|t} &= p_{i,t+\tau|t}\tilde{\sigma}_{i,\varepsilon}(q_i), \end{aligned} \tag{3.16}$$

where $\tilde{\sigma}_{i,\varepsilon}(q_i) := \sum_{t=r+\tau+1}^{T}(\Delta L_{i,t} - \mu_{i,t|t-1})^2/(p_{i,t|t-1}+1)$ for a given value of q_i. These are Kalman filter predictions which concentrate the variance of the

[4] Note that one always imposes an information set when evaluating a test statistic based on expectations.

state equation out of the recursions as derived, for example, in (Harvey, 1990, Chapter 3). Per Assumption 3.2, the Kalman filter is the optimal estimator of $\mu_{i,t|t-1}$ and $v_{i,t|t-1}$ [see e.g. Simon (2006)].

The recursions in (3.16) can be used to formulate the predictive likelihood function

$$\log \tilde{f}_{\theta_i} = -\frac{p-1}{2}\log 2\pi - \frac{p-1}{2}\log\left(\tilde{\sigma}_{i,\varepsilon}\right) - \frac{1}{2}\sum_{t=r+\tau}^{T}\log(p_{i,t|t-1}+1),$$

which can be maximised with respect to the signal-to-noise ratio q. Denote the solution to the maximisation problem $\arg\max_{\theta_i \in \boldsymbol{\Theta}} \log \tilde{f}$ by $\tilde{\theta}_i := \tilde{q}_i$. Assumption 3.4 guarantees that $\tilde{\theta}_i \to \theta_i$ for $p \to \infty$. The maximisation problem cannot be solved in closed form, but the solution can be efficiently computed using standard numerical procedures. One does not need to include the initial conditions $\mu_{i,0}$ and $v_{i,0}$ in the parameter vector, as the Kalman filter will converge exponentially fast under Assumption 3.4. In most applications, it suffices to set $\mu_{i,0} = \Delta L_{i,0}$ and $v_{i,0} = (1+q_i)\tilde{\sigma}_{i,\varepsilon}$.

To test the local null hypothesis in (3.2), we define the following $p \times 1$ dimensional vector of test statistics for pointwise conditional predictive ability:

$$S_{i,t+\tau|t} = \frac{\mu_{i,t+\tau|t}}{\sqrt{v_{i,t+\tau|t}}} \qquad \text{for any } i \in \mathsf{N}, \tag{3.17}$$

for all $t = r, ..., T-\tau$. The test statistic is defined on the entire real line, i.e. $S_{i,t+\tau|t} \in \mathbb{R}$. Using established properties of the Kalman filter, we can summarise the properties of the test statistic

Proposition 3.1. *For any $i \in \mathsf{N}$, suppose Assumptions 3.1-3.5 hold such that $\tilde{\theta}_i \xrightarrow{p} \theta_i$ as $p \to \infty$. Then for any $S_{i,t+\tau|t}$ and $i \in \mathsf{N}$, under $\mathcal{H}_{i,0}$ in (3.2)*

$$\lim_{p\to\infty} \mathbb{P}\left[|S_{i,t+\tau|t}| > k_\alpha\right] = \alpha,$$

and under the alternative in (3.3)

$$\lim_{p\to\infty} \mathbb{P}\left[|S_{i,t+\tau|t}| > k_\alpha\right] = 1,$$

where k_α is the critical value of the standard normal distribution at the significance level α.

The test is size-α under the null hypothesis and the behaviour of the test statistic under the alternative hypothesis is easily characterised as the

application of the Kalman filter eliminates the possibility of non-identical distributions. In case of a rejection, the sign of the test statistic indicates which model outperformed the other.

3.2.4.2 *Multivariate Test*

We now discuss two approaches to construct a multivariate version of the test in which the CPA hypotheses are examined jointly for all $n > 1$. If Assumption 3.5 does not hold and one compares two forecasting methods across $n > 1$ variables, it is insufficient to form a conclusion on which method outperformed the other based on the univariate tests outlined above. To avoid Type II errors, one should instead evaluate the n loss differentials jointly.

Assumption 3.6-A is necessary to ensure the parameters of the likelihood function are estimated consistently. If the size of the cross-section is small, maximum likelihood estimation of the parameter vector is still feasible and a multivariate Kalman filter can be used to obtain the cross-sectional vector $\boldsymbol{\mu}_{t+\tau|t}$ with covariance matrix $\mathbf{V}_{t+\tau|t}$. The recursions in (3.16) become

$$\begin{aligned} \boldsymbol{\mu}_{t+\tau|t} &= \boldsymbol{\mu}_{t|t-1} + \frac{p_{t|t-1}}{p_{t|t-1}+1}(\Delta\mathbf{L}_t - \boldsymbol{\mu}_{t|t-1}), \\ \mathbf{V}_{t+\tau|t} &= p_{t+\tau|t}\tilde{\boldsymbol{\Sigma}}_\varepsilon. \end{aligned} \tag{3.18}$$

with $\tilde{\boldsymbol{\Sigma}}_\varepsilon := \sum_{t=r+\tau+1}^{T}(\Delta\mathbf{L}_t - \boldsymbol{\mu}_{t|t-1})(\Delta\mathbf{L}_t - \boldsymbol{\mu}_{t|t-1})'/(p_{t|t-1}+1)$ and $p_{t+\tau|t}$ given by Equation (3.16). In the multivariate case, the parameter vector still corresponds to $\theta := q$ [see (Harvey, 1990, Chapter 8) for details]. Concentrating the covariance matrices $\boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\Sigma}_\eta = q\boldsymbol{\Sigma}_\varepsilon$ out of the likelihood function enables the construction of the test in multivariate settings, since it reduces the number of parameters to be estimated from $n(n+1)$ to 1. Nonetheless, the Kalman filter explicitly models the predicted covariance between the level components of the n loss differentials at each time step, which can be exploited to form a multivariate $n \times 1$ dimensional local test statistic of the form

$$\mathbf{S}_{t+\tau|t} = \mathbf{V}_{t+\tau|t}^{-1/2}\boldsymbol{\mu}_{t+\tau|t} \qquad \text{for } t \in \mathsf{T}, \tag{3.19}$$

where $\mathbf{S}_{t+\tau|t} = (S^M_{1,t+\tau|t'}, ..., S^M_{n,t+\tau|t})'$ and $\mathbf{V}_{t+\tau|t}^{-1/2}$ denotes the matrix square root of $\mathbf{V}_{t+\tau|t}^{-1}$. As in the univariate case, the test statistic standardises the conditional mean of the loss differential. In contrast to the univariate case, it also takes into account the covariance between the different loss differentials.

The two are identical only if $\mathbf{V}_{t+\tau|t}$ is diagonal. Finally, to jointly assess the relative predictive ability of two methods across n loss differentials, we can use

$$\bar{S}_{t+\tau|t} = \boldsymbol{\mu}'_{t+\tau|t} \mathbf{V}^{-1}_{t+\tau|t} \boldsymbol{\mu}_{t+\tau|t} \qquad \text{for } t \in \mathsf{T}. \tag{3.20}$$

Analogous to the univariate case above, to test the null hypothesis in (3.2) and (3.9) when dependence is present, we can define the following Proposition:

Proposition 3.2. *Suppose Assumptions 3.1-3.4 and 3.6 hold such that* $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ *as* $p \to \infty$*. Then for all* $t \in \mathsf{T}$

A. *under the null hypothesis in (3.2),*

$$\lim_{p\to\infty} \mathbb{P}\left[|S^M_{i,t+\tau|t}| > k_\alpha\right] = \alpha,$$

and under the alternative in (3.3),

$$\lim_{p\to\infty} \mathbb{P}\left[|S^M_{i,t+\tau|t}| > k_\alpha\right] = 1,$$

where k_α *is the critical value of a standard normal distribution at the significance level* α*.*

B. *under the null hypothesis in (3.9),*

$$\lim_{p\to\infty} \mathbb{P}\left[\bar{S}_{t+\tau|t} > q_\alpha\right] = \alpha,$$

and under the alternative in (3.10)

$$\lim_{p\to\infty} \mathbb{P}\left[\bar{S}_{t+\tau|t} > q_\alpha\right] = 1,$$

where q_α *is the critical value of a* χ^2_n *distribution with n degrees of freedom at the significance level* α*.*

The elements of $\mathbf{S}_{t+\tau|t}$ allow for inference on whether it is possible to predict that one of the two forecasting methods forming the i-th loss differential will be more accurate than the other. The multivariate null hypothesis (3.9) is rejected by the χ^2_n test of $\bar{S}_{t+\tau|t}$ if one element of $\boldsymbol{\mu}_{t+\tau|t}$ is statistically different from zero – controlling for false discovery and dependence. By conditioning on dependencies when formulating the test, we eliminate the need to correct for size distortions due to correlations between variables

when evaluating the test results.[5] In order to make the test statistic interpretable, we can multiply it by the sign of $\frac{1}{n}\sum_{i=1}^{n}\mu_{i,t+\tau|t}$, which indicates the forecasting method in favour of which the test rejects.

3.2.5 *Pointwise Total Predictive Ability Test*

3.2.5.1 *Univariate Test*

We now introduce the pointwise Total Predictive Ability test. Under Assumptions 3.1-3.3, the expectation of the loss differential conditional on $\mathcal{F}_T$ corresponds to

$$\mathbb{E}[\Delta L_{i,t+\tau} \mid \mathcal{F}_T] = \mathbb{E}[\alpha_{t+\tau} \mid \mathcal{F}_T] = \mu_{i,t+\tau|T}. \tag{3.21}$$

The conditional mean $\mu_{i,t+\tau|T}$ can again be obtained recursively as

$$\begin{aligned} \mu_{i,t-1|T} &= \mu_{i,t|t-1} + J_{i,t-1}\left(\mu_{i,t|T} - \mu_{i,t|t-1}\right), \\ v_{i,t-1|T} &= J_{i,t-1}v_{i,t|t-1} + J_{i,t-1}^2\left(v_{i,t|T} - v_{i,t|t-1}\right), \end{aligned} \tag{3.22}$$

for $t = r+\tau, ..., T$, with $J_{i,t} = \left(1 - \frac{p_{i,t|t-1}}{p_{i,t|t-1}+1}\right)\frac{p_{i,t|t-1}}{p_{i,t+1|t}}$ and initial conditions $\mu_{i,T|T} = \mu_{i,T+1|T}$ and $v_{i,T|T} = \left(1 - \frac{p_{i,t|t-1}}{p_{i,t|t-1}+1}\right)v_{i,T|T-1}$ given by the recursions in (3.16). That is, the conditional mean can be computed with a Kalman smoother using the QML estimate $\tilde{q}$. The difference between the expectations $\mu_{i,t+\tau|t}$ used in the CPA test and $\mu_{i,t+\tau|T}$ is that the former predicts the conditional mean using only information available up to point t, while the latter recovers the mean of the loss differential conditional on all previous and future observations. To test the local null hypothesis in (3.5), we define the following $n \times 1$ dimensional vector of test statistics for pointwise total predictive ability:

$$S_{i,t+\tau|T} = \frac{\mu_{i,t+\tau|T}}{\sqrt{v_{i,t+\tau|T}}} \qquad \text{for any } i \in \mathsf{N}, \tag{3.23}$$

for all $t = r, ..., T-\tau$. Under the null hypothesis in (3.5), the size and power of the test can be summarised as

[5] The consequences of falsely imposing the independence assumption when comparing multiple individual tests have been studied at length in Spreng and Urga (2022) and are not the focus of this paper.

Proposition 3.3. *Suppose Assumptions 3.1-3.5 hold such that* $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ *as* $p \to \infty$. *Then for any* $S_{i,t+\tau|T}$ *and* $i \in \mathsf{N}$, *under* $\mathcal{H}^*_{i,0}$ *in (3.5),*

$$\lim_{p\to\infty} \mathbb{P}\left[|S_{i,t+\tau|T}| > k_\alpha\right] = \alpha,$$

and under the alternative $\mathcal{H}^*_{i,A}$,

$$\lim_{p\to\infty} \mathbb{P}\left[|S_{i,t+\tau|T}| > k_\alpha\right] = 1,$$

where k_α *is the critical value of the standard normal distribution at the significance level* α.

3.2.5.2 *Multivariate Test*

In the multivariate case, the vector of conditional means and its covariance matrix is obtained as

$$\begin{aligned}
\boldsymbol{\mu}_{t-1|T} &= \boldsymbol{\mu}_{t|t-1} + J_{t-1}\left(\boldsymbol{\mu}_{t|T} - \boldsymbol{\mu}_{t|t-1}\right), \\
\mathbf{V}_{t-1|T} &= J_{t-1}\mathbf{V}_{t|t-1} + J^2_{t-1}\left(\mathbf{V}_{t|T} - \mathbf{V}_{t|t-1}\right),
\end{aligned}$$

As can be seen, $\mathbf{V}_{t|T}$ not only depends on all past and future realisations of its diagonal elements but also on past and future cross-correlations between measurements. This allows us to formulate a test statistic for the null hypothesis in (3.13), which says that no method could have outperformed the other at time t:

$$\mathbf{S}_{t+\tau|T} = \mathbf{V}^{-1/2}_{t+\tau|T}\boldsymbol{\mu}_{t+\tau|T}. \tag{3.24}$$

where $\mathbf{S}_{t+\tau|T} = (S^M_{1,t+\tau|T}, ..., S^M_{n,t+\tau|T})'$. Additionally, we can compute the following joint test statistic

$$\bar{S}_{t+\tau|T} = \boldsymbol{\mu}'_{t+\tau|T}\mathbf{V}^{-1}_{t+\tau|T}\boldsymbol{\mu}_{t+\tau|T} \qquad \text{for } t \in \mathsf{T}. \tag{3.25}$$

Proposition 3.4 demonstrates the behaviour of the test under the null and the alternative hypothesis.

Proposition 3.4. *Suppose Assumptions 3.1-3.4 and 3.6 hold such that* $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ *as* $p \to \infty$. *Then for all* $t \in \mathsf{T}$

A. *under the null hypothesis in (3.5), for any* $S^M_{i,t+\tau|T} \in \mathbf{S}_{t+\tau|T}$,

$$\lim_{p\to\infty} \mathbb{P}\left[|S^M_{i,t+\tau|T}| > k_\alpha\right] = \alpha,$$

and under the alternative in (3.6)

$$\lim_{p\to\infty} \mathbb{P}\left[|S^M_{i,t+\tau|T}| > k_\alpha\right] = 1,$$

where k_α *is the critical value of a standard normal distribution at the significance level* α.

B. *under the null hypothesis in (3.12),*

$$\lim_{p\to\infty} \mathbb{P}\left[\bar{S}_{t+\tau|T} > q_\alpha\right] = \alpha,$$

and under the alternative in (3.13)

$$\lim_{p\to\infty} \mathbb{P}\left[\bar{S}_{t+\tau|T} > q_\alpha\right] = 1,$$

where q_α *is the critical value of a* χ^2_n *distribution with* n *degrees of freedom at the significance level* α.

3.3 MONTE-CARLO SIMULATION

To analyse the size and power properties of the tests, we simulate 3 different scenarios. All results are obtained for 5000 Monte-Carlo iterations. The models are parameterised differently for size and power simulations. In this section, we conduct the simulations using Gaussian error terms. To address the issue of non-normality, we repeat all simulations presented herein using a t-distribution with 5 degrees of freedom in Appendix C.2.2. The first simulation generates forecast loss differentials as a process with a time-varying mean.

$$\Delta\mathbf{L}_t = \mathbf{c}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{MC1}$$

where $\Delta\mathbf{L}_t = (\Delta L_{1,t}, ..., \Delta L_{n,t})'$ is an $n \times 1$ vector and $\boldsymbol{\Sigma}$ is an $n \times n$ matrix that controls the dependence between loss differentials. To generate $\boldsymbol{\Sigma}$, we rely on the vine method proposed in Lewandowski et al. (2009). They propose a way to generate a correlation matrix $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}(\gamma)$, whose off-diagonal elements are decreasing in $\gamma \in \mathbb{R}^+$, such that $\boldsymbol{\Sigma} \to \mathbf{I}_n$ as $\gamma \to \infty$. That is, low values

of γ imply high dependence between variables. Next, we simulate an actual forecasting scenario in which the forecast variables, $\mathbf{y}_t$, have time-varying means

$$\mathbf{y}_t = \mathbf{c}_t + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{MC2}$$

The elements of $\mathbf{y}_t$ are forecast individually, with the two competing i-th forecasts corresponding to,

$$\hat{y}^{(1)}_{i,t+1} = \hat{c}_{i,t},$$
$$\hat{y}^{(2)}_{i,t+1} = 0.$$

The forecasts are estimated over a rolling window of length r. That is, $\hat{c}_{i,t} = \sum_{j=t-R+1}^{t} y_{i,j}$. One forecast is miss-specified, while the other assumes the correct model. In the fourth simulation, we assume the forecast variables obey a factor structure

$$\mathbf{y}_t = \boldsymbol{\beta}_t x_t + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{MC3}$$

where the common factor $x_t = 0.5x_t + \nu_t$ follows an AR(1) process with Gaussian errors, and $\boldsymbol{\beta}_t$ is an $n \times 1$ vector of time-varying factor-loadings. To forecast each $y_{i,t}$, we assume x_{t+1} is observed at time $t+1$ and use a rolling window OLS estimator to compute

$$\hat{y}^{(1)}_{i,t+1} = \hat{\beta}_{i,t} x_{t+1},$$
$$\hat{y}^{(2)}_{i,t+1} = 0.$$

As before, the first model is correctly specified, while the second one is miss-specified.

3.3.1 *Size*

To analyse the size of our test statistics, we need to ensure that the expectation of the loss differential is zero at each point in time. We conduct simulations for $n = \{1, 2, 5\}$ and, in the multivariate case, $\gamma = \{1, 10^8\}$. The case of $\gamma = 1$ resembles strong dependence between variables, and $\gamma = 10^8$ independence. In our first simulation, we simply set $c_{i,t} = 0$ for all t and i. For the actual forecasting scenarios, to satisfy the null hypothesis, we solve $E\left[(y_{i,t+1} - \hat{y}^{(1)}_{i,t+1})^2\right] = E\left[(y_{i,t+1} - \hat{y}^{(2)}_{i,t+1})^2\right]$ for all $t = r, ..., T-1$. In the case

Table 3.1: Average Size

		MC 1	MC 2				MC 3			
	p/r:	0	50	100	150	200	50	100	150	200
Panel A: $n = 1$										
$S_{t+1\|t}$	50	0.027	0.03	0.033	0.034	0.035	0.03	0.032	0.032	0.034
	100	0.031	0.027	0.033	0.031	0.035	0.027	0.031	0.031	0.032
	150	0.032	0.024	0.03	0.03	0.031	0.022	0.03	0.031	0.032
	200	0.03	0.022	0.027	0.029	0.03	0.024	0.027	0.029	0.031
$S_{t+1\|T}$	50	0.05	0.047	0.049	0.051	0.051	0.042	0.045	0.044	0.046
	100	0.049	0.043	0.047	0.044	0.048	0.037	0.041	0.042	0.04
	150	0.052	0.035	0.046	0.044	0.042	0.03	0.04	0.041	0.04
	200	0.044	0.03	0.039	0.043	0.039	0.03	0.037	0.038	0.039
Panel B: $n = 2, \gamma = 1$										
$S_{t+1\|t}$	50	0.028	0.029	0.038	0.037	0.036	0.036	0.044	0.041	0.039
	100	0.03	0.025	0.032	0.036	0.036	0.028	0.035	0.041	0.041
	150	0.027	0.023	0.028	0.032	0.032	0.028	0.034	0.036	0.04
	200	0.027	0.022	0.026	0.028	0.032	0.023	0.027	0.035	0.036
$S_{t+1\|T}$	50	0.049	0.046	0.053	0.053	0.051	0.051	0.064	0.054	0.052
	100	0.05	0.036	0.046	0.046	0.044	0.042	0.048	0.053	0.049
	150	0.041	0.034	0.036	0.044	0.04	0.036	0.048	0.047	0.048
	200	0.04	0.028	0.035	0.037	0.039	0.029	0.035	0.049	0.048
Panel C: $n = 2, \gamma \to \infty$										
$S_{t+1\|t}$	50	0.034	0.033	0.036	0.04	0.038	0.035	0.042	0.043	0.045
	100	0.025	0.03	0.029	0.036	0.04	0.031	0.032	0.032	0.042
	150	0.027	0.027	0.023	0.033	0.032	0.023	0.029	0.033	0.034
	200	0.027	0.02	0.028	0.032	0.03	0.028	0.034	0.035	0.037
$S_{t+1\|T}$	50	0.06	0.05	0.048	0.05	0.054	0.054	0.061	0.061	0.053
	100	0.042	0.044	0.032	0.049	0.05	0.037	0.038	0.035	0.05
	150	0.044	0.038	0.03	0.045	0.035	0.033	0.039	0.039	0.037
	200	0.036	0.03	0.042	0.039	0.039	0.039	0.051	0.046	0.04
Panel D: $n = 5, \gamma = 1$										
$S_{t+1\|t}$	50	0.036	0.038	0.05	0.049	0.051	0.046	0.061	0.059	0.053
	100	0.03	0.025	0.033	0.042	0.047	0.031	0.043	0.056	0.052
	150	0.028	0.021	0.028	0.033	0.041	0.027	0.036	0.044	0.048
	200	0.028	0.018	0.025	0.032	0.034	0.024	0.032	0.038	0.046
$S_{t+1\|T}$	50	0.069	0.064	0.072	0.069	0.073	0.067	0.083	0.083	0.065
	100	0.05	0.036	0.046	0.05	0.053	0.04	0.056	0.069	0.061
	150	0.041	0.029	0.038	0.037	0.043	0.032	0.046	0.054	0.061
	200	0.041	0.022	0.036	0.038	0.035	0.028	0.042	0.046	0.053
Panel E: $n = 5, \gamma \to \infty$										
$S_{t+1\|t}$	50	0.036	0.037	0.052	0.049	0.052	0.034	0.045	0.047	0.045
	100	0.03	0.028	0.032	0.042	0.043	0.025	0.034	0.041	0.044
	150	0.03	0.02	0.031	0.031	0.04	0.02	0.027	0.035	0.038
	200	0.03	0.019	0.026	0.031	0.037	0.019	0.025	0.03	0.034
$S_{t+1\|T}$	50	0.067	0.063	0.079	0.07	0.077	0.041	0.056	0.058	0.05
	100	0.049	0.04	0.046	0.049	0.054	0.028	0.041	0.043	0.041
	150	0.043	0.029	0.038	0.037	0.041	0.025	0.032	0.035	0.036
	200	0.048	0.028	0.033	0.039	0.043	0.023	0.03	0.029	0.032

Notes: The table reports the average size of the pointwise CPA test ($S_{t+1|t}$) and the pointwise TPA test ($S_{t+1|T}$). The column headers denote the size for the respective Monte-Carlo Simulation (1, 2, and 3). The rows correspond to the size for different values of p and the columns to the size for different values of r. Panel A reports the simulation results for $n = 1$, Panel B reports the results for the multivariate χ^2 test with two loss differentials ($n = 2$) and strong dependence ($\gamma = 1$). Panel C reports the results for $n = 2$ and independence between loss differentials ($\gamma = 10^8$), Panel D for $n = 5$ variables and strong dependence, and Panel E for $n = 5$ independent variables.

of the second simulation, the parameters that satisfy this condition are given by

$$c_{i,t+1} = \frac{\frac{1}{r}\left(\sum_{j=t-r+1}^{t} c_{i,j}\right)^2 + \sigma_{i,i}^2}{2\sum_{j=t-r+1}^{t} c_{i,j}}, \qquad \text{for all } i \in \mathsf{N},$$

where $\sigma_{i,i}^2$ is the (i,i)-th element of $\mathbf{\Sigma}$. For $t = 1, ..., r$ we draw $c_{i,t}$ from a normal distribution, $c_{i,t} \sim \mathcal{N}(0, 0.2)$. Moving to the third simulation, the parameter values for $t = r, ..., T-1$ are obtained as:

$$\beta_{i,t+1} = \frac{\frac{\left(\sum_{j=t-r+1}^{t} \beta_{i,j} x_j^2\right)^2}{\sum_{j=t-r+1}^{t} x_j^2} + \sigma_{i,i}^2}{2\sum_{j=t-r+1}^{t} \beta_{i,j} x_j^2} \tag{3.26}$$

This equation is derived in Giacomini and Rossi (2010). The difference between their simulation and ours is that we let $\beta_t \sim \mathcal{N}(0, 0.2)$ for $t = 1, ..., r$.

We conduct all simulations for $p = \{50, 100, 150, 200\}$ and a nominal size of $\alpha = 5\%$. In simulations 2 and 3, we focus on one-step-ahead forecasts and set the estimation window to $r = \{50, 100, 150, 200\}$. Table 3.1 reports the results of the size simulations, averaged across p. The column headers denote the nominal size for the respective Monte-Carlo Simulation. The rows correspond to the empirical size for different values of p and the columns to the empirical size for different values of r. Panel A reports the simulation results for $n = 1$. In most cases the size is no more than two percentage points below the nominal level, even for a small estimation and prediction window. For the second and third simulation, the size increases the larger the estimation window is. The size of the TPA test is closer to the nominal size across all simulations. Panel B reports the results for the multivariate χ^2 test with two loss differentials and strong dependence. As before, the size of the TPA test is closer to the nominal size, and neither test is oversized nor are they markedly undersized. Panel C reports the results for two variables under independence. There are no discernable differences between the simulations, suggesting the test maintains its size properties regardless of the degree of dependence between variables. Panel D shows the size for $n = 5$ variables and strong dependence, and Panel E for $n = 5$ independent variables. The TPA test is now oversized when the forecasting window is small, due to the increasing n/p ratio. The CPA test, however, still displays remarkably good size properties, regardless of window size. The size simulations indicate that, on average, both tests have a low Type I error even in small samples, provided the number of variables is not

too large. The averages reported in Table 3.1 may obscure some temporal differences in the size. Hence, we also plot the size of the TPA and CPA statistics against one another. Figure 3.1 displays these scatter plots together with an overlayed contour plot, to indicate how clustered the size is. The rows correspond to the respective Monte-Carlo Simulation. Panel (i) contains the simulation results for the univariate test, Panel (ii) the results for $n = 2$ variables and Panel (iii) the results for $n = 5$ variables. The rows within the panels correspond to the degree of dependence between variables. In the first simulation, the size is distributed very densely around the nominal level of 5%, regardless of the number of variables and the degree of dependence. This changes somewhat for the second and third simulation in which the size is more scattered, although it is still just under the 5% mark in most instances.

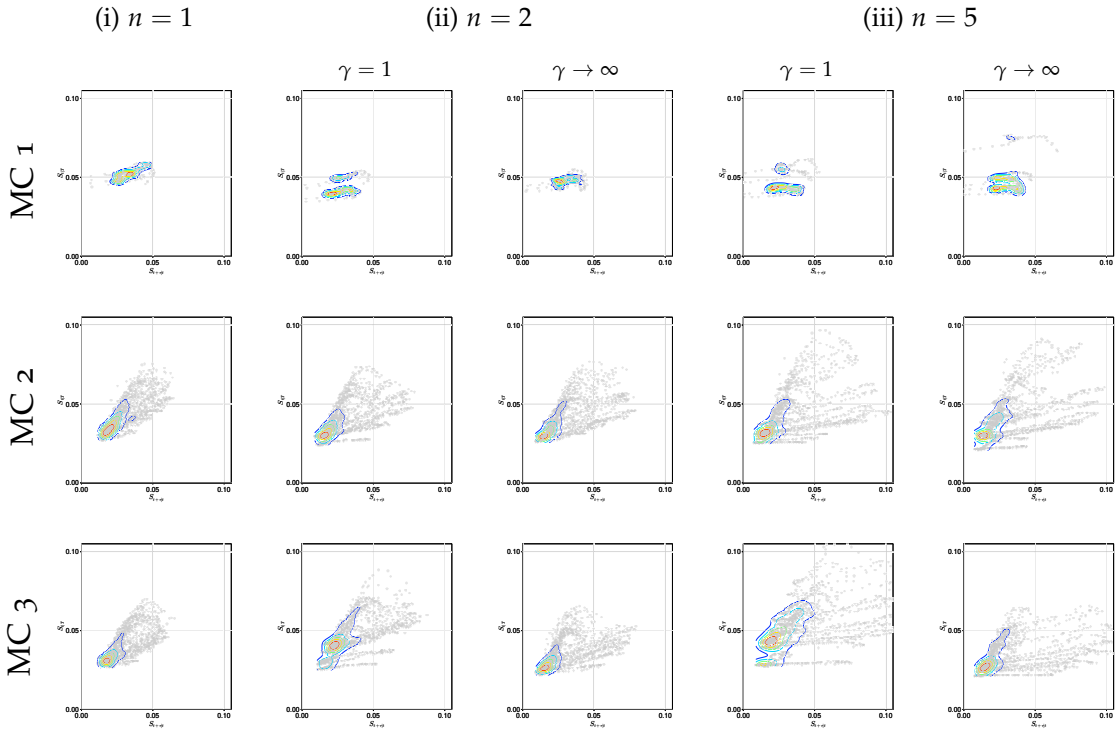


Figure 3.1: Size Contours

Note: The figure plots the size of the TPA test on the vertical axis against the size of the CPA test on the horizontal axis and displays a contours that encircles predominant clusters. The rows correspond to the respective Monte-Carlo Simulation. Panel (i) contains the simulation results for the univariate test, Panel (ii) the results for $n = 2$ variables and Panel (iii) the results for $n = 5$ variables. The columns in Panel (ii) and (iii) correspond to high dependence ($\gamma = 1$) and independence ($\gamma \to \infty$).

3.3.2 *Power Properties*

To examine the power properties of the test, we parameterise the equations for simulations 1 to 3 in a way that creates different forecasting scenarios. In the first simulation, we simply set $c_t = c$ where $c = \{0, 0.1, ..., 1\}$ is increased from 0 to 1 in steps of 0.1. For the second simulation, we draw $\mathbf{c}_t$ from $\mathcal{N}(c, \Sigma)$, where $c = \{0, 0.1, ..., 1\}$ is increased from 0 to 1. Thereby, we gradually widen the accuracy gap between the two forecasts and introduce dependence into the level component of the forecast variable. In the third simulation, we draw each factor loading $\beta_{i,t}$ from $\mathcal{N}(2c, 0.2)$, where $c = \{0, 0.1, ..., 1\}$ is increased from 0 to 1, which increases the dependence between variables and the difference between forecasting models. As above, we conduct each simulation for $n = \{1, 2, 5\}$ and dependence $\gamma = \{1.0, 10^8\}$. We set the out-of-sample window to $p = 100$ and the estimation window to $r = 100$. We first present the results for $n = 1$, which allows us to contrast our tests with the pointwise rejection accuracy of the Giacomini and Rossi (2010) fluctuation test. We emphasise, however, that, unlike our test, the fluctuation test is consistent across all p, not for each p. Therefore, the test size at each time step is zero. Figure 3.2 displays surface plots of the power of each test. The rows of the figure correspond to the respective Monte Carlo simulation and the columns to the respective test. For each Subfigure, the x-axis corresponds to the time periods, the y-axis to the different values of c and the z-axis to the power of the tests. In the first simulation, the power of the pointwise CPA test is steeply upwards-sloping. During the first few periods, the rejection frequency is lower, reflecting the fact that one cannot predict from the start with high accuracy which method will yield better forecasts. Conversely, the pointwise TPA test has high power also for the first periods, as it conditions on $\mathcal{F}_T$. The power curve of the fluctuation test is less steep, but converges to one as c increases. This changes in the second and third simulation, where the GR test has visibly lower pointwise power of up to 18% and 65%, respectively. As before, the power surface of the TPA test increases at the same rate across all p, unlike the CPA test, which has lower power at the start of the forecasting period – particularly for the second simulation. Next, we simulate the power of the test using $n = 2$ forecast variables, for which the Giacomini and Rossi (2010) fluctuation test is no longer feasible. The results are reported in Figure 3.3, in which Panel (i) contains the power surfaces of the CPA test, and Panel (ii) the power surfaces of the TPA test. The first column of each Panel corresponds to the high-dependence scenario and the second panel to the low-dependence case. Overall, the results are very similar to the univariate case, although

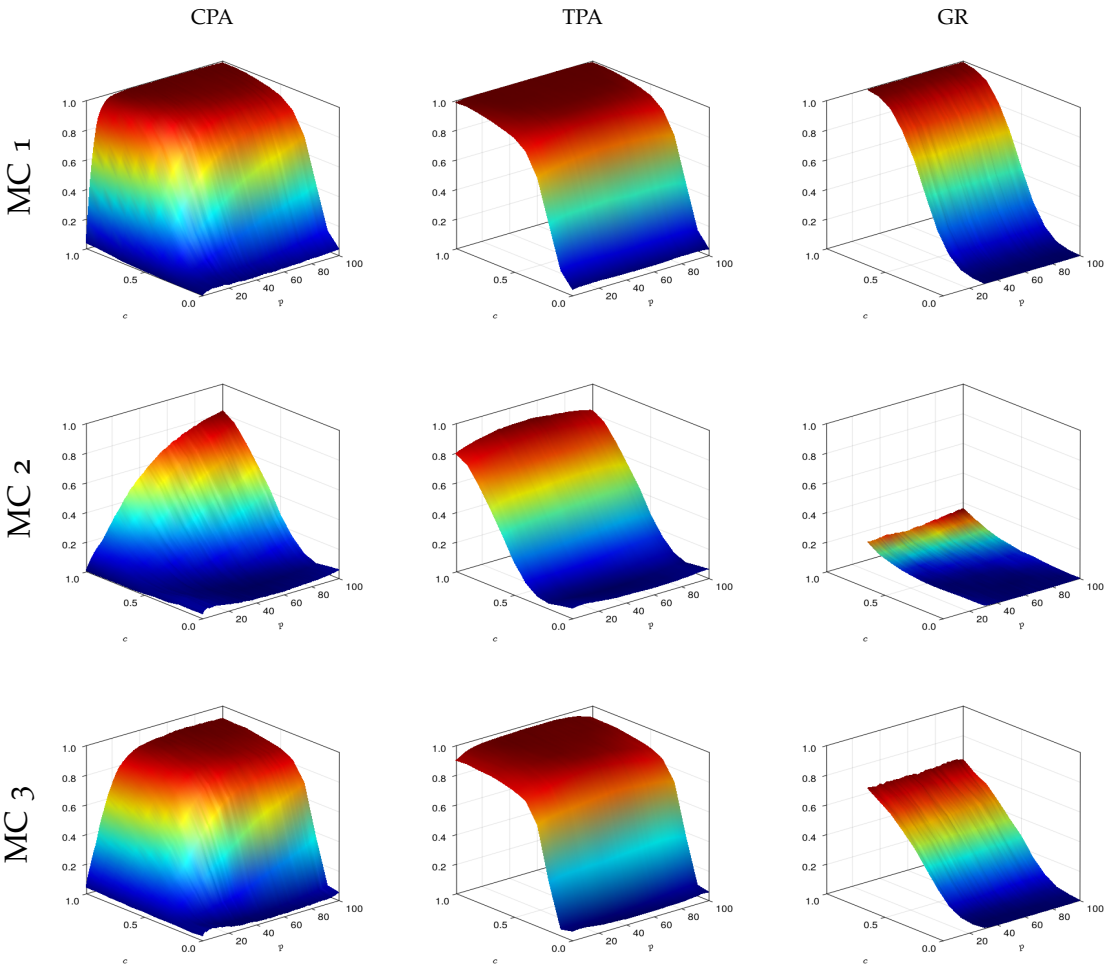


Figure 3.2: Power Surface, $n = 1$

Note: The columns of the figure display the power surface of the CPA and TPA tests as well as the power of the Giacomini and Rossi (2010) fluctuation test at each point in time for $n = 1$ variable. The rows correspond to the respective Monte-Carlo Simulation. For each figure, the x-axis corresponds to the time periods, the y-axis to the different values of c, and the z-axis to the power of the tests.

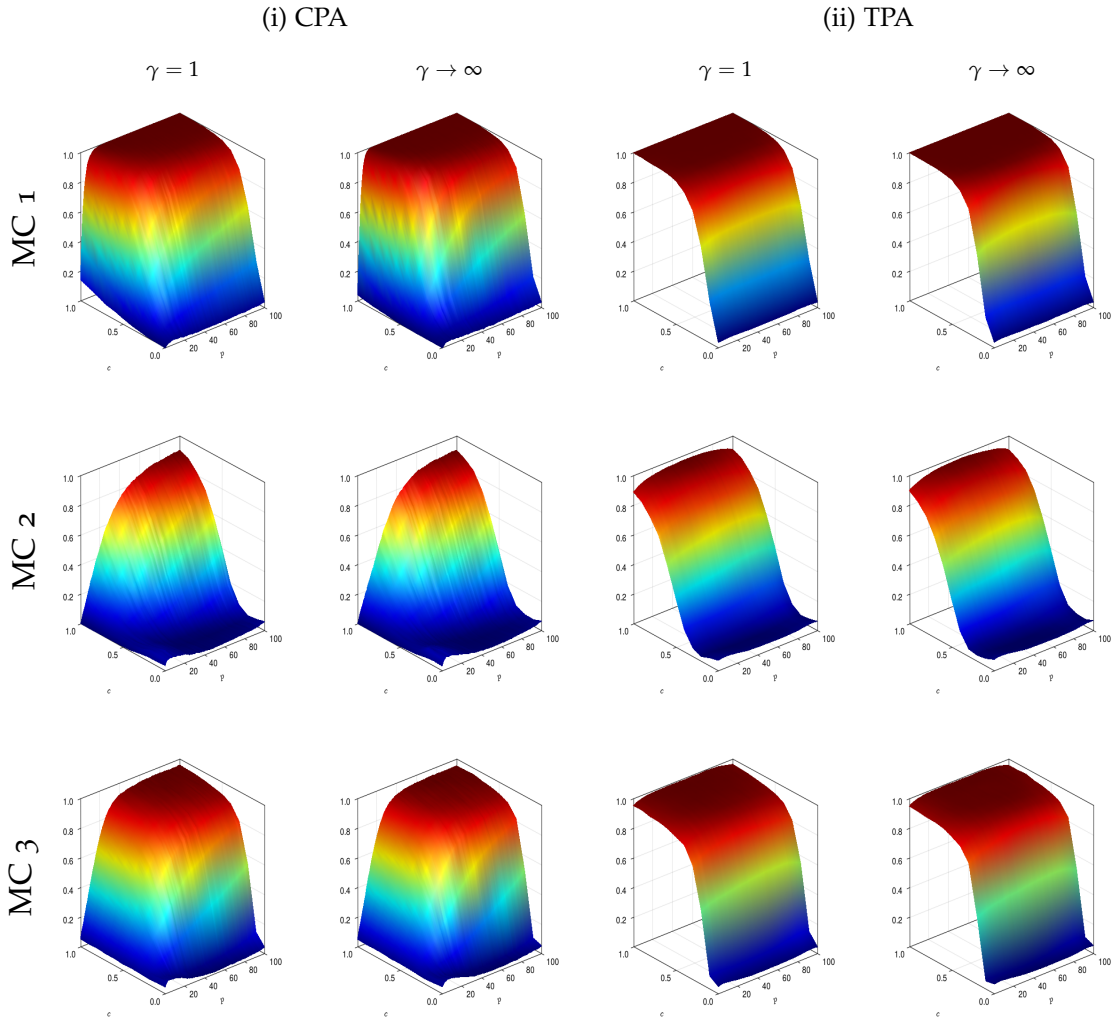


Figure 3.3: Power Surface, $n = 2$

Note: The columns of the figure display the power surface of the CPA test in Panel (i), and the power of the TPA test in Panel (ii) for $n = 2$ variables. In each Panel, the first column contains the high-dependence case and the second column the low-dependence case. The rows correspond to the respective Monte-Carlo Simulation. For each figure, the x-axis corresponds to the time periods, the y-axis to the different values of c and the z-axis to the power of the tests.

the power curves of both tests are slightly steeper and closer to one in the second and third scenario. There are virtually no differences between the different dependence cases, which is reassuring given that the simulations are otherwise identically. The results therefore illustrate that introducing dependence does not distort the results of the test.

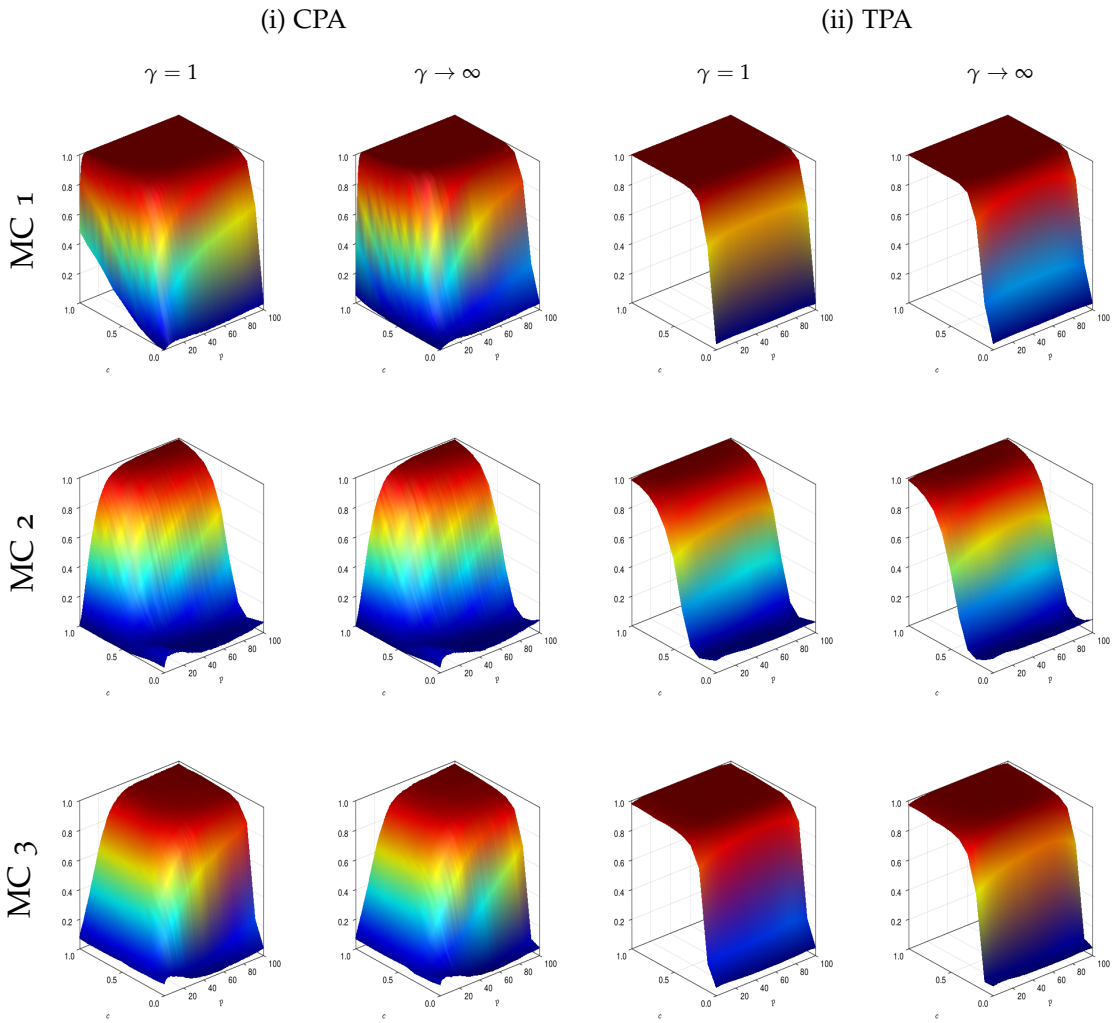


Figure 3.4: Power Surface, $n = 5$

Note: The columns of the figure display the power surface of the CPA test in Panel (i), and the power of the TPA test in Panel (ii) for $n = 5$ variables. In each Panel, the first column contains the high-dependence case and the second column the low-dependence case. The rows correspond to the respective Monte-Carlo Simulation. For each figure, the x-axis corresponds to the time periods, the y-axis to the different values of c and the z-axis to the power of the tests.

Finally, we set the number of variables to $n = 5$ (see Figure 3.4). This leads to a further increase in the power of the tests, most visibly in the second

simulation. Compared to the $n = 1$ variable case, the TPA test also displays markedly higher power in the first and third simulation. Generally, there is no difference between the two dependence scenarios, with the exception of the first simulation, in which the power of the CPA test is somewhat higher for low values of p.

In Appendix C.2.2, we repeat these simulations using a t-distribution with 5 degrees of freedom. The results indicate that the change in distribution has little effect on the power of our tests. In fact, the power of the fluctuation test is noticeably more diminished.

3.4 EMPIRICAL APPLICATION

We illustrate the performance of our test through a comparison of intraday volatility forecasts. In a seminal paper, Engle and Sokalska (2012) note that, although a large body of literature covers volatility forecasts at daily or lower frequencies, few papers address the issue of intraday forecasting. However, the latter is of high practical relevance when it comes to order submission strategies or the computation of intraday risk measures. Since Engle and Sokalska (2012), several papers have (i) proposed new models for intraday volatility (e.g. Rossi and Fantazzini, 2015; Stroud and Johannes, 2014), (ii) compared the predictive ability of such models (e.g. Khashanah and Shao, 2022), or (iii) introduced tests for the presence of periodicity in intraday volatility (e.g. Andersen et al., 2019). Our empirical application connects these lines of research by evaluating the time-varying predictive ability of two established volatility models. Predictive ability tests that evaluate out-of-sample performance across forecasting periods will fail to capture important intraday performance variation. Additionally, if a test is based on an average measure such as the Mean Squared Error (MSE) its rejection may be outlier driven. Our method can provide crucial insights into breaks in intraday forecasting performance. For example, if the aim is to forecast volatility to determine optimal end-of-day order submissions it is important to ascertain when precise breaks in predictive ability occur.

3.4.1 *Data*

We obtain 1,050,000 tick-by-tick observations on NASDAQ 100 index values from Refinitiv Eikon between 31/10/2022 and 28/11/2022. We do not apply any further cleaning procedures to the data. As the purpose of this exercise

is to forecast volatility, we instead augment our volatility models with an ARMA(1, 1) process to model microstructure noise directly, rather than removing it.[6]

Each day in our sample is denoted by $i = 1, ..., n$ and contains $t = 1, ..., T$ minutes and each minute contains an irregularly spaced number of ticks $h = 1, ..., \varsigma$. Let $p_{i,t,h}$ denote the log price of an asset. We define the continuously compounded minutely log-return of the asset as

$$r_{i,t} = \begin{cases} p_{i,t,\varsigma} - p_{i,t-1,\varsigma} & \text{if } t > 1 \\ p_{1,t,\varsigma} - p_{i-1,T,\varsigma} & \text{if } t = 1 \text{ and } i > 1 \\ p_{1,1,\varsigma} - p_{1,1,1} & \text{if } t = 1 \text{ and } i = 1 \end{cases}$$

Note that this definition includes overnight returns, which are generally of larger magnitude than the remaining $T - 1$ returns. Many papers intentionally drop overnight returns (e.g. Engle and Sokalska, 2012) as they, and their induced volatility, are more challenging to forecast. Instead of removing them, we compare the performance of the models with and without overnight returns. At high frequencies, the measurement of the true unobserved volatility becomes problematic as realised volatility is likely biased (Hansen and Lunde, 2006). Following Engle and Sokalska (2012) and Khashanah and Shao (2022), we use squared one-minute returns as a volatility proxy.

3.4.2 *Volatility Models*

We employ two established volatility models: the Markov Switching Multifractal (MSM) model of Calvet and Fisher (2001, 2002, 2004) and a GARCH(1, 1) model. The ARMA-MSM model takes the following form

$$\begin{aligned} r_{i,t} &= \alpha_i + \beta_i r_{i,t-1} + \theta_i \epsilon_{i,t-1} + \epsilon_{i,t} \\ \epsilon_{i,t} &= \sigma_{i,t} \eta_{i,t} \\ \sigma_{i,t}^2 &= \sigma_i^2 \prod_{k=1}^{\bar{k}} M_{k,i,t}. \end{aligned} \tag{3.27}$$

The $M_{k,i,t}$ are latent state variables that must be computed recursively through Bayesian updating. We can stack the state variables in a volatility

[6] It is common practice in many papers to purge raw data of microstructure noise by defining continuously compounded returns as the residuals of an MA(1) process.

state vector $M_{i,t} = (M_{1,i,t}, ..., M_{\bar{k},i,t})$. The minute-$t$ multiplier $M_{k,i,t}$ is drawn from a fixed distribution M with probability $\gamma_{i,k}$, and remains at its previous value with probability $1 - \gamma_{i,k}$: $M_{k,i,t} = M_{k,i,t-1}$. M is a binomial random variable that takes the values $m_{i,0}$ and $2 - m_{i,0}$, each with probability 0.5, and $m_{i,0} \in (0, 2]$. The marginal distribution of all $M_{k,i,t}$ is the same, their transition probabilities are different, however, and given by

$$\gamma_{i,k} = 1 - (1 - \gamma_{i,1})^{b_i^{k-1}}, \tag{3.28}$$

where $\gamma_{i,1} \in (0, 1)$ and $b_i \in (1, \infty)$. At each day, we estimate the parameter vector $\psi_i = (\alpha_i, \beta_i, \theta_i, m_{i,0}, \sigma_i, b_i, \gamma_{i,1})'$ via maximum likelihood (Calvet and Fisher, 2004). The volatility state vector $M_{i,t}$ takes $d = 2^{\bar{k}}$ possible values. The transition probabilities of the Markov chain $M_{i,t}$ are given by the matrix $A_i = (a_{qj})_{1 \leq q,j \leq d}$ with components $a_{qj} = \mathbb{P}[M_{i,t+1} = m_i^j | M_{i,t} = m_i^q]$. Although $M_{i,t}$ is latent, we can still obtain transition probabilities $\Pi_{i,t}^j = \mathbb{P}[M_{i,t} = m_i^j | x_1, ..., x_T]$ over the unobserved states. The conditional probability vector $\Pi_{i,t} = (\Pi_{i,t}^1, ..., \Pi_{i,t}^d)$ can be computed recursively as:

$$\Pi_{i,t} = \frac{\omega(\epsilon_{i,t}) \odot \Pi_{i,t-1} A_i}{(\omega(\epsilon_{i,t}) \odot \Pi_{i,t-1} A_i) \iota'},$$

where ι is a d-dimensional vector of ones, $\odot$ is the Hadamard product and $\omega(\epsilon_{i,t})$ is the conditional density of $\eta_{i,t}$. We re-estimate the models on each day and use the resulting parameters to forecast volatility throughout the next day:

$$\hat{\sigma}^2_{i+1|i,t+\tau} = \sum_{j=1}^{d} \sigma^2_{i+1|i,t} \left(m^j_{i+1|i} \right) \left(\Pi_{i+1|i,t} A^{\tau}_{i+1|i} \right)_j$$

Note that the subscript $i + 1|i$ reflects the fact that parameters used in the forecast estimation at day $i + 1$ are estimated using day-i data. It is different from the conditional expectations formed in the forecast evaluation step. The volatility function of the ARMA-GARCH(1, 1) takes the following form:

$$\sigma^2_{i,t} = \omega_i + \phi_i \sigma^2_{i,t-1} + \kappa_i \epsilon^2_{i,t-1}. \tag{3.29}$$

and its τ-minutes-ahead forecasts on day $i+1$, based on parameter estimates of day i, are generates as:

$$\begin{aligned}\tilde{\sigma}^2_{i+1|i,t+\tau} = \hat{\omega}_{i|i}\left[1 + (\hat{\kappa}_{i|i} + \hat{\phi}_{i|i}) + \cdots + (\hat{\kappa}_{i|i} + \hat{\phi}_{i|i})^{\tau-1}\right] \\ + (\hat{\kappa}_{i|i} + \hat{\phi}_{i|i})^{\tau}\left[\frac{\hat{\omega}_{i|i}}{1-\hat{\phi}_{i|i}} + \hat{\kappa}_{i|i}\sum_{j=1}^{\infty}\hat{\phi}^{j-1}_{i|i}\epsilon^2_{i,t-j}\right].\end{aligned} \tag{3.30}$$

We use the following loss differential

$$\Delta L_{i+1,t+\tau} = \left(r^2_{i+1,t+\tau} - \hat{\sigma}^2_{i+1|i,t+\tau}\right)^2 - \left(r^2_{i+1,t+\tau} - \tilde{\sigma}^2_{i+1|i,t+\tau}\right)^2, \tag{3.31}$$

which is evaluated conditional on the time-t and time-T σ-field of each day, respectively. As the market closes at 16:00, we can treat trading days as the cross-sectional dimension, and minutes-of-the-day as the time dimension in our evaluation. Hence, the vector of loss-differentials used in the multivariate evaluation is $\Delta\mathbf{L}_{t+\tau} = (\Delta L_{2,t+\tau}, ..., \Delta L_{n,t+\tau})'$. *Ex-ante*, one would expect the loss differentials across different days to be highly dependent – if one model outperforms the other during the market opening hours on day $t = (2, .., n-1)$, it is also more likely to deliver better predictions on the n-th day. Therefore, the results of the multivariate test can differ from those of the univariate test.

3.4.3 *Empirical Results*

We focus on the case of $\tau = 1$. Figure 3.5 presents the intraday test statistics on the 20 days in the sample between 9:30 and 16:00, excluding overnight returns. Subfigure (a) shows the univariate CPA tests and Subfigure (b) the univariate TPA tests. The shaded regions contain the individual test statistics, and the red lines correspond to the critical values. If the test statistic falls below the negative critical value, the test rejects in favour of the MSM model, and, *vice-versa*, in favour of the GARCH model if it exceeds the positive critical value. In both cases, there are several periods during which the test rejects, predominantly clustered at the start of the trading day. The tests offer no decisive indication as to which method has had, or will have, better predictive ability. After the first trading hour, there is a lower overall number of rejections – favouring the GARCH model on some days and the MSM model on others. In contrast, the joint multivariate evaluation is more conclusive, as shown in Subfigure (c) and (d). Looking first at the CPA test, frequent rejections in favour of either model during the first trading

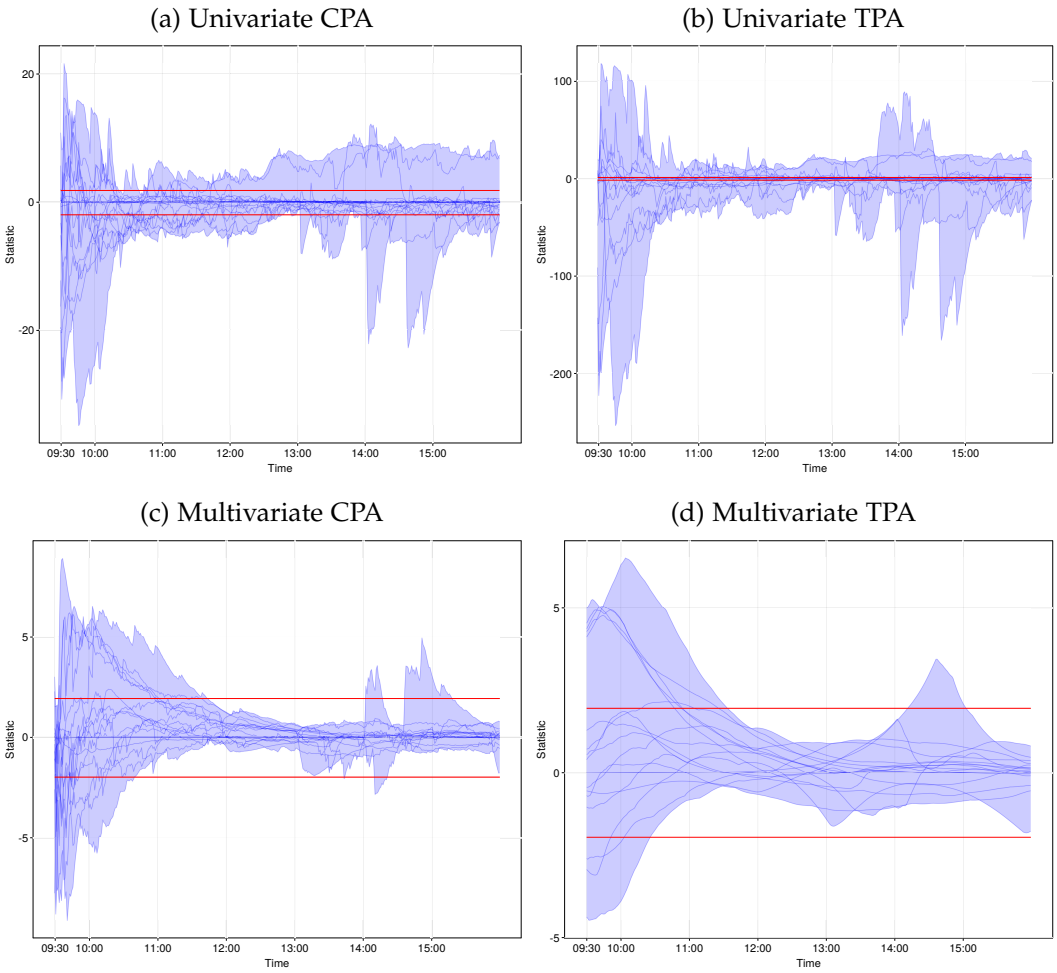


Figure 3.5: Individual Test Results Excluding Overnight Returns

Note: The figure reports the results of the test when overnight returns are excluded. The shaded blue area contains the time-series of test statistics for each trading day between 9:30 and 16:00. The red lines are the 5% critical values of the test. If the blue area, and the lines within it, fall above the positive critical value, the GARCH model has superior predictive ability for the respective time period. If the blue area falls below the negative critical value, the MSM model has superior predictive ability. The top row reports the univariate test results, and the bottom row the multivariate test results.

hour still leave some ambiguity regarding the choice of forecasting method. However, with few short-lived exceptions during the rest of the trading days, it is not possible to say if one method will outperform the other or not. The TPA test paints a clearer picture: on five days, the GARCH model would have been the better choice until 11:00. Until approximately 10:00, the MSM model would have been preferred on three different days. The differences between the univariate and the multivariate evaluation reflect the strong dependencies between the loss differentials. Our application illustrates that failure to account for dependence can lead to noisier and more ambiguous results. Including overnight returns – an uncommon practice in many academic papers – leads to fundamentally different results (see Figure 3.6). Both univariate tests reject their null hypotheses more often in favour of the MSM model, and the TPA test still rejects on most days with no clear indication as to which method to choose. This changes somewhat by considering the multivariate tests reported in Subfigures (c) and (d). As in the application without overnight returns, the overwhelming majority of rejections occurs during the fist hour of the trading day for both tests. However, during the first hour itself, it is not obvious which model to choose based solely on Figure 3.6. The multivariate χ^2 test of the CPA and TPA hypotheses can provide a more conclusive result. In Figure 3.7, we report the test statistic and critical values of the test, multiplied by the sign of the mean of the level components of the loss-differentials. Subfigure (a) shows the CPA statistic (black line) and the TPA statistic (blue line) for the forecasting exercise without overnight returns. The results provide a clear indication that the GARCH model is the preferred choice, especially during market opening hours. This finding is reversed when overnight returns are included (see Subfigure (b)). Now, both tests clearly reject in favour of the MSM model during the first hour of the day.

There are several takeaways from our empirical application. The first is that the relative predictive ability of volatility models is highly time-varying. If one model is found to perform better than another on average, this may well be due to specific and constraint intraday periods. Second, the forecasting performance of volatility models across trading days is highly dependent, which can alter the conclusions drawn from evaluation procedures. Third, a simple GARCH model performs remarkably well at an intraday level if one removes overnight returns from the sample. Fourth, an MSM model is vastly superior when it comes to forecasting overnight returns.

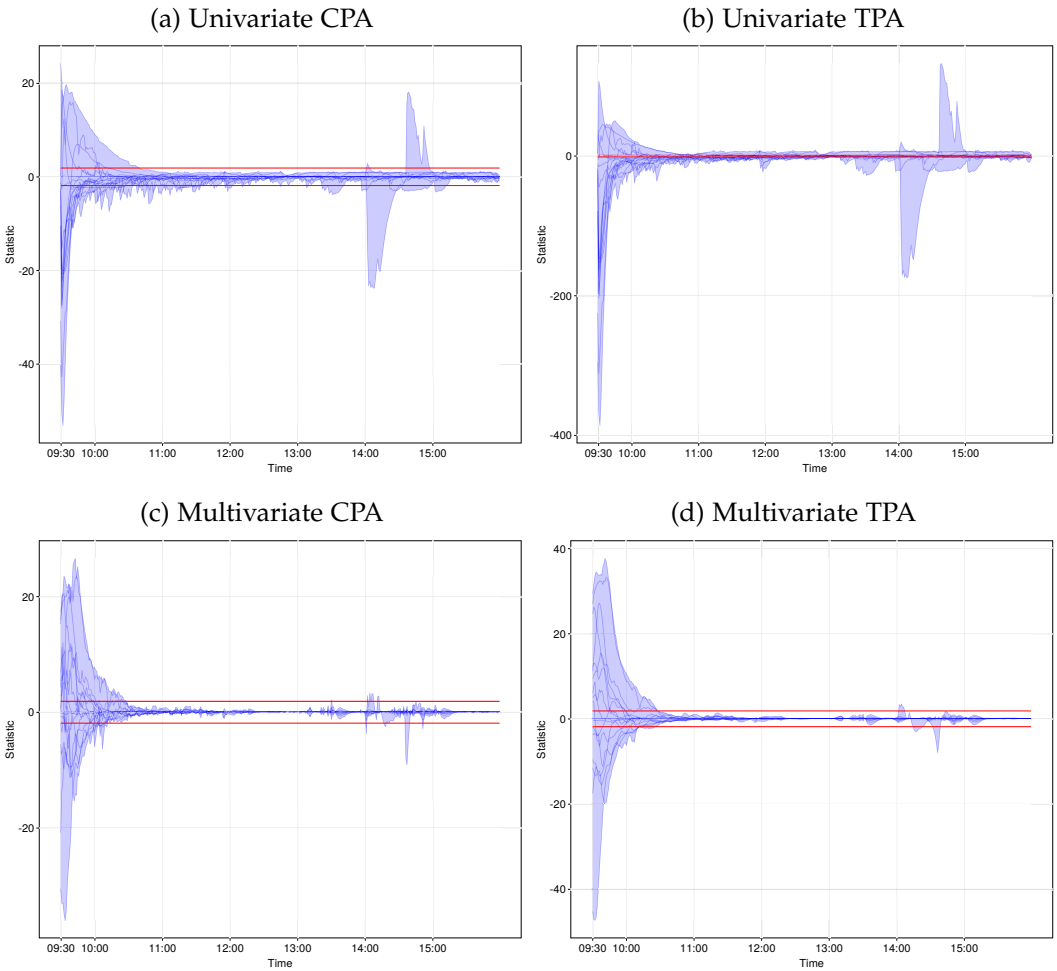


Figure 3.6: Individual Test Results Including Overnight Returns

Note: The figure reports the results of the test when overnight returns are included. The shaded blue area contains the time-series of test statistics for each trading day between 9:30 and 16:00. The red lines are the 5% critical values of the test. If the blue area, and the lines within it, fall above the positive critical value, the GARCH model has superior predictive ability for the respective time period. If the blue area falls below the negative critical value, the MSM model has superior predictive ability. The top row reports the univariate test results, and the bottom row the multivariate test results.

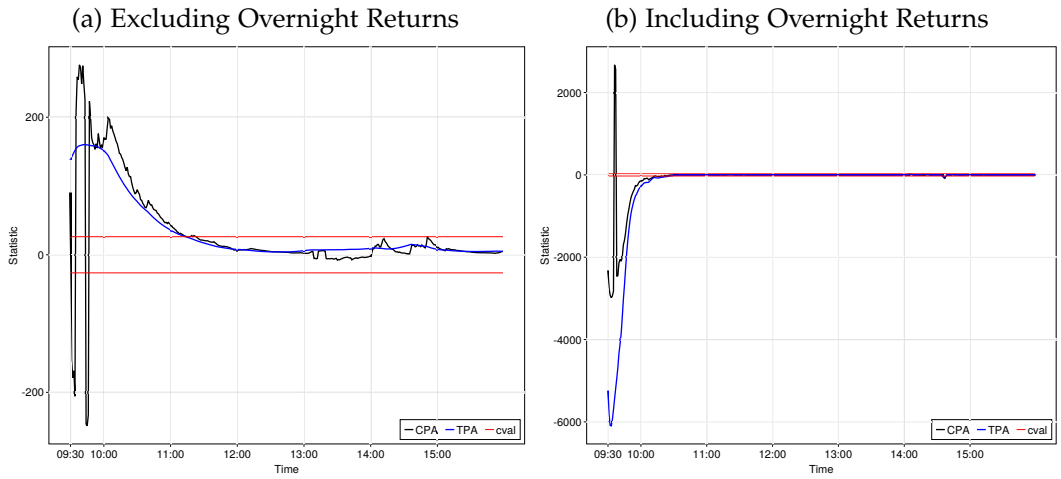


Figure 3.7: χ^2 Test Results

Note: The figure displays the results of the χ^2 test, computed jointly across trading days between 9:30 and 16:00. The blue line corresponds to the TPA tes and the black line to the CPA test. The red lines are the 5% critical values of the test. If the test statistic falls below the negative critical value, the MSM model has superior predictive ability. If it exceeds the positive critical value, the GARCH model has superior predictive ability.

3.5 CONCLUSION

In this paper, we proposed new tests to compare the predictive ability of different forecasting methods in a time-varying manner. First, we developed a time-varying version of the Conditional Predictive Ability (CPA) test of Giacomini and White (2006). The CPA null hypothesis is that, based on past information, one cannot predict if one method will outperform the other. Second, we introduced a new type of hypothesis that asks the question if – based on all available information in the sample – one method could have outperformed the other at any point. We refer to this as Total Predictive Ability (TPA) hypothesis. Furthermore, we proposed multivariate versions of both tests that account for cross-dependence between variables. The tests are pointwise consistent, meaning they reject accurately at each point in time, instead of across time periods. The size and power properties of the tests were verified in multiple Monte-Carlo simulations. We applied the test in an intraday volatility forecasting exercise that compared the Markov-Switching Multifractal (MSM) model of Calvet and Fisher (2002, 2004) to a GARCH model. Our empirical application demonstrates that there is considerable time variation in the forecasting performance of the two models, which can not be uncovered by global tests. During the first hour of the trading day, differences in predictive ability are most pronounced. The testing framework

proposed in this paper can identify such breakpoints and thereby periods during which a model provides more accurate forecasts.

CONCLUSION AND FUTURE RESEARCH

This thesis has contributed to the econometric forecasting literature in several ways. The first chapter addressed the long-standing issue of whether macroeconomic fundamentals can predict foreign exchange rates. Specifically, we showed that accounting for time-variation in the relationship between exchange rates and macroeconomic fundamentals leads to improved predictive ability. This was demonstrated by contrasting a factor model with time-varying loadings to both a constant parameter model and the random walk. In-sample, time-varying factor loadings led to a significant increase in the percentage of explained variation in exchange rates. Furthermore, incorporating these instabilities resulted in improved out-of-sample forecasts. Compared to the benchmarks, our model had better global predictive ability, produced better forecasts of sign changes in exchange rates, and yielded superior forecasts during the financial crisis. We also provided a novel real-time database of macroeconomic variables, comprised of 272 datasets, each containing over 100 variables. Future research can build upon this chapter and analyse the novel dataset by employing a tensor factor model, a burgeoning area of current research (Chen et al., 2022; Feng et al., 2020; Zhou et al., 2022; Babii et al., 2022).

The second chapter proposed an intersection-union test for multivariate predictive ability that combines the p-values of established univariate forecast accuracy tests. The test is valid under arbitrary dependence structures and requires only few broad assumptions. Through a battery of simulations, we demonstrated that evaluating multiple univariate tests with methods that fail to account for dependencies between them leads to size distortions. In contrast, our method exhibits desirable size and power properties across various scenarios. A practical application involving a dataset of 84 daily exchange rates further underscored the practical relevance of our test. A natural progression for this line of research is to directly consider panel datasets as opposed to individual time series. We can also expand our ap-

proach into other areas where approaches using combined p-values had previously been suggested (Bergamelli et al., 2019).

The third chapter proposed tests that evaluated the relative predictive ability of models pointwise for each out-of-sample period in both univariate and multivariate settings. The first test can be viewed as a time varying analogue to the Conditional Predictive Ability test of Giacomini and White (2006). Additionally, we presented a novel hypothesis called Total Predictive Ability hypothesis, which is used to test whether one method could have outperformed the other at any moment, considering all the information available in the sample. Currently, these tests operate under the assumptions underlying a Kalman filter, which could be relaxed by moving towards a Monte-Carlo testing procedure, thereby allowing for more flexibility with respect to the disturbances in the model. As an alternative approach, the computation of a likelihood function by the Kalman filter can be used to formulate a likelihood ratio test.

BIBLIOGRAPHY

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue Ratio Test for the Number of Factors. *Econometrica*, 81(3):1203–1227.

Alessi, L., Barigozzi, M., and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, 80(23-24):1806–1813.

Andersen, T. G., Thyrsgaard, M., and Todorov, V. (2019). Time-Varying Periodicity in Intraday Volatility. *Journal of the American Statistical Association*, 114(528):1695–1707. Publisher: Taylor & Francis.

Babii, A., Ghysels, E., and Pan, J. (2022). Tensor Principal Component Analysis.

Bacchetta, P. and van Wincoop, E. (2004). A Scapegoat Model of Exchange-Rate Fluctuations. 94(2):9.

Bacchetta, P. and van Wincoop, E. (2009). On the Unstable Relationship between Exchange Rates and Macroeconomic Fundamentals. *NBER Working Paper*, w15008.

Bacchetta, P. and van Wincoop, E. (2013). On the unstable relationship between exchange rates and macroeconomic fundamentals. *Journal of International Economics*, 91(1):18–26.

Bai, J. (2003). Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71(1):135–171.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1):191–221.

Barigozzi, M., Hallin, M., Soccorsi, S., and von Sachs, R. (2021). Time-varying general dynamic factor models and the measurement of financial connectedness. *Journal of Econometrics*, 222(1):324–343.

Bates, B. J., Plagborg-Møller, M., Stock, J. H., and Watson, M. W. (2013). Consistent factor estimation in dynamic factor models with structural instability. *Journal of Econometrics*, 177(2):289–304.

Bekiros, S. D. (2014). Exchange rates and fundamentals: Co-movement, long-run relationships and short-run dynamics. *Journal of Banking & Finance*, 39:117–134.

Bergamelli, M., Bianchi, A., Khalaf, L., and Urga, G. (2019). Combining p-values to test for multiple structural breaks in cointegrated regressions. *Journal of Econometrics*, 211(2):461–482.

Bernard, C., Jiang, X., and Wang, R. (2014). Risk aggregation with dependence uncertainty. *Insurance: Mathematics and Economics*, 54:93–108.

Boivin, J. and Ng, S. (2005). Understanding and Comparing Factor-Based Forecasts. Technical Report w11285, National Bureau of Economic Research, Cambridge, MA.

Borup, D., Eriksen, J. N., Kjær, M. M., and Thyrsgaard, M. (2022). Predicting Bond Return Predictability. *Management Science*, in press.

Byrne, J. P., Korobilis, D., and Ribeiro, P. J. (2018). On the Sources of Uncertainty in Exchange Rate Predictability: Uncertainty in Exchange Rates. *International Economic Review*, 59(1):329–357.

Cai, T., Liu, W., and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):349–372.

Calvet, L. and Fisher, A. (2001). Forecasting multifractal volatility. *Journal of Econometrics*, 105(1):27–58.

Calvet, L. and Fisher, A. (2002). Multifractality in Asset Returns: Theory and Evidence. *The Review of Economics and Statistics*, 84(3):381–406. Publisher: The MIT Press.

Calvet, L. E. and Fisher, A. J. (2004). How to Forecast Long-Run Volatility: Regime Switching and the Estimation of Multifractal Processes. *Journal of Financial Econometrics*, 2(1):49–83.

Carriero, A., Galvão, A. B., and Kapetanios, G. (2019). A comprehensive evaluation of macroeconomic forecasting methods. *International Journal of Forecasting*, 35(4):1226–1239.

Chen, R., Yang, D., and Zhang, C.-H. (2022). Factor Models for High-Dimensional Tensor Time Series. *Journal of the American Statistical Association*, 117(537):94–116. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2021.1912757.

Cheung, Y.-W. and Chinn, M. D. (2001). Currency traders and exchange rate dynamics: a survey of the US market. *Journal of International Money and Finance*, 20:439–471.

Cheung, Y.-W., Chinn, M. D., and Marsh, I. W. (2004). How do UK-based foreign exchange dealers think their market operates? *International Journal of Finance & Economics*, 9(4):289–306.

Choi, I. and Jeong, H. (2019). Model selection for factor analysis: Some new criteria and performance comparisons. *Econometric Reviews*, 38(6):577–596.

Clark, T. and McCracken, M. (2013). Advances in Forecast Evaluation. In *Handbook of Economic Forecasting*, volume 2, pages 1107–1201. Elsevier.

Clark, T. E. and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105:85–110.

Clark, T. E. and McCracken, M. W. (2015). Nested forecast model comparisons: A new approach to testing equal accuracy. *Journal of Econometrics*, 186(1):160–177.

Clark, T. E. and West, K. D. (2006). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics*, 135(1-2):155–186.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311.

Coroneo, L. and Caruso, A. (2022). Does Real-Time Macroeconomic Information Help to Predict Interest Rates? *Journal of Money, Credit and Banking*, forthcoming.

Demetrescu, M., Georgiev, I., Rodrigues, P. M., and Taylor, A. R. (2022). Testing for episodic predictability in stock returns. *Journal of Econometrics*, 227(1):85–113.

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics*, 33(1):1–1.

Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3). arXiv:math/0410072.

Doz, C., Giannone, D., and Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, 164(1):188–205.

Doz, C., Giannone, D., and Reichlin, L. (2012). A Quasi-Maximum Likelihood Approach for Large, Approximate Dynamic Factor Models. *The Review of Economics and Statistics*, 4(94):1014–1024.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Number 38 in Oxford statistical science series. Oxford University Press, Oxford, 2nd ed edition.

Engel, C. (2014). Exchange Rates and Interest Parity. In *Handbook of International Economics*, volume 4, pages 453–522. Elsevier.

Engel, C., Mark, N. C., and West, K. D. (2015). Factor Model Forecasts of Exchange Rates. *Econometric Reviews*, 34(1-2):32–55.

Engel, C. and West, K. D. (2005). Exchange Rates and Fundamentals. *Journal of Political Economy*, 113(3):33.

Engle, R. F. and Sokalska, M. E. (2012). Forecasting intraday volatility in the US equity market. Multiplicative component GARCH. *Journal of Financial Econometrics*, 10(1):54–83.

Feng, L., Liu, Y., Chen, L., Zhang, X., and Zhu, C. (2020). Robust block tensor principal component analysis. *Signal Processing*, 166:107271.

Fisher, R. A. (1934). *Statistical Methods for Research Workers*, volume 5. Oliver and Boyd, Edinburgh.

Fratzscher, M., Rime, D., Sarno, L., and Zinna, G. (2015). The scapegoat theory of exchange rates: the first tests. *Journal of Monetary Economics*, 70:1–21.

Georgiev, I., Harvey, D. I., Leybourne, S. J., and Taylor, A. M. R. (2018). Testing for parameter instability in predictive regression models. *Journal of Econometrics*, 204(1):101–118.

Giacomini, R. and Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, 25(4):595–620.

Giacomini, R. and White, H. (2006). Tests of Conditional Predictive Ability. *Econometrica*, 74(6):1545–1578.

Hallin, M. and Liška, R. (2007). Determining the Number of Factors in the General Dynamic Factor Model. *Journal of the American Statistical Association*, 102(478):603–617.

Hansen, P. R. and Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*, 131(1):97–121.

Harvey, A. C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1 edition.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the Cross-Section of Expected Returns. *Review of Financial Studies*, 29(1):5–68.

Heard, N. A. and Rubin-Delanchy, P. (2018). Choosing between methods of combining p-values. *Biometrika*, 105(1):239–246.

Hillebrand, E., Mikkelsen, J. G., Spreng, L., and Urga, G. (2020). Exchange Rates and Macroeconomic Fundamentals: Evidence of Instabilities from Time-Varying Factor Loadings. *CREATES Working Paper*, 19.

Kapetanios, G. (2010). A Testing Procedure for Determining the Number of Factors in Approximate Factor Models With Large Datasets. *Journal of Business & Economic Statistics*, 28(3):397–409.

Khashanah, K. and Shao, C. (2022). Short-term volatility forecasting with kernel support vector regression and Markov switching multifractal model. *Quantitative Finance*, 22(2):241–253.

Kost, J. T. and McDermott, M. P. (2002). Combining dependent P-values. *Statistics & Probability Letters*, 60(2):183–190.

Kouwenberg, R., Markiewicz, A., Verhoeks, R., and Zwinkels, R. C. J. (2017). Model Uncertainty and Exchange Rate Forecasting. *Journal of Financial and Quantitative Analysis*, 52(1):341–363.

Laurent, S., Rombouts, J. V., and Violante, F. (2013). On loss functions and ranking forecasting performances of multivariate volatility models. *Journal of Econometrics*, 173(1):1–10.

Leitch, G. and Tanner, J. E. (1991). Economic Forecast Evaluation: Profits Versus the Conventional Error Measures. *American Economic Review*, 81(3):12.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

Li, J., Liao, Z., and Quaedvlieg, R. (2022). Conditional Superior Predictive Ability. *The Review of Economic Studies*, 89(2):843–875.

Liu, Y. and Xie, J. (2019). Accurate and Efficient P -value Calculation Via Gaussian Approximation: A Novel Monte-Carlo Method. *Journal of the American Statistical Association*, 114(525):384–392.

Liu, Y. and Xie, J. (2020). Cauchy Combination Test: A Powerful Test With Analytic p -Value Calculation Under Arbitrary Dependency Structures. *Journal of the American Statistical Association*, 115(529):393–402.

Lux, T. (2022). Inference for Nonlinear State Space Models: A Comparison of Different Methods applied to Markov-Switching Multifractal Models. *Econometrics and Statistics*, 21:69–95.

Mariano, R. S. and Preve, D. (2012). Statistical tests for multiple forecast comparison. *Journal of Econometrics*, 169(1):123–130.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Meese, R. A. and Rogoff, K. (1983). Empirical exchange rate models of the seventies. *Journal of International Economics*, 14(1-2):3–24.

Mikkelsen, J. G., Hillebrand, E., and Urga, G. (2019). Consistent estimation of time-varying loadings in high-dimensional factor models. *Journal of Econometrics*, 208(2):535–562.

Obstfeld, M. and Rogoff, K. (2001). The Six Major Puzzles in International Macroeconomics: Is There a Common Cause? In *NBER Macroeconomics Annual 2000*, volume 15, pages 339–412. MIT Press, Cambridge, Mass.

Odendahl, F., Rossi, B., and Sekhposyan, T. (2022). Evaluating forecast performance with state dependence. *Journal of Econometrics*.

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, (25):379–410.

Pesaran, H. and Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics Letters*, 58(1):17–29.

Pesaran, H. M. and Timmermann, A. (1992). A Simple Nonparametric Test of Predictive Performance. *Journal of Business & Economic Statistics*, 10(4):461–465.

Pozzi, L. and Sadaba, B. (2020). Detecting Scapegoat Effects on the Relationship Between Exchange Rates and Macroeconomic Fundamentals: A New Approach. *Macroeconomic Dynamics*, 24(4):951–994.

Qu, R., Timmermann, A., and Zhu, Y. (2021). Comparing forecasting performance in cross-sections. *Journal of Econometrics*, page S0304407621002256.

Robins, J. M., van der Vaart, A., and Ventura, V. (2000). Asymptotic Distribution of *P* Values in Composite Null Models. *Journal of the American Statistical Association*, 95(452):1143–1156.

Rossi, B. (2006). Are Exchange Rates Really Random Walks? Some Evidence Robust to Parameter Instability. *Macroeconomic Dynamics*, 10(1):20–38.

Rossi, B. (2013). Exchange Rate Predictability. *Journal of Economic Literature*, 51(4):1063–1119.

Rossi, B. (2021). Forecasting in the Presence of Instabilities: How We Know Whether Models Predict Well and How to Improve Them. *Journal of Economic Literature*, 59(4):1135–90.

Rossi, B. and Inoue, A. (2012). Out-of-Sample Forecast Tests Robust to the Choice of Window Size. *Journal of Business & Economic Statistics*, 30(3):432–453.

Rossi, E. and Fantazzini, D. (2015). Long Memory and Periodicity in Intraday Volatility. *Journal of Financial Econometrics*, 40(4):922–961.

Sarno, L. and Valente, G. (2009). Exchange Rates and Fundamentals: Footloose or Evolving Relationship? *Journal of the European Economic Association*, 7(4):786–830.

Shumway, R. H. and Stoffer, D. S. (1982). An Approach to Time Series Smoothing and Forecasting using the EM Algorithm. *Journal of Time Series Analysis*, 3(4):253–264.

Simon, D. (2006). *Optimal State Estimation*. John Wiley & Sons, Inc.

Spreng, L. and Urga, G. (2022). Combining p-values for Multivariate Predictive Ability Testing. *Journal of Business and Economic Statistics*, forthcoming.

Stroud, J. R. and Johannes, M. S. (2014). Bayesian Modeling and Forecasting of 24-Hour High-Frequency Volatility. *Journal of the American Statistical Association*, 109(508):1368–1384. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2014.937003.

Su, L. and Wang, X. (2017). On time-varying factor models: Estimation and testing. *Journal of Econometrics*, 198(1):84–101.

Timmermann, A. (2008). Elusive return predictability. *International Journal of Forecasting*, 24(1):1–18.

Tippett, L. H. C. (1931). *The Methods of Statistics*. Williams and Norgate, London.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, 1 edition.

Vovk, V. and Wang, R. (2020). Combining *p* -values via averaging. *Biometrika*, 107(4):791–808.

West, K. D. (1996). Asymptotic Inference about Predictive Ability. *Econometrica*, 64(5):1067.

White, H. (1994). *Estimation, Inference and Specification Analysis.* Cambridge University Press.

White, H. and Domowitz, I. (1984). Nonlinear Regression with Dependent Observations. *Econometrica*, 52(1):143.

Zhou, P., Lu, C., and Lin, Z. (2022). Chapter 6 - Tensor principal component analysis. In Liu, Y., editor, *Tensors for Data Processing*, pages 153–213. Academic Press.

APPENDIX A

CHAPTER 1

A.1 COMPUTATION OF TIME-VARYING FACTOR LOADINGS

A.1.1 *Kalman Filter Algorithm*

This section describes the Kalman filter algorithm used to compute the time-varying loadings. Generally, the Kalman filter is applied to a linear state space model, such as

$$\begin{aligned} y_t &= Z_t\alpha_t + \varepsilon_t, & \varepsilon_t &\sim N(0, H_t) \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t, & \eta_t &\sim N(0, Q_t) \end{aligned}$$

where $\alpha_1 \sim N(a_1, P_1)$. It computes a prediction $\alpha_{t+1|t} = \mathbb{E}(\alpha_{t+1}|Y_t)$ and the corresponding variance $P_{t+1|t} = \mathbb{V}(\alpha_{t+1}|Y_t)$ recursively through:

$$\begin{aligned} v_t &= y_t - Z_t\alpha_{t|t-1} \\ \Sigma_t &= Z_tP_{t|t-1}Z_t' + H_t \\ K_t &= T_tP_{t|t-1}Z_t'F_t^{-1} \\ L_t &= T_t - K_tZ_t \\ \alpha_{t+1|t} &= T_t\alpha_{t|t-1} + K_tv_t \\ P_{t+1|t} &= T_tP_{t|t-1}L_t' + R_tQ_tR_t' \end{aligned}$$

A.1.2 *Algorithm to Compute Time-Varying Loadings*

We now turn to a factor model with time-varying loadings model, which can be written as:

$$\begin{aligned} X_t &= \Lambda_t F_t + \varepsilon_t \qquad \text{for } t = 1, ..., T, \\ x_{it} &= \lambda_{it}' F_t + \varepsilon_{it}, \qquad \text{for } i = 1, ..., n \text{ and } t = 1, .., T, \end{aligned} \tag{A.1}$$

where λ_{it} are the stationary time-varying factor loadings, collected in the $n \times r$ Matrix Λ_t, $F_t = (f_{1t}, ..., f_{rt})'$ is a $r \times 1$ vector of factors and $X_t = (x_{1t}, ..., x_{nt})'$ is a $n \times 1$ matrix of dependent variables. We can rewrite the model as

$$x_{it} = \bar{\lambda}_i' F_t + \xi_{it}' F_t + \varepsilon_{it},$$

where $\xi_{it} = \lambda_{it} - \bar{\lambda}_i$ is the time-varying mean zero part of the loadings. We have $\mathbb{E}[x_{it}] = \bar{\lambda}_i' F_t$ and $\mathbb{V}[x_{it}] = \mathbb{E}[(x_{it} - \bar{\lambda}_i' f_t)^2] = \Omega_i$. Hence, we can rewrite (A.1) as:

$$\begin{aligned} x_{it} &= \bar{\lambda}_i' F_t + u_{it} \\ X_i &= F\bar{\lambda}_i + u_i \end{aligned} \tag{A.2}$$

where $u_{it} = \xi_{it}' F_t + \varepsilon_{it}$ and the second equation is written in matrix form for each n. Clearly, mean and variance remain unchanged, since $\mathbb{E}[\xi_{it}] = 0$ for all i and t. The Generalised Least Squares (GLS) estimator for $\bar{\lambda}_i$ is:

$$\tilde{\lambda}_i = (F'\Omega_i^{-1}F)^{-1}F'\Omega_i^{-1}X_i. \tag{A.3}$$

Because the matrix Ω_i is positive definite, it can be Cholesky decomposed into $\Omega_i^{-1} = C_i'\Sigma_i^{-1}C_i$, where Σ_i is diagonal with elements σ_{ii}.

We can pre-multiply A.2 by C:

$$\begin{aligned} C_i X_i &= C_i F\bar{\lambda}_i + C_i u_i \\ X_i^* &= F^*\bar{\lambda}_i + u_i^* \end{aligned} \tag{A.4}$$

such that u_i^* are now uncorrelated over time with variance Σ_i. This results in the following GLS estimator for $\bar{\lambda}_i$:

$$\tilde{\lambda}_i = (F^{*\prime}\Sigma_i^{-1}F^*)^{-1}F^{*\prime}\Sigma_i^{-1}X_i^* \tag{A.5}$$

with $\mathbb{V}[\tilde{\lambda}_i] = (F^{*\prime}\Sigma_i^{-1}F^*)^{-1}$. Since

$$X_i^* - F^*\bar{\lambda}_i = C_i(X_i - F\bar{\lambda}_i) = v_i, \tag{A.6}$$

it follows that applying the same Kalman filter to the measurement equations

$$\begin{aligned} x_{it} &= \xi_{it}F_t' + \varepsilon_{it} = u_{it} \\ F_t &= \Xi_t'F_t' + \epsilon_{it} = \omega_{it} \end{aligned} \tag{A.7}$$

gives C_iX_i and C_iF. To see this, consider the following Kalman filter recursions, applied separately to each of the $i = 1, ..., n$ variables:

$$\begin{aligned} x_{it}^* &= x_{it} - \xi_{i,t|t-1}'F_t \\ \tilde{F}_t^* &= F_t - \Xi_{t|t-1}'F_t \\ \Sigma_{it} &= F_tP_{i,t|t-1}F_t' + \sigma_{i,\varepsilon} \\ K_{it} &= T_iP_{i,t|t-1}F_{it}'\Sigma_{it}^{-1} \\ \xi_{i,t+1|t} &= T_i\xi_{i,t|t-1} + K_{it}x_{it}^* \\ \Xi_{t+1|t} &= T_i\Xi_{t|t-1} + K_{it}F_t^* \\ P_{i,t+1|t} &= (TP_{i,t|t-1} - K_tF_tP_{i,t|t-1})T' + Q_i \end{aligned}$$

where Q_i is the covariance matrix of the state equation and $\sigma_{i,\varepsilon}$ is the variance of the measurement equation. Note that x_{it}^* and F_t^* are the *innovations* generated by this filter, which can be used to construct the GLS estimator A.5 and, ultimately, yield $\tilde{\lambda}_i$. From there, the residual $v_{it} = x_{it}^* - F_t^*\tilde{\lambda}_i'$ is easily calculated, which allows for the computation of the likelihood function for $(\theta_i, \bar{\lambda}_i)$:

$$\mathcal{L}_T = c_i + \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}v_i'\Sigma_i^{-1}v_i \tag{A.8}$$

Maximising the likelihood function gives an estimate $\tilde{\theta}_i$ of the parameter vector θ. Mikkelsen et al. (2019) prove the consistency of such a maximum likelihood estimator for the time-varying factor model (A.1) when using the principal component estimates $\tilde{F}_t$ in place of the true latent factors F_t.

Once $\tilde{\theta}$ is computed, one can apply the Kalman filter recursion to the state space formulation:

$$\begin{aligned} x_{it} &= \tilde{F}_t(\xi_{it} + \tilde{\lambda}_i) + \varepsilon_{it} \\ \xi_{i,t+1} &= T_i\xi_{it} + \eta_{it} \end{aligned} \tag{A.9}$$

to compute $\xi_{i,t+1|t} = \mathbb{E}[\xi_{i,t+1}|Y_t]$ and $P_{i,t+1|t} = \mathbb{V}[\xi_{i,t+1}|Y_t]$ by applying the following recursion separately to the n variables:

$$\begin{aligned}
\omega_{it} &= x_{it} - \tilde{F}_t(\xi_{i,t|t-1} + \tilde{\lambda}_i) \\
\Sigma_{it} &= \tilde{F}_t P_{i,t|t-1} \tilde{F}_t' + \sigma_{i\varepsilon} \\
K_{it} &= T_i P_{i,t|t-1} \tilde{F}_t' \Sigma_{it}^{-1} \\
L_{it} &= T_i - K_{it} \tilde{F}_t \\
\xi_{i,t+1|t} &= T_i \xi_{i,t|t-1} + K_{it} \omega_{it} \\
P_{i,t+1|t} &= T_i P_{i,t|t-1} L_{it}' + Q_i
\end{aligned}$$

The state smoother then calculates $\hat{\xi}_{i,t|T} = \mathbb{E}[\xi_{i,t}|Y_T]$ and $V_{i,t|T} = \mathbb{V}[\xi_{i,t}|Y_T]$ through the backwards recursion:

$$\begin{aligned}
r_{i,t-1} &= \tilde{F}_t' \Sigma_{it}^{-1} \omega_{it} + L_{it}' r_{it} \\
N_{i,t-1} &= \tilde{F}_t' \Sigma_{it}^{-1} \tilde{F}_t + L_{it}' N_{it} L_{it} \\
\hat{\xi}_{i,t|T} &= \xi_{i,t|t-1} + P_{i,t|t-1} r_{i,t-1} \\
V_{i,t|T} &= P_{i,t|t-1} - P_{i,t|t-1} N_{i,t-1} P_{i,t|t-1}.
\end{aligned}$$

This results in the smoothed time-varying factor loadings estimates $\tilde{\lambda}_{it} = \tilde{\lambda}_i + \hat{\xi}_{i,t|T}$ used in the paper.

A.2 DATA SUMMARY

Table A.1: Main Economic Indicators, 2022:02 Vintage

Series	Cat.	Unit	AUS	CAN	DNK	FRA	DEU	ITA	JPN	MEX	NZL	NOR	SWE	CHE	GBR	BRA	IND	ZAF
Industrial production	1	Idx.	✗	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓
Composite leading indicator	1	Idx.	✓*	✓	✓*	✓	✓	✓	✓*	✓	✗	✓*	✓	✓*	✓*	✓1*	✓2*	✓*
Production in construction	1	Idx.	✗	✓	✗	✓	✓	✓3	✗	✓	✗	✗	✓5	✗	✗	✓⋆	✗	✓6
Retail trade volume	4	Idx.	✗	✓	✓	✓	✓	✓	✓◇	✓	✗	✓	✓	✓	✓	✓	✗	✓
Consumer Price Index	7	Idx.	✗	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
Total employment	2	lvl.	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Hourly earnings, manufacturing	2	Idx.	✗	✓	✗	✗	✗	✓	✓	✓	✗	✗	✓	✗	✓	✗	✗	✗
Monetary aggregates	5	Lvl.	✓	✓	✓⋆	✗	✗	✗	✓	✓⋆	✗	✓	✓⋆	✓⋆	✓	✓4⋆	✓⋆	✓
Exports	1	Vol.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note: The table summarises the data used in the in-sample estimation based on the 2022:02 vintage of the OECD Main Economic Indicators database. Cat. refers to McCracken and Ng (2016) categories into which each series is sorted: (1) Output & Income, (2) Labour Market, (3) Housing, (4) Orders & Inventories, (5) Money & Credit, (6) Interest Rates, (7) Prices, and (8) Stock Markets. Unit refers to original unit in OECD database. Idx. stands for Index (2015=100), % denotes annualised percentages, Lvl. number of 1000 persons, and $ stands for US-$ billions. All variables not in % are transformed using first log differences (Δlog), variables in % are transformed using first differences Δ, consistent with McCracken and Ng (2016).

✓indicates consistent availability for the respective country between 1990:04 to 2021:09,

✗ indicates no availability throughout the sample.

∗ indicates the series ends in 2021:08,

⋆ indicates the series ends in 2018:12,

◇ indicates the series ends in 2021:07,

1 indicates the series starts in 1996:02,

2 indicates the series starts in 1996:05,

3 indicates the series starts in 1995:01,

4 indicates the series starts in 1994:07,

5 indicates the series starts in 1994:01,

6 indicates the series starts in 1993:01.

Table A.2: BTCO Survey: Manufacturing

Survey	Cat.	AUS	DNK	FRA	DEU	ITA	JPN	MEX	SWE	GBR	BRA
Production Tendency	1	✗	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1990-01 to 2022-01	1990-01 to 2021-03	✗
Production Future tendency	1	✗	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1990-01 to 2022-01	1990-01 to 2021-03	1990-01 to 2022-01
Finished goods stocks Level	1	✗	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	1990-01 to 2022-01	1996-01 to 2022-01	1990-01 to 2021-03	1990-01 to 2022-01
Order books Level	4	✗	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1990-01 to 2022-01	1990-01 to 2021-03	1990-01 to 2022-01
Orders inflow Tendency	4	✗	✗	✗	1990-01 to 2022-01	✗	✗	✗	✗	✗	✗
Export orders books or demand Level	4	✗	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1990-01 to 2022-01	1990-01 to 2021-03	1990-01 to 2022-01
Selling prices Future tendency	7	✗	1998-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1990-01 to 2022-01	1990-01 to 2021-03	✗
Employment Future Tendency	2	✗	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1990-01 to 2022-01	1990-01 to 2021-03	1990-01 to 2022-01
Capacity Utilisation	1	✗	✗	✗	✗	✗	✗	1998-01 to 2022-01	✗	✗	1990-01 to 2022-01
Business Activity Current	1	✗	✗	✗	✗	✗	✗	✗	✗	✗	1995-04 to 2022-01
Business Activity Future Tendency	1	✗	✗	✗	✗	✗	✗	✗	✗	✗	1995-04 to 2022-01
Confidence indicators	1	✗	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1990-01 to 2022-01	1990-01 to 2019-10	✗

Note: The table summarises the manufacturing survey data used in the in-sample and out-of-sample estimation, sourced from the OECD Business Tendency and Consumer Opinion (BTCO) Survey data. Cat. refers to McCracken and Ng (2016) categories into which each series is sorted: (1) Output & Income, (2) Labour Market, (3) Housing, (4) Orders & Inventories, (5) Money & Credit, (6) Interest Rates, (7) Prices, and (8) Stock Markets.
✗ indicates no availability throughout the sample.

Table A.3: BTCO Survey: Construction

Survey	Cat.	AUS	DNK	FRA	DEU	ITA	JPN	MEX	SWE	GBR	BRA
Business Activity Tendency	1	✗	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1990-01 to 2022-01	1990-01 to 2019-10	✗
Confidence indicators	1	✗	1998-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1990-01 to 2022-01	1990-01 to 2019-10	✗
Order books Level	4	✗	1990-11 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1996-08 to 2022-01	1990-01 to 2021-03	2010-03 to 2022-01
Employment Future tendency	2	✗	1990-11 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1996-08 to 2022-01	1990-01 to 2021-03	✗
Selling prices Future tendency	7	✗	1990-11 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1996-10 to 2022-01	1990-01 to 2021-03	2010-03 to 2022-01

Note: The table summarises the construction survey data used in the in-sample and out-of-sample estimation, sourced from the OECD Business Tendency and Consumer Opinion (BTCO) Survey data. Cat. refers to McCracken and Ng (2016) categories into which each series is sorted: (1) Output & Income, (2) Labour Market, (3) Housing, (4) Orders & Inventories, (5) Money & Credit, (6) Interest Rates, (7) Prices, and (8) Stock Markets.
✗ indicates no availability throughout the sample.

Table A.4: BTCO Survey: Retail Trade

Survey	Cat.	AUS	DNK	FRA	DEU	ITA	JPN	MEX	SWE	GBR	BRA
Business Activity Future Tendency	1	✗	2000-04 to 2022-01	1990-01 to 2022-01	1995-04 to 2022-01	1996-04 to 2022-01	✗	✗	1996-04 to 2022-01	1997-01 to 2021-03	✗
Business Activity Tendency	1	✗	2000-04 to 2022-01	1990-01 to 2022-01	1995-04 to 2022-01	1996-04 to 2022-01	✗	✗	1996-04 to 2022-01	1997-01 to 2021-03	2008-06 to 2022-01
Confidence Indicators	1	✗	2000-04 to 2022-01	1990-01 to 2022-01	1995-04 to 2022-01	1998-01 to 2022-01	✗	✗	1996-04 to 2022-01	1997-01 to 2021-03	✗
Volume of stocks Level	4	✗	2000-04 to 2022-01	1990-01 to 2022-01	2001-02 to 2022-01	1996-04 to 2022-01	✗	✗	1996-04 to 2022-01	1997-01 to 2021-03	2008-06 to 2022-01
Employment Future tendency	2	✗	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	2001-04 to 2022-01	1995-10 to 2022-01	1990-01 to 2022-01	✗
Order Intentions/Demand Future Tendency	4	✗	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1995-10 to 2022-01	1990-01 to 2020-12	✗

Note: The table summarises the retail trade survey data used in the in-sample and out-of-sample estimation, sourced from the OECD Business Tendency and Consumer Opinion (BTCO) Survey data. Cat. refers to McCracken and Ng (2016) categories into which each series is sorted: (1) Output & Income, (2) Labour Market, (3) Housing, (4) Orders & Inventories, (5) Money & Credit, (6) Interest Rates, (7) Prices, and (8) Stock Markets.
✗ indicates no availability throughout the sample.

Table A.5: BTCO Survey: Service

Survey	Cat.	AUS	DNK	FRA	DEU	ITA	JPN	MEX	SWE	GBR	BRA
Business Activity Tendency	1	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	2001-04 to 2022-01	1995-10 to 2022-01	1990-01 to 2022-01	1994-06 to 2022-01
Confidence indicator	1	✗	1990-11 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1996-08 to 2022-01	1990-01 to 2021-03	2010-03 to 2022-01
Demand evolution Tendency	1	✗	2000-04 to 2022-01	1990-01 to 2022-01	1995-04 to 2022-01	1998-01 to 2022-01	✗	✗	1996-04 to 2022-01	1997-01 to 2021-03	2008-06 to 2022-01
Demand evolution Future Tendency	1	✗	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	1998-01 to 2022-01	1996-01 to 2022-01	1990-01 to 2022-01	1995-04 to 2022-01
Employment Tendency	2	✗	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1990-01 to 2022-01	1990-01 to 2019-10	2010-07 to 2022-01
Employment Future Tendency	2	✗	2000-04 to 2022-01	1990-01 to 2022-01	1995-04 to 2022-01	1998-01 to 2022-01	✗	✗	1996-04 to 2022-01	1997-01 to 2021-03	2008-06 to 2022-01

Note: The table summarises the service sector survey data used in the in-sample and out-of-sample estimation, sourced from the OECD Business Tendency and Consumer Opinion (BTCO) Survey data. Cat. refers to McCracken and Ng (2016) categories into which each series is sorted: (1) Output & Income, (2) Labour Market, (3) Housing, (4) Orders & Inventories, (5) Money & Credit, (6) Interest Rates, (7) Prices, and (8) Stock Markets.
✗ indicates no availability throughout the sample.

Table A.6: BTCO Survey: Consumer

Survey	Cat.	AUS	DNK	FRA	DEU	ITA	JPN	MEX	SWE	GBR	BRA
Economic Situation Future Tendency	1	✗	1998-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1990-01 to 2022-01	1990-01 to 2019-10	2010-07 to 2022-01
Confidence Indicators	1	✗	1990-11 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1996-08 to 2022-01	1990-01 to 2021-03	✗
Inflation Future Tendency	7	✗	1990-11 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	1990-01 to 2022-01	✗	✗	1996-08 to 2022-01	1990-01 to 2021-03	2010-03 to 2019-07

Note: The table summarises the construction consumer survey data used in the in-sample and out-of-sample estimation, sourced from the OECD Business Tendency and Consumer Opinion (BTCO) Survey data. Cat. refers to McCracken and Ng (2016) categories into which each series is sorted: (1) Output & Income, (2) Labour Market, (3) Housing, (4) Orders & Inventories, (5) Money & Credit, (6) Interest Rates, (7) Prices, and (8) Stock Markets.
✗ indicates no availability throughout the sample.

A.3 ADDITIONAL RESULTS

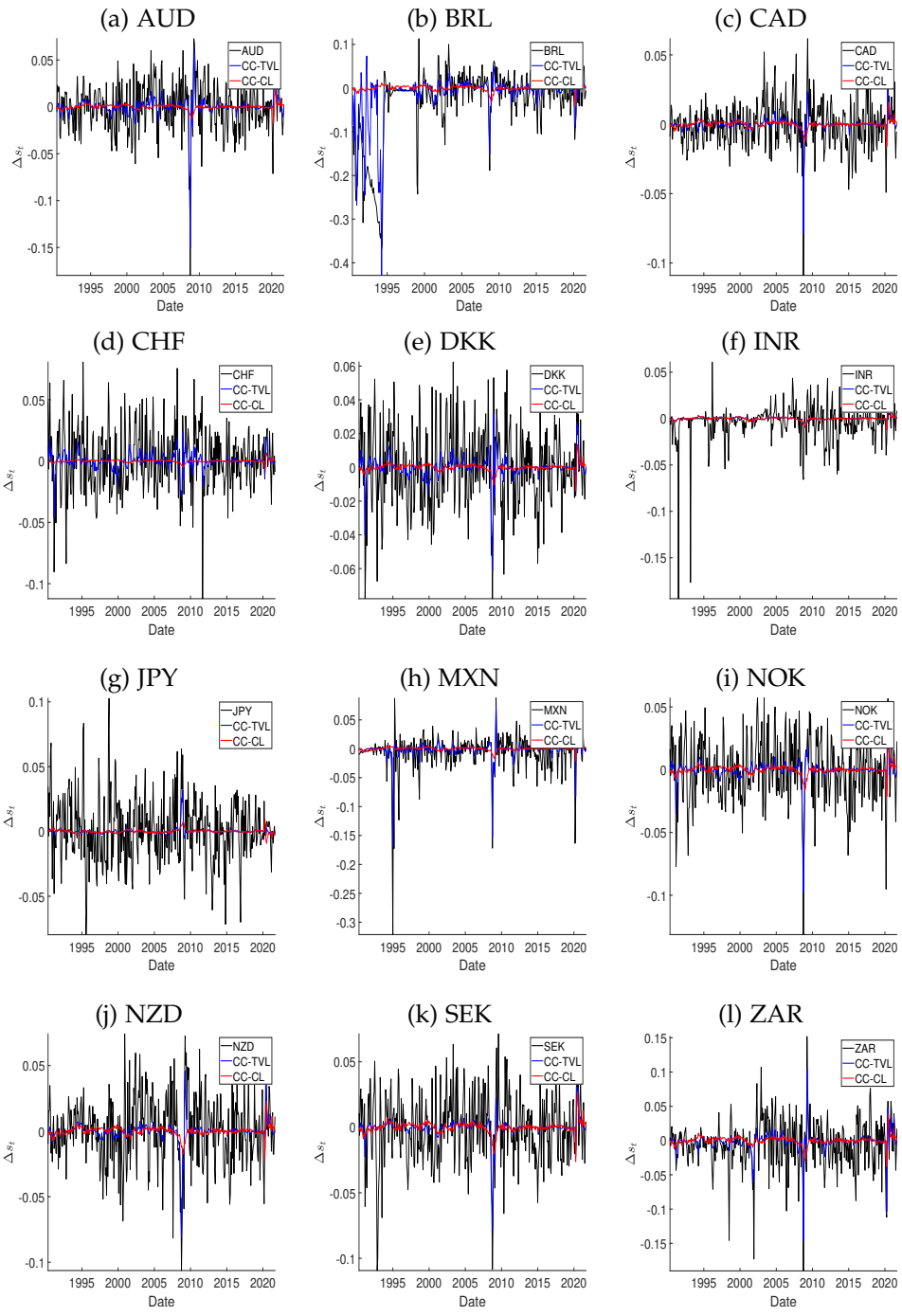


Figure A.1: In-sample Fit – Real Economy Factor

Note: The figure displays the results of a model estimated with only 1-factor, the real economy factor. The black line is the FX change, the blue line is the TVL model fit, and the red line is the CL model fit. CC-TVL stands for the Common Component of the time-varying loadings, CC-CL for the Common Component of the constant loadings model.

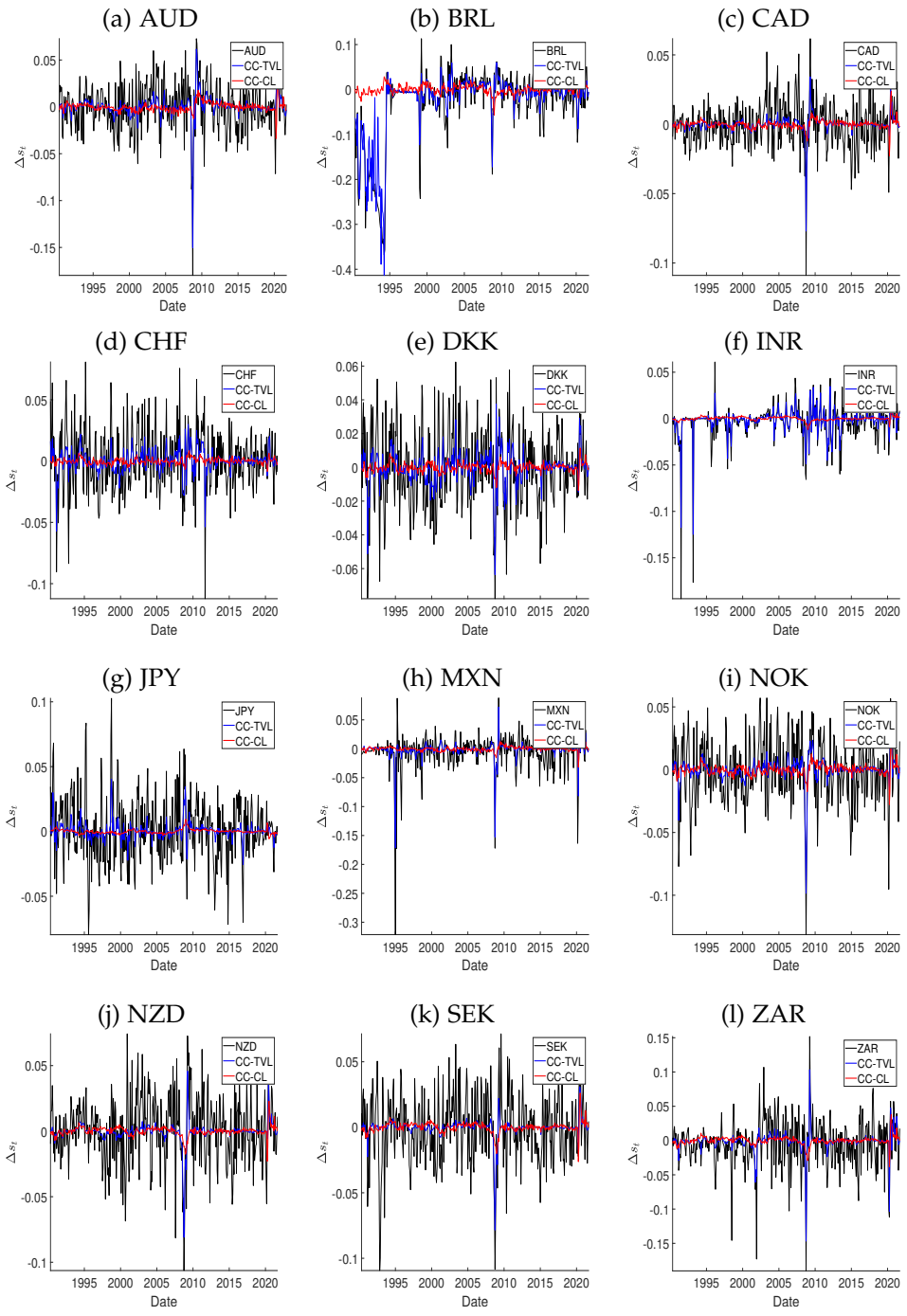


Figure A.2: In-sample Fit – 3 Factors

Note: The figure displays the results of a model estimated with only 1-factor, the real economy factor. The black line is the FX change, the blue line is the TVL model fit, and the red line is the CL model fit. CC-TVL stands for the Common Component of the time-varying loadings, CC-CL for the Common Component of the constant loadings model.

(a) GBP (b) EUR

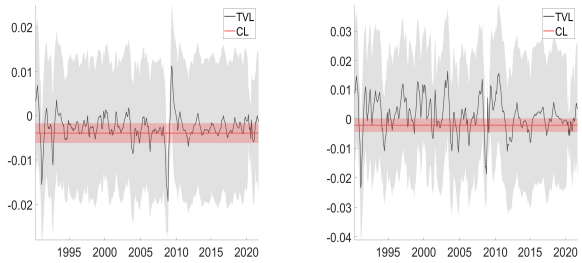


Figure A.3: Loadings – Real Economy Factor

Note: The figure displays the loadings of the 1-factor TVL model with pointwise confidence intervals together with the loadings on the 1-factor CL model.

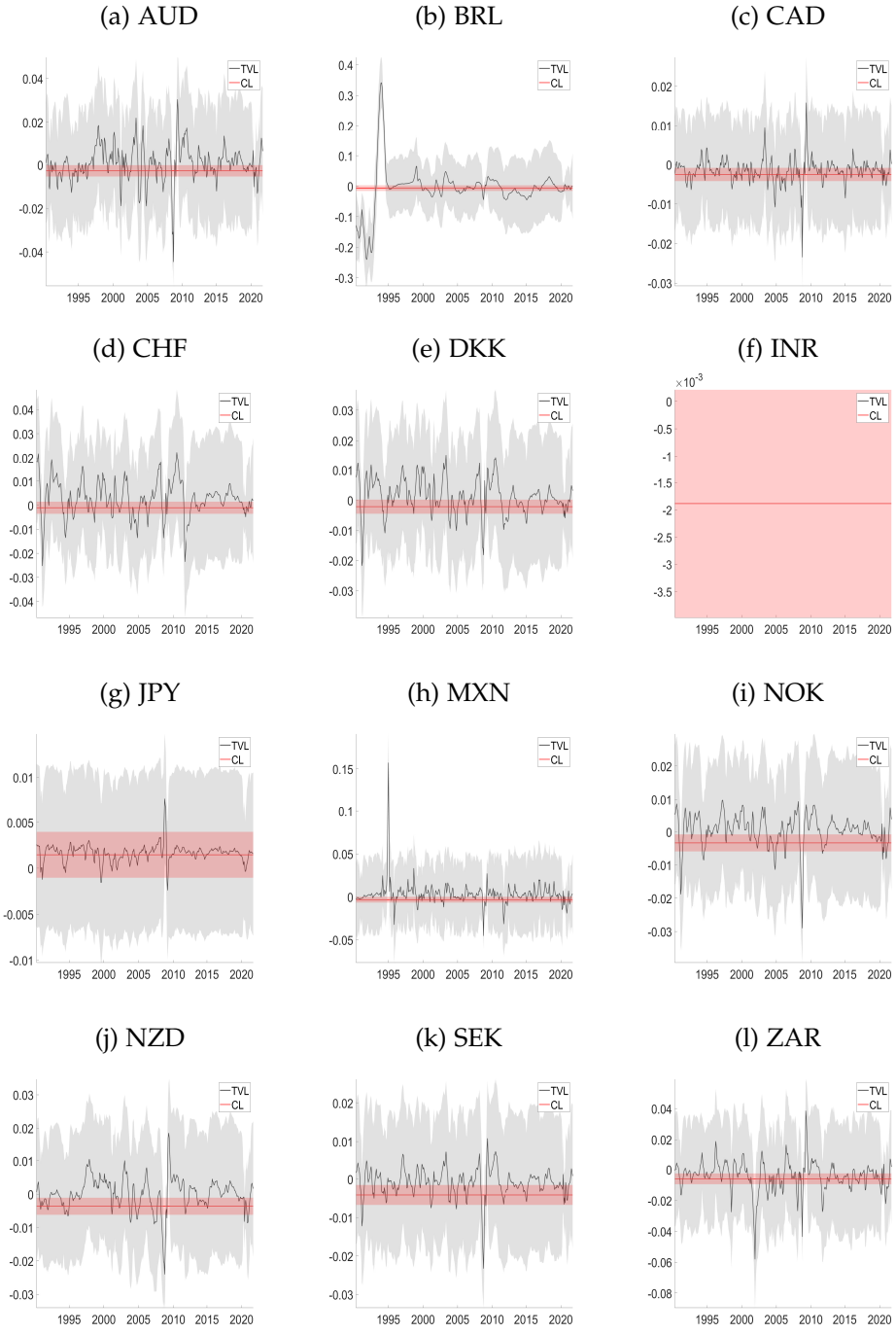


Figure A.4: Loadings – Real Economy Factor

Note: The figure displays the loadings of the 1-factor TVL model with pointwise confidence intervals together with the loadings on the 1-factor CL model.

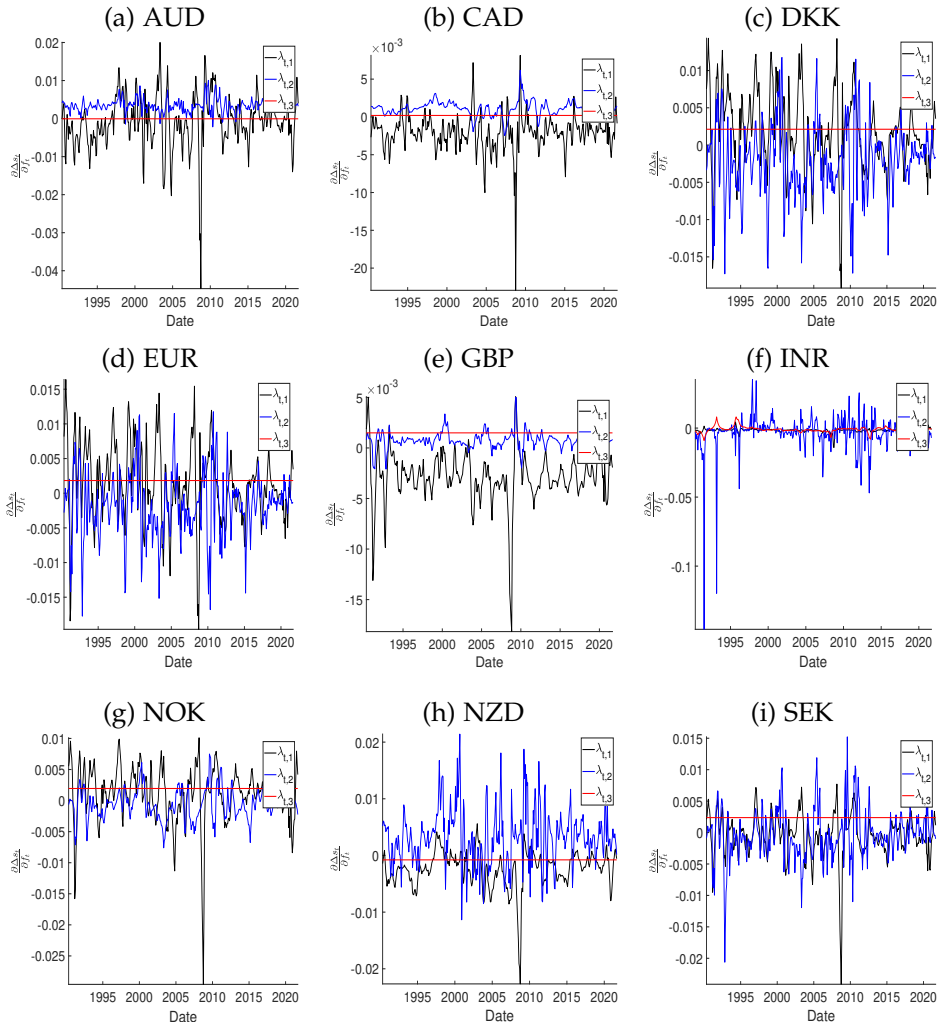


Figure A.5: Loadings – 3 Factors

Note: The figure displays the loadings of the 3-factor model. The black line are the loadings on the real economy factor, the blue line the loadings on the housing factor, and the red line the loadings on the interest rate factor.

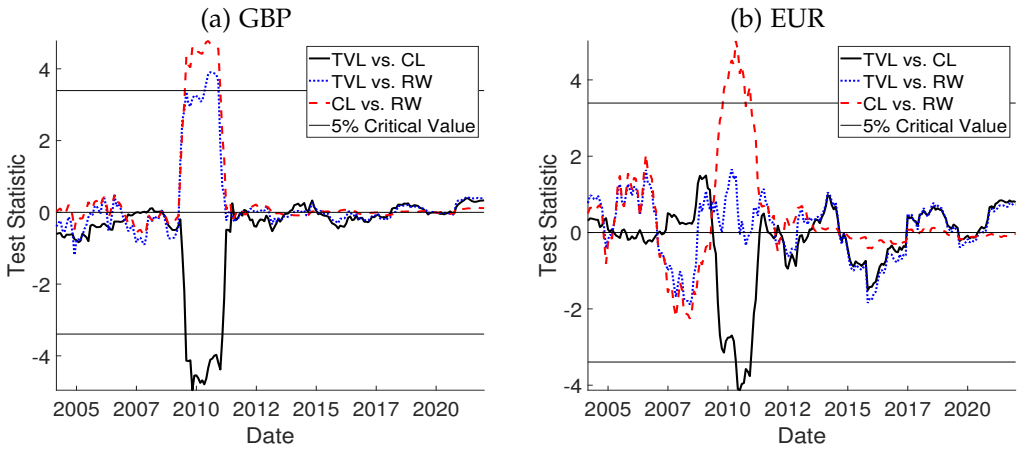


Figure A.6: Fluctuation Test, $h = 6$

Note: This figure plots the results of the Giacomini and Rossi (2010) fluctuation test. The solid blue line is the test statistic, the dotted red lines are the 5% critical values. The results are based on a rolling window of $m = 20$ and a quadratic loss function.

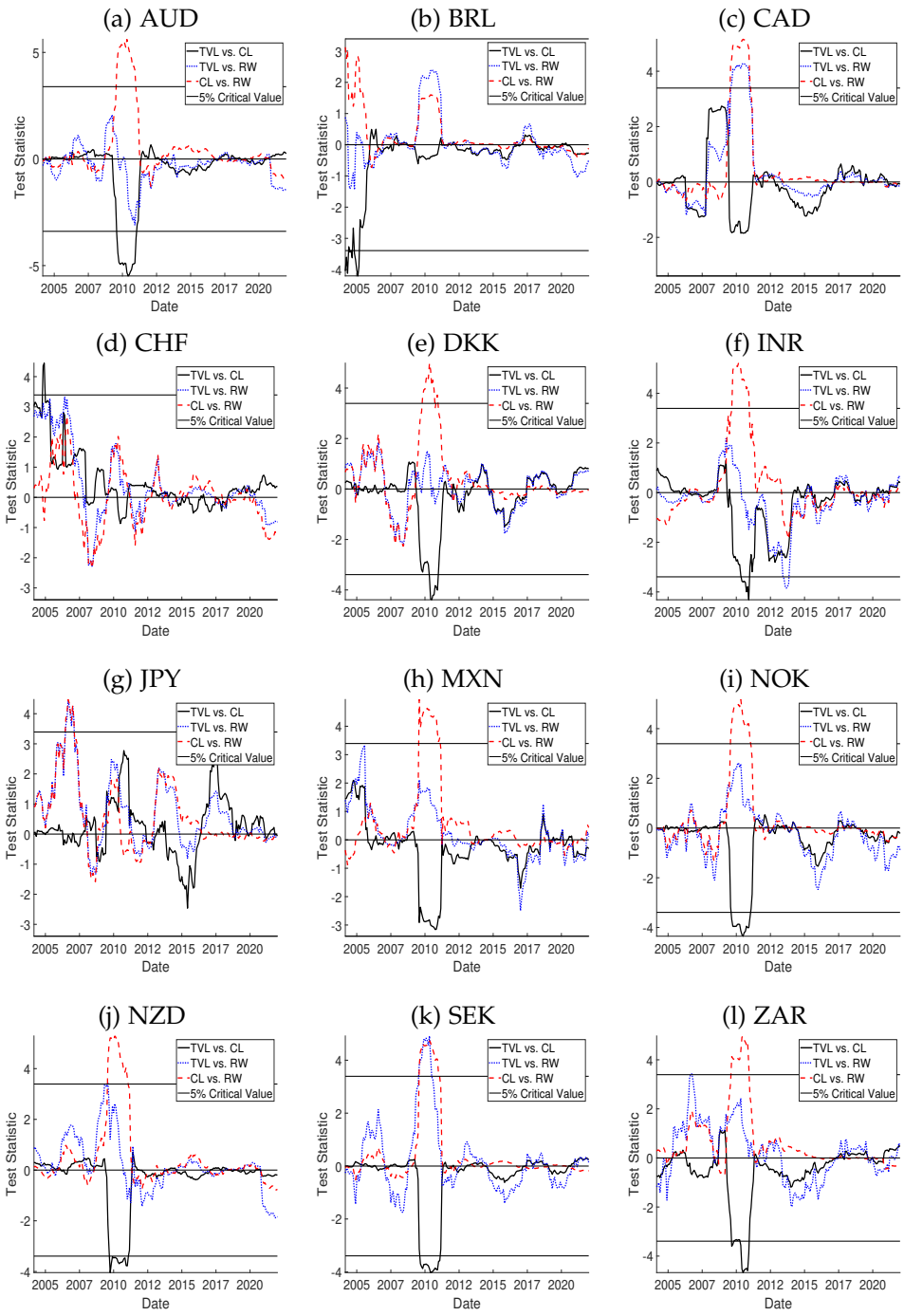


Figure A.7: Fluctuation Test, $h = 6$

Note: This figure plots the results of the Giacomini and Rossi (2010) fluctuation test. The solid blue line is the test statistic, the dotted red lines are the 5% critical values. The results are based on a rolling window of $m = 20$ and a quadratic loss function.

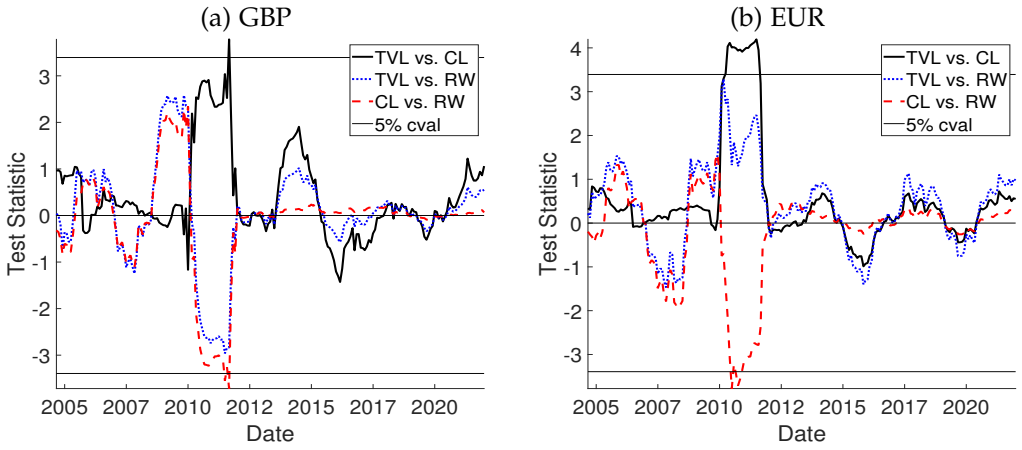


Figure A.8: Fluctuation Test, $h = 12$

Note: This figure plots the results of the Giacomini and Rossi (2010) fluctuation test. The solid blue line is the test statistic, the dotted red lines are the 5% critical values. The results are based on a rolling window of $m = 20$ and a quadratic loss function.

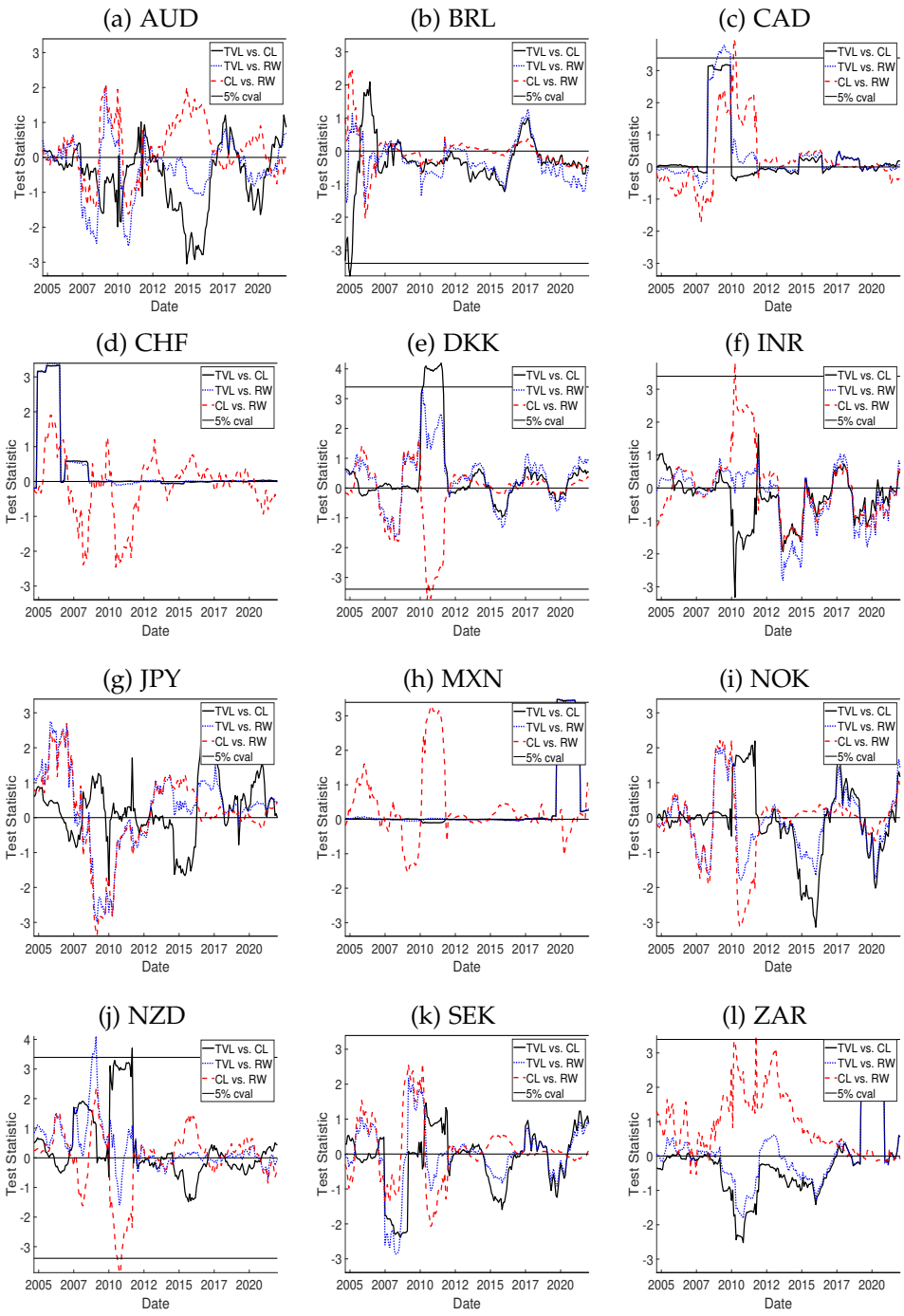


Figure A.9: Fluctuation Test, $h = 12$

Note: This figure plots the results of the Giacomini and Rossi (2010) fluctuation test. The solid blue line is the test statistic, the dotted red lines are the 5% critical values. The results are based on a rolling window of $m = 20$ and a quadratic loss function.

A.4 ROBUSTNESS CHECKS

A.4.1 *In-Sample Robustness*

Table A.7: In-Sample Performance 5 Factor Model

Currency	R^2		Hit Rate	
	TVL	CL	TVL	CL
AUD	0.60	0.07	78.78	56.76
CAD	0.54	0.06	76.39	56.76
DKK	0.65	0.02	81.70	53.05
JPY	0.46	0.00	70.56	53.05
MXN	0.53	0.03	66.84	52.79
NZD	0.62	0.07	81.70	57.82
NOK	0.53	0.04	75.86	54.11
SEK	0.61	0.04	81.70	55.17
CHF	0.56	0.01	78.51	53.85
GBP	0.28	0.05	66.31	53.85
BRL	0.94	0.06	92.31	57.29
INR	0.95	0.01	90.72	54.91
ZAR	0.74	0.06	85.15	58.62
EUR	0.68	0.02	83.55	53.85

Note: The table reports measures of in-sample fit to compare the CL and TVL model. Namely, both the squared correlations between changes in the exchange rate and the in-sample prediction of the TVL & CL model as well as the hit rate in %. The latter being the times the sign of the fitted values corresponded to the sign of the realised values. The currency abbreviations stand for Australian Dollar (AUD), Brazilian Real (BRL), Canadian Dollar (CAD), Danish Krone (DKK), Indian Rupee (INR), Mexican Peso (MXN), New Zealand Dollar (NZD), Norwegian Krone (NOK), South African Rand (ZAR), Swedish Krona (SEK), Swiss Franc (CHF), British Pound (GBP), and Euro (EUR).

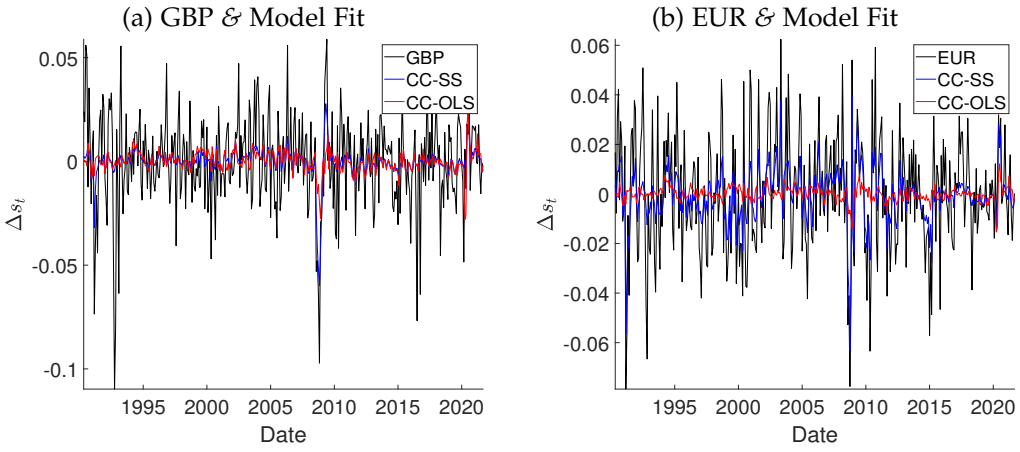


Figure A.10: GBP & EUR In-Sample Fit – 5 Factor Model

Note: The figure displays the results of a model estimated with 5 factors. The black line is the FX change, the blue line is the TVL model fit, and the red line is the CL model fit. CC-TVL stands for the Common Component of the State Space model, CC-CL for the Common Component of the CL model.

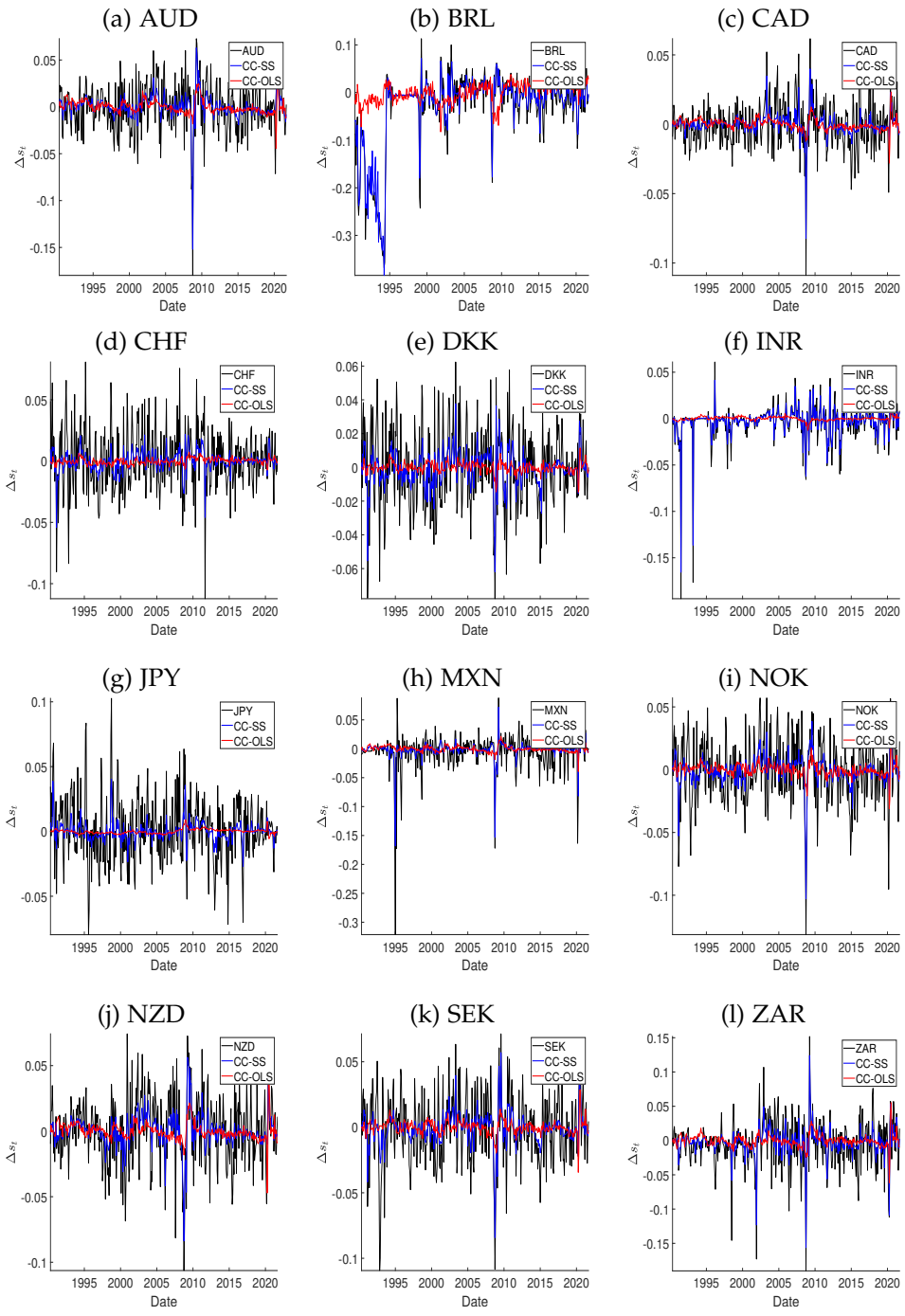


Figure A.11: In-sample Fit – 5 Factor Model

Note: The figure displays the results of a model estimated with only 1-factor, the real economy factor. The black line is the FX change, the blue line is the TVL model fit, and the red line is the CL model fit. CC-TVL stands for the Common Component of the time-varying loadings, CC-CL for the Common Component of the constant loadings model.

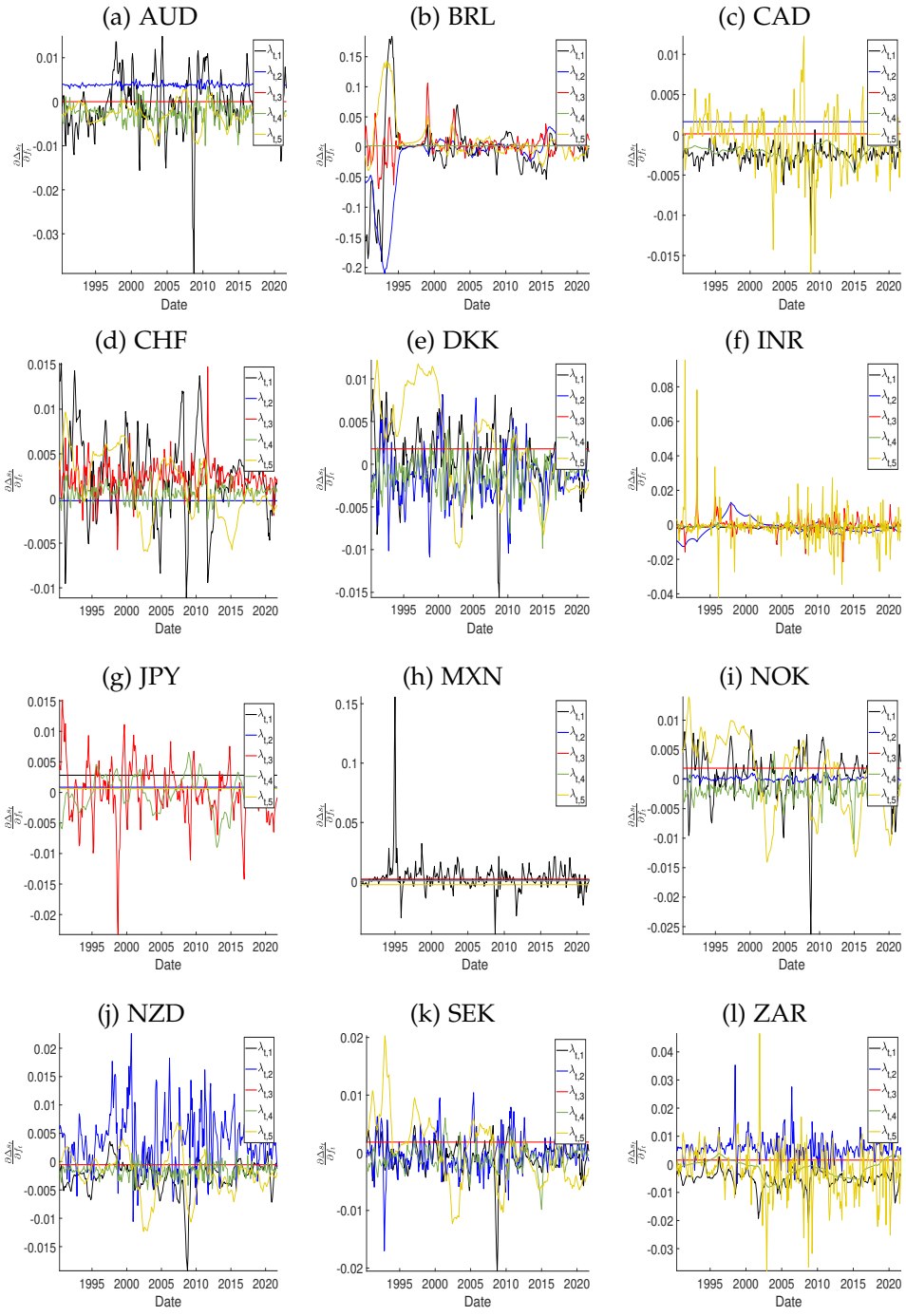


Figure A.12: Loadings – 5 Factor Model

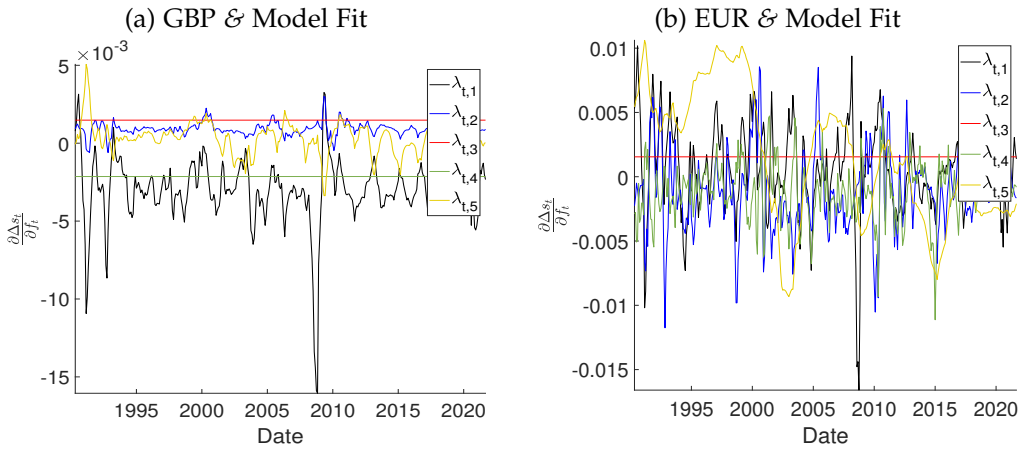


Figure A.13: GBP & EUR Loadings – 5 Factor Model

A.4.2 *Out-of-Sample Robustness*

Table A.8: Forecast Statistics, 1 Factor, $P = 238$

	RMSE		TVL vs. CL						TVL vs. RW				CL vs. RW			
			Quad.		Abs.		DA		Quad.		Abs.		Quad.		Abs.	
	TVL	CL	CPA	UPA	CPA	UPA	TVL	CL	CPA	UPA	CPA	UPA	CPA	UPA	CPA	UPA
Panel A: 1 Factor, $h = 1$																
AUD	0.965	1.013	0.057	0.019	0.037	0.022	0.122	0.342	0.136	0.048	0.236	0.090	0.346	0.149	0.511	0.255
CAD	0.990	1.014	0.044	0.013	0.018	0.004	0.790	0.982	0.658	0.378	0.366	0.203	0.404	0.280	0.369	0.248
DKK	1.008	1.010	0.971	0.871	0.554	0.470	0.617	0.373	0.893	0.636	0.903	0.985	0.330	0.158	0.365	0.176
JPY	1.019	1.018	0.408	0.842	0.762	0.537	0.369	0.368	0.349	0.168	0.349	0.150	0.102	0.032	0.364	0.152
MXN	0.995	1.021	0.168	0.064	0.020	0.034	0.590	0.695	0.824	0.571	0.613	0.414	0.387	0.198	0.426	0.247
NZD	0.965	1.019	0.073	0.023	0.048	0.041	0.095	0.983	0.107	0.114	0.115	0.122	0.298	0.126	0.284	0.237
NOK	0.980	1.007	0.126	0.053	0.135	0.064	0.031	0.768	0.323	0.188	0.441	0.264	0.149	0.178	0.168	0.093
SEK	0.972	1.014	0.040	0.040	0.158	0.125	0.011	0.280	0.062	0.126	0.164	0.371	0.409	0.263	0.410	0.282
CHF	1.007	1.005	0.301	0.545	0.357	0.961	0.672	0.660	0.522	0.276	0.319	0.159	0.054	0.193	0.081	0.044
GBP	0.973	1.015	0.145	0.090	0.296	0.288	0.293	0.333	0.329	0.260	0.082	0.700	0.206	0.453	0.319	0.508
BRL	0.992	1.034	0.001	0.001	0.000	0.000	0.323	0.984	0.634	0.540	0.643	0.440	0.001	0.000	0.000	0.000
INR	0.992	1.004	0.474	0.567	0.818	0.968	0.333	0.444	0.417	0.742	0.895	0.819	0.881	0.625	0.865	0.702
ZAR	0.998	1.019	0.272	0.107	0.511	0.279	0.511	0.872	0.630	0.900	0.950	0.767	0.269	0.200	0.408	0.187
EUR	1.005	1.010	0.934	0.714	0.365	0.226	0.414	0.281	0.946	0.773	0.795	0.636	0.361	0.173	0.434	0.225
$\sum$	9	0	3	5	5	5	2	0	0	1	0	0	1	2	1	2
Panel B: 1 Factor, $h = 6$																
AUD	1.002	1.016	0.301	0.197	0.344	0.166	0.828	0.786	0.120	0.385	0.539	0.776	0.089	0.037	0.108	0.035
CAD	1.023	1.027	0.568	0.378	0.604	0.668	0.421	0.344	0.130	0.063	0.054	0.017	0.110	0.055	0.043	0.016
DKK	0.998	1.006	0.159	0.150	0.184	0.165	0.362	0.352	0.592	0.304	0.539	0.327	0.177	0.102	0.182	0.093
JPY	1.012	1.004	0.163	0.181	0.182	0.183	0.880	0.740	0.057	0.026	0.044	0.035	0.300	0.231	0.387	0.370
MXN	1.003	1.015	0.317	0.132	0.116	0.039	0.752	0.999	0.232	0.189	0.185	0.189	0.070	0.108	0.015	0.047
NZD	1.002	1.022	0.358	0.218	0.304	0.203	0.895	0.589	0.800	0.663	0.975	0.906	0.145	0.055	0.114	0.039
NOK	1.002	1.011	0.296	0.186	0.291	0.173	0.206	0.344	0.118	0.496	0.127	0.832	0.182	0.081	0.127	0.046
SEK	1.006	1.030	0.297	0.217	0.318	0.138	0.234	0.277	0.271	0.294	0.212	0.798	0.270	0.130	0.102	0.034
CHF	1.001	1.002	0.619	0.470	0.270	0.976	0.362	0.297	0.345	0.450	0.536	0.345	0.166	0.328	0.394	0.342
GBP	1.014	1.025	0.297	0.202	0.348	0.181	0.880	0.852	0.277	0.134	0.353	0.169	0.164	0.058	0.124	0.062
BRL	1.010	1.020	0.030	0.026	0.010	0.004	0.074	0.136	0.096	0.049	0.127	0.063	0.013	0.004	0.001	0.000
INR	0.996	1.005	0.198	0.136	0.331	0.173	0.693	0.852	0.344	0.303	0.524	0.422	0.380	0.298	0.409	0.183
ZAR	0.999	1.015	0.513	0.215	0.390	0.186	0.164	0.117	0.107	0.595	0.124	0.055	0.091	0.035	0.101	0.033
EUR	1.000	1.006	0.266	0.283	0.300	0.308	0.678	0.451	0.997	0.950	0.983	0.901	0.177	0.104	0.180	0.096
$\sum$	3	0.000	1.000	1.000	1.000	2.000	0.000	0.000	0.000	2.000	1.000	2.000	1.000	3.000	3.000	8.000
Panel C: 1 Factor, $h = 12$																
AUD	1.000	1.000	0.813	0.675	0.316	0.964	0.989	0.933	0.653	0.791	0.971	0.811	0.389	0.967	0.296	0.950
CAD	1.004	1.007	0.275	0.152	0.758	0.741	0.141	0.191	0.749	0.448	0.462	0.363	0.485	0.233	0.415	0.262
DKK	1.005	0.992	0.250	0.115	0.148	0.104	0.496	0.162	0.441	0.391	0.357	0.321	0.279	0.113	0.190	0.074
JPY	1.004	1.001	0.296	0.179	0.223	0.099	0.829	0.957	0.612	0.328	0.305	0.149	0.523	0.714	0.113	0.155
MXN	1.001	1.006	0.376	0.370	0.202	0.159	0.838	0.650	0.456	0.230	0.171	0.279	0.262	0.156	0.045	0.013
NZD	0.999	0.997	0.308	0.515	0.398	0.827	0.252	0.334	0.659	0.449	0.483	0.227	0.358	0.520	0.373	0.744
NOK	0.997	0.997	0.946	0.739	0.413	0.284	0.219	0.124	0.081	0.086	0.401	0.200	0.439	0.205	0.274	0.831
SEK	0.998	1.001	0.272	0.239	0.239	0.407	0.055	0.165	0.510	0.257	0.375	0.528	0.835	0.774	0.579	0.603
CHF	0.999	0.997	0.479	0.201	0.427	0.243	0.511	0.609	0.894	0.640	0.488	0.918	0.421	0.190	0.361	0.507
GBP	0.996	0.995	0.246	0.390	0.439	0.326	0.868	0.839	0.740	0.502	0.993	0.914	0.801	0.505	0.907	0.788
BRL	1.000	1.003	0.249	0.210	0.224	0.176	0.466	0.555	0.916	0.983	0.868	0.955	0.606	0.336	0.257	0.317
INR	1.001	1.000	0.451	0.839	0.655	0.977	0.005	0.002	0.659	0.817	0.934	0.934	0.955	0.982	0.595	0.945
ZAR	0.998	1.002	0.170	0.136	0.623	0.293	0.008	0.235	0.110	0.069	0.156	0.151	0.564	0.548	0.381	0.645
EUR	0.999	0.992	0.206	0.257	0.162	0.242	0.296	0.098	0.560	0.536	0.996	0.963	0.282	0.116	0.199	0.080
$\sum$	7	6	0	0	0	0	2	1	0	0	0	0	0	0	1	1

Note: Columns 2 and 3 report the Root Mean square Error (RMSE) of TVL and CL forecasts divided by the RMSE of forecasts by a Random Walk (RW). A value < 1 implies the respective model has a smaller RMSE than the RW. Columns 4 to 9 compare the TVL and CL models and report the p-values of the Conditional Predictive Ability (CPA) and Unconditional Predictive Ability (UPA) test of Giacomini and White (2006) using a quadratic (Quad.) and an absolute (Abs.) loss function. Further, they report the p-values of the Pesaran and Shin (1998) nonparametric Direction Accuracy (DA) test for TVL and CL. Columns 10 to 13 compare TVL and RW, reporting p-values of the CPA and UPA test for quadratic and absolute loss differentials. Columns 14 to 17 compare CL and RW, reporting p-values of the same tests. The rows denoted by $\sum$ report the total number of p-values $\leq 5\%$ and in the case of the second and third columns, the number of RMSE smaller than those of a random walk. Results are shown for forecast horizons of $h = 1, 6, 12$. The currency abbreviations stand for Australian Dollar (AUD), Brazilian Real (BRL), Canadian Dollar (CAD), Danish Krone (DKK), Indian Rupee (INR), Mexican Peso (MXN), New Zealand Dollar (NZD), Norwegian Krone (NOK), South African Rand (ZAR), Swedish Krona (SEK), Swiss Franc (CHF), British Pound (GBP), and Euro (EUR).

Table A.9: Forecast Statistics, 3 Factors, $P = 200$

	RMSE		TVL vs. CL						TVL vs. RW				CL vs. RW			
			Quad.		Abs.		DA		Quad.		Abs.		Quad.		Abs.	
	TVL	CL	CPA	UPA	CPA	UPA	TVL	CL	CPA	UPA	CPA	UPA	CPA	UPA	CPA	UPA
Panel A: 3 Factors, $h = 1$																
AUD	0.984	1.009	0.320	0.129	0.646	0.524	0.011	0.176	0.683	0.364	0.863	0.559	0.616	0.435	0.784	0.945
CAD	1.131	1.009	0.602	0.363	0.538	0.444	0.238	0.468	0.346	0.331	0.547	0.371	0.506	0.511	0.870	0.675
DKK	1.008	1.013	0.789	0.760	0.286	0.139	0.049	0.806	0.839	0.709	0.390	0.508	0.202	0.189	0.255	0.359
JPY	1.019	1.023	0.311	0.796	0.295	0.463	0.022	0.246	0.236	0.331	0.272	0.219	0.108	0.129	0.515	0.371
MXN	0.982	1.015	0.148	0.081	0.103	0.031	0.016	0.243	0.232	0.317	0.120	0.164	0.533	0.338	0.627	0.471
NZD	0.985	1.017	0.260	0.110	0.702	0.479	0.032	0.211	0.576	0.463	0.738	0.608	0.454	0.273	0.561	0.843
NOK	0.991	1.008	0.397	0.269	0.296	0.179	0.003	0.478	0.702	0.608	0.948	0.895	0.081	0.366	0.071	0.093
SEK	0.966	1.010	0.027	0.031	0.004	0.015	0.003	0.080	0.030	0.080	0.050	0.142	0.434	0.479	0.414	0.289
CHF	1.134	1.010	0.199	0.074	0.136	0.050	0.701	0.621	0.155	0.056	0.080	0.036	0.011	0.246	0.082	0.473
GBP	0.988	1.025	0.356	0.165	0.954	0.823	0.005	0.090	0.578	0.621	0.231	0.599	0.186	0.199	0.505	0.465
BRL	1.020	1.200	0.013	0.015	0.005	0.001	0.004	0.850	0.927	0.735	0.859	0.889	0.001	0.000	0.000	0.000
INR	0.997	1.018	0.230	0.426	0.586	0.515	0.501	0.618	0.356	0.909	0.774	0.459	0.439	0.214	0.027	0.012
ZAR	0.985	1.022	0.046	0.012	0.124	0.038	0.047	0.475	0.378	0.515	0.443	0.366	0.265	0.198	0.296	0.172
EUR	1.010	1.013	0.825	0.860	0.677	0.346	0.015	0.877	0.802	0.656	0.625	0.760	0.248	0.201	0.282	0.425
$\sum$	8	0	3	3	2	4	11	0	1	0	1	1	2	1	2	2
Panel B: 3 Factors, $h = 6$																
AUD	0.998	1.010	0.425	0.243	0.463	0.300	0.021	0.860	0.127	0.587	0.299	0.104	0.097	0.145	0.089	0.776
CAD	1.005	1.015	0.355	0.161	0.146	0.194	0.258	0.595	0.042	0.507	0.169	0.534	0.096	0.158	0.159	0.178
DKK	1.005	1.002	0.637	0.634	0.610	0.896	0.206	0.500	0.107	0.428	0.105	0.642	0.091	0.633	0.321	0.398
JPY	1.001	1.007	0.094	0.100	0.333	0.243	0.703	0.421	0.291	0.761	0.496	0.819	0.075	0.103	0.353	0.204
MXN	1.003	1.012	0.239	0.111	0.182	0.096	0.570	0.995	0.368	0.463	0.954	0.955	0.099	0.098	0.039	0.098
NZD	0.993	1.014	0.325	0.233	0.598	0.306	0.565	0.591	0.465	0.357	0.467	0.432	0.193	0.129	0.111	0.400
NOK	0.992	1.004	0.377	0.331	0.547	0.362	0.056	0.522	0.158	0.174	0.307	0.403	0.098	0.497	0.160	0.290
SEK	1.006	1.019	0.666	0.404	0.520	0.267	0.092	0.116	0.114	0.372	0.126	0.502	0.272	0.188	0.108	0.048
CHF	0.998	0.996	0.607	0.413	0.741	0.844	0.028	0.105	0.356	0.759	0.575	0.643	0.324	0.428	0.529	0.485
GBP	1.016	1.015	0.438	0.936	0.562	0.833	0.555	0.684	0.308	0.231	0.248	0.088	0.276	0.177	0.388	0.175
BRL	1.077	1.037	0.451	0.348	0.484	0.315	0.001	0.174	0.031	0.007	0.084	0.019	0.190	0.045	0.551	0.243
INR	0.993	1.005	0.211	0.058	0.291	0.341	0.130	0.532	0.328	0.169	0.852	0.725	0.076	0.288	0.042	0.577
ZAR	1.004	1.016	0.448	0.210	0.392	0.241	0.385	0.994	0.350	0.303	0.597	0.323	0.018	0.018	0.053	0.017
EUR	1.004	1.002	0.661	0.837	0.598	0.731	0.208	0.565	0.128	0.578	0.171	0.838	0.087	0.644	0.289	0.399
$\sum$	5	0	0	0	0	0	3	0	2	1	0	1	1	2	2	2
Panel C: 3 Factors, $h = 12$																
AUD	1.001	1.003	0.067	0.410	0.201	0.376	0.469	0.940	0.167	0.831	0.506	0.285	0.791	0.496	0.543	0.827
CAD	0.998	1.003	0.239	0.088	0.372	0.150	0.106	0.614	0.543	0.566	0.678	0.452	0.822	0.619	0.557	1.000
DKK	1.011	0.996	0.092	0.149	0.189	0.106	0.246	0.201	0.055	0.108	0.061	0.172	0.079	0.400	0.372	0.300
JPY	0.996	0.999	0.343	0.163	0.557	0.779	0.755	0.268	0.463	0.294	0.579	0.692	0.909	0.739	0.753	0.537
MXN	1.001	1.003	0.790	0.429	0.483	0.162	0.958	0.988	0.265	0.385	0.412	0.261	0.422	0.195	0.069	0.019
NZD	1.005	1.000	0.319	0.112	0.331	0.173	0.988	0.713	0.070	0.212	0.245	0.283	0.819	0.924	0.586	0.834
NOK	1.004	0.998	0.291	0.284	0.318	0.228	0.362	0.468	0.081	0.412	0.322	0.261	0.352	0.552	0.687	0.943
SEK	1.003	0.999	0.531	0.272	0.716	0.448	0.250	0.435	0.104	0.502	0.243	0.485	0.349	0.780	0.481	0.989
CHF	1.002	0.998	0.484	0.190	0.445	0.221	0.205	0.134	0.266	0.707	0.364	0.853	0.331	0.565	0.350	0.256
GBP	0.997	0.995	0.592	0.492	0.829	0.701	0.738	0.994	0.646	0.630	0.973	0.905	0.747	0.505	0.942	0.965
BRL	1.012	0.991	0.587	0.286	0.524	0.285	0.002	0.176	0.730	0.447	0.991	0.891	0.386	0.333	0.177	0.105
INR	0.997	0.997	0.898	0.971	0.427	0.464	0.469	0.319	0.839	0.555	0.869	0.924	0.346	0.387	0.620	0.436
ZAR	1.001	1.006	0.032	0.043	0.299	0.150	0.606	0.999	0.608	0.646	0.815	0.644	0.067	0.021	0.102	0.205
EUR	1.008	0.995	0.037	0.182	0.212	0.167	0.290	0.290	0.070	0.185	0.124	0.320	0.080	0.348	0.353	0.300
$\sum$	4	9	2	1	0	0	1	0	0	0	0	0	0	1	0	1

Note: Columns 2 and 3 report the Root Mean square Error (RMSE) of TVL and CL forecasts divided by the RMSE of forecasts by a Random Walk (RW). A value < 1 implies the respective model has a smaller RMSE than the RW. Columns 4 to 9 compare the TVL and CL models and report the p-values of the Conditional Predictive Ability (CPA) and Unconditional Predictive Ability (UPA) test of Giacomini and White (2006) using a quadratic (Quad.) and an absolute (Abs.) loss function. Further, they report the p-values of the Pesaran and Shin (1998) nonparametric Direction Accuracy (DA) test for TVL and CL. Columns 10 to 13 compare TVL and RW, reporting p-values of the CPA and UPA test for quadratic and absolute loss differentials. Columns 14 to 17 compare CL and RW, reporting p-values of the same tests. The rows denoted by $\sum$ report the total number of p-values $\leq 5\%$ and in the case of the second and third columns, the number of RMSE smaller than those of a random walk. Results are shown for forecast horizons of $h = 1, 6, 12$. The currency abbreviations stand for Australian Dollar (AUD), Brazilian Real (BRL), Canadian Dollar (CAD), Danish Krone (DKK), Indian Rupee (INR), Mexican Peso (MXN), New Zealand Dollar (NZD), Norwegian Krone (NOK), South African Rand (ZAR), Swedish Krona (SEK), Swiss Franc (CHF), British Pound (GBP), and Euro (EUR).

Table A.10: Forecast Statistics, 3 Factors, $P = 260$

	RMSE		TVL vs. CL						TVL vs. RW				CL vs. RW			
			Quad.		Abs.		DA		Quad.		Abs.		Quad.		Abs.	
	TVL	CL	CPA	UPA	CPA	UPA	TVL	CL	CPA	UPA	CPA	UPA	CPA	UPA	CPA	UPA
Panel A: 3 Factors, $h = 1$																
AUD	0.964	1.006	0.139	0.047	0.261	0.116	0.004	0.405	0.221	0.093	0.133	0.080	0.806	0.552	0.435	0.743
CAD	0.987	1.005	0.295	0.129	0.583	0.549	0.016	0.207	0.527	0.331	0.478	0.334	0.657	0.661	0.712	0.509
DKK	1.000	1.018	0.431	0.325	0.454	0.199	0.141	0.726	0.421	0.986	0.475	0.612	0.383	0.148	0.577	0.323
JPY	1.013	1.013	0.764	0.998	0.407	0.987	0.319	0.856	0.632	0.360	0.193	0.924	0.421	0.219	0.408	0.892
MXN	0.990	1.017	0.609	0.358	0.727	0.425	0.001	0.323	0.883	0.650	0.733	0.434	0.239	0.145	0.177	0.921
NZD	0.958	1.020	0.055	0.018	0.118	0.083	0.009	0.493	0.161	0.118	0.212	0.168	0.487	0.249	0.732	0.642
NOK	0.973	1.010	0.024	0.022	0.014	0.017	0.001	0.767	0.260	0.152	0.521	0.316	0.087	0.302	0.105	0.051
SEK	0.950	1.011	0.008	0.009	0.004	0.014	0.001	0.253	0.006	0.024	0.044	0.124	0.324	0.447	0.288	0.207
CHF	1.022	1.010	0.610	0.364	0.486	0.348	0.740	0.666	0.344	0.168	0.393	0.221	0.481	0.308	0.769	0.515
GBP	0.990	1.026	0.325	0.158	0.593	0.534	0.502	0.369	0.857	0.663	0.203	0.451	0.161	0.228	0.235	0.171
BRL	0.964	1.145	0.011	0.004	0.002	0.001	0.020	0.102	0.550	0.308	0.507	0.390	0.005	0.002	0.002	0.001
INR	1.003	1.015	0.539	0.634	0.984	0.845	0.127	0.464	0.587	0.916	0.536	0.344	0.201	0.111	0.035	0.030
ZAR	0.992	1.018	0.115	0.041	0.347	0.138	0.213	0.804	0.615	0.642	0.773	0.566	0.400	0.160	0.160	0.217
EUR	1.000	1.018	0.338	0.355	0.551	0.243	0.216	0.579	0.375	0.995	0.391	0.687	0.389	0.148	0.578	0.323
$\sum$	8	0	3	6	3	3	7	0	1	1	1	0	1	1	2	2
Panel B: 3 Factors, $h = 6$																
AUD	0.996	1.009	0.149	0.199	0.511	0.344	0.027	0.810	0.340	0.163	0.121	0.056	0.093	0.161	0.062	0.734
CAD	1.007	1.015	0.030	0.118	0.725	0.526	0.375	0.634	0.035	0.446	0.151	0.171	0.079	0.086	0.144	0.094
DKK	1.004	1.007	0.820	0.560	0.864	0.611	0.443	0.884	0.559	0.457	0.471	0.179	0.085	0.175	0.165	0.062
JPY	1.228	1.004	0.344	0.307	0.364	0.302	0.021	0.238	0.359	0.308	0.316	0.315	0.009	0.167	0.059	0.980
MXN	1.002	1.003	0.486	0.887	0.972	0.796	0.666	0.386	0.685	0.682	0.311	0.891	0.167	0.320	0.795	0.865
NZD	1.005	1.016	0.500	0.350	0.849	0.535	0.531	0.269	0.430	0.389	0.791	0.565	0.176	0.117	0.119	0.282
NOK	1.003	1.009	0.366	0.224	0.554	0.342	0.205	0.740	0.181	0.578	0.208	0.250	0.177	0.191	0.120	0.060
SEK	1.003	1.021	0.340	0.244	0.025	0.116	0.477	0.622	0.222	0.556	0.777	0.824	0.303	0.174	0.100	0.028
CHF	1.019	1.000	0.632	0.347	0.831	0.605	0.295	0.560	0.675	0.365	0.789	0.482	0.352	0.996	0.401	0.543
GBP	1.011	1.016	0.624	0.311	0.968	0.734	0.672	0.791	0.346	0.231	0.157	0.058	0.306	0.141	0.263	0.112
BRL	1.000	1.045	0.415	0.219	0.483	0.176	0.000	0.040	0.974	0.974	0.646	0.303	0.152	0.063	0.103	0.102
INR	1.000	1.009	0.377	0.114	0.527	0.284	0.021	0.236	0.249	0.917	0.314	0.736	0.039	0.063	0.093	0.189
ZAR	1.001	1.017	0.146	0.054	0.181	0.082	0.829	0.948	0.541	0.842	0.985	0.963	0.012	0.004	0.037	0.008
EUR	1.003	1.006	0.669	0.547	0.617	0.575	0.386	0.891	0.688	0.566	0.640	0.291	0.095	0.210	0.187	0.082
$\sum$	1	0	1	0	1	0	4	1	1	0	0	0	3	1	1	2
Panel C: 3 Factors, $h = 12$																
AUD	1.000	1.001	0.462	0.529	0.355	0.529	0.060	0.956	0.336	0.893	0.364	0.138	0.861	0.647	0.596	0.736
CAD	1.002	1.002	0.561	0.991	0.216	0.975	0.534	0.442	0.698	0.666	0.700	0.979	0.789	0.630	0.583	0.995
DKK	1.000	0.993	0.301	0.247	0.375	0.176	0.278	0.500	0.118	0.951	0.454	0.871	0.268	0.210	0.288	0.124
JPY	1.556	0.999	0.313	0.302	0.305	0.294	0.369	0.623	0.372	0.318	0.390	0.307	0.166	0.747	0.073	0.607
MXN	1.006	1.000	0.173	0.063	0.208	0.076	0.778	0.237	0.396	0.161	0.155	0.067	0.452	0.807	0.862	0.938
NZD	1.000	0.997	0.053	0.351	0.101	0.538	0.396	0.858	0.185	0.896	0.531	0.923	0.621	0.463	0.552	0.677
NOK	1.010	0.998	0.619	0.428	0.342	0.742	0.077	0.706	0.083	0.514	0.268	0.820	0.756	0.652	0.384	0.900
SEK	0.999	0.999	0.378	0.798	0.683	0.487	0.106	0.579	0.145	0.876	0.591	0.437	0.411	0.731	0.391	0.884
CHF	1.111	0.996	0.445	0.304	0.386	0.325	0.209	0.542	0.616	0.336	0.606	0.418	0.578	0.257	0.365	0.131
GBP	0.995	0.992	0.437	0.203	0.216	0.204	0.911	0.714	0.600	0.408	0.974	0.877	0.572	0.295	0.939	0.815
BRL	1.008	1.020	0.503	0.362	0.520	0.630	0.032	0.370	0.639	0.374	0.960	0.914	0.167	0.072	0.677	0.573
INR	0.997	0.998	0.454	0.661	0.945	0.714	0.449	0.278	0.254	0.387	0.364	0.483	0.336	0.620	0.227	0.299
ZAR	1.011	1.005	0.439	0.588	0.414	0.609	0.669	0.877	0.545	0.322	0.666	0.385	0.091	0.038	0.175	0.371
EUR	1.000	0.993	0.347	0.272	0.368	0.273	0.376	0.388	0.115	0.989	0.603	0.935	0.251	0.191	0.310	0.137
$\sum$	3	9	0	0	0	0	1	0	0	0	0	0	0	1	0	0

Note: Columns 2 and 3 report the Root Mean square Error (RMSE) of TVL and CL forecasts divided by the RMSE of forecasts by a Random Walk (RW). A value < 1 implies the respective model has a smaller RMSE than the RW. Columns 4 to 9 compare the TVL and CL models and report the p-values of the Conditional Predictive Ability (CPA) and Unconditional Predictive Ability (UPA) test of Giacomini and White (2006) using a quadratic (Quad.) and an absolute (Abs.) loss function. Further, they report the p-values of the Pesaran and Shin (1998) nonparametric Direction Accuracy (DA) test for TVL and CL. Columns 10 to 13 compare TVL and RW, reporting p-values of the CPA and UPA test for quadratic and absolute loss differentials. Columns 14 to 17 compare CL and RW, reporting p-values of the same tests. The rows denoted by $\sum$ report the total number of p-values $\leq 5\%$ and in the case of the second and third columns, the number of RMSE smaller than those of a random walk. Results are shown for forecast horizons of $h = 1, 6, 12$. The currency abbreviations stand for Australian Dollar (AUD), Brazilian Real (BRL), Canadian Dollar (CAD), Danish Krone (DKK), Indian Rupee (INR), Mexican Peso (MXN), New Zealand Dollar (NZD), Norwegian Krone (NOK), South African Rand (ZAR), Swedish Krona (SEK), Swiss Franc (CHF), British Pound (GBP), and Euro (EUR).

APPENDIX B

CHAPTER 2

B.1 PROOFS

Proof of Theorem 2.1: To show

$$\mathbb{P}\left[P_{r,n} \in C_{r,n} \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1]\right] \leq \alpha, \qquad \forall\, \alpha \in (0,1), \tag{A1}$$

note first that:

$$n^{-1}\left(\sum_{i=1}^{n} p_i^{-r}\right)^{1/r} = n^{\frac{1-r}{r}}\left(\frac{1}{n}\sum_{i=1}^{n} p_i^{-r}\right)^{1/r}.$$

We then have

$$\begin{aligned}
&\mathbb{P}\left[n^{-1}\left(\sum_{i=1}^{n} p_i^{-r}\right)^{1/r} \geq \frac{1}{\alpha}\frac{r}{r-1} \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1]\right] \\
&= \mathbb{P}\left[n\left(\sum_{i=1}^{n} p_i^{-r}\right)^{-1/r} \leq \alpha\frac{r-1}{r} \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1]\right] \\
&= \mathbb{P}\left[\left(\frac{1}{n}\sum_{i=1}^{n} p_i^{-r}\right)^{-1/r} \leq \alpha n^{\frac{1-r}{r}}\frac{r-1}{r} \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1]\right] \\
&\leq \alpha
\end{aligned}$$

The term $\left(\frac{1}{n}\sum_{i=1}^{n} p_i^r\right)^{1/r}$ fulfils the definition of a Kolmogorov-Nagumo average:

Definition 1 (Kolmogorov-Nagumo)**.** For a strictly continuous and monotonic function $\psi : [0,1] \mapsto \mathbb{R}$, the Kolmogorov-Nagumo average is defined as

$$M(\mathbf{P}) = \psi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\psi(p_i)\right),$$

where $M : \bigcup_{n=1}^{\infty} I^n \mapsto \mathbb{R}$ for a closed and bounded interval $I \subset \mathbb{R}$.

As $\psi(\cdot) = p_i^{-r}$ is strictly continuous and monotonic, we obtain

$$\mathbb{P}\left[M(\mathbf{P}) \leq \alpha n^{\frac{1-r}{r}}\frac{r-1}{r} \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1]\right] \leq \alpha$$

Hence, to prove the first part of Theorem 1, it is sufficient to show that

$$\mathbb{P}\left[M(\mathbf{P}) \leq \alpha \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1]\right] \leq \frac{r}{r-1} n^{\frac{r-1}{r}}\alpha.$$

For this purpose, we can apply Theorem 2 of Vovk and Wang (2020) which states that for any Kolmogorov-Nagumo average with stricly decreasing continuous $\psi : [0,1] \to [-\infty,\infty]$ with $\psi(0) = \infty$ and $\psi(\alpha) \geq 0$ for any $\alpha \in (0,1)$:

$$\mathbb{P}\left[M(\mathbf{P}) \leq \alpha\right] \leq \inf_{v\in(0,\psi(\alpha)]} \frac{\int_{\psi(\alpha)-v}^{\psi(\alpha)+(n-1)v} \psi^{-1}(u)du}{v}.$$

The function $\psi(\cdot) = u^{-r}$ is strictly decreasing for $r = (1,\infty)$. Thus, we have

$$\inf_{v\in(0,\psi(\alpha)]} \frac{\int_{\psi(\alpha)-v}^{\psi(\alpha)+(n-1)v} \psi^{-1}(u)du}{v} = \frac{\int_0^{\alpha^{-r}n} u^{-1/r}du}{\alpha^{-r}} = \frac{r}{r-1} n^{\frac{r-1}{r}}\alpha,$$

which completes the first part of the proof.
To prove the second part of Theorem 2.1, we demonstrate

$$\limsup_{n\to\infty} \mathbb{P}\left[P_{r,n} \in C_{r,n} \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1]\right] = \alpha, \qquad \forall\ (p_1,...,p_n) \in \mathcal{P}^n.$$

Which is equivalent to proving the following condition:

$$\sup_{\mathbf{P}\in\mathcal{P}^n}\left\{\mathbb{P}\left[n^{1-1/r}\frac{r}{r-1}M(\mathbf{P})\leq\alpha\;\middle|\;\bigcap_{i\in\mathsf{N}}p_i\sim\mathsf{Un}[0,1]\right]\right\}=\alpha,\qquad as\ n\to\infty. \tag{A2}$$

for the set of all p-values $\mathcal{P}$. Define the left-continuous p-quantile of a random variable as:

$$q_p(X):=\sup\{x\in\mathbb{R}:\mathbb{P}[X\leq x]<p\},\quad p\in(0,1],$$

and the right-continuous p-quantile as $q_p^+(X)$. If

$$\inf\left\{q_\alpha\left(n^{1-1/r}\frac{r}{r-1}M(\mathbf{P})\right)\;\middle|\;\bigcap_{i\in\mathsf{N}}p_i\sim\mathsf{Un}[0,1]\right\}=\alpha,$$

can be demonstrated, condition (A2) is satisfied. Following Vovk and Wang (2020), we can use Theorem 4.6 of Bernard et al. (2014) according to which:

$$\begin{aligned}
&\inf\left\{q_\alpha\left(n^{1-1/r}\frac{r}{r-1}M(\mathbf{P})\right)\;\middle|\;\bigcap_{i\in\mathsf{N}}p_i\sim\mathsf{Un}[0,1]\right\}\\
&=n^{1-1/r}\frac{r}{r-1}\inf\left\{\alpha q_1\left(M(\mathbf{P})\right)\;\middle|\;\bigcap_{i\in\mathsf{N}}p_i\sim\mathsf{Un}[0,1]\right\}\\
&=n^{1-1/r}\frac{r}{r-1}\alpha\left(n^{-1}\sup\left\{q_0^+\left(\sum_{i=1}^n p_i^{-r}\right)\;\middle|\;\bigcap_{i\in\mathsf{N}}p_i\sim\mathsf{Un}[0,1]\right\}\right)^{-1/r}
\end{aligned}$$

By reformulating the final step of the proof of Proposition 5 Vovk and Wang (2020) it can be seen that:

$$\lim_{n\to\infty}\left(n^{-1}\sup\left\{q_0^+\left(\sum_{i=1}^n p_i^{-r}\right)\;\middle|\;\bigcap_{i\in\mathsf{N}}p_i\sim\mathsf{Un}[0,1]\right\}\right)n^{1-r}=\left(\frac{r}{r-1}\right)^r.$$

We therefore obtain:

$$\inf\left\{q_\alpha\left(n^{1-1/r}\frac{r}{r-1}M(\mathbf{P})\right)\;\middle|\;\bigcap_{i\in\mathsf{N}}p_i\sim\mathsf{Un}[0,1]\right\}=\alpha,\qquad \text{as } n\to\infty,$$

which verifies condition A2 and completes the proof. □

Proof of Theorem 2.2. The i-th p-value of a sub-test is defined as $p_i = 1 - F_i(\hat{s}_i)$ and without loss of generality we write $p_i = 1 - F_i(|\hat{s}_i|)$, where $\hat{s}_i$ is the test statistic associated with each of the $i = 1, ..., n$ sub-tests. Each $\hat{s}_i$ is distributed according to $F_i(\cdot)$ under the null hypothesis $\mathcal{H}_{i,0}$. Assume the set of sub-tests which reject their sub-hypothesis is $R = \{i \in \{1, ..., n\} : \mu_i \in \mathsf{M}_{i,A}\}$. The cardinality of the set, $1 \leq R_0 \leq n$, corresponds to the number of tests that reject their sub-hypothesis. The number of tests that do not reject is given by the cardinality of the set $S = \{i \in \{1, ..., n\} : \mu_i \in \mathsf{M}_{i,0}\}$, $S_0 = n - R_0$. To prove Theorem 2, it suffices to show that under the alternative

$$P_{r,n} = n^{-1} \left(\sum_{i=1}^{n} p_i^{-r} \right)^{1/r} = n^{-1} \left(\sum_{i=1}^{n} (1 - F_i(|\hat{s}_i|))^{-r} \right)^{1/r} \geq \frac{r}{\alpha(r-1)}.$$

As $x^{1/r}$ is a real concave function, the finite version of Jensen's inequality holds for the combination $\left(n^{-1} \sum_{i=1}^{n} p_i^{-r}\right)^{1/r}$. Thus, we have:

$$P_{r,n} = n^{\frac{1-r}{r}} \left(\frac{1}{n} \sum_{i=1}^{n} p_i^{-r} \right)^{1/r} \geq n^{\frac{1-2r}{r}} \sum_{i=1}^{n} (1 - F_i(|\hat{s}_i|))^{-1}$$

Sort the terms $(1 - F_i(|\hat{s}_i|))$ from small to large. Then

$$\begin{aligned} P_{r,n} &\geq n^{\frac{1-2r}{r}} \sum_{i=1}^{n} (1 - F_i(|\hat{s}_i|))^{-1} \\ &= n^{\frac{1-2r}{r}} \left\{ \sum_{i=1}^{R_0} (1 - F_i(|\hat{s}_i|))^{-1} + \sum_{j=R_0+1}^{n} \left(1 - F_j(|\hat{s}_j|)\right)^{-1} \right\} \end{aligned}$$

Under Assumption 2.3, we have that $(1 - F_j(|\hat{s}_j|)) \geq 1 - \alpha$, as $\hat{s}_j < c_j$, i.e the test statistics is smaller than the critical value of the test. Therefore,

$$= n^{\frac{1-2r}{r}} \left\{ \sum_{i=1}^{R_0} (1 - F_i(|\hat{s}_i|))^{-1} + O_p(1) \right\}$$

By Assumption 2.3 (ii), under the alternatives $\mathcal{H}_{i,A}$, for $i = 1, ..., R_0$ the sub-tests associated with $\hat{s}_i$ produce test statistics that are larger than any critical value c_i. Thus,

$$\begin{aligned} P_{r,n} &\geq n^{\frac{1-2r}{r}} \sum_{i=1}^{R_0} \lim_{\hat{s}_i \to \infty} (1 - F_i(|\hat{s}_i|))^{-1} + O_p(1) \\ &\to \infty. \end{aligned}$$

□

Proof of Proposition 2.1: The proof is based on several of the steps in the proof of Theorem 3 in Liu and Xie (2020) which we adapted for our test statistic. The i-th p-value of a sub-test is defined as $p_i = 1 - F_i(\hat{s}_i)$ and without loss of generality we write $p_i = 1 - F_i(|\hat{s}_i|)$, where $\hat{s}_i$ is the test statistic associated with each of the $i = 1, ..., n$ sub-tests. The vector of test statistics is distributed according to $\mathbf{S} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Assume the set of sub-tests which reject their sub-hypothesis is $R = \{i \in \{1, ..., n\} : \mu_i \in \mathsf{M}_{i,A}\}$. The cardinality of the set $R_0 = n^\gamma$, $\gamma \in [0, 0.5]$, corresponds to the number of tests that reject their sub-hypothesis. Under the alternative, $\mu_i = \sqrt{2\delta \log n}$ for all $i \in \mathsf{N}$ and $\delta > -2\sqrt{\gamma(2r-1)/r} + \gamma - 1/r + 2$.

To prove the Proposition, we show:

$$P_{r,n} = n^{-1} \left(\sum_{i=1}^{n} p_i^{-r} \right)^{1/r} = n^{-1} \left(\sum_{i=1}^{n} (1 - F_i(|\hat{s}_i|))^{-r} \right)^{1/r} \geq \frac{r}{\alpha(r-1)}.$$

By applying Jensen's inequality and proceeding as in the proof of Theorem 2.2 above, we obtain

$$\begin{aligned} P_{r,n} &\geq n^{\frac{1-2r}{r}} \left\{ \sum_{i=1}^{R_0} (1 - F_i(|\hat{s}_i|))^{-1} \right\} + o_p(1) \\ &\geq n^{\frac{1-2r}{r}} \left(1 - F_i(\max_{i \in R} |\hat{s}_i|) \right)^{-1} + O_p(1) \end{aligned}$$

Clearly, the last term is positive, so it suffices to show $(1 - F_i(\max_{i \in R} |\hat{s}_i|))^{-1} \to \infty$. The lower bound for a standard normal distribution is:

$$1 - F_i(x) \geq \frac{1}{\sqrt{2\pi}} \frac{x}{x^2 + 1} \exp\{-x^2/2\}$$

We have $\mathbf{S} = \boldsymbol{\mu} + \mathbf{Z}$ for $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and under Assumption 2.4 (i) [see Cai et al. (2014, Lemma 6)]:

$$\max_{i \in R} |\hat{s}_i| \geq \max_{i \in R} |Z_i + \mu_i| \geq \mu_1 + \max_{i \in R} |Z_i| = \mu_1 + \sqrt{2 \log R_0} + o_p(1)$$

Thus, we obtain:

$$\begin{aligned}
P_{r,n} \geq & n^{\frac{1-2r}{r}} \left(\exp\left\{ (\max_{i\in R} |\hat{s}_i|)^2/2 \right\} \left(\sqrt{2\pi} \max_{i\in R} |\hat{s}_i| + \frac{\sqrt{2\pi}}{\max_{i\in R} |\hat{s}_i|} \right) \right) \\
\geq & n^{\frac{1-2r}{r}} \exp\left\{ (\mu_1 + \sqrt{2\log R_0})^2/2 \right\} \\
& \times \left(\sqrt{2\pi}(\mu_1 + \sqrt{2\log R_0}) + \frac{\sqrt{2\pi}}{\mu_1 + \sqrt{2\log R_0}} \right) + o_p(1) \\
= & n^{\frac{1-2r}{r}} \exp\left\{ (\sqrt{2\delta\log n} + \sqrt{2\gamma\log n})^2/2 \right\} \\
& \times \left(\sqrt{2\pi}(\sqrt{2\delta\log n} + \sqrt{2\gamma\log n}) + \frac{\sqrt{2\pi}}{\sqrt{2\delta\log n} + \sqrt{2\gamma\log n}} \right) + o_p(1) \\
= & \frac{\sqrt{\pi} n^{2\sqrt{\gamma\delta}+\gamma+\delta+1/r-2}(2(2\sqrt{\gamma\delta}+\gamma+\delta)\log n + 1)}{(\sqrt{\gamma}+\sqrt{\delta})\log n} + o_p(1)
\end{aligned}$$

Since δ, γ, and r are real valued and $\delta > -2\sqrt{\gamma(2r-1)/r} + \gamma - 1/r + 2$, $\gamma \in [0, 0.5]$ and $r \in (1, \infty)$ we have $2\sqrt{\gamma\delta} + \gamma + \delta + 1/r > 2$. Therefore, we obtain $P_{r,n} \to \infty$ as $n \to \infty$. □

Proof of Corollary 2.1. The range $r \in (1, (2 - 2\sqrt{\gamma\delta} - \gamma - \delta)^{-1})$ can be derived by noting that Proposition 2.1 holds if $2\sqrt{\gamma\delta} + \gamma + \delta + 1/r > 2$ (see proof of Proposition 2.1). Since $r > 1$, we obtain the bounds $r \in (1, (2 - 2\sqrt{\gamma\delta} - \gamma - \delta)^{-1})$. The range $\delta \in (\gamma - 2\sqrt{\gamma} + 1, \gamma - 2\sqrt{2}\sqrt{\gamma} + 2)$ is obtained by solving $(2 - 2\sqrt{\gamma\delta} - \gamma - \delta)^{-1} > 1$ for δ conditional on $\gamma \in [0, 0.5]$. □

Proof of Proposition 2.2: A p-value can be defined as $\mathbb{P}[p_i \leq \epsilon] \leq \epsilon$ for $\epsilon \in (0, 1)$. From the proof of Theorem 2.1 it can be seen that

$$\begin{aligned}
& \mathbb{P}\left[n^{\frac{1-r}{r}} \left(\frac{1}{n} \sum_{i=1}^{n} p_i^{-r} \right)^{1/r} \geq \frac{r}{\alpha(r-1)} \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1] \right] \\
= & \mathbb{P}\left[n \frac{r}{r-1} \left(\sum_{i=1}^{n} p_i^{-r} \right)^{-1/r} \leq \alpha \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1] \right] \\
= & \mathbb{P}\left[\frac{r}{r-1} \frac{1}{P_{r,n}} \leq \alpha \,\middle|\, \bigcap_{i\in\mathsf{N}} p_i \sim \mathsf{Un}[0,1] \right] \\
\leq & \alpha.
\end{aligned}$$

As it is possible that $\frac{r}{r-1}\frac{1}{P_{r,n}} > 1$, a p-value for the test statistic is given by $h(P_{r,n}) = \frac{r}{r-1}\frac{1}{P_{r,n}} \wedge 1$. □

B.2 THE GIACOMINI-WHITE TEST IN A MULTIVARIATE SET-UP

This appendix provides theoretical details and derivations of the multivariate version of the GW test [see also Borup et al. (2022), who use GW in a multivariate setting to examine bond return predictability]. The conditional predictive ability test of GW is testing for serial correlation in the loss differential for the following reason: GW formulate the null hypothesis conditional on the time-t information set available to the forecaster, i.e. $\mathcal{H}_0 : \mathbb{E}[\Delta L_{i,R,t+1} \mid \mathscr{F}_t] = 0$. Therefore, the loss differential is a martingale difference sequence (MDS) under the null hypothesis. That is, the null can be rewritten as the following moment condition: $\mathcal{H}_0 : \mathbb{E}[\tilde{h}_t \Delta L_{i,R,t+1}] = 0$, for all $\mathscr{F}_t$-measurable functions $\tilde{h}_t$. GW suggest using a subset of such functions as a test function, h_t, which consists of lagged values of $\Delta L_{i,R,t+1}$. In this case one essentially tests for the presence of serial correlation in the loss differential. Under homoskedasticity, the null hypothesis becomes:

$$\Delta L_{R,t+1} = \mu + \beta \Delta L_{R,t} + \epsilon_{t+1}, \qquad \mathcal{H}_0 : \mu = 0 \cap \beta = 0.$$

Their framework is widely applicable and, in contrast to the unconditional approach, indicates which forecasting method may be more accurate in the future. The test preserves estimation error under the null hypothesis and does not require the estimation window, R, to converge to infinity. The test is consistent for a large out-of-sample window $p \to \infty$. In a multivariate setting, the test also incorporates serial correlation between loss differentials. This requires only a single additional assumption and is easily done with knowledge of multivariate MDS central limit theorems and is useful to exemplify the curse of dimensionality in a multivariate forecast evaluation environment.

Suppose that we are now interested in comparing $m \geq 1$ models against some benchmark and seek to forecast $\vartheta \geq 2$ variables, collected in the vector $\mathbf{Y}_t$. The quantity $n = m + \vartheta$ is equivalent to the number of sub-tests in the IU framework. This results in $i = 1, ..., m$ vectors each including ϑ loss differentials: $\Delta L_{i,R,t+\tau} = (\Delta L^{(1)}_{R,t+1}, ..., \Delta L^{(\vartheta)}_{R,t+1})'$. Now define the vector $\Delta \mathbf{L}_{R,t+1} = (\overline{\Delta L}_{1,R,t+1}, ..., \overline{\Delta L}_{m,R,t+1})'$ where $\overline{\Delta L}_{i,R,t+1} = \frac{1}{\vartheta} \sum_{i=1}^{\vartheta} L^{(i)}_{R,t+1}$. That is, $\Delta \mathbf{L}_{R,t+1}$ is an $m \times 1$ vector containing the cross-sectional averages for each model across ϑ variables. If we have $\vartheta = 1$ and $m > 1$, the framework reduces to $\mathbf{L}_{R,t+1} = (\Delta L_{1,R,t+1}, ..., \Delta L_{m,R,t+1})'$, i.e. a comparison which is not based on averages. Likewise, if $m = 1$ while $\vartheta > 1$, we can write $\mathbf{L}_{R,t+1} = (\Delta L^{(1)}_{R,t+1}, ..., \Delta L^{(\vartheta)}_{R,t+1})'$. If $m = 1$ and $\vartheta = 1$, we apply the univariate GW test. The rationale behind this set-up is the following: We seek to

investigate if the models have equal predictive ability relative to a benchmark across different variables, based on the information set available to the forecaster. If this is indeed the case, the expectation $\mathbb{E}[\Delta \mathbf{L}_{R,t+1} \mid \mathscr{F}_t]$ is a vector of zeros conditional on the σ-field $\mathscr{F}_t$, regardless of whether we use averages or not. The null hypothesis is therefore the following moment condition based on a $q \times 1$ vector of test function vectors, $\mathbf{h}_t$:

$$\mathcal{H}_0 : \mathbb{E}[\mathbf{h}_t \otimes \Delta \mathbf{L}_{R,t+1}] = 0. \tag{B.1}$$

Denote $\underset{qm\times 1}{\mathbf{Z}_{R,t+1}} = \mathbf{h}_t \otimes \Delta \mathbf{L}_{R,t+1}$, $\bar{\mathbf{Z}}_{R,m} = p^{-1}\sum_{t=R}^{T-1} \mathbf{Z}_{R,t+1}$, and $\underset{qm\times qm}{\hat{\mathbf{\Omega}}_m} = p^{-1}\sum_{t=R}^{T-1} \mathbf{Z}_{R,t+1}\mathbf{Z}'_{R,t+1}$. In the original GW test, we have $\vartheta = m = 1$, meaning the dimensions of the matrices and vectors in the multivariate extension are increased by m times. GW impose the following assumptions:

Assumption B.1. (i) All elements of $\{\mathbf{V}_t\}$ and $\{\mathbf{h}_t\}$ are α-mixing of size $-\frac{b}{(b-1)}, b > 1$ or ϕ-mixing of size $-\frac{b}{2b-1}, b \geq 1$, (ii) $\mathbb{E}\|\mathbf{Z}_{R,t+1,i}\|^{2(b+\delta)} < \infty$ for some $\delta > 0, i = 1, ..., qm$, and for all t, and (iii) $\hat{\mathbf{\Omega}}_m \equiv p^{-1}\sum_{t=R}^{T-1} \mathbb{E}[\mathbf{Z}_{R,t+1}\mathbf{Z}'_{R,t+1}]$ is uniformly positive definite.

In addition we require:

Assumption B.2. $\frac{qm}{p} \to 0$ as $p \to \infty$ and $qm \to \bar{c}$, where $\bar{c} \in \mathbb{N}^+$.

Assumption B.1 is imposed in GW; (i) quantifies the dependence of the data and allows for considerable heterogeneity, (ii) is a standard moment bound assumption that ensures no single observations dominate the asymptotic distribution, and (iii) ensures the covariance matrix of the test statistic is positive definite. GW emphasize that this matrix is not computed at the probability limits of the model parameters which ensures the test remains valid for nested models. A noteworthy difference to the univariate setting is that $\mathbf{\Omega}_m$ now also contains the covariances between the cross-sectional averages of the ϑ forecast loss differentials, computed across time periods. If we use lagged values of $\Delta \mathbf{L}_{R,t+1}$ as a test function, we can think of the test as one for both serial correlation and cross-sectional correlation between the averages of the loss differentials generated for each variable by each model. Finally, we add Assumption B.2 which implies the dimensionality of the loss differential and the test functions is small relative to the sample size. If we allow $qm/p \to \bar{c} \in (0, \infty]$ it is well known that $\hat{\mathbf{\Omega}}_m \overset{p}{\not\to} \mathbf{\Omega}_m$, which is a required condition for the MDS Central Limit Theorem (CLT). Furthermore,

if $qm > p$, $\hat{\mathbf{\Omega}}_m$ is non-invertible. The multivariate GW test statistic looks as follows:

$$T^h_{R,m} = p\bar{\mathbf{Z}}'_{R,m}\hat{\mathbf{\Omega}}_m^{-1}\bar{\mathbf{Z}}_{R,m}. \tag{B.2}$$

The dimensions of the vectors and the covariance matrix of this statistic are larger compared to the univariate version as both are a multiple m of the dimension of the test function. What is more, the matrix $\hat{\mathbf{\Omega}}_m^{-1}$ includes the covariance between loss differentials, the test is also evaluating dependence in the cross-section of forecasts, rather than only serial correlation. The Wald-type test statistic remains a scalar, wherefore it is true that:

Theorem B.1. *Suppose Assumptions B.1-B.2 hold. Then, under* $\mathcal{H}_0$, $T^h_{R,p} \xrightarrow{d} \chi^2_{qm}$ *as* $p \to \infty$.

Proof. To prove this Theorem, we modify the proof in Giacomini and White (2006) slightly in order to show that a Martingale Difference Sequence (MDS) CLT can still be applied to the test statistic. We first demonstrate that $\{\Delta\mathbf{L}_{R,t+1}\}$ is α-mixing of size $\frac{-b}{b-1}, b > 1$ or ϕ-mixing of size $\frac{-b}{2b-1}, b \geq 1$. White and Domowitz (1984, Lemma 2.1) show that if a variable $L_t = \psi(\mathbf{V}_t, \mathbf{V}_{t-1}, ..., \mathbf{V}_{t-\tau})$ is $\mathscr{F}_t$-measurable onto $\mathbb{R}^v$ where τ and v are finite integers, and $\mathbf{V}_t$ is defined as above, L_t is mixing of the same size as $\mathbf{V}_t$. Consider first the case where $\vartheta > 1$ and $m > 1$. By Assumption B.1, $\mathbf{V}_t$ is mixing, hence the same is true for $\Delta L^{(i)}_{r,t+1}$ and $\overline{\Delta L}_{i,R,t+\tau} = \frac{1}{\vartheta}\sum_{i=1}^{\vartheta}\Delta L^{(i)}_{R,t+\tau}$. Consequently, $\Delta\mathbf{L}_{R,t+1} = (\overline{\Delta L}_{1,R,t+\tau}, ..., \overline{\Delta L}_{m,R,t+\tau})'$ is mixing of the same size as $\mathbf{V}_t$. As $\mathbf{Z}_{R,t+1} = \psi(\mathbf{h}_t, \mathbf{V}_t, ..., \mathbf{V}_{t-\tau})$, it is also mixing of the same size as $\mathbf{V}_t$ and so is $\mathbf{Z}_{R,t+1}\mathbf{Z}'_{R,t+1}$. It is easy to see that the same is true if $m > 1$ and $\vartheta = 1$ such that $\Delta\mathbf{L}_{R,t+1} = (\Delta L_{1,R,t+1}, ..., \Delta L_{m,R,t+1})'$ and if $m = 1$ while $\vartheta > 1$ where $\Delta\mathbf{L}_{R,t+1} = (\Delta L^{(1)}_{R,t+1}, ..., \Delta L^{(\vartheta)}_{R,t+1})'$. The remainder of the proof is equivalent to the proof of Theorem 1 in Giacomini and White (2006). □

Analogously to GW, we can specify the alternative $\mathcal{H}_{A,h} = \mathbb{E}[\bar{\mathbf{Z}}'_{R,p}\bar{\mathbf{Z}}_{R,p}] \geq \delta > 0$ and obtain the following result:

Theorem B.2. *Suppose Assumptions B.1-B.2 hold. Then under* $\mathcal{H}_{A,h}$, $\mathbb{P}[T^h_{R,p} > c] \to 1$ *for any constant* $c \in \mathbb{R}$ *as* $p \to \infty$.

Proof. Having established in the previous proof that $\{\mathbf{Z}_{R,t+1}\}$ is mixing of the same size as $\mathbf{V}_t$ and $\hat{\mathbf{\Omega}}_m \xrightarrow{p} \mathbf{\Omega}_m$ by Assumptions B.1 (ii) and B.2, the proof of Theorem 2 in Giacomini and White (2006) is still valid here: All the conditions of White (1994, Theorem 8.13) are satisfied and $\mathbb{P}[T^h_{r,p} > c] \to 1$ for any $c \in \mathbb{R}$ as $p \to \infty$. □

As in the univariate case, if $\Delta \mathbf{L}_{R,t+1}$ is correlated with elements in $\mathscr{F}_t$ that are not included in $\mathbf{h}_t$, the test will have no power. That is, if one excludes such elements, the test may incorrectly fail to reject a false null hypothesis. If one follows the suggestion of GW and uses lagged values of $\Delta \mathbf{L}_{R,t+1}$ in the test function, then $q \geq m$, hence $qm/p \geq m^2/p$. Therefore, Assumption B.2 may already be violated for small values of m, unless p is large. Similar issues are encountered for the MP test. That is, the covariance matrices used in both tests will only be consistent when m is small. Hence, they quickly encounter inconsistency problems as m increases. Vice-verse, tests that rely on $m \to \infty$ in the presence of cross-sectional dependence are inconsistent in a small m environment.

APPENDIX C

CHAPTER 3

C.1 PROOFS

To demonstrate the Propositions in the paper, we proceed from a general homogeneous Kalman filter and, without loss of generality, focus on the univariate case $n = 1$. For $t = 1, ..., T$, the corresponding state space model is

$$\begin{aligned} y_t &= \alpha_t + \varepsilon_t \\ \alpha_t &= \alpha_{t-1} + \eta_t \end{aligned}$$

where the error terms have variances denoted by σ_η and σ_ϵ. The Kalman filter recursions are

$$\begin{aligned} f_t &= p_{t|t-1} + 1 & a_{t|t} &= a_{t|t-1} + \epsilon_t k_t \\ k_t &= p_{t|t-1}/f_t & a_{t+1|t} &= a_{t|t} \\ \epsilon_t &= y_t - a_{t|t-1} & v_{t|t} &= (1-k_t)v_{t|t-1} \\ p_{t+1|t} &= (1-k_t)p_{t|t-1} + \tilde{q} & v_{t+1|t} &= p_{t+1|t}\tilde{\sigma}_\varepsilon \end{aligned}$$

Define $a_{t+1} := a_{t+1|t}$ and let

$$\Psi_{t,j} := \begin{cases} \prod_{i=j}^{t}(1-k_i) & \text{if } t > j \\ 1 & \text{if } t = j \end{cases}$$

We can define the prediction error of the filter as

$$\begin{aligned} x_{t+1} &\equiv \alpha_{t+1} - a_{t+1|t} \\ &= \alpha_t + \eta_t - a_{t|t-1} - k_t(\Delta L_t - a_{t|t-1}) \\ &= \alpha_t + \eta_t - a_{t|t-1} - k_t(\alpha_t + \epsilon_t - a_{t|t-1}) \\ &= (1 - k_t)x_t + \eta_t - k_t\epsilon_t \\ &= \Psi(t,0)x_0 + \sum_{i=1}^{t} \Psi_{t,i}(\eta_i - k_i\varepsilon_i) \end{aligned}$$

Under the null hypothesis of equal predictive ability, $\alpha_{t+1} = 0$. Therefore, the Kalman filter prediction can be written as:

$$a_{t+1|t} = \sum_{i=0}^{t} \Upsilon_{ti}$$

where $\Upsilon_{ti} := \Psi_{t,i}(k_i\varepsilon_i - \eta_i)$ for $i > 1$ and $\Upsilon_{t0} := \Psi(t,0)x_0$. That is, the Kalman filter prediction can be written as the sum of η_t, ε_t. Under Assumption 3.2-A, both disturbances are *i.i.d.* Gaussian as well as independent from one another. If the starting values of the Kalman filter are chosen as $a_0 = y_0$, then $x_0 = 0$. Thus, by the closure property of Gaussian distributions under linear transformations, the Kalman filter predictions are normally distributed, with mean zeros and variance $v_{t+1|t}$ under the null hypothesis (3.2). It is easy to verify that this generalises to the multivariate case (3.9).

Consequently, for the test statistics in (3.23), (3.24), and (3.25) we obtain the following distributions for the pointwise CPA test

$$\begin{aligned} &S_{i,t+\tau|t} \sim \mathcal{N}(0,1), \\ &S^M_{i,t+\tau|t} \sim \mathcal{N}(0,1) \qquad \text{for all } i \in \mathbb{N}, \\ &\bar{S}_{t+\tau|t} \sim \chi^2_n. \end{aligned} \tag{C.1}$$

Proposition 3.1, 2-A, and 2-B follow accordingly.

The Kalman smoother equations are

$$\begin{aligned} \alpha_{t-1|T} &= \alpha_{t|t-1} + J_{t-1}\left(\alpha_{t|T} - \alpha_{t|t-1}\right), \\ v_{t-1|T} &= J_{t-1}v_{t|t-1} + J^2_{t-1}\left(v_{t|T} - v_{t|t-1}\right), \end{aligned}$$

with $\alpha_{T|T} = \alpha_{T|T-1}$. By the closure property of the Gaussian distribution under linear transformation, the Kalman smoother estimates are also normally distributed. Therefore,

$$\begin{aligned} &S_{i,t+\tau|T} \sim \mathcal{N}(0,1), \\ &S^{M}_{i,t+\tau|T} \sim \mathcal{N}(0,1) \qquad \text{for all } i \in \mathsf{N}, \\ &\bar{S}_{t+\tau|T} \sim \chi^2_n, \end{aligned} \tag{C.2}$$

and Proposition 3.3, 2-A, and 2-B follow.

C.2 RESULTS UNDER NON-NORMALITY

The theory underlying the Kalman filter is derived under the Gaussian assumption and, nonetheless, has proved to be successful in many scenarios. Under the assumption that the noise terms are *i.i.d*, no general (asymptotic) results for the distribution of the filtered and smoothed states are a available. However, the Kalman filter will still be the optimal linear estimator for the states. If the distribution of the error terms is known, Durbin and Koopman (2012, Section 9.3) can be used. In this section, we explore how relaxing the normality assumption (Assumption 3.2-A) to i.i.d. impacts the results of our test.

C.2.1 *Theoretical Results*

If $\sigma_\eta = 0$, i.e. the states are constant across time, the Kalman filter equations become:

$$\begin{aligned} p_{t+1} &= 1/t \\ a_{t+1} &= \frac{t-1}{t} a_t + \frac{1}{t} y_t \\ v_{t+1} &= \sigma_\varepsilon / t \end{aligned}$$

So

$$\begin{aligned} a_{t+1} &= \sum_{i=0}^{t} Y_{ti} \\ Y_{t0} &= \left(\frac{t-1}{t}\right)^t \mu_0 \\ Y_{ti} &= \sum_{j=i}^{t} \left(\frac{t-1}{t}\right)^{t-j} \frac{1}{t} \eta_{i-1} + \left(\frac{t-1}{t}\right)^{t-i} \frac{1}{t} \varepsilon_i \end{aligned}$$

In the general case of the previous section above, the first two moments are

$$\mathbb{E}[Y_{ti}] = 0$$
$$\mathbb{E}[Y_{ti}^2] = (\Psi_{t,i})^2 (\sigma_\eta + k_i^2 \sigma_\epsilon) = (\Psi_{t,i})^2 (q + k_i^2)\sigma_\epsilon$$

Following the arguments above, asymptotically, we have $Y_{t0} = o_p(1)$ and

$$\mathbb{E}[Y_{ti}] = 0$$
$$\mathbb{E}[Y_{ti}^2] = \left(\frac{t-1}{t}\right)^{2(t-i)} \frac{1}{t^2}\sigma_\varepsilon$$

It is clear that $\mathbb{V}[a_{t+1}] \leq \sum_{i=0}^{t} \mathbb{E}[Y_{ti}^2]$. Thus

$$\frac{\sup_{0\leq i\leq T} \mathbb{E}[Y_{ti}^2]}{v_{t+1}} = \frac{1}{t} \to 0 \qquad \text{as } t \to \infty$$

meaning the Lindeberg condition is satisfied and

$$\frac{a_{T+1}}{\sqrt{v_{T+1}}} \xrightarrow{d} \mathcal{N}(0,1).$$

That is, the distribution of the last Kalman filter prediction converges to a normal distribution as $T \to \infty$. In fact, backwards substitution yields

$$\alpha_{t+1} = \frac{1}{t}\sum_{i=0}^{t} y_i.$$

We obtain $J_t = 1$ and thus

$$\alpha_{t|T} = \frac{1}{T}\sum_{i=0}^{T} y_i \qquad \text{for all } t = 1, ..., T.$$

Hence, if the variance of the state is constant, the final Kalman filter prediction is asymptotically normally distributed, and all smoothed states are asymptotically normal.

C.2.2 *Simulations*

In this section, we repeat the simulations in the main part of the paper using a t-distribution with $\nu = 5$ degrees of freedom to abstract from normality.

C.2.2.1 *Size*

Table C.1 reports the results of the size simulations in the main text using a t-distribution. As can be seen, the empirical size is virtually identical to the size under normality.

C.2.2.2 *Power*

Figures C.1, C.2, and C.3 display the power simmulation results in the main text under t-distributed innovations. Interestingly, looking at Figure C.1, the Giacomini and Rossi (2010) test appears to suffer in a greater reduction in power than our tests, compared to the normally distributed scenarios. However, their power, too, is somewhat diminished in the second simulation. Altogether, our tests still exhibit high power across different simulations for all values of n under consideration, with only minor differences to the normally distributed case.

Table C.1: Average Size, t-distribution

		MC 1	MC 2				MC 3			
	p/r:	0	50	100	150	200	50	100	150	200
Panel A: $n = 1$										
$S_{t+1\|t}$	50	0.031	0.031	0.034	0.035	0.036	0.03	0.033	0.034	0.036
	100	0.03	0.028	0.032	0.033	0.034	0.025	0.031	0.032	0.033
	150	0.029	0.024	0.028	0.033	0.034	0.024	0.028	0.029	0.034
	200	0.03	0.023	0.023	0.03	0.031	0.023	0.027	0.031	0.031
$S_{t+1\|T}$	50	0.054	0.046	0.05	0.052	0.053	0.044	0.049	0.046	0.05
	100	0.046	0.045	0.046	0.049	0.046	0.034	0.041	0.04	0.043
	150	0.044	0.034	0.043	0.047	0.048	0.034	0.038	0.038	0.043
	200	0.045	0.032	0.033	0.04	0.043	0.03	0.036	0.04	0.04
Panel B: $n = 2, \gamma = 1$										
$S_{t+1\|t}$	50	0.028	0.033	0.04	0.037	0.038	0.034	0.043	0.041	0.041
	100	0.03	0.027	0.03	0.034	0.039	0.028	0.036	0.037	0.037
	150	0.028	0.022	0.029	0.032	0.034	0.027	0.032	0.037	0.038
	200	0.027	0.02	0.025	0.029	0.032	0.022	0.031	0.035	0.034
$S_{t+1\|T}$	50	0.051	0.051	0.06	0.051	0.054	0.048	0.06	0.06	0.052
	100	0.051	0.04	0.041	0.042	0.048	0.038	0.049	0.048	0.044
	150	0.042	0.03	0.042	0.04	0.039	0.036	0.04	0.049	0.047
	200	0.037	0.028	0.035	0.039	0.04	0.027	0.041	0.046	0.045
Panel C: $n = 2, \gamma \to \infty$										
$S_{t+1\|t}$	50	0.028	0.03	0.041	0.037	0.036	0.03	0.037	0.036	0.038
	100	0.028	0.025	0.029	0.035	0.036	0.026	0.029	0.036	0.035
	150	0.028	0.023	0.027	0.033	0.033	0.023	0.026	0.029	0.034
	200	0.028	0.021	0.027	0.029	0.031	0.022	0.025	0.028	0.031
$S_{t+1\|T}$	50	0.048	0.049	0.058	0.051	0.051	0.037	0.046	0.046	0.047
	100	0.045	0.038	0.042	0.043	0.046	0.034	0.039	0.04	0.041
	150	0.044	0.031	0.04	0.045	0.042	0.029	0.033	0.031	0.04
	200	0.039	0.026	0.037	0.037	0.036	0.027	0.034	0.035	0.035
Panel D: $n = 5, \gamma = 1$										
$S_{t+1\|t}$	50	0.036	0.037	0.051	0.048	0.048	0.047	0.061	0.061	0.059
	100	0.033	0.026	0.033	0.041	0.043	0.032	0.044	0.052	0.054
	150	0.03	0.021	0.03	0.033	0.04	0.026	0.035	0.043	0.05
	200	0.031	0.019	0.024	0.029	0.036	0.022	0.033	0.037	0.045
$S_{t+1\|T}$	50	0.071	0.061	0.077	0.07	0.068	0.068	0.087	0.079	0.08
	100	0.054	0.038	0.04	0.047	0.047	0.041	0.057	0.07	0.065
	150	0.043	0.028	0.041	0.039	0.044	0.035	0.044	0.056	0.054
	200	0.046	0.026	0.033	0.036	0.039	0.024	0.042	0.047	0.051
Panel E: $n = 5, \gamma \to \infty$										
$S_{t+1\|t}$	50	0.038	0.038	0.048	0.05	0.05	0.034	0.047	0.048	0.045
	100	0.031	0.026	0.033	0.041	0.044	0.025	0.033	0.041	0.043
	150	0.028	0.02	0.027	0.034	0.042	0.018	0.027	0.036	0.04
	200	0.028	0.018	0.025	0.029	0.035	0.019	0.024	0.031	0.036
$S_{t+1\|T}$	50	0.074	0.06	0.069	0.079	0.07	0.047	0.056	0.057	0.056
	100	0.048	0.043	0.044	0.052	0.049	0.029	0.037	0.041	0.038
	150	0.045	0.03	0.039	0.041	0.047	0.02	0.029	0.037	0.037
	200	0.041	0.021	0.034	0.035	0.042	0.018	0.026	0.032	0.03

Notes: The table reports the average size of the pointwise CPA test ($S_{t+1|t}$) and the pointwise TPA test ($S_{t+1|T}$). The column headers denote the size for the respective Monte-Carlo Simulation (1, 2, and 3). The rows correspond to the size for different values of p and the columns to the size for different values of r. Panel A reports the simulation results for $n = 1$, Panel B reports the results for the multivariate χ^2 test with two loss differentials ($n = 2$) and strong dependence ($\gamma = 1$). Panel C reports the results for $n = 2$ and independence between loss differentials ($\gamma = 10^8$), Panel D for $n = 5$ variables and strong dependence, and Panel E for $n = 5$ independent variables.

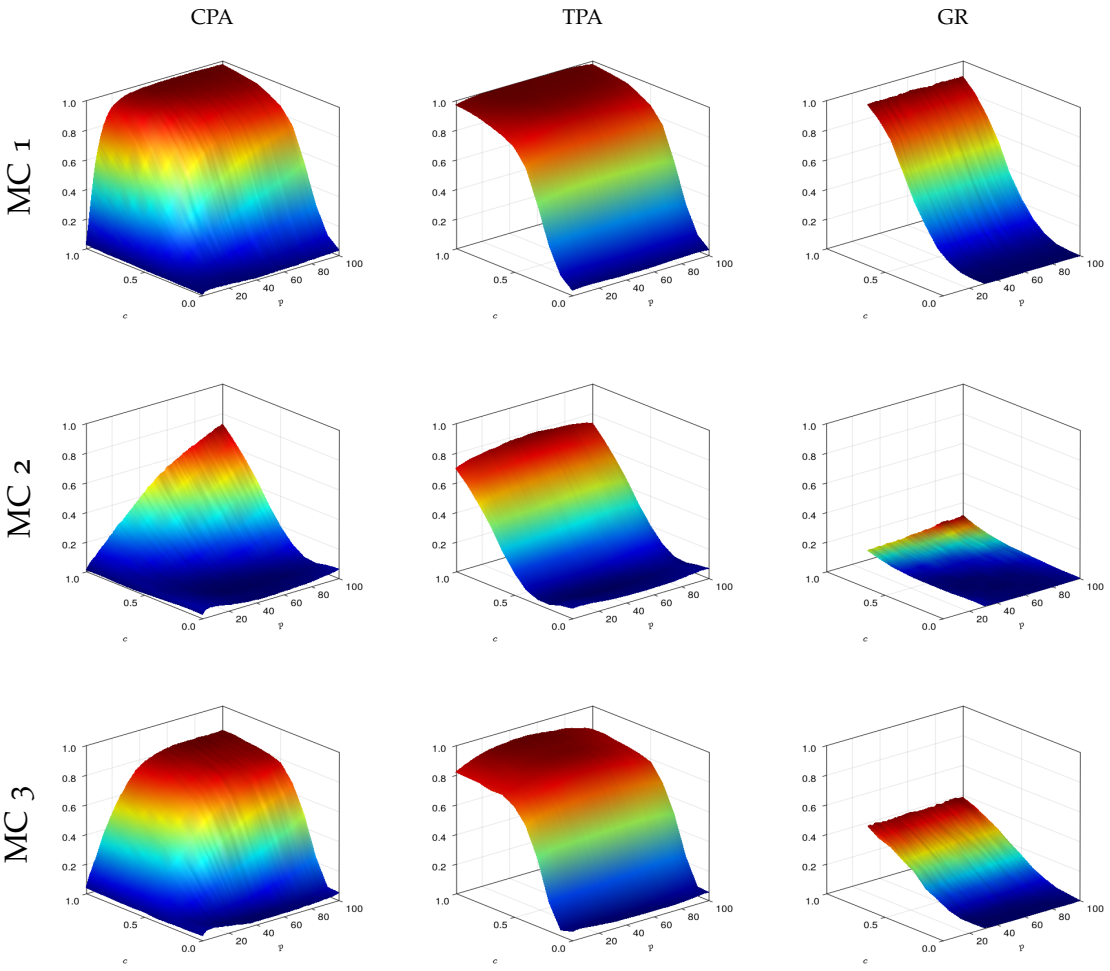


Figure C.1: Power Surface, t-distribution $n = 1$

Note: The columns of the figure display the power surface of the CPA and TPA tests as well as the power of the Giacomini and Rossi (2010) fluctuation test at each point in time for $n = 1$ variable. The rows correspond to the respective Monte-Carlo Simulation. For each figure, the x-axis corresponds to the time periods, the y-axis to the different values of c, and the z-axis to the power of the tests.

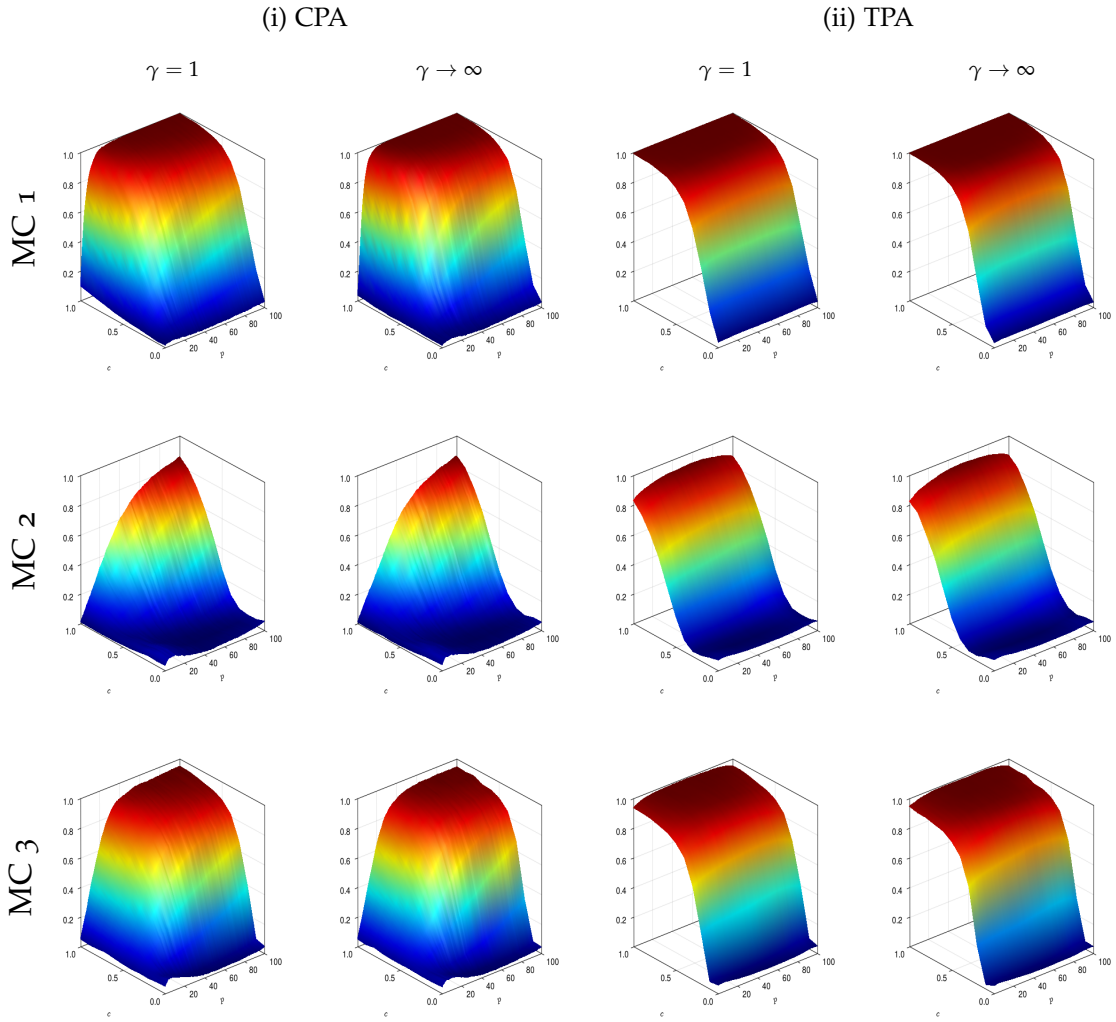


Figure C.2: Power Surface, t-distribution, $n = 2$

Note: The columns of the figure display the power surface of the CPA test in Panel (i), and the power of the TPA test in Panel (ii) for $n = 2$ variables. In each Panel, the first column contains the high-dependence case and the second column the low-dependence case. The rows correspond to the respective Monte-Carlo Simulation. For each figure, the x-axis corresponds to the time periods, the y-axis to the different values of c and the z-axis to the power of the tests.

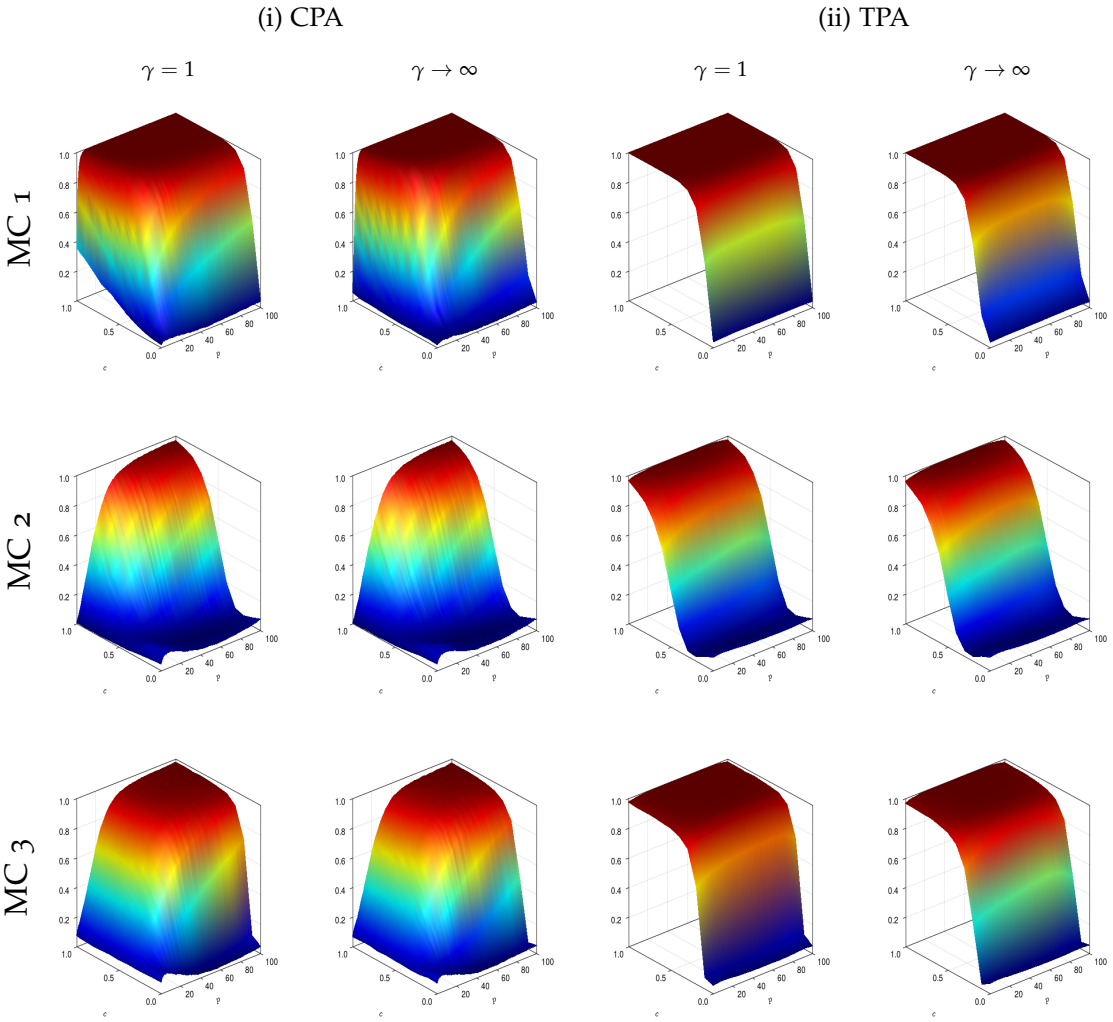


Figure C.3: Power Surface, t-distribution, $n = 5$

Note: The columns of the figure display the power surface of the CPA test in Panel (i), and the power of the TPA test in Panel (ii) for $n = 5$ variables. In each Panel, the first column contains the high-dependence case and the second column the low-dependence case. The rows correspond to the respective Monte-Carlo Simulation. For each figure, the x-axis corresponds to the time periods, the y-axis to the different values of c and the z-axis to the power of the tests.