DISCUSSION PAPER SERIES

DP15217

COMPARING FORECAST PERFORMANCE WITH STATE DEPENDENCE

Florens Odendahl, Barbara Rossi and Tatevik Sekhposyan

MONETARY ECONOMICS AND FLUCTUATIONS

ISSN 0265-8003

COMPARING FORECAST PERFORMANCE WITH STATE DEPENDENCE

Florens Odendahl, Barbara Rossi and Tatevik Sekhposyan

Discussion Paper DP15217
Published 28 August 2020
Submitted 26 August 2020

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

COMPARING FORECAST PERFORMANCE WITH STATE DEPENDENCE

Abstract

We propose a novel forecast comparison methodology to evaluate models' relative forecasting performance when the latter is a state-dependent function of economic variables. In our bench¬mark case, the relative forecasting performance, measured by the forecast loss differential, is modeled via a threshold model. Importantly, we allow the threshold that triggers the switch from one state to the next to be unknown, leading to a non-standard test statistic due to the presence of a nuisance parameter. Existing tests either assume a constant out-of-sample forecast performance or use non-parametric techniques robust to time-variation; consequently, they may lack power against state-dependent predictability. Importantly, our approach is applicable to point forecasts as well as predictive densities. Monte Carlo results suggest that our proposed test statistics perform well in finite samples and have better power than existing tests in selecting the best forecasting model in the presence of state dependence. Our test statistics uncover "pockets of predictability" in U.S. equity premia forecasts; the pockets are a state-dependent function of stock market volatility. Models using economic predictors perform significantly worse than a simple mean forecast in periods of high volatility, but, in periods of low volatility, the use of economic predictors may lead to small forecast improvements.

JEL Classification: C52, C53, G17

Keywords: State Dependence, Forecast evaluation, Pockets of Predictability

Florens Odendahl - florens.odendahl@banque-france.fr
Banque de France

Barbara Rossi - barbara.rossi@upf.edu
ICREA-Universitat Pompeu Fabra, Univ. Pompeu Fabra, ICREA, Barcelona School of Economics, CREI and CEPR

Tatevik Sekhposyan - tatevik.sekhposyan@gmail.com
Texas A&M

Acknowledgements
The views expressed herein are those of the authors and should not be attributed to the Banque de France or the Eurosystem. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 615608) and the Spanish Ministry of Economy and Competitiveness, Grant ECO2015-68136-P and FEDER, UE. The Barcelona GSE acknowledges financial support from the Spanish ministry of the Economy and Competitiveness through the Severo Ochoa Programme for Centers of Excellence in R&D (SEV-2015-0063).

Comparing Forecast Performance with State Dependence

Florens Odendahl[1], Barbara Rossi[2], and Tatevik Sekhposyan[3]

[1]*Banque de France*[*]
[2]*ICREA-UPF, Barcelona GSE and CREI*[†]
[3]*Texas A&M University*[‡]

June 17, 2020

Abstract

We propose a novel forecast comparison methodology to evaluate models' relative forecasting performance when the latter is a state-dependent function of economic variables. In our benchmark case, the relative forecasting performance, measured by the forecast loss differential, is modeled via a threshold model. Importantly, we allow the threshold that triggers the switch from one state to the next to be unknown, leading to a non-standard test statistic due to the presence of a nuisance parameter. Existing tests either assume a constant out-of-sample forecast performance or use non-parametric techniques robust to time-variation; consequently, they may lack power against state-dependent predictability. Importantly, our approach is applicable to point forecasts as well as predictive densities. Monte Carlo results suggest that our proposed test statistics perform well in finite samples and have better power than existing tests in selecting the best forecasting model in the presence of state dependence. Our test statistics uncover "pockets of predictability" in U.S. equity premia forecasts; the pockets are a state-dependent function of stock market volatility. Models using economic predictors perform significantly worse than a simple mean forecast in periods of high volatility, but, in periods of low volatility, the use of economic predictors may lead to small forecast improvements.

Keywords: State Dependence, Forecast Evaluation, Predictive Ability Testing, Pockets of Predictability.
JEL codes: C52, C53, G17.

[*] Address: 31 rue Croix des Petits Champs, 75001 Paris, France. Tel.: +33-1429-26306, email: florens.odendahl@banque-france.fr.

[†] Address: c/Ramon Trias Fargas 25/27, Barcelona 08005, Spain; tel.: +34-93-542-1655; email: barbara.rossi@upf.edu.

[‡] Address: 3060 Allan Building, 4228 TAMU, College Station, TX 77843, USA; tel.: +1-979-862-8857; email: tsekhposyan@tamu.edu. The views expressed herein are those of the authors and should not be attributed to the Banque de France or the Eurosystem. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 615608) and the Spanish Ministry of Economy and Competitiveness, Grant ECO2015-68136-P and FEDER, UE. The Barcelona GSE acknowledges financial support from the Spanish ministry of the Economy and Competitiveness through the Severo Ochoa Programme for Centers of Excellence in R&D (SEV-2015-0063).

1 Introduction

In practice, decision-makers face an abundance of candidate models to produce predictions and, starting with Diebold and Mariano (1995) and West (1996), the literature has proposed a variety of forecast comparison tests to guide forecasters in choosing the model. However, usually, no single model emerges as the best overall; typically forecasting performances are prone to instabilities or depend on the sample. One possible explanation is that the economic mechanisms that generate the data are unstable such that a given model is better in some periods and worse in others, resulting in a relative forecasting performance that is state-dependent, or, more generally, non-linear.

We, therefore, propose a new forecast comparison test that has power against the alternative of state dependence in competing models' relative forecasting performance. The state dependence is assumed to take the parametric form of a threshold model, i.e. the relative forecasting performance is a nonlinear function of an economic observable and a respective threshold. Importantly, we allow the value of the threshold to be unknown and estimated alongside the testing procedure. Existing tests either focus on constant relative out-of-sample performance (Giacomini and White, 2006) or use non-parametric techniques to detect time-varying deviations from equal performance (Giacomini and Rossi, 2010; Amisano and Giacomini, 2007); both approaches may lack power against the alternative of parametric state dependence.

Our paper is the first to model state dependence in the form of a threshold model directly on the relative forecasting performance. While Hansen's (1996) test detects non-linearities *in-sample*, our test instead addresses forecasters' need to test whether the *out-of-sample* forecasting performance of competing models is equal against the alternative that it might be unequal and state-dependent. Testing in the presence of an unknown threshold requires non-standard statistics, since the nuisance parameter (the threshold) is present only under the alternative; therefore, the standard Wald, Likelihood ratio, and Lagrange multiplier tests do not have the usual asymptotic chi-square distribution (Davies, 1977, 1987). While in some cases there might be an economic justification for selecting a specific threshold value and treat it as known, this is not generally the case, and allowing for an unknown threshold makes our approach broadly applicable. In detail, differently from Hansen (1996): (i) we apply the threshold model to the relative predictive performance, measured by the forecast loss differential, and (ii) we test for a zero expected forecast performance differential, while Hansen (1996) leaves the expected value unspecified under the null hypothesis. Consequently, under the alternative, we jointly test whether the out-of-sample average relative forecasting performance is different from zero as well as whether it is state-dependent.

The are several reasons why we apply a threshold model on the loss differentials instead of on the variable that is forecasted directly. First, our approach allows the forecaster to impose the null hypothesis directly on the object of interest, namely the loss differential. Second, the procedure applies to both the case when the forecast model is known as well as the case when it is not known. The latter case is important when considering survey forecasts, e.g. when comparing the widely used Survey of Professional Forecasters (SPF) or Greenbook projections. Third, in some cases, to produce forecasts with a threshold model the indicator variable might have to be predicted as well, which complicates the forecasting procedure and might make the threshold model unattractive.

Our paper contributes to the recent literature on forecast comparison tests (Diebold and Mariano, 1995; West, 1996; Clark and McCracken, 2001; Clark and West, 2006, 2007; Giacomini and White, 2006; Giacomini and Rossi, 2010). In particular, Giacomini and White (2006) (GW henceforth) showed the validity of the asymptotic Normal distribution for the out-of-sample equal predictive ability test proposed by Diebold and Mariano (1995) (DM henceforth) when the underlying forecasting models are estimated using a rolling window estimation scheme and the data satisfies some mixing properties.[1] Following their framework, our testing procedure similarly relies on a rolling window estimation scheme to preserve the parameter estimation error asymptotically. Hence, we compare forecasting methods rather than forecasting models. However, while GW focus on the null hypothesis of an equal out-of-sample predictive ability, on average, *unconditionally* or *conditionally* on economic variables, our test allows for state dependence of the conditioning variables, i.e. we test for deviations from the null hypothesis in sub-samples identified by state variables, and therefore our approach is more general. Importantly, we do not require the conditioning variable itself to explain the relative forecasting performance but only to indicate the state, i.e. the magnitude of the superior predictive ability within a regime can be independent of the conditioning variable. Our paper is also related to Giacomini and Rossi (2010), who allow the forecast performance to be prone to instabilities, using a non-parametric time-variation approach based on the rolling window estimation of a local GW test. As a result, their test has good power against smooth and persistent changes but, as we show, it might lack power against the discrete and weakly dependent switches of a threshold model.

We demonstrate the usefulness of our methodology in comparing models that predict U.S. equity premia from 1940 to 2017. As noted in Paye and Timmermann (1995) and Rapach and Wohar (2006), financial return predictability is typically time-varying and appears only in sub-samples.[2] Instabilities in forecasting performances in other financial variables are widespread as well: Paye and Timmermann (2006), for instance, cannot reject the presence of structural breaks in stock return predictive regressions and Rossi (2006, 2013) finds similar results for exchange rate returns. As summarized in Timmermann (2008), "... there appear to be pockets in time where there is modest evidence of local predictability; (...) the best forecasting method can be expected to vary over time, and there are likely to be periods of model breakdown where no approach seems to work".

We indeed find evidence of state dependence in stock market return predictability and show the usefulness of our test statistic for detecting pockets of predictability. Furthermore, our approach can shed light on which factors create such pockets of predictability, which so far, have been unknown. More in detail, our benchmark model is an in-sample mean, re-estimated in real-time in rolling windows, whereas the competitor models use the financial variables from Goyal and Welch's (2008) comprehensive dataset of predictors. We find evidence of state dependence in the relative forecasting performance, where the state dependence is a function of uncertainty, measured by stock market volatility: in periods of high stock market volatility, the economic model tends to underperform relative to the model with financial predictors. However, in periods of low stock market volatility, economic models using either long-term government bond yields

[1] Hereafter, we refer to the DM test under the conditions of Giacomini and White (2006) as the GW test.

[2] See Goyal and Welch (2003, 2008) for a related discussion.

or the spread as predictors lead to small, positive forecast improvements. On the other hand, the GW and Fluctuation tests cannot reject the null hypothesis of equal forecasting ability and would fail to uncover such pockets of predictability.[3]

The paper is organized as follows. Section 2 formalizes our null hypothesis, introduces our test statistics, and describes the challenges that arise when testing for state dependence in relative forecasting performance. Section 3 evaluates size and power of our proposed procedure in finite samples via Monte Carlo simulations, and Section 4 investigates the existence of pockets of predictability in financial data. Section 5 concludes.

2 Testing for State-Dependence: Methodology

We first describe the model and the null hypothesis. Then, we introduce the necessary notation, the technical assumptions, and the test statistic in Section 2.2.

2.1 The General Framework

Let $\widehat{f}^{(1)}_{t+h|t}(A_t, A_{t-1}, ..., A_{t-R+1}; \widehat{\beta}^{(1)}_{t,R})$ and $\widehat{f}^{(2)}_{t+h|t}(A_t, A_{t-1}, ..., A_{t-R+1}; \widehat{\beta}^{(2)}_{t,R})$ denote two measurable functions, which provide the forecasts of two competing models, labeled (1) and (2), where t denotes the forecast origin, h denotes the forecast horizon, and the vector of stochastic processes $A_t = (Y_t, Z_t)$ contains the variable of interest Y_t and the column vector of predictors Z_t. In turn, $\widehat{\beta}^{(i)}_{t,R}$ denotes the vector of estimated parameters at time t of model (i) using a rolling window estimation scheme of size $R \leq \bar{R} < \infty$ and data $A_t, ..., A_{t-R+1}$.[4] Henceforth, we simply write $\widehat{f}^{(1)}_{t+h|t}$ and $\widehat{f}^{(2)}_{t+h|t}$. Importantly, note that the function $\widehat{f}^{(i)}_{t+h|t}$ can denote either a point or a density forecast.

Let $L_{t+h|t}(Y_{t+h}, \widehat{f}^{(i)}_{t+h|t})$ denote a loss function, which evaluates the prediction $\widehat{f}^{(i)}_{t+h|t}$ of Y_{t+h}. The loss functions we allow for are quite general and encompass the Mean Squared Forecast Error (MSFE), asymmetric losses (such as the lin-lin loss), as well as the log score and Continuous Rank Probability Score (CRPS) for density forecasts. We define the loss differential as

$$\Delta L_{t+h|t} \equiv L_{t+h|t}(Y_{t+h}, \widehat{f}^{(1)}_{t+h|t}) - L_{t+h|t}(Y_{t+h}, \widehat{f}^{(2)}_{t+h|t}). \tag{1}$$

Note that the loss differential is a function of the estimated parameters $\widehat{\beta}^{(i)}_{t,R}$ and the rolling window size R. As we assume the parameters are estimated over a rolling and finite window size, the loss differential compares forecasting methods rather than forecasting models.

We allow the loss differential to evolve over time according to a nonlinear model (Teräsvirta, 2006):

$$\Delta L_{t+h|t} = \Psi(X_t, S_t; \varphi) + u_{t+h}, \tag{2}$$

where X_t and S_t are explanatory variables, φ is a vector of parameters, u_t is an error term and Ψ is allowed to be a nonlinear function. The nonlinear model in equation (2) encompasses several interesting cases for $\Psi(X_t, S_t; \varphi) = X_t'\mu + X_t'\theta \cdot G(S_t; \gamma)$. In particular, it includes threshold models (Tong, 1990), where $G(S_t; \gamma) = \mathbb{1}(S_t \geq \gamma)$. In the latter, the parameter changes if S_t is above the

[3] Note that the forecasting gains using the financial predictors are small and that any large deviations from equal predictive ability in favor of the economic models would imply strong violations of the rational expectations hypothesis.

[4] The window size R is assumed to be the same across the two models for notational convenience only.

threshold γ. This is the model we consider in detail in our paper. Section 2.6 provides a discussion of Markov-switching as well as exponential and logistic Smooth Transition Autoregressive Models (STAR) models, which are alternative ways to model nonlinearities.

That is, we aim at testing two forecasting models' equal predictive ability while being able to detect possible additive nonlinearities in the form of a threshold model. For this purpose, we let the loss differential depend on a vector of economic observables X_t, a threshold γ and a threshold indicator variable S_t, such that:

$$\Delta L_{t+h|t} = X_t'\mu + X_t'\theta \cdot \mathbb{1}(S_t \leq \gamma) + u_{t+h}. \tag{3}$$

In eq. (3), μ and θ denote the parameters of interest, the vector X_t is a k_1 dimensional column vector that denotes economic observables and a constant, S_t denotes the economic observable that introduces the state dependence, γ denotes the unknown threshold, $\mathbb{1}(\cdot)$ denotes the indicator function and u_t is the error term.[5] In Appendix A we discuss the possibility of several candidate variables for S_t and how to extend the testing procedure to account for that. For the remainder, S_t is assumed to be a scalar. Potential serial correlation can be accounted for by including lags of $\Delta L_{t+h|t}$, which are allowed, but not required, to also be a function of the threshold indicator. S_t is a stochastic process and assumed to be continuous. The timing t of X_t and S_t is merely a notational convention, and both variables are allowed to represent economic observables that realize in $t+h$.

Our null hypothesis of equal predictive ability at each point in time is:

$$\mathrm{E}\big(\Delta L_{t+h|t}\big) = 0 \;\; \forall t, \tag{4}$$

versus the alternative

$$\mathrm{E}\big(\Delta L_{t+h|t}|X_t, S_t\big) = X_t'\mu + X_t'\theta \cdot \mathbb{1}(S_t \leq \gamma). \tag{5}$$

Under eq. (3), the null and alternative hypothesis involve μ and θ and become $\mathrm{H}_0 : \mu = \theta = 0$ and $\mathrm{H}_\mathrm{A} : \mu \neq 0,\ \theta \neq 0$ respectively. Note that the null hypothesis defined in eq. (4) holds conditionally on X_t and S_t, and, therefore, by the law of iterated expectations, also *unconditionally*. Our test has power against either μ or θ or both jointly deviating from zero under the alternative, i.e. either a constant non-equal predictive ability or a state-dependent (or nonlinear) predictive ability or both.[6] Importantly, we allow the nuisance parameter γ to be unknown. Therefore, testing for the null hypothesis described in equation (4) is subject to the problem of a nuisance parameter that is present only under the alternative, which makes standard asymptotic inference invalid (Davies, 1977, 1987; Hansen, 1996).

Before describing our proposed test statistics, we want to emphasize two points. First, although the assumption of an unknown γ comes at the cost of non-standard inference, it brings the large benefit that it allows the researcher to test over a range of threshold values, instead of having to choose an arbitrary value. This is particularly important in practice because an ad-hoc choice for γ can be detrimental to the power of detecting state dependence. In practice, we recommend

[5] Both X_t and S_t can also contain variables realized at $t+h$, depending on the specific economic relationship considered.

[6] Note that the case of $\mu = \theta \neq 0$ is a valid alternative and merely represents the joint presence of a non-equal and nonlinear non-equal predictive ability.

to formulate γ in terms of the empirical distribution function $\Xi_n(\cdot)$ of Ξ_t such that the indicator becomes $\mathbb{1}(\Xi_n(S_t) < \gamma)$, with $\gamma \in \Gamma = [0,1]$ and $\Xi_n^{-1}(\gamma)$ provides the threshold in units of S_t (Hansen, 1996). This is particularly useful when implementing the model in statistical programs, as it allows formulating a unit-free grid for γ. Following Hansen (1996) and others, we restrict γ to be away from the boundaries and choose, for instance, $\Gamma = [0.15, 0.85]$.

Second, we want to introduce the following specification of (3), which is of particular interest in the forecast comparison case, as it specifies state dependence that is a function solely of S_t and does not depend on any additional observables X_t:

$$\Delta L_{t+h|t} = \mu + \theta \cdot \mathbb{1}(S_t \leq \gamma) + u_{t+h}. \tag{6}$$

The specification in eq. (6) is of special interest as it encompasses the standard Diebold and Mariano (1995) and unconditional Giacomini and White (2006) tests for equal predictive ability as special cases, and, unlike the latter, is capable of detecting periods of unequal performance that depend on S_t.

2.2 Test Statistics and Assumptions

The threshold value, γ, is an element of the compact set Γ. Let $Q_t(\gamma)$ be a k dimensional column vector that contains the explanatory variables of the threshold model as described in equation (3), i.e. $Q_t(\gamma) = \left[X_t', \left(X_t \cdot \mathbb{1}(S_t \leq \gamma)\right)'\right]'$, and let $Q_t = \sup_{\gamma\in\Gamma} |Q_t(\gamma)|$. Let $\widehat{\psi}(\gamma) = [\widehat{\mu}(\gamma)', \widehat{\theta}(\gamma)']'$ denote the vector of OLS parameter estimates under the alternative, and let $\widehat{u}_{t+h} = \Delta L_{t+h|t} - Q_t(\gamma)'\widehat{\psi}(\gamma)$ denote the error term under the alternative. The score under the alternative is then given by $\widehat{s}_{t+h}(\gamma) = Q_t(\gamma)\widehat{u}_{t+h}(\gamma)$. Let H_r denote a restriction matrix that corresponds to the null hypothesis defined in eq. (4). For instance, for the model described in eq. (6) we have that $H_r = I_2$, where I_2 is a two-dimensional identity matrix. R denotes the rolling window estimation size, h the forecast horizon, T the total sample size, and $P = T - R - h$ denotes the out-of-sample size, i.e. the number of observations of $\Delta L_{t+h|t}$. Let $\widehat{V}_P(\gamma) = \frac{1}{P}\sum_{t=R}^{T-h} \widehat{s}_{t+h}(\gamma)\widehat{s}_{t+h}(\gamma)'$ denote the variance-covariance matrix of the score, let $V(\gamma) = \mathrm{E}\left(s_{t+h}(\gamma)\, s_{t+h}(\gamma)'\right)$ be finite and positive definite for $s_{t+h}(\gamma) = Q_t(\gamma)u_{t+h}$, and let $\widehat{V}_P^*(\gamma) = M_P(\gamma,\gamma)^{-1}\widehat{V}_P(\gamma)M_P(\gamma,\gamma)^{-1}$ be the robust estimator of the variance-covariance matrix of $\widehat{\psi}$, with $M_P(\gamma,\gamma) = \frac{1}{P}\sum_{t=R}^{T-h} Q_t(\gamma)Q_t(\gamma)'$, and $M(\gamma_1,\gamma_2) = \mathrm{E}\big(Q_t(\gamma_1)Q_t(\gamma_2)'\big)$.

We consider the following test statistics, based on Hansen (1996) and Andrews and Ploberger (1994), which we collectively refer to as the DM$^{\text{NL}}$ test:

$$\text{DM}^{\text{NL}}\text{:}\ \ g_\Gamma(W_P) = \begin{cases} \sup_{\gamma\in\Gamma} W_P(\gamma) & (\text{“sup-W”}) \\ \int_\Gamma W_P(\gamma)\mathrm{d}w(\gamma) & (\text{“ave-W”}) \\ \ln\left(\int_\Gamma \exp\left(\frac{1}{2}W_P(\gamma)\right)\mathrm{d}w(\gamma)\right) & (\text{“exp-W”}) \end{cases} \tag{7}$$

where $w(\gamma)$ is a weighting function[7] over $\gamma \in \Gamma$, ln$(\cdot)$ denotes the natural logarithm and $W_P(\gamma)$ is defined as

$$W_P(\gamma) = P\widehat{\psi}(\gamma)'H_r\left[H_r'\widehat{V}_P^*(\gamma)H_r\right]^{-1}H_r'\widehat{\psi}(\gamma). \tag{8}$$

Henceforth, we let $g_\Gamma\left(W_P(\gamma)\right)$ denote either of the three above mentioned functions, i.e. sup-W, exp-W,

[7] Throughout the paper we use an equal weighting, i.e. $w(\gamma) = \gamma$.

and ave-W. We derive the limiting distribution of $\mathrm{DM}^{\mathrm{NL}}$ under the following assumptions:

Assumption A1 *(i) (A_t, X_t, S_t) is strictly stationary and absolutely regular with mixing coefficients $\eta(m) = O(m^{-\delta})$ for some $\delta > v/(v-1)$ and $v > 1$. (ii) The estimation window size (R) is finite and the estimation scheme is a rolling window estimation.*

Assumption A2 *For $r > v > 1$, $E|Q_t|^{4r} < \infty$, $E|u_t|^{4r} < \infty$, and $\inf_{\gamma\in\Gamma} det\left(M(\gamma,\gamma)\right) > 0$.*

Assumption A3 *Let $r > v$ and let S_t have a density function $g(S_t)$ such that $\sup_{s\in\mathbb{R}} g(s) = \bar{g} < \infty$.*

Assumption A4 *$f^{(i)}_{t+h|t}(.)$ is a measurable function of leads and lags of A_t, for $i = 1, 2$.*

A1 limits the dependence and time-variation allowed in the loss differential under the null. A2 ensures that the explanatory variables in (3) have at least $4r + \varepsilon$, $\varepsilon > 0$, finite moments and that the variance-covariance matrix of X_t and S_t is non-singular for all γ. In A3, we follow Theorem 3 of Hansen (1996) and assume that the density function of S_t is bounded. A4 is an assumption on the functional form of the point forecast itself, and ensures measurability of $\Delta L_{t+h|h}$.

2.3 Point Forecasts

In the case of point forecasts, the asymptotic distribution in eq. (7) can be described as follows.

Proposition 1 *(Point forecast comparison) Let $g_\Gamma(W_p)$ be either $\sup_{\gamma\in\Gamma} W_P(\gamma)$, $\int_\Gamma W_P(\gamma)dw(\gamma)$ or $ln\left(\int_\Gamma exp(\frac{1}{2}W_P(\gamma))dw(\gamma)\right)$, where Γ is compact and $W_P(\gamma) = P\widehat{\psi}(\gamma)' H_r \left[H_r' \widehat{V}^*_P(\gamma) H_r\right]^{-1} H_r' \widehat{\psi}(\gamma)$, and $\widehat{\psi}(\gamma) = \left[\widehat{\mu}(\gamma)', \widehat{\theta}(\gamma)'\right]'$ is estimated from eq. (3). Then, under A1 to A4 and H_0 defined in eq. (4): $E\left(\Delta L_{t+h|t}\right) = 0$ for all $t = R+h, ..., T$ and*

$$\lim_{P\to\infty} g_\Gamma\left(W_P(\gamma)\right) \underset{d}{\to} g_\Gamma\left(\chi^2(\gamma)\right), \tag{9}$$

where $\chi^2(\gamma)$ is a chi-square distribution with degrees of freedom rank(H_r), and $g_\Gamma\left(\chi^2(\gamma)\right)$ can be completely characterized by its covariance kernel $K(\gamma_1, \gamma_2)$. The test rejects H_0 defined in eq. (4) when $g_\Gamma\left(W_P(\gamma)\right) > \phi_\alpha$, where ϕ_α is the critical value (for a nominal size of α) that can be simulated according to Algorithm 1 below.

Proof of Proposition 1. According to Theorem 3.49 in White (2001), if A_t is α-mixing (or strong mixing) with coefficients of size $-\delta$, $\delta > 0$, so is any measurable function of a finite number of leads and lags of A_t. Under A1(i), $\delta > \nu/(\nu-1)$ and $\nu > 1$, such that $\delta > 0$, and as absolute regularity implies α-mixing, A1(i) implies that any measurable function of a finite number of leads and lags of A_t is absolutely regular. By A1(ii) and A4, $\Delta L_{t+h|t}$ and X_t are measurable functions of a finite number of leads and lags of A_t, and thus, under A1(i), they are absolutely regular with coefficients of size $-\delta$. Consequently, $(\Delta L_{t+h|t}, X_t)$ is strictly stationary and absolutely regular with mixing coefficients $\eta(m) = O\left(m^{-\delta}\right)$ for some $\delta > \nu/(\nu-1)$ and $\nu > 1$, and thus satisfying assumption 1(i) in Hansen (1996). Further, A2 implies that assumptions 1(ii)-(iii) in Hansen (1996) hold. Thus, under A1 to A4, the result follows from Theorem 3 of Hansen (1996).

2.4 Density Forecasts

In the case of density forecasts, let $\widehat{f}^{(i)}_{t+h|t}(Y_{t+h}; A_t, A_{t-1}, ..., A_{t-R+1}, \widehat{\beta}^{(i)}_{t,R})$ denote the h-step-ahead predictive density at t, with $i = 1, 2$, and $L_{t+h|t}(Y_{t+h}, f^{(i)}_{t+h|t})$ denote the loss function. For example, the log score case implies $L_{t+h|t}(Y_{t+h}, f^{(i)}_{t+h|t}) = \log\big(f^{(i)}_{t+h|t}(Y_{t+h})\big)$ and $\Delta L_{t+h|t} \equiv \log\big(f^{(1)}_{t+h|t}(Y_{t+h})\big) - \log\big(f^{(2)}_{t+h|t}(Y_{t+h})\big)$.[8] Furthermore, let $\widehat{f}^{(i)}_{t+h|t}(Y_{t+h})$ denote the estimate of $f^{(i)}_{t+h|t}(Y_{t+h})$. For density forecasts, we specify the following additional assumption.

Assumption A5 *$f^{(i)}_{t+h|t}$ is a measurable function of leads and lags of A_t, for $i = 1, 2$.*

A5 ensures that $\Delta L_{t+h|t}$ is a measurable function of A_t, and refers to the functional form of the density itself. Then, the asymptotic distribution of the DM$^{\text{NL}}$ test of equal forecasting performance in eq. (7) for density forecasts can be described as follows.

Proposition 2 *(Density forecast comparison) Let $g_\Gamma(W_p)$ be either $\sup_{\gamma\in\Gamma} W_P(\gamma)$, $\int_\Gamma W_P(\gamma)dw(\gamma)$ or $ln\big(\int_\Gamma exp(\frac{1}{2}W_P(\gamma))dw(\gamma)\big)$, where Γ is compact and $W_P(\gamma) = P\widehat{\psi}(\gamma)' H_r \left[H_r' \widehat{V}^*_P(\gamma) H_r\right]^{-1} H_r' \widehat{\psi}(\gamma)$, and $\widehat{\psi}(\gamma) = \left[\widehat{\mu}(\gamma)', \widehat{\theta}(\gamma)'\right]'$ is estimated from eq. (3). Then, under A1 to A3, A5, and H_0 defined in eq. (4): $E\left(\Delta L_{t+h|t}\right) = 0$ for all $t = R+h, ..., T$ and*

$$\lim_{P\to\infty} g_\Gamma\left(W_P\right) \underset{d}{\to} g_\Gamma\left(\chi^2\left(\gamma\right)\right). \tag{10}$$

As in Proposition 1, $\chi^2(\gamma)$ is a chi-square distribution with degrees of freedom rank(H_r), and $g_\Gamma\left(\chi^2(\gamma)\right)$ can be completely characterized by its covariance kernel $K(\gamma_1, \gamma_2)$. The test rejects H_0 defined eq. (4) when $g_\Gamma\left(W_P(\gamma)\right) > \phi_\alpha$, where ϕ_α is the critical value (for a nominal size of α) that can be simulated according Algorithm 1 described below.

Proof of Proposition 2. According to Theorem 3.49 in White (2001), if A_t is α-mixing (or strong mixing) with coefficients of size $-\delta$, $\delta > 0$, so is any measurable function of a finite number of leads and lags of A_t. Under A1(i), $\delta > \nu/(\nu-1)$ and $\nu > 1$, such that $\delta > 0$, and as absolute regularity implies α-mixing, A1(i) implies that any measurable function of a finite number of leads and lags of A_t is absolutely regular. By A1(ii) and A5, $\Delta L_{t+h|t}$ and X_t are measurable functions of a finite number of leads and lags of A_t, and thus under A1(i) they are absolutely regular with coefficients of size $-\delta$. Consequently, $(\Delta L_{t+h|t}, X_t)$ is strictly stationary and absolutely regular with mixing coefficients $\eta\left(m\right) = O\left(m^{-\delta}\right)$ for some $\delta > \nu/(\nu-1)$ and $\nu > 1$, and thus satisfying assumption 1(i) in Hansen (1996). Further, A2 implies that assumptions 1(ii)-(iii) in Hansen (1996) hold. Thus, under A1 to A3, and A5, the result follows from Theorem 3 of Hansen (1996).

Potential serial correlation in the error term in eq. (3), i.e. in u_{t+h}, can be controlled for by either including lags of $\Delta L_{t+h|t}$ or by explicitly modeling the time dependence in u_{t+h}.

[8] "log(.)" here denotes the natural logarithm.

2.5 Practial Implementation

The asymptotic distribution in eq. (24) is not nuisance parameter free and cannot be tabulated except for special cases.[9] Therefore, we follow Hansen (1996) to propose an algorithm that can be used to simulate the critical values and which we report here for the readers' convenience.

Simulation Algorithm 1 (Hansen, 1996) Let $\widehat{s}_{t+h}(\gamma), M(\gamma,\gamma), \widehat{V}_P^*(\gamma)$, and H_r be as defined in Section 2.2. Then, for each $j = 1, ..., J$ do the following steps:

1. Draw a set of standard Normal random variates $\{v_{tj}\}_{t=1}^P$;
   (a) Calculate $\widehat{\lambda}_P^j(\gamma) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} \widehat{s}_{t+h}(\gamma) v_{tj}$;
   (b) Using $\widehat{\lambda}_P^j(\gamma)$, calculate $W_P^j(\gamma) = \widehat{\lambda}_P^j(\gamma)' M(\gamma,\gamma)^{-1} H_r \left[H_r' \widehat{V}_P^*(\gamma) H_r\right]^{-1} H_r' M(\gamma,\gamma)^{-1} \widehat{\lambda}_P^j(\gamma)$;
   (c) Repeat (a)-(b) for all $\gamma \in \Gamma$;
2. Compute $W_P^j = g_\Gamma\left(W_P^j(\gamma)\right)$.

After J iterations, we obtain a set of $\{W_P^j\}_{j=1}^J$ draws from the asymptotic distribution, which we can use to construct critical values and p-values. In particular, the approximate p-value is given by $\widehat{p}(J) = \frac{1}{J}\sum_{j=1}^J \mathbb{1}(W_P > W_P^j)$, where W_P denotes the value of the test statistic computed using the actual data.

2.6 State Dependence via Markov Switching and STAR Models

An alternative modeling approach for state dependence are Markov switching models (Hamilton, 1989). Differently from the threshold model, the regime changes in the Markov switching model depend upon an unobserved (latent) Markov chain, S_t. Testing in the presence of Markov switching also requires non-standard statistics as it is subject to two problems. The first problem is again the presence of nuisance parameters that are only identified under the alternative; in this case, the state-to-state transition probabilities and the coefficients that switch. The second problem is that, under the null, the score with respect to the restricted parameters is identically zero, which violates the regularity conditions that are imposed to derive the asymptotic chi-square distribution of the finite dimensional LR (Wald, LM) statistic by a first-order approximation. Therefore, the procedure proposed in Hansen (1996), which deals with a nuisance parameter present only under the alternative, does not readily apply to the case of Markov switching models. Instead, Hansen (1992); Garcia (1998); Cho and White (2007); Carrasco et al. (2014a) and Qu and Zhuo (2017) provide a variety of solutions that address both problems.

We propose a test for equal predictive ability in the presence of Markov switching based on Carrasco et al. (2014) in Appendix B and investigate its size properties as well. However, the test, like all the tests for Markov switching listed above, relies crucially on a correctly specified distribution under the null hypothesis.[10] A misspecified likelihood under the null will generally lead to size distortions. For instance, consider the case where the true but unknown distribution is a Student's t with no Markov switching. The researcher assumes a Gaussian distribution with

[9] See Hansen (1996) for a discussion.

[10] Under the assumption of normality, the power of the test of Carrasco et al. (2014a) relies on serial correlation in the error terms, instead of other deviations from the distribution specified under the null.

no Markov switching under the null and a Markov switching model with regime dependent, conditional Gaussian distributions under the alternative. Then, despite the absence of Markov switching in the data generating process, the mixture property of the Markov switching model under the alternative may approximate the Student's t distribution better than the Gaussian model under the null hypothesis. Unreported results show that this leads to an over-rejection of the null hypothesis of no Markov switching.

While the assumption of Normality may be justified when applying tests for Markov switching models directly on economic observables, the distribution of a loss differential is generally unknown and may exhibit fatter tails than a Normal distribution (e.g. when using a quadratic loss). Consequently, the above-described problem is more severe in the case of forecast comparisons, and testing for Markov switching in this framework may be very sensitive to the choice of the parametric distribution. In contrast, and as outlined in Section 2.2, threshold models do not rely as heavily on the parametric assumptions on the error terms u_t, and testing is, therefore, more robust in practice.

Another popular non-linear model is the Smooth Transition Autoregressive (STAR) Model, using either a logistic or an exponential function. Luukkonen et al. (1988), Teräsvirta (1994) and Granger and Teräsvirta (1993), among others, have proposed testing procedures to test for linearity in the STAR model. However, since their testing procedures also rely on the assumption of Normality, they will suffer from a similar problem as Markov Switching models in the case of comparing forecasting performances.

3 Monte Carlo Simulation Analysis

We generate data according to two data generating processes (DGPs), and then simulate a point and also a density forecast comparison for each of the two DGPs. In particular, we consider the case of non-nested forecasting models under DGP1 and the case of nested forecasting models under DGP2. The two point forecast comparisons, applied to the data generated by DGP1 and DGP2 respectively, are labeled PF1 and PF2. The two density forecast comparisons, applied to the data generated by DGP1 and DGP2 respectively, are labeled DF1 and DF2. In both cases, the forecast horizon is one (h=1), and the number of Monte Carlo replications is 5,000. The total sample size, the rolling window estimation size, and the out-of-sample size are denoted by T, R and P respectively.

Point Forecast Comparison 1 (PF1):

The underlying data for PF1 is generated by

$$y_{t+h} = \nu + \delta_1 z_{t,1} + \delta_2 z_{t,2} + e_{t+h}, \tag{11}$$

where $\nu = \delta_1 = \delta_2 = 1$, $e_t \sim_{\text{iid}} \text{N}(0,1)$, $z_{t,1} \sim_{\text{iid}} \text{N}(0,1)$ and $z_{t,2} \sim_{\text{iid}} \text{N}(0,1)$. The parameter vector $\widehat{\beta}_t^{(j)} = [\widehat{\nu}_{t,j}, \widehat{\delta}_{t,j}]$ denotes the OLS estimator $\widehat{\beta}_t^{(j)} = \left(\sum_{i=t-R+1}^{t} z_{i-h}^{(j)\prime} z_{i-h}^{(j)}\right)^{-1} \sum_{i=t-R+1}^{t} z_{i-h}^{(j)\prime} y_i$, where $z_t^{(j)} = [1, z_{t,j}]$.The two point forecasts, both of which are misspecified, are denoted by: $\widehat{f}_{t+h|t}^{(1)} = z_t^{(1)} \widehat{\beta}_t^{(1)}$, and $\widehat{f}_{t+h|t}^{(2)} = z_t^{(2)} \widehat{\beta}_t^{(2)}$. As the misspecification of the two models is symmetric, it is straightforward to show that they have the same predictive ability in expectation. That is, the loss differential, given by

$$\Delta L_{t+h|t} = \left(y_{t+h} - \widehat{f}^{(1)}_{t+h|t}\right)^2 - \left(y_{t+1} - \widehat{f}^{(2)}_{t+h|t}\right)^2, \tag{12}$$

is zero in expectation: $\mathrm{E}(\Delta L_{t+h|t}) = 0$ for all $t = R+h, ..., T$.

Point Forecast Comparison 2 (PF2):
The underlying data for PF1 is generated by

$$y_t = \beta + e_t, \tag{13}$$

with $e_t \sim_{\text{iid}} \mathrm{N}(0,1)$ and β a constant parameter. Let $\widehat{\beta}_t = \frac{1}{R}\sum_{i=t-R+1}^{t} y_i$ denote the OLS estimate of β. The two point forecasts are $\widehat{f}^{(1)}_{t+h|t} = 0$, and $\widehat{f}^{(2)}_{t+h|t} = \widehat{\beta}_t$ respectively. For $\beta = \frac{1}{\sqrt{R}}$, the expected squared forecast error difference is zero in expectation, i.e. the loss differential

$$\Delta L_{t+h|t} = \left(y_{t+h} - \widehat{f}^{(1)}_{t+h|t}\right)^2 - \left(y_{t+h} - \widehat{f}^{(2)}_{t+h|t}\right)^2, \tag{14}$$

is zero in expectation: $\mathrm{E}(\Delta L_{t+h|t}) = 0$ for all $t = R+h, ..., T$.

Density Forecast Comparison 1 (DF1):
The data for DF1 is generated by the process in eq. (11). The two competing density forecasts are both based on a normal density, given by $\phi(x|\tau, \sigma^2)$, where x denotes the value at which the density is evaluated, τ denotes the conditional mean forecasts, and σ^2 the conditional variance forecast.[11] The two conditional means of the normal densities are the same as the point forecasts in PF1, i.e. $\widehat{\tau}^{(1)}_{t+h|t} = z^{(1)}_t \widehat{\beta}^{(1)}_t$, and $\widehat{\tau}^{(2)}_{t+h|t} = z^{(2)}_t \widehat{\beta}^{(2)}_t$, with $\widehat{\beta}^{(j)}_t = \left(\sum_{i=t-R+1}^{t} z^{(j)\prime}_{i-h} z^{(j)}_{i-h}\right)^{-1} \sum_{i=t-R+1}^{t} z^{(j)\prime}_{i-h} y_i$ and $z^{(j)}_t = [1, z_{t,j}]$. In turn, the variance forecasts is based on the in-sample estimate of the error variance: $\widehat{\sigma}^{2^{(j)}}_{t+h|t} = \frac{1}{R}\sum_{i=t-R+1}^{t}\left(y_i - z^{(j)}_{i-h}\widehat{\beta}^{(j)}_t\right)^2$. The two density forecasts, both of which are misspecified, are denoted by: $\widehat{f}^{(1)}_{t+h|t} = \phi\left(y_{t+h}|\widehat{\tau}^{(1)}_{t+h|t}, \widehat{\sigma}^{2^{(1)}}_{t+h|t}\right)$, and $\widehat{f}^{(2)}_{t+h|t} = \phi\left(y_{t+h}|\widehat{\tau}^{(2)}_{t+h|t}, \widehat{\sigma}^{2^{(2)}}_{t+h|t}\right)$. The loss differential is then given by

$$\Delta L_{t+h|t} = \log\left(\widehat{f}^{(1)}_{t+h|t}(y_{t+h})\right) - \log\left(\widehat{f}^{(2)}_{t+h|t}(y_{t+h})\right), \tag{15}$$

and is zero in expectation: $\mathrm{E}(\Delta L_{t+h|t}) = 0$ for all $t = R+h, ..., T$.

Density Forecast Comparison 2 (DF2):
The data for DF2 is generated by the process in eq. (13). The two competing density forecasts are again both based on a normal density. The two conditional means of the normal densities are the same as the point forecasts in PF2, i.e. $\widehat{\tau}^{(1)}_{t+h|t} = 0$, and $\widehat{\tau}^{(2)}_{t+h|t} = \widehat{\beta}_t$, with $\widehat{\beta}_t = \frac{1}{R}\sum_{i=t-R+1}^{t} y_i$. In turn, the variance forecasts is based on the in-sample estimate of the error variance: $\widehat{\sigma}^{2^{(1)}}_{t+h|t} = \frac{1}{R}\sum_{i=t-R+1}^{t} y_i^2$ and $\widehat{\sigma}^{2^{(2)}}_{t+h|t} = \frac{1}{R}\sum_{i=t-R+1}^{t}\left(y_i - \widehat{\beta}_t\right)^2$. The two density forecasts, both of which are misspecified, are denoted by: $\widehat{f}^{(1)}_{t+h|t} = \phi\left(y_{t+1}|\widehat{\tau}^{(1)}_{t+h|t}, \widehat{\sigma}^{2^{(1)}}_{t+h|t}\right)$, and $\widehat{f}^{(2)}_{t+h|t} = \phi\left(y_{t+1}|\widehat{\tau}^{(2)}_{t+h|t}, \widehat{\sigma}^{2^{(2)}}_{t+h|t}\right)$.

[11] We deviate from the standard notation of μ for the mean of a Normal density to not confuse the reader with the μ defined in eq. (3).

Then, the loss differential is given by

$$\Delta L_{t+h|t} = \log\left(\widehat{f}^{(1)}_{t+h|t}(y_{t+1})\right) - \log\left(\widehat{f}^{(2)}_{t+h|t}(y_{t+1})\right), \tag{16}$$

and is zero in expectation: $\mathrm{E}(\Delta L_{t+h|t}) = 0$ for all $t = R+h, ..., T$.

3.1 Size Results

We generate time series of $\Delta L_{t+h|t}$ as described in (12), (14), (15) and (16) for several values of R and P: $R = [25, 50, 100]$ and $P = [50, 100, 200, 1000]$. Then, we estimate the following model on the loss differential:

$$\Delta L_{t+h|t} = \mu + \theta \cdot \mathbb{1}\{S_t \leq \gamma\} + u_t, \tag{17}$$

where $S_t \sim_{\text{iid}} \mathrm{N}(0,1)$ and we treat γ as unknown.

Table 1 shows the point forecast results for the null hypothesis defined in eq. (4) for the three different test statistics: sup-W, exp-W and ave-W. Overall, the ave-W has the best size properties in the Monte Carlo study and delivers size results that are good for $P > 50$ and $R > 25$ for both the nested and the non-nested cases.

Table 1: Size Results for Threshold in Mean Model — Point Forecasts

Panel A. ave-W

	PF1								PF2							
	Size 5 %				Size 10 %				Size 5 %				Size 10 %			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.088	0.078	0.064	0.065	0.161	0.142	0.127	0.119	0.059	0.046	0.035	0.028	0.116	0.086	0.076	0.066
50	0.072	0.065	0.062	0.052	0.137	0.129	0.120	0.116	0.069	0.049	0.039	0.028	0.133	0.099	0.088	0.067
100	0.074	0.060	0.053	0.052	0.143	0.118	0.110	0.113	0.077	0.059	0.044	0.032	0.147	0.108	0.091	0.068

Panel B. exp-W

	PF1								PF2							
	Size 5 %				Size 10 %				Size 5 %				Size 10 %			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.106	0.079	0.064	0.061	0.181	0.148	0.124	0.117	0.089	0.051	0.044	0.036	0.150	0.096	0.087	0.075
50	0.093	0.069	0.063	0.054	0.163	0.133	0.123	0.114	0.101	0.062	0.044	0.032	0.169	0.119	0.093	0.074
100	0.093	0.067	0.057	0.053	0.168	0.125	0.118	0.108	0.109	0.065	0.049	0.034	0.187	0.124	0.097	0.078

Panel C. sup-W

	PF1								PF2							
	Size 5 %				Size 10 %				Size 5 %				Size 10 %			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.125	0.090	0.070	0.059	0.208	0.166	0.137	0.111	0.119	0.064	0.057	0.042	0.197	0.122	0.107	0.086
50	0.114	0.083	0.070	0.060	0.193	0.155	0.127	0.114	0.133	0.077	0.056	0.039	0.214	0.143	0.108	0.086
100	0.120	0.077	0.067	0.055	0.201	0.149	0.124	0.110	0.143	0.086	0.055	0.043	0.224	0.152	0.116	0.087

Note: The table displays empirical rejection frequencies of the null hypothesis $\mathrm{H}_0 : \mu = \theta = 0$ for the $\mathrm{DM}^{\mathrm{NL}}$ test for point forecasts evaluated with the MSFE loss function. Size 5% and 10% denote the nominal size. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. Panel A to C show the results for the three $\mathrm{DM}^{\mathrm{NL}}$ tests: the sup-W, exp-W and ave-W. The results are based on 5,000 MC replications.

The results of the exp-W test in Table 1 are similar to the ave-W; however, size distortions are slightly bigger in small samples than in the ave-W case. While the sup-W test works well in large samples ($P > 100$), it somewhat over-rejects in smaller samples. In the case of PF2 and for small samples, the under-rejections are not too surprising and mirror the results in Giacomini and White (2006)[12].

A similar picture emerges when looking at the results for density forecasts, given in Table 2. In particular, the empirical rejection frequencies are close to the nominal size for the ave-W and the exp-W test even for moderate sample sizes, such as $P > 50$ and $R > 25$. Again, there is a slight under-rejection for the case of nested models for large samples.

Table 2: Size Results for Threshold in Mean Model — Density Forecasts

Panel A. ave-W

	DF1								DF2							
	Size 5 %				Size 10 %				Size 5 %				Size 10 %			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.071	0.045	0.028	0.024	0.134	0.087	0.069	0.057	0.071	0.045	0.040	0.039	0.135	0.085	0.081	0.092
50	0.084	0.061	0.044	0.041	0.153	0.118	0.088	0.080	0.078	0.049	0.036	0.025	0.149	0.102	0.081	0.068
100	0.077	0.068	0.053	0.045	0.145	0.126	0.104	0.095	0.081	0.054	0.045	0.028	0.148	0.111	0.089	0.065

Panel B. exp-W

	DF1								DF2							
	Size 5 %				Size 10 %				Size 5 %				Size 10 %			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.105	0.058	0.039	0.030	0.170	0.109	0.083	0.064	0.127	0.070	0.057	0.041	0.193	0.123	0.102	0.091
50	0.114	0.073	0.053	0.044	0.188	0.128	0.101	0.088	0.124	0.067	0.050	0.034	0.195	0.127	0.099	0.076
100	0.104	0.080	0.058	0.048	0.180	0.136	0.112	0.096	0.122	0.074	0.055	0.033	0.189	0.134	0.105	0.077

Panel C. sup-W

	DF1								DF2							
	Size 5 %				Size 10 %				Size 5 %				Size 10 %			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.133	0.080	0.054	0.040	0.214	0.146	0.107	0.081	0.167	0.098	0.074	0.046	0.247	0.165	0.133	0.090
50	0.145	0.089	0.066	0.049	0.231	0.154	0.122	0.102	0.159	0.099	0.068	0.042	0.240	0.164	0.123	0.091
100	0.139	0.097	0.069	0.052	0.221	0.160	0.130	0.106	0.158	0.094	0.069	0.043	0.234	0.162	0.124	0.090

Note: The table displays empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test for density forecasts evaluated with the log score. Size 5% and 10% denote the nominal size. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. Panel A to C show the results for the three DM^{NL} tests: the sup-W, exp-W and ave-W. The results are based on 5,000 MC replications.

3.2 Power Results

In order to assess power, we specify three different alternatives for the loss differential defined in equations (12) and (14). The first alternative investigates the power of the proposed test statistics for detecting state dependence. The second alternative investigates the empirical rejection frequency when both $\mu \neq 0$ and $\theta \neq 0$. The third alternative investigates power against a constant deviation from the null of equal predictive ability, i.e. power to detect $\mu \neq 0$.

[12] Results are not reported here.

In order to conduct the power analysis we proceed as follows. Let $\Delta L^{(0)}_{t+h|t}$ be the loss differential obtained from one Monte Carlo draw of either (12) or (14), normalized by its sample standard deviation (to ensure that the magnitude of the alternative is constant relative to the variation in $\Delta L_{t+h|t}$). For all simulations we used $S_t \sim_{i.i.d.} \mathrm{N}(0,1)$ and $\gamma = 0$. In particular, we define the loss differential under the first alternative, Alternative (1), as

$$\Delta L^{(1)}_{t+h|t}(c) = \Delta L^{(0)}_{t+h|t} + \mu_c + \theta_c \cdot \mathbb{1}(s_t \leq \gamma), \tag{18}$$

where the s_t are realizations of the stochastic process S_t, and $c = 1, 2, ..., 14$ such that $\mu_1 = 0, \mu_2 = 0.085, \mu_2 = 0.170, ..., \mu_{14} = 1.10$, and $\theta_c = -2\mu_c$. Note that $\gamma = 0$ implies that $\mathrm{E}(S_t \leq \gamma) = \frac{1}{2}$. Therefore, it follows that $\mathrm{E}_t \Delta L^{(1)}_{t+1|t} = \mu_c + \mathrm{E}(S_t \leq \gamma)\theta_c = \mu_c - \frac{1}{2}2\mu_c = 0$, i.e. the overall sample has a zero mean and the magnitude of the regime switching coefficient is 0.17 times the standard deviation of $\Delta L^{(0)}_{t+h|t}$, and so forth. In the case where $c = 1$, $\mu_1 = \theta_1 = 0$ implies that the joint null, defined in equation (4), holds.

For Alternative (2), the values of μ_c are unchanged but $\theta_i = -\mu_i$, which implies that $\mathrm{E}_t \Delta L_{t+h|t} \neq 0$. In other words, Alternative (2) is a case where both state dependence and a constant deviation are present:

$$\Delta L^{(2)}_{t+h|t}(c) = \Delta L^{(0)}_{t+h|t} + \mu_c + \theta_c \cdot \mathbb{1}(s_t \leq \gamma). \tag{19}$$

Alternative (3) considers constant deviation from the null hypothesis, i.e. $\theta_c = 0 \; \forall c$:

$$\Delta L^{(3)}_{t+h|t}(c) = \Delta L^{(0)}_{t+h|t} + \mu_c, \tag{20}$$

with $\mu_1 = 0, \mu_2 = 0.085, \mu_2 = 0.170, ..., \mu_{14} = 1$.

Note that the s_t are re-drawn for each Monte Carlo iteration of each alternative; we suppressed the respective subscripts for notational convenience.

We then estimate the model in eq. (6), treating γ as unknown, and we test the null hypothesis defined in eq. (4) using the $\mathrm{DM}^{\mathrm{NL}}$ test defined in eq. (7). Figures 1 to 3 show the size-adjusted power results for the three different alternatives, defined in equations (18) to (20) for PF1.[13] We compare its performance with the Diebold and Mariano (1995) (DM) and the Giacomini and Rossi (2010) Fluctuation test. Figure 1 shows results for Alternative (1), i.e. state dependence without a constant deviation. As we can see, size-adjusted power increases quickly with the magnitude of the alternative for the sup-W, ave-W, and exp-W test. In turn, the DM and Fluctuation tests have no power to detect the lack of equal predictive ability arising from the state dependence in the relative forecasting performance, and their power remains flat around the nominal size.[14]

Figure 2 shows results for Alternative (2), i.e. the case of a constant deviation and state dependence. The sup-W, ave-W, and exp-W tests show again good size-adjusted power properties, and due to the presence of a constant deviation, the DM and Fluctuation rejection frequencies also increase as a function of the alternative's magnitude.

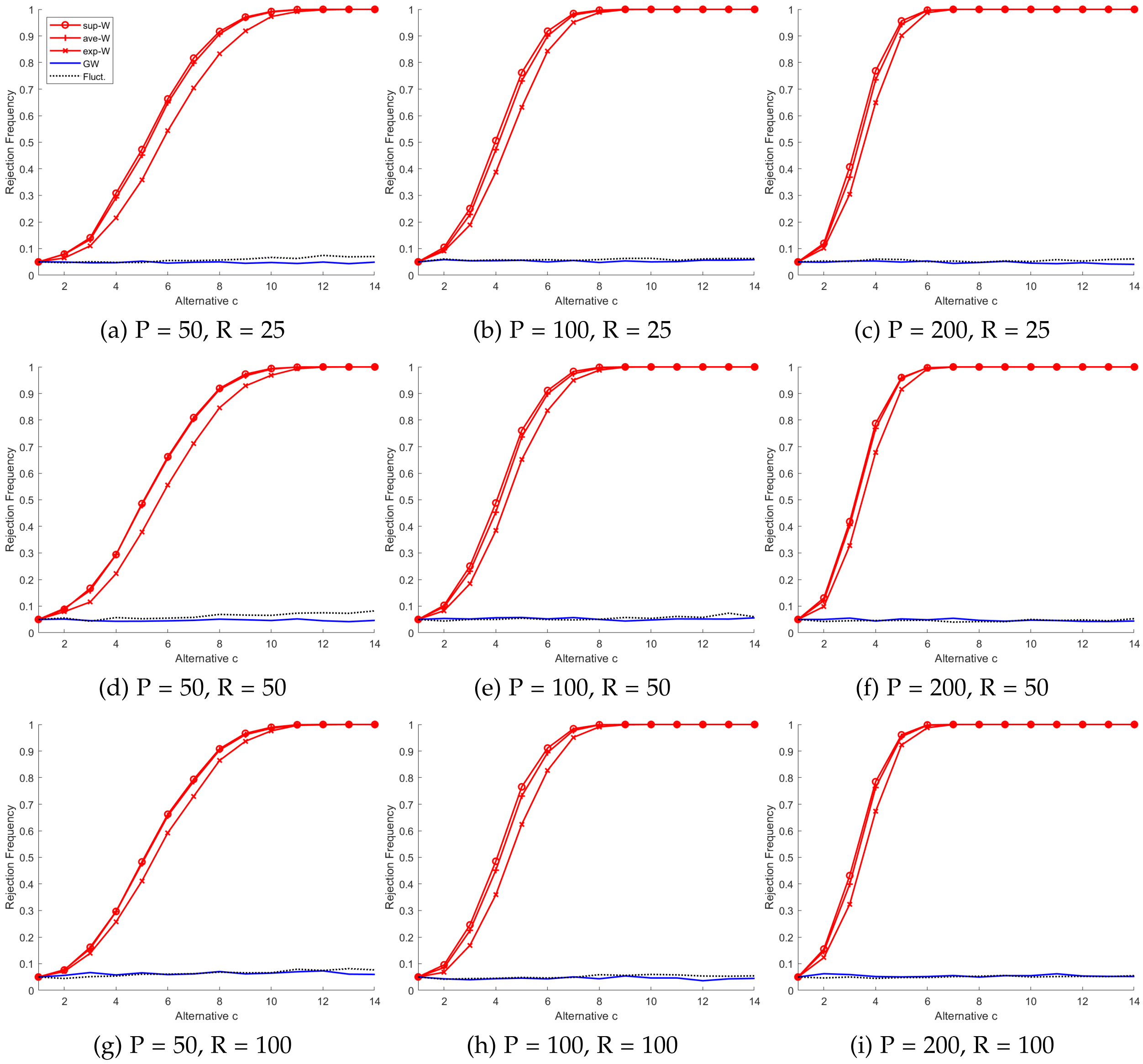
Figure 3 shows results for Alternative (3), i.e. a constant deviation without state-dependence. As expected, the DM test tends to be the most powerful test in this scenario; however, the

[13] Results for PF1, DF1 and DF2 are shown in Appendix C.

[14] Note that the Fluctuation test might have better power in cases where S_t is a persistent variable.

size-adjusted power of the sup-W, ave-W, and exp-W is very similar to that of the DM test.

Figure 1: Size-Adjusted Power Results for PF1, Alternative (1): State Dependence



(a) P = 50, R = 25

(b) P = 100, R = 25

(c) P = 200, R = 25

(d) P = 50, R = 50

(e) P = 100, R = 50

(f) P = 200, R = 50

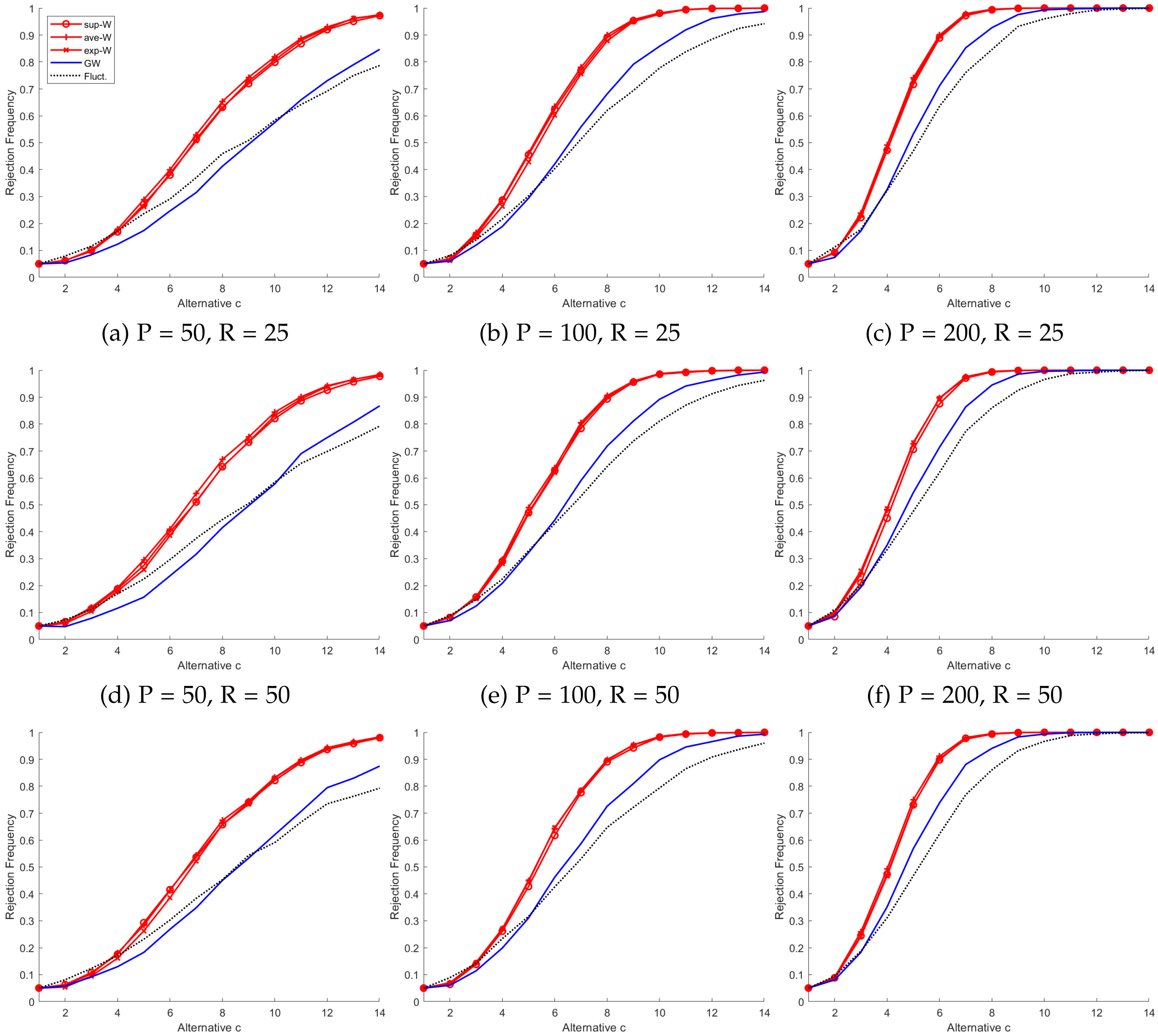
(g) P = 50, R = 100

(h) P = 100, R = 100

(i) P = 200, R = 100

Note: On the y-axis the figures displays size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test under Alternative (1) for point forecasts evaluated with the MSFE loss function. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. The solid lines with markers “o”, “x” and “+” display the sup-W, the exp-W, and the ave-W test results respectively. The dashed line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the solid line displays the results of the GW test. The nominal level is 5%. The results are based on 3,000 MC replications.

Figure 2: Size-Adjusted Power Results for PF1, Alternative (2): State Dependence and Constant Deviation



(a) P = 50, R = 25

(b) P = 100, R = 25

(c) P = 200, R = 25

(d) P = 50, R = 50

(e) P = 100, R = 50

(f) P = 200, R = 50

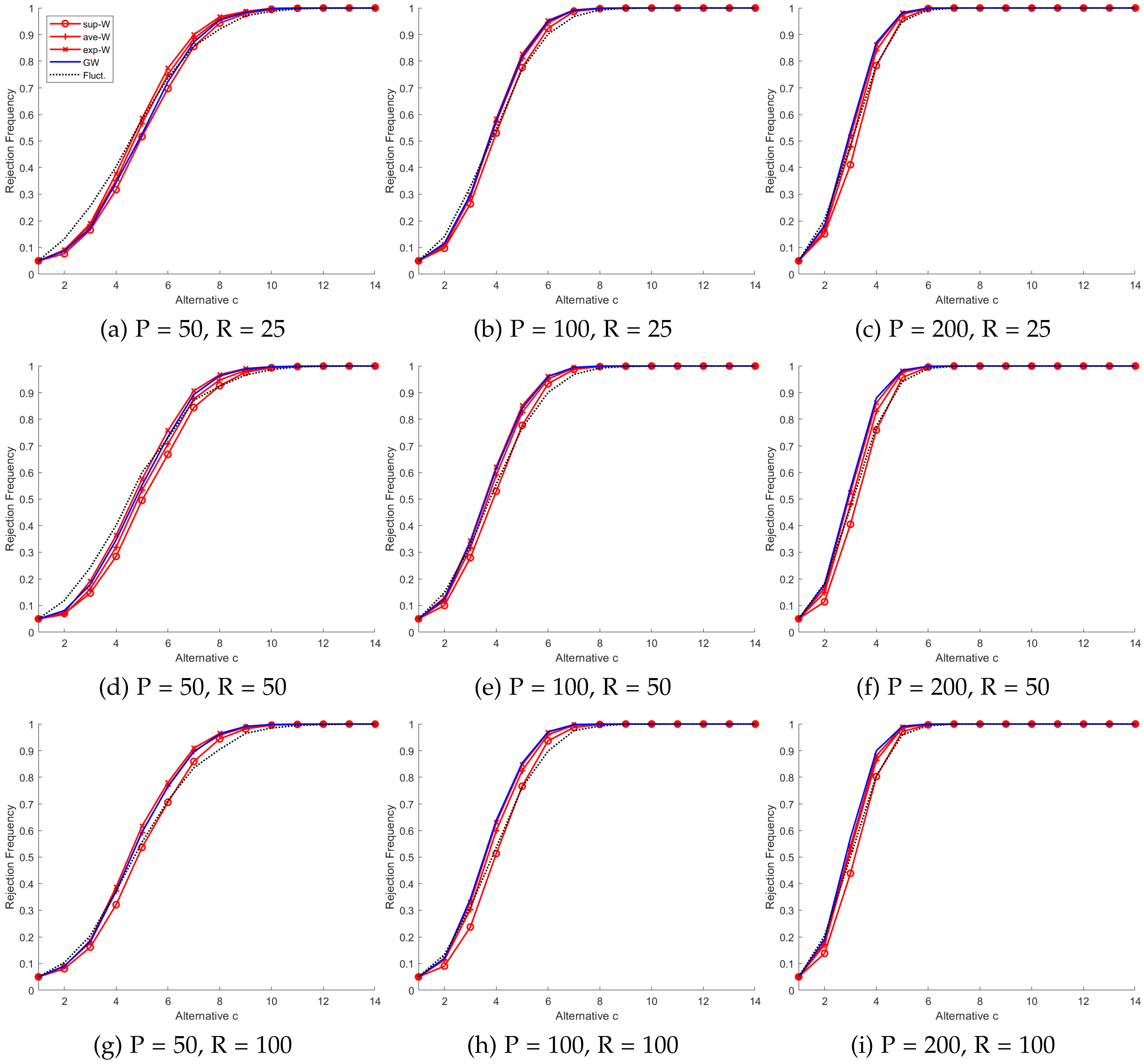
(g) P = 50, R = 100

(h) P = 100, R = 100

(i) P = 200, R = 100

Note: On the y-axis the figures displays size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test under Alternative (2) for point forecasts evaluated with the MSFE loss function. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. The solid lines with markers “o”, “x” and “+” display the sup-W, the exp-W, and the ave-W test results respectively. The dashed line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the solid line displays the results of the GW test. The nominal level is 5%. The results are based on 3,000 MC replications.

Figure 3: Size-Adjusted Power Results for PF1, Alternative (3): Constant Deviation



(a) P = 50, R = 25

(b) P = 100, R = 25

(c) P = 200, R = 25

(d) P = 50, R = 50

(e) P = 100, R = 50

(f) P = 200, R = 50

(g) P = 50, R = 100

(h) P = 100, R = 100

(i) P = 200, R = 100

Note: On the y-axis the figures displays size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test under Alternative (3) for point forecasts evaluated with the MSFE loss function. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. The solid lines with markers "o", "x" and "+" display the sup-W, the exp-W, and the ave-W test results respectively. The dashed line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the solid line displays the results of the GW test. The nominal level is 5%. The results are based on 3,000 MC replications.

4 Empirical Application: Uncovering Pockets of Predictability in Equity Premia

Financial return predictability is typically time-varying and elusive. As noted by Paye and Timmermann (1995), Rapach and Wohar (2006); Rapach and Zhou (2010), the predictability of stock market returns appears only when focusing on special sub-samples; Goyal and Welch (2003, 2008) similarly find that predictors that successfully forecast equity premia, U.S. returns or dividend price ratios typically change over time. Instabilities are widespread: Paye and Timmermann (2006), for example, cannot reject the presence of structural breaks in stock return predictive regressions in several countries and Rossi (2006, 2013) finds similar results for exchange rate returns. As summarized in Timmermann (2008), "... there appear to be pockets in time where there is modest evidence of local predictability; (...) the best forecasting method can be expected to vary over time, and there are likely to be periods of model breakdown where no approach seems to work". It is then inevitable that one must confront instabilities when evaluating financial models' predictive ability in an attempt to track their "local" forecasting performance.

As discussed in Timmermann (2008) and Paye and Timmermann (2006), the predictability of equity premia could be caused by market inefficiencies. If that is the case, then rational investors will take the opportunity to trade and make profits. However, if a large number of investors engage in taking advantage of the predictability, their behavior will eventually eliminate the predictability altogether. This implies the existence of short windows of time in which equity premia are predictable, but eventually disappear.

In what follows, we attempt to uncover pockets of predictability in U.S. equity premia out-of-sample. We use several of the economic predictors considered in Goyal and Welch (2008): the book to market ratio (calculated as the ratio of the book value and the market value of the Dow Jones Industrial Average and labeled "BookToMarket"); the consumption, wealth and income ratio proposed by Lettau and Ludvigson (2001) labeled "CAY"); the default yield spread (calculated as the difference between BAA and AAA-rated corporate bond yields and labeled "DFY"); the investment to capital ratio (labeled "Inv/K"); the long term government bond yield (labeled "LongYield"); and the term spread (calculated as the difference between the long term yield on government bonds and the Treasury bill and labeled "Spread").[15] Thus, the economic models are as follows:

$$\mathrm{E}_{t-1} r_t = \nu + \delta z_{t-1},$$

where z_{t-1} is the lagged economic predictor and ν is the intercept. All models are estimated in a window of past twenty years of data, producing a series of rolling one-year-ahead out-of-sample forecasts. As the benchmark model, we focus on the historical mean, also calculated using a rolling window of past returns over the previous twenty years.

We estimate the "local" forecasting performance using the nonlinear model in the loss differences, where the loss difference is the difference in the squared out-of-sample forecast errors of the benchmark minus that of the economic model:

$$\mathrm{E}_t \Delta L_{t+1|t} = \mu + \theta \cdot \mathbb{1}\left(s_t \leq \gamma\right), \tag{21}$$

[15] The data are from A. Goyal's website: http://www.hec.unil.ch/agoyal/

where s_t is the S&P 500's variance, a frequently used proxy for uncertainty in financial markets, computed as the sum of squared daily returns of the S&P 500 and taken from Goyal and Welch (2008). When the loss differential is positive, the economic model is better than the benchmark (the historical mean). The idea, formalized in equation (21), is to capture Timmermann's 2008 ”pockets of predictability”, where the pockets of predictability depend on the volatility of returns and, hence, on the uncertainty in financial markets. That is, the relative performance of the models' changes over time depending on whether the volatility of returns is higher (or lower) than an unknown threshold value.

Table 3 reports the results. For each predictor, listed in the first column, we report the p-values for the sup-W, ave-W, and exp-W test. In addition, we report the test statistics of the Diebold and Mariano (1995)/Giacomini and White (2006) (DM/GW) test and Fluctuation test, as well as the in-sample and out-of-sample sizes. For the two cases where the DM^{NL} test rejects the null hypothesis of equal performance, we report the estimated parameters of the model defined in equation (21) in Table 4. In addition, Table 4 reports the results of t-tests on the parameters, the result of a Wald test on the sum of the parameters, the estimated threshold parameter, and the frequencies of the regimes.

Overall, our results show evidence of pockets of predictability when forecasting using two predictors: the long yield and the spread. In both cases, the estimate of ν is negative, indicating that loss difference is negative when the volatility of the returns is higher than the threshold value, in which case the benchmark has a better predictive ability than the economic model. However, when the return volatility is sufficiently small, the loss difference becomes positive. That is, the long yield and the spread are capable of predicting the returns when the volatility is small, while the opposite is true when volatility is high.

Notice that in none of these cases the DM/GW finds that the model with the economic predictor is significantly different than the benchmark. This is because our proposed test is more powerful to detect pockets of predictability when there are instabilities associated with nonlinear behavior. Notice that the Giacomini and Rossi (2010) Fluctuation test statistic is never bigger than the critical value either; hence, even though the Fluctuation test is robust to instabilities in the relative forecasting performance, nevertheless it is less powerful than the test proposed in this paper and, in these data, never finds evidence that the predictive ability appears sporadically over time.

For the predictors for which we found that the economic model performs sometimes better than the benchmark, i.e. the long yield and the spread, Figure 4 reports the loss differences ($\Delta L_{t+1|t}$) over time, together with the stock market variance z_t that triggers the regime switching. Shaded areas depict periods where the benchmark has a lower squared forecast error than the economic model.[16] Periods that are not shaded indicate times in which the economic model performs better than the benchmark. The figure shows that, for both predictors, there are several pockets of predictability, where the model predicts slightly better than the benchmark. Further, these pockets persist for several periods and are interrupted by periods where the economic model performs much worse than the benchmark, causing the average performance of the model to be poor over the entire sample. The pockets of predictability, hence, correspond to tranquil times, where the forecast improvements of the economic models relative to the benchmark are

[16] That is, the loss difference is negative.

small in magnitude; the overall poor forecasting performance of the models is associated with highly volatile times.

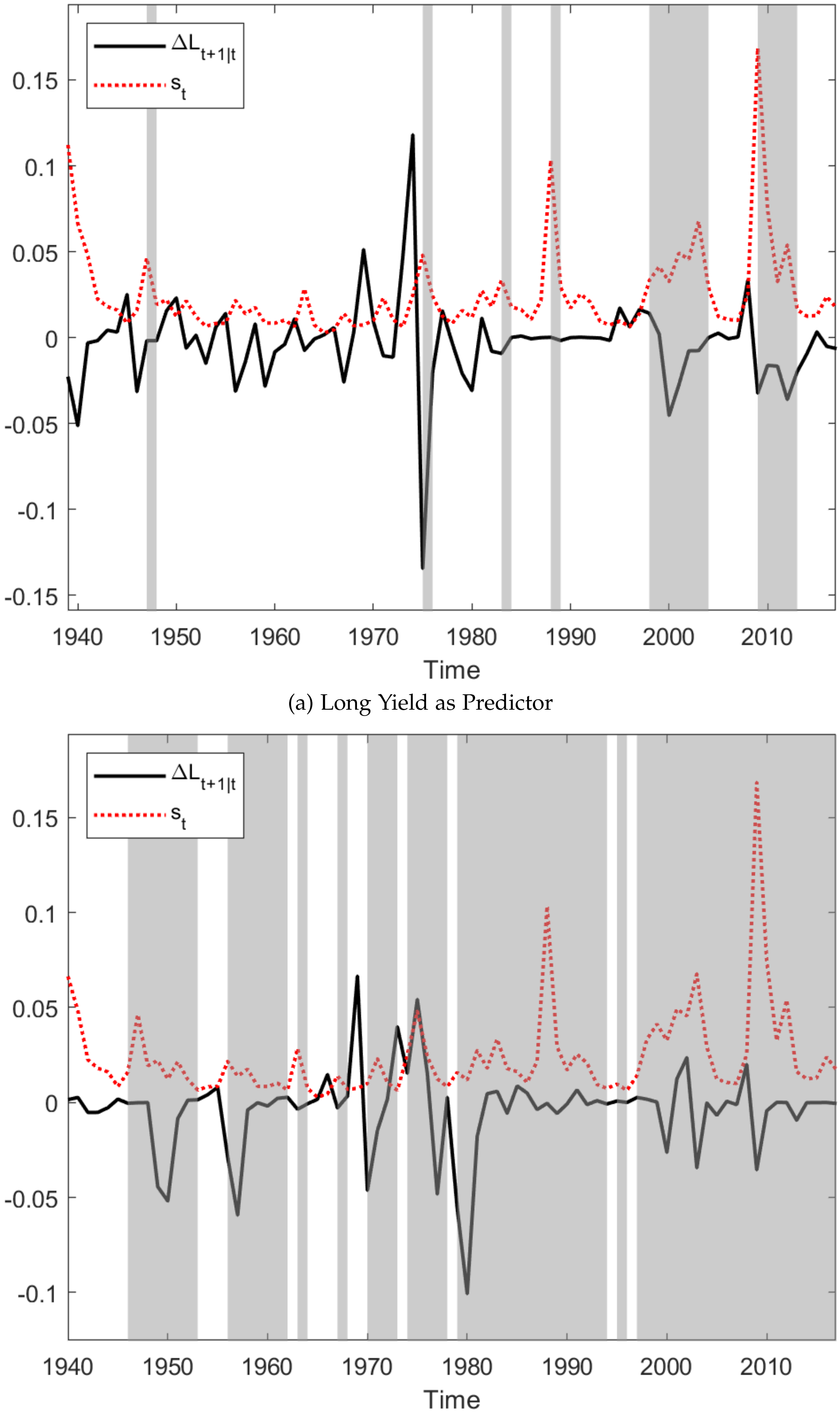
An interesting implication of our empirical findings, thus, is the following: since volatility in financial markets indicates uncertainty, periods of low uncertainty are associated with economic predictability. Periods of high uncertainty, instead, associated with economic predictors performing substantially poorly: in fact, so poorly that any gain previously achieved is washed away.

Our results are linked with the recent financial literature that explains the existence of time-varying risk premia with disaster risk, in particular Barro (2006). In his paper, Barro (2006) explains the equity premium using the probability of a rare disaster. On the empirical side, Berkman et al. (2011) have shown that crisis indices, which proxy for perceived disaster probability, impact stock market returns, and are positively correlated with earning-price ratios and the dividend yield. Our paper also relates equity premia predictability to risk and uncertainty, but does so from a completely different point of view, namely using a non-linear model directly capturing changes in out-of-sample predictive ability over time related to switches in the volatility of equity premia themselves.

Interestingly, our result that the predictability is present in low-volatility scenarios is reminiscent of Ismailov and Rossi (2018), who found, in the very different context of international markets, that exchange rate returns are more predictable by economic models in times of low exchange rate uncertainty. In our analysis, what matters for predicting the equity premium is uncertainty in stock markets, which we measure by the volatility in the S&P 500. Farmer et al. (2019) also aims at investigating the presence of pockets of in-sample predictability in U.S. equity returns. Farmer et al. (2019) find evidence of pockets of predictability in equity returns in both the early-2000s and the mid-2010s; they also find evidence of predictability for the Treasury bill rate in both late-2000s and the mid-2010s. Although their methodology is very different, as they employ a time-varying parameter model estimated non-parametrically while we model directly the forecast loss differential, their results are similar to ours. Our results, however, differ from Rapach et al. (2010), who found that return predictability is correlated with the business cycle, and more similar to Farmer et al. (2019), who found only a weak link between the two. Our results suggest instead that return predictability is correlated with the volatility of the financial cycle, a proxy for uncertainty.

Finally, note that, in this paper, we focus on detecting 'pockets of predictability' in historical data and linking it to the time-variation in an economic threshold variable. This does not necessarily imply that it is possible to detect pockets of predictability in real-time. For readers interested in the latter, Inoue and Rossi (2015) and Harvey et al. (2020) propose real-time monitoring procedures to detect structural changes. They suggest sequentially repeating t-tests over short time periods and control the overall rejection rates. For example, in their application to predictive regressions, Harvey et al. (2020) find that the one-month ahead equity premium had been predictable at several points in time and that such episodes could have been detected in real-time by their methodology.

Figure 4: Delta losses and threshold variable



(a) Long Yield as Predictor

(b) Spread as Predictor

Note: The figure shows the estimated $\Delta L_{t+1|t}$ (solid line) together with the stock market uncertainty measure s_t (dashed line), which triggers the regime switching. Non-shaded areas indicate periods where the economic model performed better than the benchmark, i.e. they show the pockets of predictability.

Table 3: Testing for State Dependence: Empirical Results

Variable Name	DM^{NL} p-values			Alternative Statistics		Sample Sizes	
	sup-W	ave-W	exp-W	DM/GW	Fluct.	R	P
DFY	0.327	0.538	0.410	0.522	1.322	20	79
Inflation	0.320	0.270	0.283	-1.055	1.293	20	84
Stock Var	0.524	0.624	0.646	-1.097	1.554	20	113
LongYield	**0.048**	**0.069**	**0.050**	-1.347	0.932	20	79
Spread	**0.025**	**0.098**	**0.038**	-1.351	0.451	20	78
T-bill	0.122	**0.063**	**0.077**	-1.528	0.536	20	78
BookToMkt	0.790	0.778	0.791	-0.510	1.135	20	77
InvtoK	0.773	0.742	0.741	0.138	2.583	20	51
CAY	0.593	0.591	0.605	-0.348	1.314	20	53

Note: The critical value for the one-sided DM-GW test at the 5% significance level is 1.64. The one-sided Fluctuation test is implemented using a window size equal to one-third of the out-of-sample portion of the sample; its critical value at the 5% significance level is 2.770 (see Giacomini and Rossi (2010), Table 1). The columns R and P show the in-sample and out-of-sample size respectively. Bold numbers indicate significance at the 10% level.

Table 4: Model Results given the Presence of State Dependence

Variable Name	Parameter Estimates				Parameter Tests			Regime Characteristics	
	$\hat{\mu}$	$\hat{\theta}$	$\hat{\mu}+\hat{\theta}$	$\hat{\gamma}$	$\hat{\mu}=0$	$\hat{\theta}=0$	$\hat{\mu}+\hat{\theta}=0$	$\bar{s}$	$\text{P}(s_t<\hat{\gamma})$
LongYield	-0.023	0.025	0.002	0.029	$\mathbf{-3.033}$	**2.748**	0.551	0.025	0.215
Spread	-0.007	0.017	0.010	0.008	$\mathbf{-2.141}$	**2.978**	**4.629**	0.024	0.821

Note: The columns $\hat{\mu}, \hat{\theta}, \hat{\mu}+\hat{\theta}$, and $\hat{\gamma}$ show the parameter estimates associated the largest test statistic $W_P(\gamma)$. The columns $\hat{\mu}=0, \hat{\theta}=0$, and $\hat{\mu}+\hat{\theta}=0$ show the values of the statistic when using a t-test or a Wald test respectively, for testing the hypothesis that the parameters, or their sum, are equal to zero. The critical values of a Wald test, with one restriction, at the 5% and 10% level are 2.706 and 3.842. Bold numbers indicate a rejection at the 10% level. The column $\bar{s}$ shows the average value of the conditioning variable s_t, and $\text{P}(s_t<\hat{\gamma})$ shows the relative frequency of being in the regime where both μ and θ are present.

5 Conclusion

We propose a forecast comparison test robust to state dependence, where the states are a function of economic observables and allow the threshold that indicates switching between the states to be unknown. Due to the unidentified nuisance parameter under the null, the asymptotic distribution is non-standard and cannot be tabulated in general. However, the asymptotic distribution can be simulated with negligible computational costs. The testing framework assumes that the parameters of the competing forecasting models are estimated using a rolling window scheme, i.e. we compare forecasting methods rather than forecasting models, and we allow for nested and non-nested models. Results from a Monte Carlo study indicate good size and power properties of the test statistics for moderate sample sizes.

In an empirical application, we document the existence of state dependence in the relative forecasting performance in models that predict stock returns. In particular, simpler models perform better during times of high uncertainty, measured by stock market volatility, whereas models with the spread or long-run yield as predictors have a better forecasting performance during times of low volatility. Hence, our results link predictability in returns to low uncertainty, and have important implications for models of “tail events” or “rare disasters”. Existing tests, such as the GW and Fluctuation test, cannot detect these “pockets of predictability” as they lack power against state dependence.

References

Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, 25:177–190.

Andrews, D. and Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62:1383–1414.

Barro, R. J. (2006). Rare disasters and asset markets in the twentieth century. *Quarterly Journal of Economics*, 121:823–866.

Berkman, H., Jacobsen, B., and Lee, J. B. (2011). Time varying rare disaster risk and stock returns. *Journal of Financial Economics*, 101:313–332.

Carrasco, M., Hu, L., and Ploberger, W. (2014a). Optimal test for markov switching parameters. *Econometrica*, 82:765–784.

Carrasco, M., Hu, L., and Ploberger, W. (2014b). Supplement to optimal test for markov switching parameters. *Econometrica*, 82:765–784.

Cho, J. S. and White, H. (2007). Testing for regime switching. *Econometrica*, 75:1671–1720.

Clark, T. and McCracken, M. (2001). Tests of Equal Forecast Accuracy and Encompassing for Nested Models. *Journal of Econometrics*, 105:85–110.

Clark, T. and West, K. (2006). Using Out-of-sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis. *Journal of Econometrics*, 135:155–186.

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

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64:247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74:33–43.

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

Farmer, L., Schmidt, L., and Timmermann, A. (2019). Pockets of predictability. *mimeo*.

Garcia, R. (1998). Asymptotic null distribution of the likelihood ratio test in markov switching models. *International Economic Review*, 39:763–788.

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

Giacomini, R. and White, H. (2006). Test of conditional predictive ability. *Econometrica*, 74:1545–1578.

Goyal, A. and Welch, I. (2003). Predicting the equity premium with dividend ratios. *Management Science*, 49:639–654.

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

Granger, C. W. J. and Teräsvirta, T. (1993). *ModellingNon-linear Economic Relationships*. Oxford: Oxford University Press.

Hamilton, J. D. (1989). A new approach to the economic analysis of non-stationary time series and the business cycle. *Econometrica*, 57:357–384.

Hansen, B. E. (1992). The likelihood ratio test under non-standard conditions: Testing the markov switching model of gnp. *Journal of Applied Econometric*, 7:61–82.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64:413–430.

Harvey, D., Leybourne, S., and Sollis R. and, Taylor, M. (2020). Real-time detection of regimes of predictability in the u.s. equity premium. *mimeo*.

Inoue, A. and Rossi, B. (2015). Recursive predictability tests for real time data. *Journal of Business and Economic Statistics*, 23:336–345.

Ismailov, A. and Rossi, B. (2018). Uncertainty and deviations from uncovered interest rate parity. *Journal of International Money and Finance*, 88:242–259.

Lettau, M. and Ludvigson, S. (2001). Resurrecting the (c)capm: A cross sectional test when risk premia are time-varying. *Journal of Political Economy*, 109:1238–1287.

Luukkonen, R., Saikkonen, P., and Teräsvirta, T. (1988). Testing linearity against transition autoregressive models. *Biometrika*, 75:491–499.

Paye, B. and Timmermann, A. (1995). Predictability of stock returns: Robustness and economic significance. *Journal of Finance*, 50:1201–1228.

Paye, B. and Timmermann, A. (2006). Instability of return prediction models. *Journal of Empirical Finance*, 13:274–315.

Qu, Z. and Zhuo, F. (2017). Likelihood ratio based tests for markov regime switching.

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

Rapach, D. and Wohar, M. (2006). Structural breaks and predictive regression models of aggregate u.s. stock returns. *Journal of Financial Econometrics*, 4:238–274.

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

Rossi, B. (2006). Are exchange rates really random walks? some evidence robust to parameter instability. *Macroeconomic Dynamics*, 10:20–38.

Rossi, B. (2013). Are exchange rates predictable. *Journal of Economic Literature*, 51:1063–1119.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, 89:208–218.

Teräsvirta, T. (2006). Forecasting economic variables with nonlinear models. In Elliott, G., C. G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, pages 414–457. North-Holland.

Timmermann, A. (2008). Elusive return predictability. *International Journal of Forecasting*, 24:1–8.

Tong, H. (1990). *Non-Linear Time Series. A Dynamical System Approach*. Oxford University Press, Oxford.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64:1067–1084.

White, H. (2001). *Asymptotic Theory for Econometricians*. Emerald Group Publishing Limited.

A The Case of Multiple Threshold Variables

For now, we have treated the threshold variable S_t as known, and only the threshold γ as unknown. As noted by Hansen (1996), in practice, the researcher might have several potential threshold variables S_t at hand and needs to decide which variable to include. This case can be naturally accommodated in the framework described above and we sketch the procedure in the following.

Let D denote a finite set of index numbers, from 1 to $\bar{d}$, for candidate threshold variables, such that $S_t(d)$, $d \in D$, denotes the candidate threshold variable indexed by d. Eq. (3) then becomes

$$\Delta L_{t+h|t} = X_t'\mu + X_t'\theta \cdot \mathbb{1}(S_t(d) \leq \gamma) + u_{t+h}. \tag{22}$$

Conditional on a value $(\gamma, d) \in (\Gamma \text{ x } D)$, the estimation of eq. (22) is analogue to that in the model described in eq. (3). Further, and to simplify notation, let all terms of Section 2.2 that are a function of γ be defined analogously as a function of (γ, d). Further, let

$$\mathrm{DM}^{\mathrm{NL}}_{\Gamma,\mathrm{D}}: \quad g_{\Gamma,D}(W_p) = \begin{cases} \sup_{d\in D}\sup_{\gamma\in\Gamma} W_P(\gamma,d) \\ \frac{1}{D}\sum_D \int_\Gamma W_P(\gamma,d)\mathrm{d}w(\gamma,d) \\ \ln\left(\frac{1}{D}\sum_D \int_\Gamma \exp(\frac{1}{2}W_P(\gamma,d))\mathrm{d}w(\gamma,d)\right) \end{cases} \tag{23}$$

denote the statistic that takes the supremum over both $\gamma \in \Gamma$ and $d \in D$. It is straightforward to show that the test statistic in eq. (23) has as an asymptotic distribution for point and density forecasts that is analogue to that derived in Proposition 1 and Proposition 2. We now state the necessary assumptions and then the corollary that accommodates the case of testing for a threshold model when there is more than one candidate threshold variable.

Assumption A.A1 *(i) For all $d \in D$, where D is a finite set of index numbers, $(A_t, X_t, S_t(d))$ is strictly stationary and absolutely regular with mixing coefficients $\eta(m) = O(m^{-\delta})$ for some $\delta > v/(v-1)$ and $v > 1$. (ii) The estimation window size (R) is finite and the estimation scheme is a rolling window estimation.*

Assumption A.A2 *For $r > v > 1$, $E|Q_t|^{4r} < \infty$, $E|u_t|^{4r} < \infty$, $\inf_{d\in D}\inf_{\gamma\in\Gamma}\ det\big(M(\gamma,\gamma,d,d)\big) > 0$.*

Assumption A.A3 *Let $r > v$ and let S_t have a density function $g(S_t)$ such that $\sup_{s\in\mathbb{R}^{\bar{d}}} g(s) = \bar{g} < \infty$.*

Assumption A.A4 *$f^{(i)}_{t+h|t}(.)$ is a measurable function of leads and lags of A_t, for $i = 1, 2$.*

Corollary 1 *Let $g_{\Gamma,D}(W_p)$ be one of the statistics defined in eq. (23) . Then, under A.A1 to A.A4 and H_0 defined in eq. (4): $E\left(\Delta L_{t+h|t}\right) = 0$ for all $t = R+h, ..., T$, we have*

$$\lim_{P\to\infty} g_{\Gamma,D}\big(W_P(\gamma,d)\big) \rightarrow g_{\Gamma,D}\big(\chi^2(\gamma,d)\big), \tag{24}$$

where $\chi^2(\gamma)$ is a chi-square distribution with degrees of freedom rank(H_r), and $g_{\Gamma,D}\big(\chi^2(\gamma,d)\big)$ can be completely characterized by its covariance kernel $K(\gamma_1,\gamma_2,d_1,d_2)$.

Given A.A1 to A.A4, the proof of Corollary 1 follows from Proposition 1. The algorithm to simulate the critical values is similar to the algorithm described in Section 2.5, and is given below.

Simulation Algorithm 2. For each $j = 1, ..., J$, do the following steps:

1. Draw a set of standard Normal random variates $\{v_{tj}\}_{t=1}^{P}$;
   (a) Select a threshold variable $S_t(d)$, $d \in D$.
      i. Calculate $\widehat{\lambda}_P^j(\gamma, d) = \frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} \widehat{s}_{t+h}(\gamma, d) v_{tj}$;
      ii. Using $\widehat{\lambda}_P^j(\gamma, d)$, calculate:
      $W_P^j(\gamma, d) = \widehat{\lambda}_P^j(\gamma, d)' M(\gamma, \gamma, d, d)^{-1} H_r \left[H_r' \widehat{V}_P^*(\gamma, d) H_r\right]^{-1} H_r' M(\gamma, \gamma, d, d)^{-1} \widehat{\lambda}_P^j(\gamma, d)$;
      iii. Repeat (i)-(ii) for all $\gamma \in \Gamma$;
   (b) Repeat (a) for all $d \in D$;
2. Compute $W_P^j = g_\Gamma\left(W_P^j(\gamma, d)\right)$.

After J iterations, we obtain a set of $\{W_P^j\}_{j=1}^J$ draws from the asymptotic distribution, which we can use to construct critical values and p-values. In particular, the approximate p-value is given by $\widehat{p}(J) = \frac{1}{J} \sum_{j=1}^J \mathbb{1}(W_P > W_P^j)$, where W_P denotes the value of the test statistic computed using the actual data.

B Forecast Comparison under Markov Switching

In this section, we discuss a test for equal predictive ability in the presence of Markov switching changes in the relative forecasting performance. The test is inspired by Carrasco et al. (2014a) (CHP hereafter).

The loss differential is modeled as:

$$\Delta L_{t+h|t} = \mu + \mu_t + u_{t+h}, \tag{25}$$

where $\mu_t = \mu_S S_t$, S_t is a stationary geometric ergodic two-state univariate first-order Markov chain, μ_S is the magnitude of the change and u_{t+h} is mean zero and satisfies assumption A6 below, let $\theta \equiv \left\{\mu, \sigma^2\right\}$, and let $\theta_0 \equiv \left\{\mu_0, \sigma_0^2\right\}$ denote the parameters under the null, where μ_0 is a parameter of interest and $\sigma_0^2 > 0$ is left unspecified.

Our null hypothesis is described in eq. (4), and is such that $H_0 : E_t\left(\Delta L_{t+h|t}\right) = 0$. Under the model in eq. (25), the null hypothesis can be reparameterized as: $\mu = \mu_S = 0$. Our null hypothesis is different from CHP in two ways: the first is that the latter only test $\mu_S = 0$; the second is that our objective is to test the forecast loss differential, that depends on the estimates of the forecasting models' parameters. Under the alternative, $E_t\left(\Delta L_{t+h|t}\right) \neq 0$, which again can be caused by either Markov switching or constant and unequal forecast performance.

To derive the asymptotic distribution of the test, we require the following additional assumptions:

Assumption B1 *The latent variable μ_t is defined as $\mu_t = \mu_S S_t$, where μ_S is a finite scalar constant, S_t is a stationary geometric ergodic finite-state univariate first-order Markov chain with $var(S_t) = 1$ and covariance $cov\left(S_t, S_{t-i}\right) = \rho^i$, $\rho \neq 0$ and $-1 < \rho < 1$. Furthermore, μ_t is strongly exogenous to A_t, $t = 1, ..., T$, such that the joint likelihood of $A_1, ..., A_T, \mu_1, ..., \mu_T$ factorizes as $\Pi_{t=1}^T f\left(A_1, ..., A_T; \theta\right) q\left(\mu_t | \mu_{t-1}, ..., \mu_1; \rho\right)$ and the values of $\mu_t + \mu$ belong to some compact set containing μ.*

Assumption B2 *Let the conditional log-density of $\Delta L_{t+h|t}$ be Normal and be denoted by ℓ_t under the null hypothesis. Let N_0 be a neighborhood around θ_0, where θ_0 is an interior point of N_0; the information matrix $\mathcal{I}(\theta_0) = E_0\left(||\ell_t^{(1)}(\theta_0)\,\ell_t^{(1)}(\theta_0)'||^{20}\right)$ is nonsingular.*

For convenience, we maintain Assumption A1. Assumption B1 specifies the behavior of the time variation and it requires that, under the null, the distribution of the data A_t and that of η_t are mutually independent. Assumption B2 makes a convenient distributional assumption which implies that the asymptotic distribution of our test statistic is the same as in CHP — for details see the proof of Proposition 3.

The following proposition provides the result of our test of equal predictive ability in the presence of Markov switching alternatives.

Proposition 3 *Let DM^{NL}: $g(TS_P) = \sup_{\rho\in[\underline{\rho},\overline{\rho}]} TS_P(\rho)$, and $TS_P(\rho) = \frac{1}{2}\left(\max\left(0, \frac{\Gamma_P^*(\rho)}{\sqrt{\widehat{\xi}(\rho)'\widehat{\xi}(\rho)}}\right)\right)^2$, where $\Gamma_P^*(\rho) = P^{-1/2}\sum_t \eta_t^*\left(\rho,\widehat{\theta}_0\right)$, and*

$$\eta_t^*(\rho,\theta) = \frac{1}{2}\left\{\left[\ell_t^{(2)}(\theta) + \ell_t^{(1)}(\theta)\,\ell_t^{(1)}(\theta)'\right] + 2\sum_{\tau<t}\rho^{(t-\tau)}\ell_t^{(1)}(\theta)\,\ell_\tau^{(1)}(\theta)'\right\}.$$

$\widehat{\xi}(\rho)$ is the residual of a regression of $\eta_t\left(\rho,\widehat{\theta}_0\right)$ on $\ell_t^{(1)}\left(\widehat{\theta}_0\right)$ and $\widehat{\theta}_0$ is the constrained ML estimator of θ under the null. Then, under A1, B1, B2 and H_0 defined in eq. (4): $E\left(\Delta L_{t+h|t}\right) = 0$ for all $t = R+h,...,T$:

$$g(TS_P) = \sup_{\rho\in[\underline{\rho},\overline{\rho}]} TS_P(\rho) \underset{d}{\rightarrow} \sup_{\rho\in[\underline{\rho},\overline{\rho}]} \frac{1}{2}\left(\max(0,K)\right)^2, \tag{26}$$

where $K = sign(\rho)\sqrt{1-\rho^2}\sum_{i=0}^{\infty}\rho^i Z_i$, where $sign(\rho) = 1$ if $\rho > 0$, zero if $\rho = 0$ and equal to -1 if $\rho < 0$, and Z_i are iid standard Normal variables. The DM^{NL} test rejects H_0 defined in eq. 4 when $g_\Gamma(TS_P) > \phi_\alpha$, where ϕ_α is the critical value (for a nominal size of α) in Table B.1 below, where either $\underline{\rho} = -0.7, \overline{\rho} = 0.7$ or $\underline{\rho} = -0.98, \overline{\rho} = 0.98$.

Proof of Proposition 3. From a similar argument as that in the proof of Proposition 1, since the forecast errors $v_{t+h}(\widehat{\beta}_{t,R})$ are measurable functions of leads and lags of A_t, under A1(i) they are absolutely regular with coefficients of size $-\delta$. Consequently, $\Delta L_{t+h|t}$ is strictly stationary and absolutely regular with mixing coefficients $\eta(m) = O\left(m^{-\delta}\right)$ for some $\delta > \nu/(\nu-1)$ and $\nu > 1$. Under Assumptions A1, B1 and B2, the assumptions in CHP hold. In particular, let $cov()$ denote the covariance and let

$$\begin{aligned} d^*(\rho) \equiv d^*(\rho,\theta_0) &= \mathcal{I}(\theta_0)^{-1} cov\left(\eta_t^*(\rho,\theta_0), \ell_t^{(1)}(\theta_0)\right) \\ &= \mathcal{I}(\theta_0)^{-1} cov\left(\eta_t^*(\rho,\theta_0), \left(\begin{array}{cc} \ell_{\mu,t}^{(1)}(\theta_0) & \ell_{\sigma^2,t}^{(1)}(\theta_0) \end{array}\right)\right) \\ &= \mathcal{I}(\theta_0)^{-1}\left(\begin{array}{cc} cov\left(\eta_t^*(\rho,\theta_0), \ell_{\mu,t}^{(1)}(\theta_0)\right), & cov\left(\eta_t^*(\rho,\theta_0), \ell_{\sigma^2,t}^{(1)}(\theta_0)\right) \end{array}\right). \end{aligned}$$

Under normality, $\eta_t^*(\rho,\theta) = \frac{1}{2\sigma^4}\left[(u_{t+h}^2 - \sigma^2) + 2\sum_{\tau<t}\rho^{(t-\tau)}u_{t+h}u_{\tau+h}\right]$ and $\ell_{\mu,t}^{(1)}(\theta_0) = u_{t+h}/\sigma_0^2$; therefore, $cov\left(\eta_t^*(\rho,\theta_0), \ell_{\mu,t}^{(1)}(\theta_0)\right) = 0$. Furthermore, because of the Normality assumption, the matrix $\mathcal{I}(\theta_0)^{-1}$ is block diagonal. Thus, the first element of the vector $d^*(\rho)$ equals

zero. This implies that $d^*(\rho)'\ell_t^{(1)}\left(\widehat{\theta}_0\right) = 0$ since (i) we just showed that the first element of $d^*(\rho)$ is zero; and (ii) the second element of $\ell_t^{(1)}\left(\widehat{\theta}_0\right) = 0$ because this component of the score is evaluated at the constrained MLE of σ^2. Thus, $\Gamma_P^*(\rho) = P^{-1/2}\sum_t \eta_t^*\left(\rho,\widehat{\theta}_0\right) = P^{-1/2}\sum_t\left(\eta_t^*\left(\rho,\widehat{\theta}_0\right) - d^*(\rho)'\ell_t^{(1)}\left(\widehat{\theta}_0\right)\right)$, as $d^*(\rho)'\ell_t^{(1)}\left(\widehat{\theta}_0\right) = 0$. Consequently, the arguments of Lemma C.1 of Carrasco et al. (2014b) apply to eq. (26), and the results follow from Theorem 3.1 of Carrasco et al. (2014a).

Table B.1: Critical Values (ϕ_α)

α	$\rho \in [-0.7, 0.7]$	$\rho \in [-0.98, 0.98]$
1%	3.96	4.52
5%	2.45	2.99
10%	1.82	2.32

Note: The critical values are taken from table A-I of Carrasco et al. (2014b).

Table B.2 reports size results for the test $g(TS_P)$ using the data generated according to PF1 and PF2 as considered in Section 3. The table shows that the test tends to over-reject in small samples (P< 200, R < 25), but is well-sized for larger sample sizes. As before, large sample results for the nested model case of PF2 are slightly undersized.

Table B.2: Size Results for Forecast Comparison Markov Switching Test

	PF1								PF2							
	Size 5 %				Size 10 %				Size 5 %				Size 10 %			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.159	0.135	0.083	0.067	0.232	0.195	0.154	0.118	0.116	0.091	0.071	0.045	0.154	0.134	0.102	0.086
50	0.138	0.129	0.088	0.055	0.211	0.184	0.137	0.104	0.106	0.093	0.067	0.035	0.168	0.140	0.114	0.061
100	0.148	0.117	0.092	0.059	0.203	0.169	0.153	0.118	0.121	0.100	0.049	0.038	0.175	0.142	0.095	0.074

Note: The table displays empirical rejection frequencies of the null hypothesis $H_0 : \mu = \mu_S = 0$ for the DM$^{\text{CHP}}$ test. Size 5% and 10% denote the nominal size. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. DGP1 and DGP2 are based on Section 3. The results are based on 1,000 MC replications and using the critical values of Table B.1 with $\rho \in [-0.98, 0.98]$.

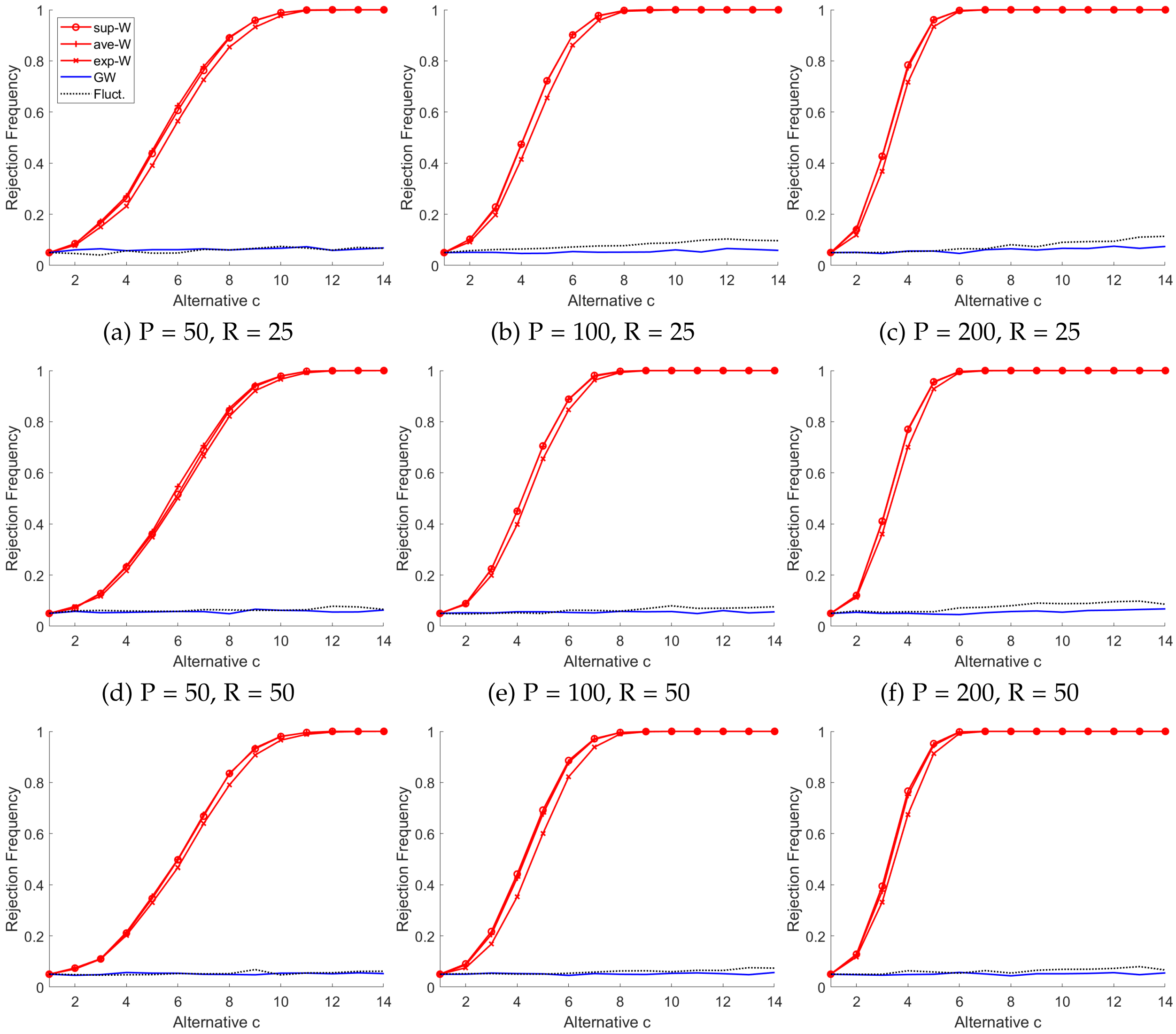
For a model that includes autoregressive components the distribution depends on the autoregressive component and can be derived from the asymptotic distribution of $\sup_{\rho\in[\underline{\rho},\bar{\rho}]} v_T(\theta_0,\rho)$. For an AR(1), similar to Carrasco et al. (2014a), the critical values can be simulated from:

$$\frac{\sqrt{1-\rho^2}\,|1-\rho\phi|}{|\rho-\phi|}\left[\sum_{i=0}^{\infty}\rho^i Z_i - \frac{(1-\phi^2)}{(1-\phi\rho)}\sum_{i=0}\phi^i Z_i\right]. \tag{27}$$

C Additional Monte Carlo Results: PF1, DF1 and DF2

C.1 Point Forecast Comparison 2

Figure C.1: Size-Adjusted Power Results for PF2, Alternative (1): State Dependence



(a) P = 50, R = 25

(b) P = 100, R = 25

(c) P = 200, R = 25

(d) P = 50, R = 50

(e) P = 100, R = 50

(f) P = 200, R = 50

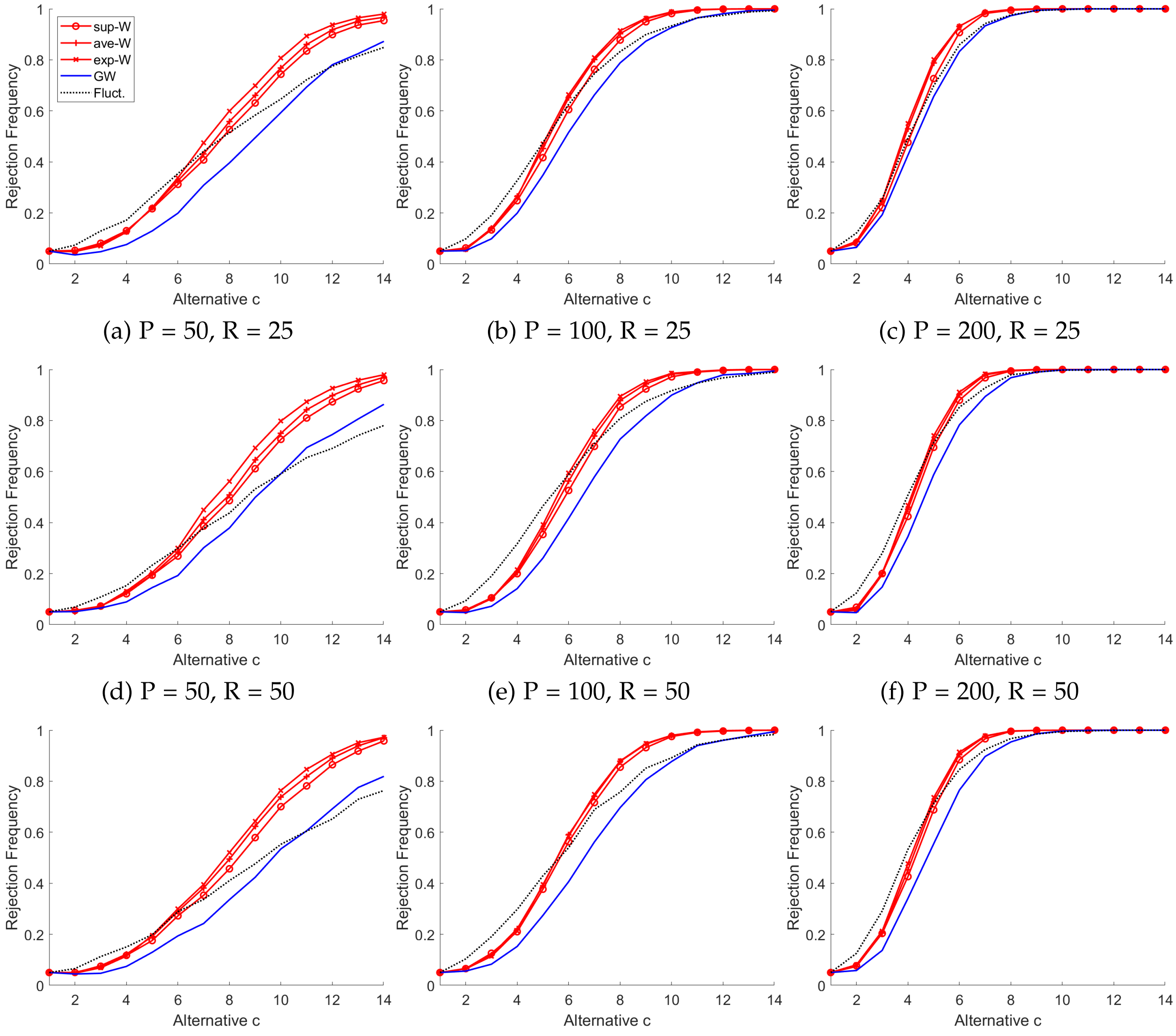
(g) P = 50, R = 100

(h) P = 100, R = 100

(i) P = 200, R = 100

Note: On the y-axis the figures displays size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test under Alternative (1) for point forecasts evaluated with the MSFE loss function. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. The solid lines with markers "o", "x" and "+" display the sup-W, the exp-W, and the ave-W test results respectively. The dashed line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the solid line displays the results of the GW test. The nominal level is 5%. The results are based on 3,000 MC replications.

Figure C.2: Size-Adjusted Power Results for PF2, Alternative (2): State Dependence and Constant Deviation



(a) P = 50, R = 25

(b) P = 100, R = 25

(c) P = 200, R = 25

(d) P = 50, R = 50

(e) P = 100, R = 50

(f) P = 200, R = 50

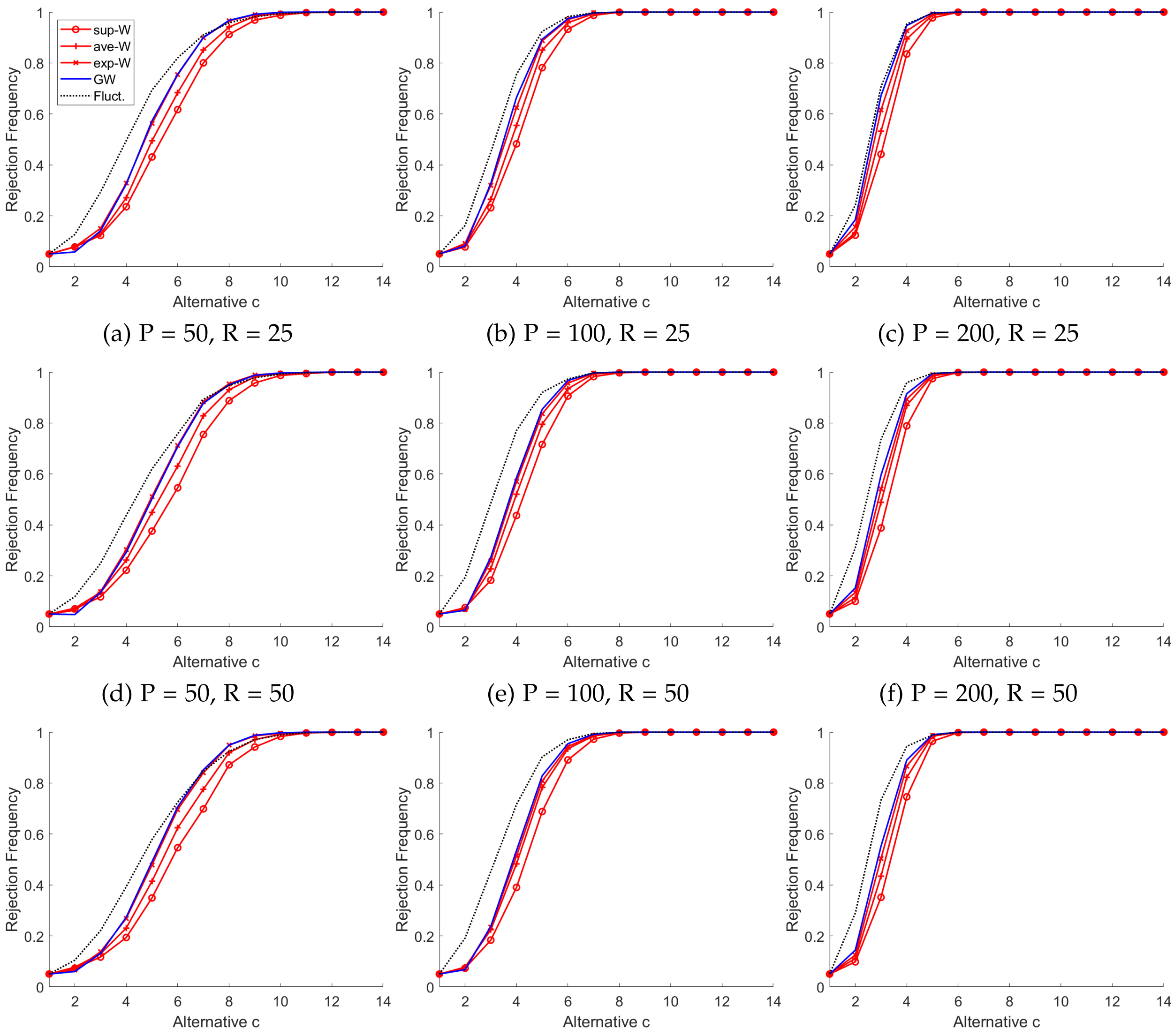
(g) P = 50, R = 100

(h) P = 100, R = 100

(i) P = 200, R = 100

Note: On the y-axis the figures displays size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test under Alternative (2) for point forecasts evaluated with the MSFE loss function. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. The solid lines with markers "o", "x" and "+" display the sup-W, the exp-W, and the ave-W test results respectively. The dashed line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the solid line displays the results of the GW test. The nominal level is 5%. The results are based on 3,000 MC replications.

Figure C.3: Size-Adjusted Power Results for PF2, Alternative (3): Constant Deviation



(a) P = 50, R = 25 (b) P = 100, R = 25 (c) P = 200, R = 25

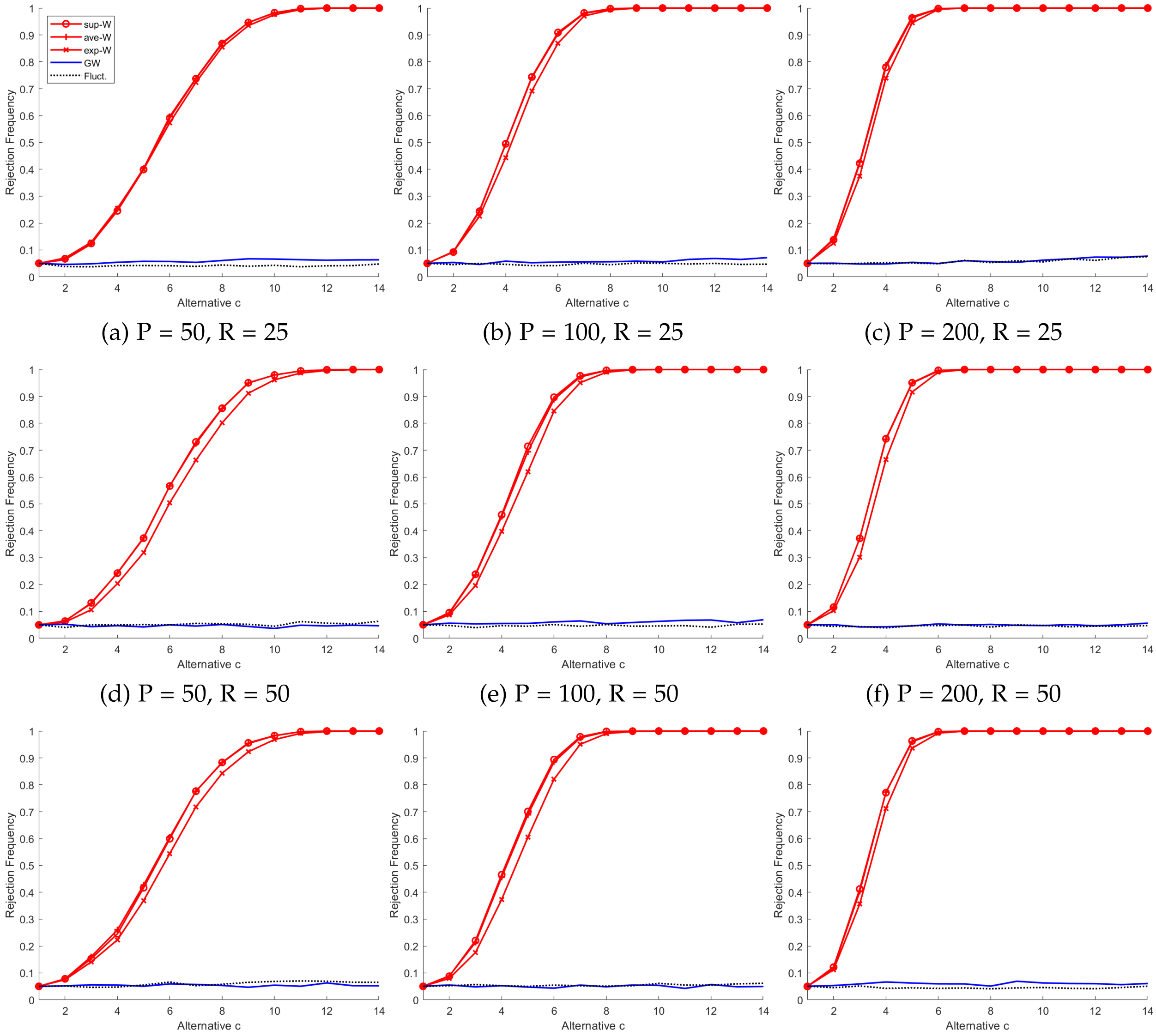
(d) P = 50, R = 50 (e) P = 100, R = 50 (f) P = 200, R = 50

(g) P = 50, R = 100 (h) P = 100, R = 100 (i) P = 200, R = 100

Note: On the y-axis the figures displays size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test under Alternative (3) for point forecasts evaluated with the MSFE loss function. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. The solid lines with markers “o”, “x” and “+” display the sup-W, the exp-W, and the ave-W test results respectively. The dashed line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the solid line displays the results of the GW test. The nominal level is 5%. The results are based on 3,000 MC replications.

C.2 Density Forecast Comparison 1

Figure C.4: Size-Adjusted Power Results for DF1, Alternative (1): State Dependence



(a) P = 50, R = 25

(b) P = 100, R = 25

(c) P = 200, R = 25

(d) P = 50, R = 50

(e) P = 100, R = 50

(f) P = 200, R = 50

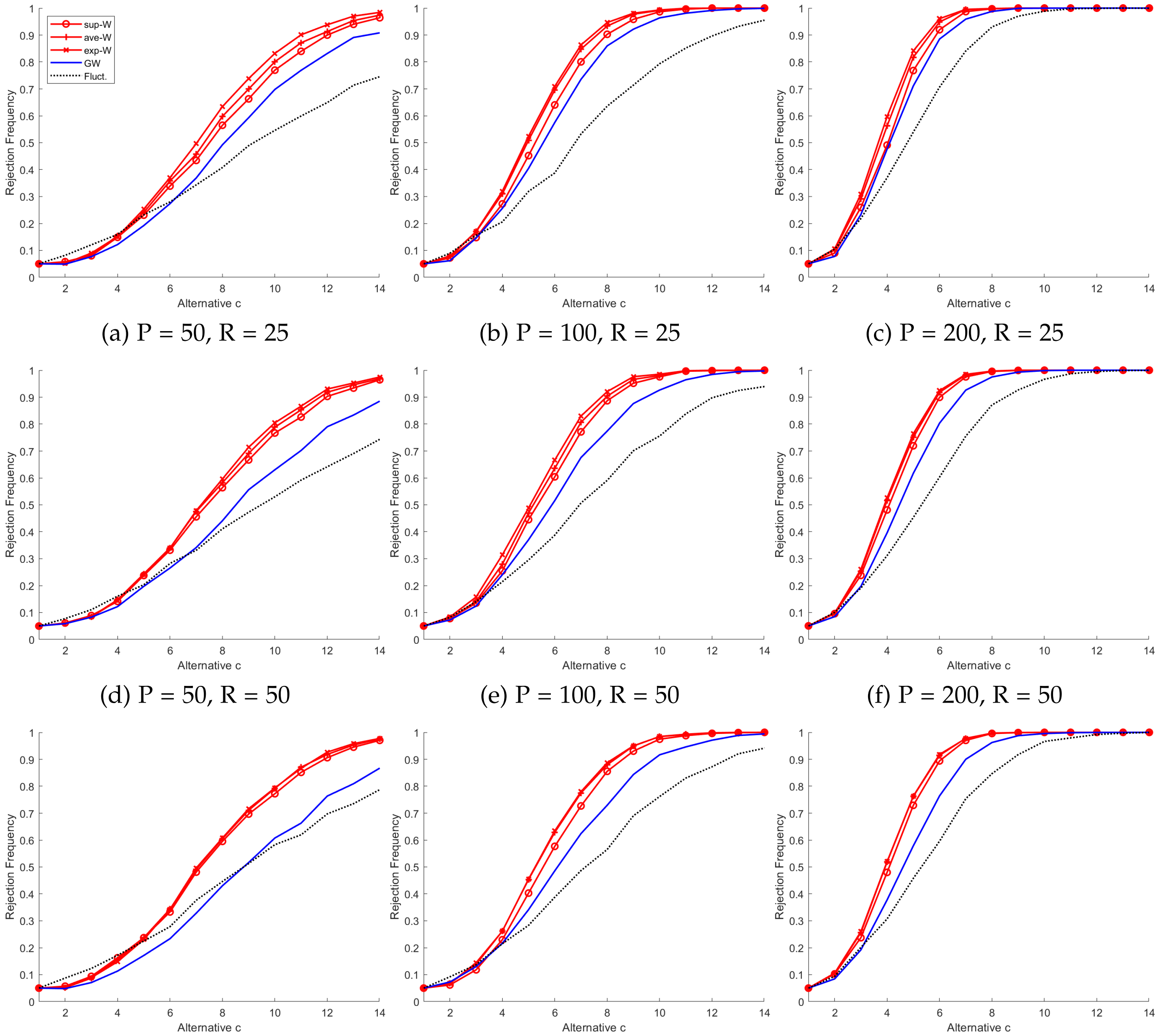
(g) P = 50, R = 100

(h) P = 100, R = 100

(i) P = 200, R = 100

Note: On the y-axis the figures displays size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test under Alternative (1) for density forecasts evaluated with the log score. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. The solid lines with markers "o", "x" and "+" display the sup-W, the exp-W, and the ave-W test results respectively. The dashed line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the solid line displays the results of the GW test. The nominal level is 5%. The results are based on 3,000 MC replications.

Figure C.5: Size-Adjusted Power Results for DF1, Alternative (2): State Dependence and Constant Deviation



(a) P = 50, R = 25

(b) P = 100, R = 25

(c) P = 200, R = 25

(d) P = 50, R = 50

(e) P = 100, R = 50

(f) P = 200, R = 50

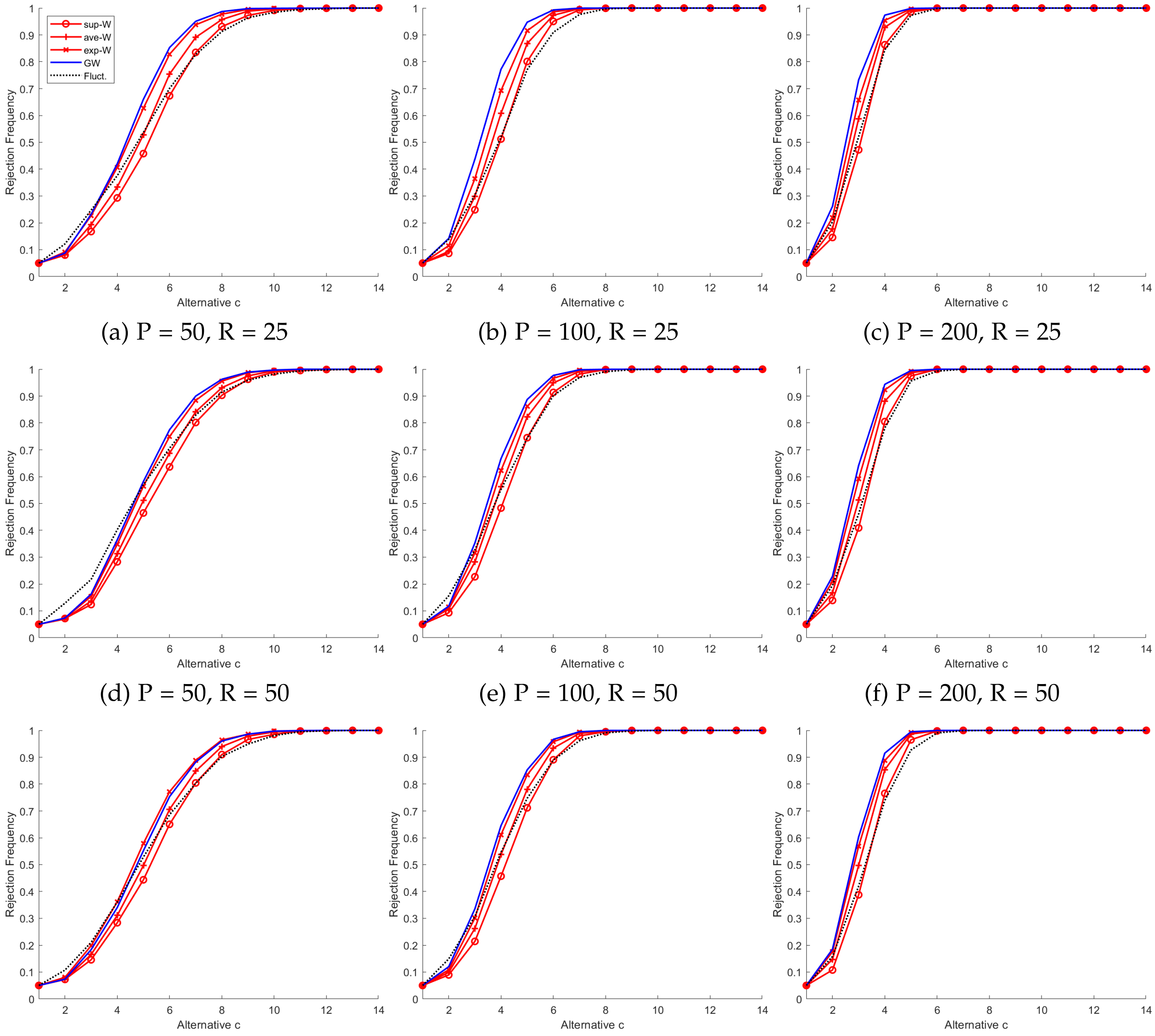
(g) P = 50, R = 100

(h) P = 100, R = 100

(i) P = 200, R = 100

Note: On the y-axis the figures displays size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test under Alternative (2) for density forecasts evaluated with the log score. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. The solid lines with markers “o”, “x” and “+” display the sup-W, the exp-W, and the ave-W test results respectively. The dashed line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the solid line displays the results of the GW test. The nominal level is 5%. The results are based on 3,000 MC replications.

Figure C.6: Size-Adjusted Power Results for DF1, Alternative (3): Constant Deviation



(a) P = 50, R = 25

(b) P = 100, R = 25

(c) P = 200, R = 25

(d) P = 50, R = 50

(e) P = 100, R = 50

(f) P = 200, R = 50

(g) P = 50, R = 100

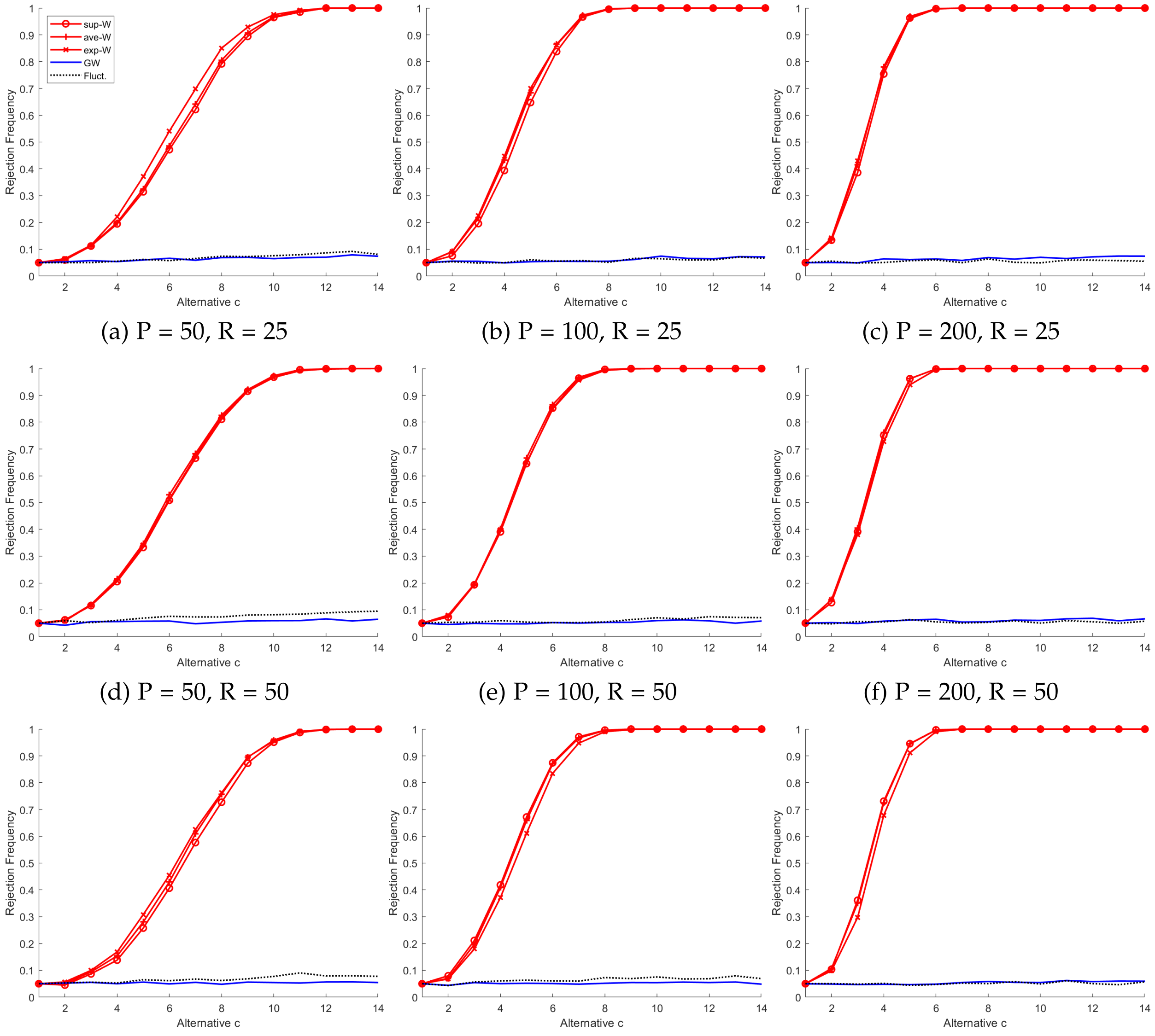
(h) P = 100, R = 100

(i) P = 200, R = 100

Note: On the y-axis the figures displays size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM$^{\text{NL}}$ test under Alternative (3) for density forecasts evaluated with the log score. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. The solid lines with markers "o", "x" and "+" display the sup-W, the exp-W, and the ave-W test results respectively. The dashed line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the solid line displays the results of the GW test. The nominal level is 5%. The results are based on 3,000 MC replications.

C.3 Density Forecast Comparison 2

Figure C.7: Size-Adjusted Power Results for DF2, Alternative (1): State Dependence



(a) P = 50, R = 25

(b) P = 100, R = 25

(c) P = 200, R = 25

(d) P = 50, R = 50

(e) P = 100, R = 50

(f) P = 200, R = 50

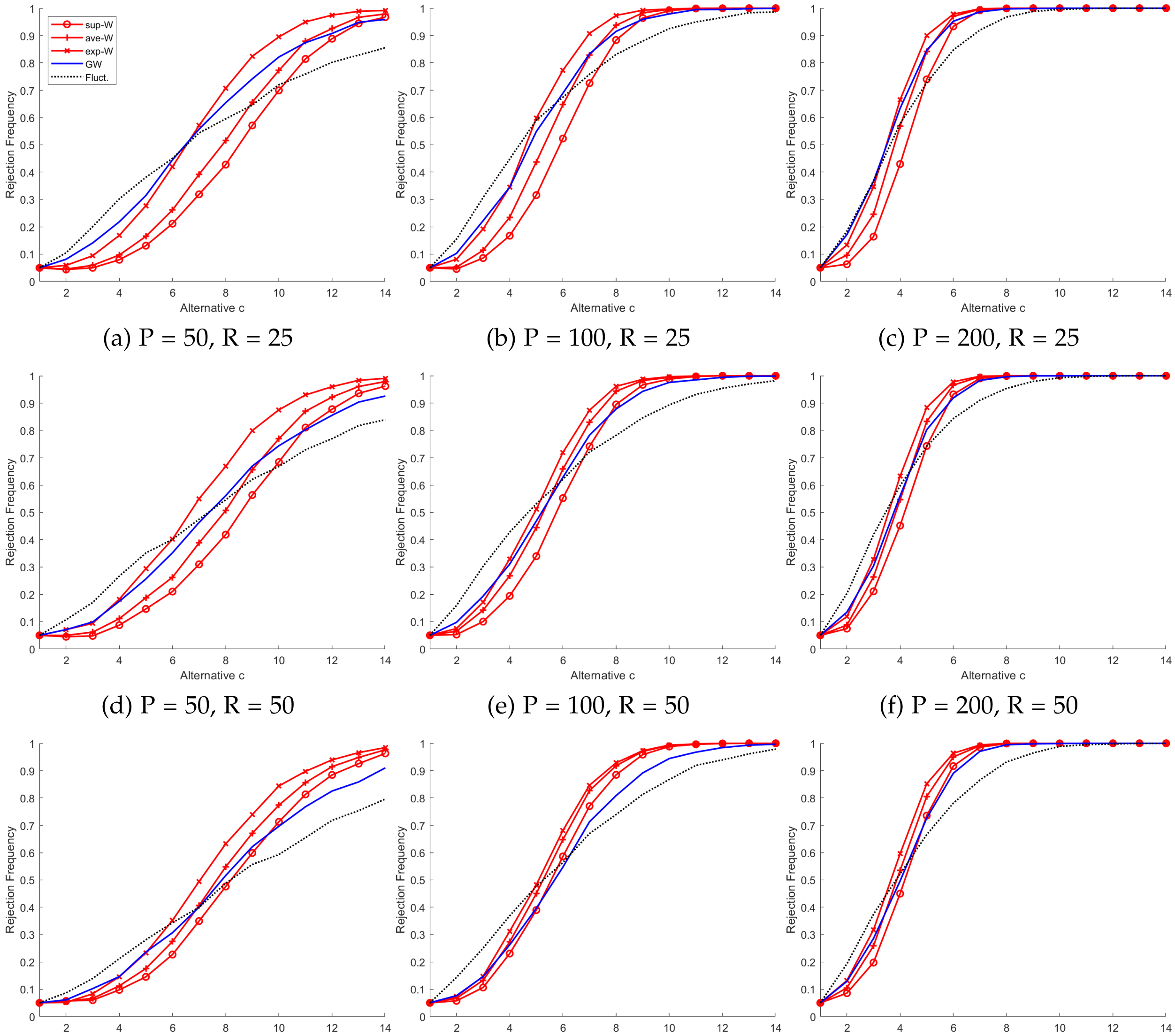
(g) P = 50, R = 100

(h) P = 100, R = 100

(i) P = 200, R = 100

Note: On the y-axis the figures displays size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test under Alternative (1) for density forecasts evaluated with the log score. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. The solid lines with markers "o", "x" and "+" display the sup-W, the exp-W, and the ave-W test results respectively. The dashed line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the solid line displays the results of the GW test. The nominal level is 5%. The results are based on 3,000 MC replications.

Figure C.8: Size-Adjusted Power Results for DF2, Alternative (2): State Dependence and Constant Deviation



(a) P = 50, R = 25

(b) P = 100, R = 25

(c) P = 200, R = 25

(d) P = 50, R = 50

(e) P = 100, R = 50

(f) P = 200, R = 50

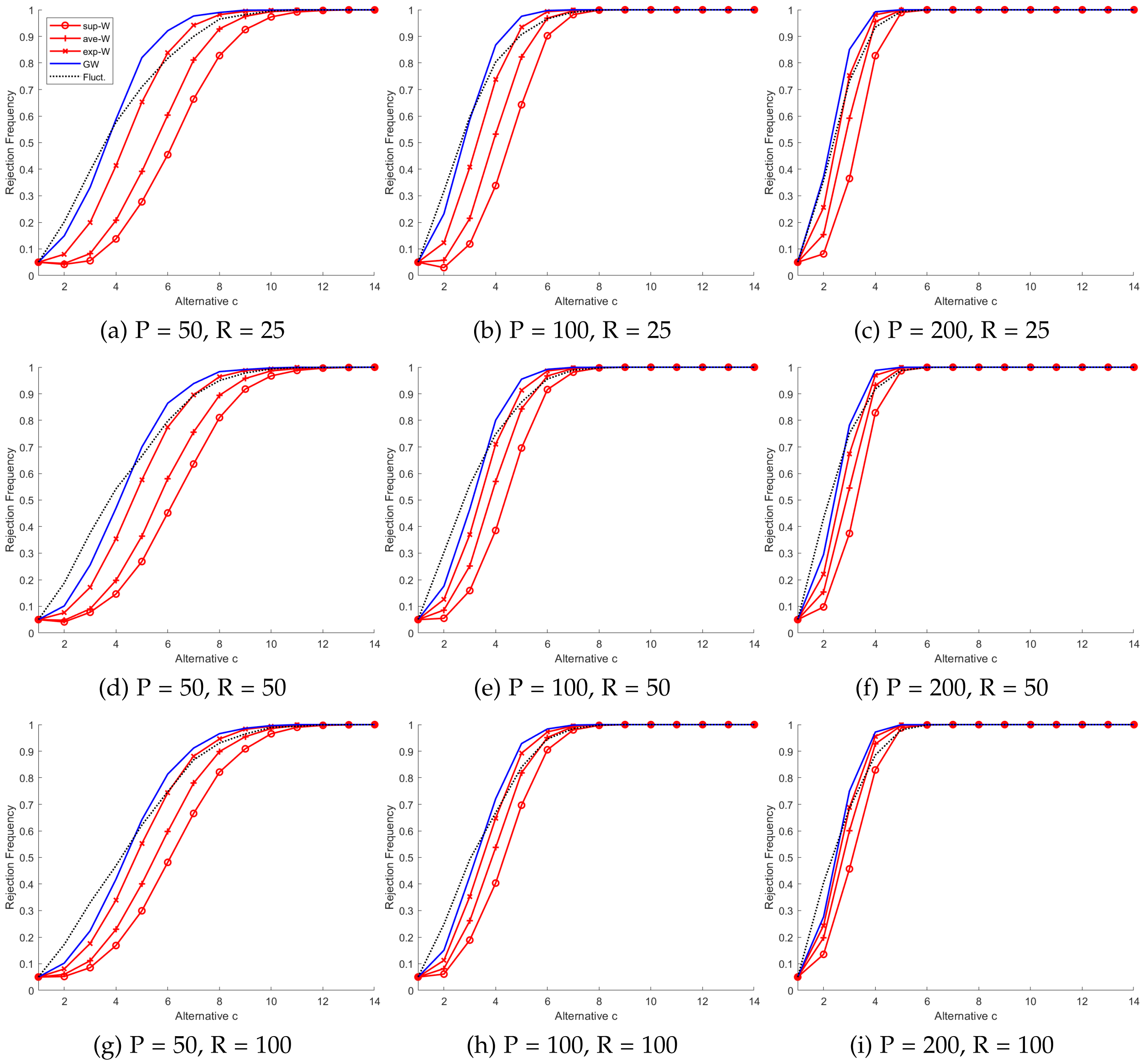
(g) P = 50, R = 100

(h) P = 100, R = 100

(i) P = 200, R = 100

Note: On the y-axis the figures displays size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test under Alternative (2) for density forecasts evaluated with the log score. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. The solid lines with markers "o", "x" and "+" display the sup-W, the exp-W, and the ave-W test results respectively. The dashed line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the solid line displays the results of the GW test. The nominal level is 5%. The results are based on 3,000 MC replications.

Figure C.9: Size-Adjusted Power Results for DF2, Alternative (3): Constant Deviation



(a) P = 50, R = 25

(b) P = 100, R = 25

(c) P = 200, R = 25

(d) P = 50, R = 50

(e) P = 100, R = 50

(f) P = 200, R = 50

(g) P = 50, R = 100

(h) P = 100, R = 100

(i) P = 200, R = 100

Note: On the y-axis the figures displays size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test under Alternative (3) for density forecasts evaluated with the log score. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. R denotes the in-sample parameter estimation window. P denotes the out-of-sample evaluation size. The solid lines with markers "o", "x" and "+" display the sup-W, the exp-W, and the ave-W test results respectively. The dashed line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the solid line displays the results of the GW test. The nominal level is 5%. The results are based on 3,000 MC replications.