



City Research Online

City St George's, University of London

Citation: Corradi, V., Fosten, J. & Gutknecht, D. (2024). Predictive ability tests with possibly overlapping models. *Journal of Econometrics*, 241(1), 105716. doi: 10.1016/j.jeconom.2024.105716

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/33295/>

Link to published version: <https://doi.org/10.1016/j.jeconom.2024.105716>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

Predictive Ability Tests with Possibly Overlapping Models*

Valentina Corradi[†]

Jack Fosten^{‡§}

Daniel Gutknecht[¶]

February 8, 2024

Abstract

This paper provides novel tests for comparing out-of-sample predictive ability of two or more competing models that are possibly overlapping. The tests do not require pre-testing, they allow for dynamic misspecification and are valid under different estimation schemes and loss functions. In pairwise model comparisons, the test is constructed by adding a random perturbation to both the numerator and denominator of a standard Diebold-Mariano test statistic. This prevents degeneracy in the presence of overlapping models but becomes asymptotically negligible otherwise. The test is shown to control the Type I error probability asymptotically at the nominal level, uniformly over all null data generating processes. A similar idea is used to develop a superior predictive ability test for the comparison of multiple models against a benchmark. Monte Carlo simulations demonstrate that our tests exhibit very good size control in finite samples reducing both over- and under-rejection relative to its competitors. Finally, an application to forecasting U.S. excess bond returns provides evidence in favour of models using macroeconomic factors.

JEL Classification: C12, C22, C53

Keywords: degeneracy, uniform inference, block bootstrap, out-of-sample evaluation, excess bond returns

*We are grateful to the Editor, Serena Ng, an Associate Editor, two anonymous referees as well as Majid Al Sadoon, Daniele Bianchi, Margaret Davenport, Matei Demetrescu, Juan Dolado, Jean-Marie Dufour, Simone Gianerini, Jesús Gonzalo, Liudas Giraitis, Emmanuel Guerre, Gary Koop, Rustam Ibragimov, Jordi Llorens-Terrazas, Dimitris Korobilis, Stuart McIntyre, James Mitchell, Valentin Patilea, Hashem Pesaran, Katerina Petrova, Anthoulla Phella, Francesco Ravazzolo, Jeff Racine, Roberto Renò, Mirco Rubin, Ovidijus Stauskas, Genaro Sucarrat, Martin Weale, and other seminar participants at the BI Norwegian Business School, at the University of Glasgow, the iCEBA 2023 in Tashkent, the NBER TSF Conference 2023 in Montreal, the Goodness-Of-Fit, Change-Point workshop 2023 at the North-West University (SA), the Florence-Paris workshop on stochastic processes and its application to financial econometrics at the University of Florence, the ICEEE meeting 2023 at the University of Cagliari, the econometrics and learning workshop at Imperial College London, the Queen Mary University of London, the University of Strathclyde, the workshop for statistical learning and econometrics at the Free University Bozen-Bolzano, the Frankfurt Econometrics workshop, the 13th Workshop in Time Series Econometrics, University of Zaragoza, the King's Business School Economics Department research workshop, and the ESCoE Conference on Economic Measurement 2023.

[†]Department of Economics, University of Surrey, School of Economics, Guildford, GU2 7XH, UK, E-mail: V.Corradi@surrey.ac.uk

[‡]King's Business School, King's College London, WC2B 4BG London, UK, E-mail: jack.fosten@kcl.ac.uk

[§]Data Analytics in Finance and Macro (DAFM) Research Centre, King's College London, UK

[¶]Faculty of Economics and Business, Goethe University Frankfurt, 60629 Frankfurt am Main, Germany. E-mail: Gutknecht@wiwi.uni-frankfurt.de.

1 Introduction

This paper proposes a unified framework for pairwise and multiple out-of-sample model comparison tests in the presence of possibly overlapping models. Under the null of equal expected predictive ability, overlapping models occur when different sets of predictors are used but only a common subset (or none) of them has predictive power. In this situation, which comprises strictly nested models as a subcase, inference becomes very challenging as the asymptotic distribution of the Diebold-Mariano test (Diebold and Mariano, 1995, DM) can depend on a variety of factors. Under quadratic loss and with martingale difference errors, ruling out dynamic misspecification, Clark and McCracken (2014) show scenarios in which the distribution of the DM statistic can either be standard normal or a functional of Brownian motion dependent on unknown nuisance parameters. When we allow for the more realistic case of dynamic misspecification, the DM statistic may even converge to zero in probability. With the exception of Clark and McCracken (2014), who propose a pre-testing two-step procedure or a conservative one-step procedure, the case of overlapping models has remained largely unexplored. This is in contrast to the case of comparing strictly nested models which has received more attention (e.g. Clark and McCracken, 2001; Corradi and Swanson, 2002; McCracken, 2007; Pitarakis, 2023; Amburgey and McCracken, 2023). We aim to provide a test which is simple to implement and does not rely on any pre-testing.

We make two new contributions in this paper. Firstly, for the case of pairwise comparisons, we introduce a single step DM-type test, which is asymptotically standard normal regardless of whether models are overlapping or not. This holds under different estimation schemes such as the fixed or the recursive scheme, for a generic loss function, and it allows for the comparison of dynamically misspecified models. Most importantly, our test asymptotically attains its nominal level uniformly over all data generating processes (DGPs) under the null hypothesis of equal predictive ability. Secondly, we also consider a test for superior predictive ability (SPA) involving multiple models, as in White (2000) and Hansen (2005), allowing for the possibility that all competing models are overlapping with the benchmark. We use a weak moment inequalities set-up, relying on the generalized moment selection (GMS) approach of Andrews and Soares (2010), and establish the asymptotic validity of bootstrap based critical values. This holds regardless of the number of competing models which overlap with the benchmark. Our approach does not rely on pre-testing, and the test achieves asymptotic non-conservative size control also in the presence of some (but not all) models that overlap.

Our tests operate by introducing an artificial random perturbation into the test statistic of the DM or SPA type tests. Specifically, for the DM test this is done in such a way that the added randomness does not contribute to the asymptotic distribution when the models are strictly non-nested or “moderately” overlapping, i.e. the variance of the original DM statistic approaches zero at a sufficiently slow rate. On the other hand, when the variance of the DM statistic collapses to zero quickly, the added randomness dominates the limiting distribution thereby solving any degeneracy issue. Inference in the DM test can be conducted using standard normal critical values, and we

establish an asymptotic size for our modified DM test that is well controlled at the nominal level.

For the SPA test, we add a different random perturbation to each pairwise comparison. Asymptotically, this does not affect the number of inferior models that give rise to a “slack” moment inequality and that do not contribute to the limiting distribution since the perturbation has mean zero. In this multiple comparison case, we show that the limiting distribution is a functional of a Gaussian process, and thus there are no ready-to-use critical values. We provide bootstrap critical values based on the block bootstrap which are simple to compute, and establish their first-order asymptotic validity.

The implementation of our test is straightforward and requires little more input than that of commonly-used DM type tests. The random perturbation is multiplied by a tuning parameter which controls its influence and vanishes to zero as the sample size grows. Following a similar approach to Schennach and Wilhelm (2017), we propose a data-driven choice based on the minimization of the power loss due to the added perturbation when models do not overlap, and on the minimization of size distortion when they do overlap. Constructing the data-driven tuning parameter only requires the estimation of objects such as the (asymptotic) score and the Hessian matrix of the objective function, which can be easily obtained from most statistical software packages. This makes our test very accessible to applied forecasters.

We illustrate the finite sample properties of our pairwise and multiple model tests using Monte Carlo simulations. The results confirm the theory in showing that our tests, indeed, control the Type I error probability accurately regardless of whether we are in the overlapping or strictly non-nested case. In the pairwise comparison context, our test achieves a rejection rate close to the nominal level, in contrast to the non-adjusted DM test which becomes very conservative in the overlapping case and too liberal in the non-nested case. Our test therefore addresses the well-known issue of over-rejection by the standard DM statistic in non-nested model comparisons (see the discussion in Coroneo and Iacono, 2020). Similar results also carry over to the multiple comparison testing case where we compare our SPA test with an unadjusted version and the SPA test of Hansen (2005). Here we find that the power of our test is similar to the others whereas we, again, make substantial improvements in controlling the size.

We then apply our test using an empirical example of forecasting U.S. excess bond returns at various maturities, using combinations of bond market variables and macroeconomic factors estimated from the FRED-MD database (McCracken and Ng, 2016). We therefore contribute to a recent and growing literature, initiated by Ludvigson and Ng (2009), which looks to assess whether factors extracted from macroeconomic series have additional explanatory power over the traditional financial series used in explaining bond risk premia such as forward spreads (Fama and Bliss, 1987), yield spreads (Campbell and Shiller, 1991) and forward rate factors (Cochrane and Piazzesi, 2005). We find that macroeconomic factors are, indeed, useful at predicting excess bond returns, especially when using a parsimonious representation involving only the first factor from the FRED-MD database. Our SPA test uncovers statistical evidence that the use of macro data produces superior forecasts and we

find that there may be different predictive methods required across different bond maturities.

While our main theoretical, simulation and empirical results cover the case of differentiable loss functions, in the supplement we also treat the case of the non-differentiable check loss function. This shows that our tests may also be applied to the evaluation of quantile forecasting models which have gained importance in the recent empirical literature (e.g. Adrian et al., 2019). In this case, even though we cannot provide a data-driven procedure as in the differentiable loss case because we cannot have a precise convergence rate for parameter estimation error (PEE), we can still provide an ad-hoc procedure that works very well in finite samples.

In placing our paper within the literature, we can trace the study of model comparisons with overlapping models back to Vuong (1989). The classical “Vuong test” has recently been extended by proposing alternatives that attain the nominal level asymptotically across all DGPs, either through simulation based techniques (Shi, 2015), sample splitting (Schennach and Wilhelm, 2017), or the addition of artificial randomness (Hsu and Shi, 2017). In the pairwise comparison case, our statistic is closest to Hsu and Shi (2017). These papers all deal with independent and identically distributed observations which rules out dynamic misspecification which our tests are able to accommodate. Moreover, none of the aforementioned papers focus on forecast comparison tests under recursive or rolling parameter estimation as we do in our paper. Finally, to the best of our knowledge, no existing Vuong-type test accommodates the comparison of multiple models.

An alternative literature proposing tests which circumvent the degeneracy of the DM statistic with overlapping models is the conditional testing procedure of Giacomini and White (2006). This approach uses a small, fixed-length rolling window of observations for estimation, ensuring that PEE does not vanish asymptotically. A recent paper by Zhu and Timmermann (2020), however, shows that the null of Giacomini and White (2006) is not valid under this rolling estimation scheme unless the data satisfy a restrictive m -dependence assumption, with m denoting the number of observations used in the rolling window.

The rest of the paper is organized as follows. Section 2 first outlines the pairwise comparison with possibly overlapping models, and establishes the limiting distribution of the suggested statistic in the case of a deterministic tuning parameter. Then, it introduces a data-driven procedure for the choice of the tuning parameter and establishes the uniform asymptotic equivalence between the statistics based on a deterministic versus data-driven tuning parameter. Section 3 outlines multiple comparison in the presence of an unknown number of overlapping models. Section 4 provides some Monte Carlo evidence, while Section 5 provides the empirical application. Section 6 concludes. The proofs of the main theorems are provided in the Appendix, while auxiliary lemmas and other theorems are relegated to the supplementary material together with an extension to non-differentiable loss functions and some additional graphs.

2 Diebold-Mariano Type Test

2.1 Set-up and Limiting Distribution

In the context of DM type tests, it is a well known fact that the contribution of PEE to the variance of the test statistic vanishes asymptotically if the same loss function is used for both estimation and out of sample evaluation. This occurs regardless of whether, as the sample size $T \rightarrow \infty$, $P/R \rightarrow 0$, or $P/R \rightarrow \pi$ with $0 < \pi < \infty$, where R denotes the number of observations in the estimation sample, while P denotes the part of the sample used for (pseudo) out-of-sample prediction so that $T = R + P$. Examples include the use of ordinary least squares (OLS) or nonlinear least squares (NLS) for model estimation and a quadratic loss function for forecast evaluation (see West, 1996), or the least absolute deviation (LAD) estimator for model estimation and the absolute deviation loss function for prediction evaluation (see McCracken, 2000).¹ On the other hand, even when different loss functions are used for estimation and evaluation, the contribution of PEE to the variance becomes asymptotically negligible whenever $P/R \rightarrow 0$ as $T \rightarrow \infty$. In both cases, under the null hypothesis of equal expected predictive accuracy, the statistic of the DM test does no longer follow a standard normal limiting distribution when models overlap because only a common subset (or none) of the regressors holds predictive power. This in turn results in misguided inference when using standard normal critical values. The objective of this paper is instead to propose a test statistic which is asymptotically normal under all null DGPs.

Hereafter, let $u_{j,t+h} = (y_{t+h} - y_{j,t+h|t})$ and $\hat{u}_{j,t+h} = (y_{t+h} - \hat{y}_{j,t+h|t})$ denote the population and estimated prediction error of model j , for $j = 1, 2$, where y_{t+h} denotes the target variable of interest and $h \geq 1$ is the forecast horizon. We denote $y_{j,t+h|t}$ as the infeasible prediction at time t for time $t+h$ if we knew the parameters of model j , and $\hat{y}_{j,t+h|t}$ the prediction at time t for time $t+h$, based on the estimated parameters. Specifically, let:

$$u_{j,t+h} = y_{t+h} - m_j \left(Z_{j,t}; \theta_j^\dagger \right)$$

where $Z_{j,t}$ is the set of predictors used by model j , and $m_j(Z_{j,t}; \cdot)$ is a (non)linear function of $Z_{j,t}$ known up to some population parameter vector θ_j^\dagger . We also let:

$$\hat{u}_{j,t+h} = y_{t+h} - m_j \left(Z_{j,t}; \hat{\theta}_{j,1:t} \right). \quad (1)$$

Here, $\hat{\theta}_{j,1:t}$ denotes the estimator of the parameter vector using the recursive estimation scheme, in which we re-estimate the parameters each period using all observations available up to time t with $t \geq R$. On the other hand, when the fixed estimation scheme is used, where we estimate θ_j^\dagger only once using the first R observations, we replace $\hat{\theta}_{j,1:t}$ in Equation (1) by $\hat{\theta}_{j,1:R}$. Finally, when the rolling

¹ PEE becomes asymptotically negligible because the derivative of the expected value of the (forecast) loss, evaluated at the population parameters, equals zero.

scheme with the most recent R observations is used, we use $\widehat{\theta}_{j,(t-R+1):t}$ instead of $\widehat{\theta}_{j,1:t}$.

Given these definitions, let \mathcal{F} denote the set of distributions defined on the support of $(y_{t+h}, Z'_{j,t})$, such that Assumptions A.1, A.2, and A.4 below hold. The corresponding probability measures and expectation for each $F \in \mathcal{F}$ will be denoted by \Pr_F and E_F in what follows. For the recursive estimation scheme define:

$$\theta_j^\dagger = \arg \min_{\theta_j \in \Theta_j} \frac{1}{t} \sum_{k=1}^{t-h} E_F (q(y_{k+h} - m_j(Z_{j,k}; \theta_j))) \quad \text{with } t \geq R$$

and:

$$\widehat{\theta}_{j,1:t} = \arg \min_{\theta_j \in \Theta_j} \frac{1}{t} \sum_{k=1}^{t-h} q(y_{k+h} - m_j(Z_{j,k}; \theta_j)) \quad \text{with } t \geq R,$$

where $q(\cdot)$ denotes the estimation loss function. For the remainder of the paper we focus on the recursive estimation scheme, but note that the results can be extended to the other schemes too.

As the main case of this paper, we suppose that the loss functions $g(\cdot)$ and $q(\cdot)$ satisfy the differentiability condition in Assumption A.3 below, and that the same loss function is used both for estimation and out of sample evaluation, $q(\cdot) = g(\cdot)$. As noted earlier, an implication of the latter condition is that the contribution of estimation error becomes asymptotically negligible (cf. West, 1996). On the other hand, when $P/R \rightarrow 0$, our set-up is also applicable even when $q(\cdot)$ and $g(\cdot)$ differ from each other. An example is when forecasting models are estimated via (ordinary or nonlinear) least squares, but the evaluation loss is asymmetric such as the linear exponential (Linex) loss of the form $g(u) = \exp(au) - au - 1$ with $a \neq 0$ (Zellner, 1986). The Linex loss function has recently been used to evaluate forecasts by Coroneo et al. (2023). Moreover, Hansen and Dumitrescu (2022) even point to a possible robustness versus efficiency trade-off when a likelihood estimator is used at the model estimation stage rather than simply setting $q(\cdot) = g(\cdot)$. Finally, in Section S3 of the supplementary material, we also examine the case where the check loss function is used for quantile regression forecasts, which is a primary example of a non-differentiable loss function.

Given some loss function for evaluation $g(\cdot)$ and $F \in \mathcal{F}$, the null hypothesis in a standard DM test is given by:

$$H_0 : E_F (g(u_{1,t+h}) - g(u_{2,t+h})) = 0. \quad (2)$$

The alternative hypothesis on the other hand is:

$$H_A : E_F (g(u_{1,t+h}) - g(u_{2,t+h})) \neq 0.$$

We collect all DGPs F that satisfy Equation (2) in the following null set:

$$\mathcal{F}_0 \equiv \{F \in \mathcal{F} : H_0 \text{ holds}\}, \quad (3)$$

while $\mathcal{F}_A = \{F \in \mathcal{F} : H_A \text{ holds}\}$ is the complement of \mathcal{F}_0 , and denotes the set of DGPs under the

alternative hypothesis. As outlined in the Introduction, the main goal of this paper is to examine the asymptotic behaviour of the DM test uniformly over the class of null DGPs \mathcal{F}_0 . To outline our asymptotic framework and the test statistic, we introduce the following notation: let $\{F_P\}_{P=1}^\infty$ denote a sequence of distributions with $F_P \in \mathcal{F}$ for all P and define:

$$\sigma_{DM}^2(F_P) \equiv \sigma_{DM,P}^2 = \text{var}_{F_P} \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} (g(u_{1,t+h}) - g(u_{2,t+h})) \right)$$

as well as $\sigma_{DM}^2 \equiv \lim_{P,R \rightarrow \infty} \sigma_{DM,P}^2$. We say that two models are *overlapping* if for some sequence with $F_P \in \mathcal{F}_0$ for all P it is the case that $\lim_{P,R \rightarrow \infty} \sigma_{DM,P}^2 = \sigma_{DM}^2 = 0$. On the other hand, we say that two models are *non-overlapping* or *strictly non-nested* when, for some sequence with $F_P \in \mathcal{F}_0$, for all P it holds that $\lim_{P,R \rightarrow \infty} \sigma_{DM,P}^2 > 0$. Note that \mathcal{F}_0 contains strictly non-nested DGPs for which $\sigma_{DM}^2 > 0$, as well as a ‘continuum’ of overlapping DGPs for which either $\lim_{P,R \rightarrow \infty} P\sigma_{DM,P}^2 = c$, for some $0 \leq c < \infty$, or $\lim_{P,R \rightarrow \infty} P\sigma_{DM,P}^2 = \infty$. In other words, we allow the variance to approach zero at rate P or at any possible faster or slower rate. Our objective is to construct a test that asymptotically attains its nominal level uniformly over \mathcal{F}_0 .

We suggest the following augmented DM statistic:

$$\widehat{DM}_P(\eta_{P,R}) = \frac{\frac{1}{\sqrt{P}} \sum_{t=R}^{T-h} (g(\widehat{u}_{1,t+h}) - g(\widehat{u}_{2,t+h})) + \eta_{P,R} \frac{1}{\sqrt{P}} \sum_{t=1}^P e_t}{\sqrt{\widehat{\sigma}_{DM,P}^2 + \eta_{P,R}^2 \frac{1}{P} \sum_{t=1}^P e_t^2}}, \quad (4)$$

where e_t are i.i.d. $N(0, 1)$ draws, $\eta_{P,R}$ is a positive, deterministic sequence with $\eta_{P,R} \rightarrow 0$ as $P, R \rightarrow \infty$, and:

$$\widehat{\sigma}_{DM,P}^2 = \frac{1}{P} \sum_{t=R+l_P}^{T-h-l_P} \sum_{\tau=-l_P}^{l_P} \omega(\tau, l_P) (g(\widehat{u}_{1,t+h}) - g(\widehat{u}_{2,t+h})) (g(\widehat{u}_{1,t+\tau+h}) - g(\widehat{u}_{2,t+\tau+h})) \quad (5)$$

is a heteroskedasticity and autocorrelation consistent (HAC) estimator of the variance, where l_P denotes the usual lag truncation parameter satisfying $l_P \rightarrow \infty$ as $P \rightarrow \infty$ and $\omega(\tau, l_P) = (1 - (|\tau|/l_P))$, see Newey and West (1987).

The parameter $\eta_{P,R}$ here plays the role of a regularisation parameter that governs the contribution of artificial noise to the test statistic. Specifically, from Equation (4), it is immediate to see that if $\eta_{P,R} = 0$, $\widehat{DM}_P(\eta_{P,R})$ collapses to a standard DM statistic with a HAC estimator of the variance, say \widehat{DM}_P^S . Clark and McCracken (2014) studied the behaviour of the latter in the possibly overlapping case, when $g(\cdot)$ is a quadratic function, and l_P in Equation (5) is set equal to zero and $\omega(\tau, l_P) = 1$. They show that if either $P/R \rightarrow 0$ or a fixed estimation scheme is used, then the DM test with sample variance estimator, say \widehat{DM}_P^{S0} , is asymptotically standard normal, regardless of whether the models are overlapping or not. More generally, however, when $\pi > 0$ or the more common recursive (or rolling) estimation scheme is used, this is no longer the case. In fact, in the latter case the

limiting distribution is standard normal for strictly non-nested models and is a ratio of functionals of Brownian motion in the overlapping case.

Remark 1: As an alternative to the studentized statistic $\widehat{DM}_P(\eta_{P,R})$ in Equation (4), we could replace $P^{-1/2} \sum_{t=1}^P e_t$ in the numerator with a single draw from the standard normal distribution, say Z , and $P^{-1} \sum_{t=1}^P e_t^2$ in the denominator with its probability limit one. In fact, in the pairwise case, the two statistics can be shown to be asymptotically equivalent uniformly over all null DGPs and they behave similarly in Monte Carlo simulations. On the other hand, the block bootstrap procedure for constructing critical values for the SPA test in Section 3 would no longer apply if we only have one sample draw Z as the added randomness component could not be resampled. Given this limitation, and since multiple model comparisons via SPA tests are increasingly common in empirical studies, we do not pursue this alternative.

In the following, we first derive the asymptotic properties for the case of $\eta_{P,R}$ being a deterministic sequence (see Theorem 1 below). Then we introduce a data-driven choice for $\eta_{P,R}$ in Subsection 2.2, and show that the two statistics are asymptotically equivalent over \mathcal{F}_0 . Moreover, as remarked above, to allow for possible dynamic misspecification, we have replaced the sample variance in the standard DM test by a HAC estimator in Equation (4), where $l_P \rightarrow \infty$ as $P \rightarrow \infty$. The key difference with respect to a standard DM statistic based on the sample variance, \widehat{DM}_P^{SO} , is that the numerator is now of a smaller probability order than the denominator. Thus, in the overlapping case, without the random perturbation, the statistic would approach zero in probability implying a test of asymptotic level zero when standard normal critical values are used.

We next state the assumptions required to derive the limiting distribution of the test statistic in Equation (4), and set the forecast horizon $h = 1$ without loss of generality. Since we also consider the general case with $J > 2$ models below, we will state all conditions for a generic (finite) number of models, J . In addition, let $\nabla^{(k)} f(\cdot)$ denote the k -th order derivative of the function $f(\cdot)$ with respect to its argument, and note that $\|\cdot\|$ represents the Euclidean norm. Whenever the function has several arguments, we also add a subscript of the argument for clarity.

Assumption A.1: For all $j = 1, \dots, J$, $(y_{t+1}, Z'_{j,t})'$ are strictly stationary, absolutely regular and β -mixing with size $-4(4 + \psi)/\psi$, $\psi > 0$.

Assumption A.2: For all $j = 1, \dots, J$ and $F \in \mathcal{F}$, $m_j(Z_{j,t}; \cdot)$ is, with probability one, twice continuously differentiable on the interior of Θ_j , with Θ_j compact, and θ_j^\dagger lies in that interior. The elements of $\sup_{\theta \in \Theta_j} \|\nabla_{\theta_j}^{(1)} m_j(Z_{j,t}; \theta)\|$ and $\sup_{\theta \in \Theta_j} \|\nabla_{\theta_j}^{(2)} m_j(Z_{j,t}; \theta)\|$ are p -dominated with $p > 2(4 + \psi)$ and ψ defined in A.1.²

²We say that a random function $f_j(Z_{j,t}; \theta)$ is p -dominated if for each element i of that function, say $f_{i,j}(Z_{j,t}; \theta)$, there exists a measurable, real-valued dominating function $D_{i,j}(Z_{j,t})$ such that $\sup_{\theta \in \Theta_j} \|f_i(Z_{j,t}; \theta)\| \leq D_{i,j}(Z_{j,t})$ a.s. and $E_F(|D_{i,j}(Z_{j,t})|^p) \leq \Delta < \infty$, see e.g. White and Domowitz (1984).

Assumption A.3: The loss function $f \in \{g, q\}$, is twice continuously differentiable on the interior of its domain, an interval of the extended real line which includes zero. The function satisfies the following conditions:

- (i) $f(u) = 0$ for $u = 0$,
- (ii) $\nabla^{(1)}f(u) \leq 0$ for $u \leq 0$ and $\nabla^{(1)}f(u) \geq 0$ for $u \geq 0$,
- (iii) $\nabla^{(2)}f(u) > 0$ for all u in the interior of the domain.

Assumption A.4: For all $j = 1, \dots, J$ and $F \in \mathcal{F}$, it holds that:

- (i) for $f \in \{g, q\}$ and all $\theta_j \in \Theta_j$, the elements of $\nabla_{\theta_j}^{(1)}f(y_{t+1} - m_j(Z_{j,t}; \theta_j))$ and $\nabla_{\theta_j}^{(2)}f(y_{t+1} - m_j(Z_{j,t}; \theta_j))$ defined in Equations (29) and (30) of the Appendix, respectively, are p -dominated with $p > 2(4 + \psi)$ and ψ as in A.1,
- (ii) the matrices:

$$V_{j,F} = \lim_{P,R \rightarrow \infty} \sup_{t \geq R} \text{var}_F \left(\frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} \nabla_{\theta_j}^{(1)}q(y_{t+1} - m_j(Z_{j,t}; \theta_j^\dagger)) \right) \quad (6)$$

and:

$$H_{j,F} = \lim_{P,R \rightarrow \infty} \sup_{t \geq R} \text{E}_F \left(\frac{1}{t} \sum_{t=1}^{t-1} \nabla_{\theta_j}^{(2)}q \left(y_{t+1} - m(Z_{j,t}; \theta_j^\dagger) \right) \right), \quad (7)$$

with $\nabla_{\theta_j}^{(1)}q(y_{t+1} - m_j(Z_{j,t}; \theta_j^\dagger))$ and $\nabla_{\theta_j}^{(2)}q(y_{t+1} - m_j(Z_{j,t}; \theta_j^\dagger))$ defined as in part (i) with $f = q$ are positive definite.

Assumption A.5: As $T \rightarrow \infty$, $P/R \rightarrow \pi$, with $0 \leq \pi < \infty$.

Assumption A.6: It holds that, as $T \rightarrow \infty$, $l_P \rightarrow \infty$, $\eta_{R,P} \rightarrow 0$ and (i) $l_P P^{-1/2} \rightarrow 0$, (ii) $(Pl_P^{-1})^{1/4} \eta_{R,P} \rightarrow \infty$ and (iii) $l_P R^{-1/3} ((\ln \ln R)(\ln \ln T))^{1/3} \rightarrow 0$ if $g(\cdot) = q(\cdot)$ with $0 \leq \pi < \infty$ or $l_P (P/R)^{-1/3} (\ln \ln T)^{1/3} \rightarrow 0$ if $g(\cdot) \neq q(\cdot)$ with $\pi = 0$.

Assumption A.7: It holds that $\mathbf{e}_t = (e_{2,t}, \dots, e_{J,t})' \stackrel{i.i.d.}{\sim} N(0_{(J-1)}, I_{(J-1) \times (J-1)})$, where $0_{(J-1)}$ denotes the $(J-1)$ dimensional zero vector and $I_{(J-1) \times (J-1)}$ the $(J-1) \times (J-1)$ dimensional identity matrix.

Assumptions A.1 and A.2 are standard conditions from the forecasting literature on the time dependence and model class (for example Corradi and Swanson, 2006, 2007) that do not warrant further discussion but for the moment conditions. In fact, Assumption A.2 imposes restrictions on the moments of the first two derivatives of $m_j(Z_{j,t}; \cdot)$. In particular, the requirement of the existence of at least $8 + 2\psi$ moments is of technical nature since in the construction of a data-driven regularisation parameter in Subsection 2.2 below we rely on an almost sure bound for PEE, which in turn builds on a strong invariance principle result from Corradi (1999) requiring at least $8 + 2\psi$

moments for the almost sure convergence of the HAC variance estimator.³ Assumption A.3 on the other hand is satisfied by twice-differentiable and convex loss functions, such as the quadratic or the Linex loss. Note that the current set-up rules out non-differentiable loss functions like the check loss used in linear quantile regression. We provide an extension to this loss function, which is frequently used in Value-at-Risk or Growth-at-Risk forecasting, in Section S3 of the supplement. Assumption A.4(i) imposes again conditions on the existence of moments that parallel Assumption A.2, but this time the conditions involve the first and second order derivatives of the models. In particular, whether these moment conditions are satisfied or not depends not only on the loss functions $g(\cdot)$ and $q(\cdot)$, but also on the underlying DGP.⁴ Part (ii) of A.4 on the other hand concerns the positive definiteness of the score and Hessian of each model j . In fact, when $g(\cdot) = q(\cdot) = (\cdot)^2$ and under strict stationarity, these quantities are given by the well-known expressions $V_{1,F} = 4E_F(u_{j,t+1}^2(Z_{j,t}Z'_{j,t}))$ and $H_{j,F} = 2E_F(Z_{j,t}Z'_{j,t})$, respectively.

Assumption A.6 places rate conditions on the truncation lag parameter l_P and the regularisation parameter $\eta_{P,R}$. In particular, A.6(i) is a standard rate condition on l_P in the strictly stationary case (e.g. Andrews, 1991), while A.6(ii) places an additional condition on the relative rate of l_P and $\eta_{P,R}$. The latter is in turn required in the proofs of Theorems 1.1 and in Theorem 2.2, respectively, while A.6(iii) is instead used only in the context of Theorem 2.2 to establish the rate at which a data-driven tuning parameter $\widehat{\eta}_{P,R}$ converges to its deterministic counterpart $\eta_{P,R}$. Finally, setting $e_t \equiv e_{2,t}$ in the pairwise case, Assumption A.7 formalizes the random perturbation added to the statistic in Equation (4). Also note that the independence of the draws across models $j = 2, \dots, J$ will be required for the multiple model SPA test introduced in Section 3 below.

We start by deriving a pointwise result for the asymptotic behaviour of our test under the null hypothesis and under any fixed alternative. Thus, recalling the definition of \mathcal{F}_A as the collection of $F \in \mathcal{F}$ such that H_A holds, note that all DGPs under the fixed alternative have to be strictly non nested with $\sigma_{DM}^2 > 0$. We have the following result.

Theorem 1.1: Let Assumptions A.1-A.7 hold. Then:

(i) Under H_0 if either $g(\cdot) = q(\cdot)$ and $0 \leq \pi < \infty$ or $g(\cdot) \neq q(\cdot)$ and $\pi = 0$, for any given $F \in \mathcal{F}_0$,

$$\widehat{DM}_P(\eta_{P,R}) \xrightarrow{d} N(0, 1),$$

³In fact, the result in Corradi (1999) could be relaxed to a bound in probability, in which case there is also scope to relax the moment requirement.

⁴For instance, suppose a linear forecast model $y_{t+1} = Z_{j,t}\theta_j^\dagger + u_{j,t+1}$, and that the DGP is in fact given by $y_{t+1} = Z_{j,t}\theta_j + u_{j,t+1}$ with $(Z_{j,t}, u_{j,t+1})' \sim N(0_2, I_{2 \times 2})$, where 0_2 denotes the zero vector and $I_{2 \times 2}$ the (2×2) dimensional identity matrix. In addition, suppose that the researcher evaluates her forecast via Linex loss setting $a = -1$ so that $\nabla_{\theta_j}^2 g(y_{t+1} - m_j(Z_{j,t}; \theta_j^\dagger)) = Z_{j,t}^2 \exp(-(y_{t+1} - Z_{j,t}\theta_j^\dagger))$, while the parameter θ_j^\dagger is estimated via OLS. In this case, choosing ψ from Assumption A.1 to be $\psi = 0.5$, we have that A.4(i) is indeed satisfied as $E\left(\left(Z_{j,t}^2 \exp(-u_{j,t+1})\right)^{10}\right) < \infty$. A similar argument can also be made for the first order derivative. On the other hand, in the above example, Assumption A.4(i) will generally not be satisfied when $u_{j,t+1}$ follows for instance a t -distribution with only a few degrees of freedom.

(ii) Under H_A and for any given $F \in \mathcal{F}_A$, there exists an $\varepsilon > 0$ such that:

$$\lim_{P,R \rightarrow \infty} \Pr_F \left(P^{-1/2} \left| \widehat{DM}_P(\eta_{P,R}) \right| > \varepsilon \right) = 1.$$

The statement in (i) establishes that for any given null DGP, the statistic $\widehat{DM}_P(\eta_{P,R})$ is asymptotically standard normal. This result is established when $g(\cdot) = q(\cdot)$, but also when $g(\cdot) \neq q(\cdot)$ with $\pi = 0$. The statement in (ii), on the other hand, establishes the power of the test against any given fixed alternative.

We next study the asymptotic behaviour of the test under a specific, deterministic sequence $\{F_P\}_{P=1}^\infty$ of local alternatives, namely:

$$H_{LA,P} : E_{F_P} (g(u_{1,t+1}) - g(u_{2,t+1})) = \frac{\delta}{\sqrt{P}}$$

for some $\delta \neq 0$, where the class of DGPs that belong to $H_{LA,P}$ is given by

$$\mathcal{F}_{LA,P} = \{F \in \mathcal{F} : H_{LA,P} \text{ holds}\}.$$

Note that $\mathcal{F}_{LA,P}$ allows for sequences that converge to strictly non-nested DGPs with $\sigma_{DM}^2 > 0$, or overlapping DGPs with $\lim_{P,R \rightarrow \infty} \sigma_{DM,P}^2 = 0$ and $\lim_{P,R \rightarrow \infty} P\sigma_{DM,P}^2 = \infty$.

Theorem 1.2: Let Assumptions A.1-A.7 hold. Then, if either $g(\cdot) = q(\cdot)$ and $0 \leq \pi < \infty$ or $g(\cdot) \neq q(\cdot)$ and $\pi = 0$, under a sequence $\{F_P\}_{P=1}^\infty$ such that $F_P \in \mathcal{F}_{LA,P}$ for all P :

(i) if $\sigma_{DM}^2 > 0$:

$$\widehat{DM}_P(\eta_{P,R}) \xrightarrow{d} N(\lambda, 1)$$

where $\lambda = \lim_{P,R \rightarrow \infty} \frac{\delta}{\sqrt{\sigma_{DM}^2 + \eta_{P,R}^2}}$.

(ii) if $\sigma_{DM}^2 = 0$ and $\delta > 0$ ($\delta < 0$):

$$\widehat{DM}_P(\eta_{P,R}) \xrightarrow{\Pr_{F_P}} +\infty(-\infty).$$

Theorem 1.2 establishes the power against local alternatives as given by $H_{LA,P}$. Note that when $\sigma_{DM}^2 > 0$, the power increases as $\eta_{P,R}$ converges to zero. In fact, power is maximized for $\lim_{P,R \rightarrow \infty} \eta_{P,R} = 0$.

At this stage, a more detailed comparison to the standard DM test under fixed asymptotics is warranted, which in turn will highlight the need for uniform inference. Specifically, to facilitate the discussion, we focus on the case of a quadratic loss function and linear models estimated via OLS.

In this case, the DM statistic is:

$$\widehat{DM}_P^{S0} = \frac{P^{-1/2} \sum_{t=R}^{T-1} (\widehat{u}_{1,t+1}^2 - \widehat{u}_{2,t+1}^2)}{\sqrt{P^{-1} \sum_{t=R}^{T-1} (\widehat{u}_{1,t+1}^2 - \widehat{u}_{2,t+1}^2)^2}} \quad (8)$$

as considered by Clark and McCracken (2014). When $Z'_{j,t}u_{j,t+1}$, $j = 1, 2$ is a martingale difference sequence and $u_{1,t+1} = u_{2,t+1}$ a.s. so that $\sigma_{DM,F}^2 = 0$ for some fixed DGP F , Clark and McCracken (2014) show that \widehat{DM}_P^{S0} , which corresponds to the case with $l_P = 0$ (and $\omega(\tau, l_P) = 1$), weakly converges to the ratio of integrals of weighted Brownian motion (cf. Theorem 1 therein), when $\pi > 0$ and either a recursive or rolling estimation scheme is used.⁵ For the special case where forecast errors are homoskedastic and models are overlapping and nested in sense that the set of predictors from one model strictly nests all predictors from the other model, we may then deduce from Table 1 in McCracken (2007), that the $1 - (\alpha/2)$ level critical values (CVs) of the limiting distribution of \widehat{DM}_P^{S0} are in fact smaller than the corresponding standard normal CVs, so that inference based on the latter leads to an asymptotically conservative test. Similarly, when the HAC estimator of the variance is used, we have that in the overlapping case the test statistic converges to zero in probability as $l_P \rightarrow \infty$ and $P \rightarrow \infty$ since the denominator collapses to zero only at rate $\sqrt{l_P/P}$, while the numerator converges to zero in probability at rate $\sqrt{1/P}$. On the other hand, in the strictly non-nested case for any fixed DGP F with $\sigma_{DM,F}^2 > 0$, we have by standard arguments that the DM test converges weakly to a standard normal limiting distribution. That being said, it is well documented in the literature on out-of-sample forecast comparison tests (e.g. Clark and McCracken, 2014; Coroneo and Iacone, 2020) that the latter distribution is not a good approximation in finite samples as the test is too liberal in many situations rejecting with much higher frequency than the nominal rate. In fact, we corroborate this finding in our Monte Carlo simulations in Section 4.

Since the actual null DGP is unknown in practice, the aforementioned discussion makes clear that a uniform approach to inference is needed so that the asymptotic level of the test equals its nominal one across all different (overlapping and strictly non-nested) DGPs. We reflect this view with our local asymptotic framework in this paper. That is, the results from the existing literature provide approximations for scenarios where either $\sigma_{DM,P}^2 = 0$ for all $F_P \in \mathcal{F}_0$ or where $\lim_{P,R \rightarrow \infty} \sigma_{DM,P}^2 = \sigma_{DM}^2 > 0$, while the goal of this paper is instead to consider also DGPs where $\sigma_{DM,P}^2$ converges to zero at any possible rate. This likely provides a better approximation for finite sample situations where models are close to being overlapping and thus difficult to discern. In particular, it can be shown that fixed asymptotics suffer from a discontinuity in the asymptotic distribution of \widehat{DM}_P^{S0} at the “tipping case” where $\sigma_{DM,P} \eta_{P,R} \rightarrow C$ for some constant $0 < C < \infty$, which suggests poor performance of the standard DM test in attaining its nominal level and low power against alternatives that are close to

⁵For the special case of nested models (all predictors from one model are contained in the set of predictors from the other) and recursive estimation, Hansen and Timmermann (2015) show that the derivation of the limiting distribution does actually largely simplify as the statistic in Equation (8) is asymptotically equivalent to a linear combination of Wald statistics whose limiting distribution can be expressed as a combination of independent χ^2 random variables.

the null hypothesis. In fact, while a uniform asymptotic control of the Type I error probability at the nominal level is not necessarily desirable per se (see Perlman and Wu, 1999), it is this aspect that mainly motivates our proposed modified DM test.

For our test statistic, normal critical values remain uniformly valid under any DGP in \mathcal{F}_0 . We formally establish this result next.

Theorem 1.3: Let Assumptions A.1-A.7 hold. If either $g(\cdot) = q(\cdot)$ and $0 \leq \pi < \infty$ or $g(\cdot) \neq q(\cdot)$ and $\pi = 0$, then:

$$\lim_{P,R \rightarrow \infty} \sup_{F \in \mathcal{F}_0} \Pr_{F} \left(\left| \widehat{DM}_P(\eta_{P,R}) \right| > z_{1-\alpha/2} \right) = \alpha,$$

where $z_{1-\alpha/2}$ denotes the $1 - \frac{\alpha}{2}$ critical value of the standard normal distribution.

2.2 Data-Driven Regularisation Parameter

So far we have derived the asymptotic properties of the test for the case in which $\eta_{P,R}$ is a deterministic sequence. We now would like to provide a data-driven counterpart of $\eta_{P,R}$, say $\widehat{\eta}_{P,R}$, and show that the statistic based on $\eta_{P,R}$ and on $\widehat{\eta}_{P,R}$ are asymptotically equivalent over any DGP in \mathcal{F} . Following Schennach and Wilhelm (2017), we construct this adaptive tuning parameter sequence $\widehat{\eta}_{P,R}$ by balancing the trade-off between power loss under the alternative and size control under the null due to the introduction of the random perturbation. Specifically, we choose $\widehat{\eta}_{P,R}$ in such a way as to maximize the local power when $\sigma_{DM}^2 > 0$, and to minimize size distortion when $\sigma_{DM}^2(F) = 0$ for some $F \in \mathcal{F}_0$.

Theorem 2.1: Let Assumptions A.1-A.7 hold. Then:

(i) Under $H_{LA,P}$ for any sequence $\{F_P\}_{P=1}^{\infty}$ with $F_P \in \mathcal{F}_{LA,P}$ for all P and $\sigma_{DM}^2 > 0$, we have for any $\alpha \in (0, 1)$:

$$\begin{aligned} & \Pr_{F_P} \left(\left| \widehat{DM}_P(\eta_{P,R}) \right| > z_{1-\alpha/2} \right) \\ & \leq \Phi \left(z_{\alpha/2} + \frac{\delta}{\sigma_{DM}} \right) + \Phi \left(z_{\alpha/2} - \frac{\delta}{\sigma_{DM}} \right) - CPL^\dagger \eta_{P,R}^2 + O(\eta_{P,R}^3) + O\left(P^{-1/2} \sqrt{\ln P}\right) \end{aligned}$$

for all $\delta \neq 0$, where $\Phi(\cdot)$ ($\phi(\cdot)$) denotes the cumulative distribution (density) function of the standard normal distribution, and:

$$CPL^\dagger = -\phi \left(z_{\alpha/2} - \frac{\delta^\dagger}{\sigma_{DM}} \right) \frac{\delta^\dagger}{\sigma_{DM}^3}, \quad (9)$$

with $\delta^\dagger = \frac{\sigma_{DM}}{2} \left(z_{\alpha/2} - \sqrt{4 + z_{\alpha/2}^2} \right)$.

(ii) Under H_0 , if $g(\cdot) = q(\cdot)$ and $0 \leq \pi < \infty$, for a given $F \in \mathcal{F}_0$ with $\sigma_{DM}^2(F) = 0$ in the recursive estimation scheme, for any $\alpha \in (0, 1)$,

$$\Pr_F \left(\left| \widehat{DM}_P(\eta_{P,R}) \right| > z_{1-\alpha/2} \right) \leq \alpha + 2\phi(z_{\alpha/2}) \frac{C_{SD}}{\eta_{P,R}} R^{-1/2} \sqrt{2 \ln \ln P} \sqrt{2 \ln \ln T} + CP^{-1/2}$$

where:

$$C_{SD} = 2\sqrt{\frac{T}{R}} \max \left\{ \text{tr} \left(H_{1,F}^{-1} V_{1,F} \right), \text{tr} \left(H_{2,F}^{-1} V_{2,F} \right) \right\} \quad (10)$$

with $V_{k,F}$ and $H_{k,F}$, $k = 1, 2$, defined in Equations (6) and (7) of Assumption A.4(ii), respectively. (iii) Under H_0 , if $g(\cdot) \neq q(\cdot)$ and $\pi = 0$, for a given $F \in \mathcal{F}_0$ with $\sigma_{DM}^2(F) = 0$ in the recursive estimation scheme, for any $\alpha \in (0, 1)$,

$$\Pr_F \left(\left| \widehat{DM}_P(\eta_{P,R}) \right| > z_{1-\alpha/2} \right) \leq \alpha + 2\phi \left(z_{\alpha/2} \right) \frac{\overline{C}_{SD}}{\eta_{P,R}} \sqrt{\frac{P}{R}} \sqrt{2 \ln \ln T} + CP^{-1/2}$$

where:

$$\overline{C}_{SD} = 2 \max \left\{ \|G_{1,F}\| \|H_{1,F}^{-1}\| \left\| V_{1,F}^{\frac{1}{2}} \right\|, \|G_{2,F}\| \|H_{2,F}^{-1}\| \left\| V_{2,F}^{\frac{1}{2}} \right\| \right\} \quad (11)$$

with $V_{k,F}$ and $H_{k,F}$, $k = 1, 2$, are defined in Equations (6) and (7) of Assumption A.4(ii), respectively, while $G_{k,F}$ is defined in Equation (31) of the Appendix.

Theorem 2.1 provides asymptotic expansions of the power function as well as of the Type I error probability of $\widehat{DM}_P(\eta_{P,R})$ when models are overlapping. Observe that while power (part (i)) is a decreasing function in $\eta_{P,R}$, the opposite holds for the Type I error probability when models are overlapping (parts (ii) and (iii)). This suggests that a data-driven tuning parameter may be chosen to balance the trade-off. Here we will focus on the case of $g(\cdot) = q(\cdot)$ for simplicity since the construction for $g(\cdot) \neq q(\cdot)$ and $\pi = 0$ is very similar. Hereafter, let $CSD^\dagger = 2\phi \left(z_{\alpha/2} \right) C_{SD}$, which in turn consists of the variance of the asymptotic score, $V_{j,F}$, and Hessian $H_{j,F}$, $k = 1, 2$, defined in Equations (6) and (7), respectively. Straightforward algebra using the first order terms from Theorem 2.1(i) and (ii) shows that:

$$\eta_{P,R}^3 = \frac{CSD^\dagger}{CPL^\dagger} R^{-1/2} \sqrt{2 \ln \ln P} \sqrt{2 \ln \ln T}$$

with CPL^\dagger defined as in Equation (9), and so:

$$\eta_{P,R} = \left(\frac{CSD^\dagger}{CPL^\dagger} R^{-1/2} \sqrt{2 \ln \ln P} \sqrt{2 \ln \ln T} \right)^{1/3}. \quad (12)$$

Therefore, to obtain the sample analogue of CSD^\dagger and of CPL^\dagger , it suffices for $j = 1, 2$ to replace $H_{j,F}^{-1}$, $V_{j,F}$ and σ_{DM} with their sample analogues $\widehat{H}_{j,P}^{-1}$, $\widehat{V}_{j,P}^{-1}$ and $\widehat{\sigma}_{DM,P}$.⁶ Unfortunately, however, note that CPL^\dagger in Equation (9) has been obtained under the assumption that $\sigma_{DM}^2 > 0$ and cannot be consistently estimated when $\widehat{\sigma}_{DM,P}^2 = o_{\Pr_F(1)}$. Hence, we redefine CPL^\dagger as:

$$CPL^{\dagger\dagger} = CPL^\dagger 1 \left\{ \sigma_{DM}^2(F) \geq \underline{s} \right\} + CPL^\dagger \sigma_{DM}^2(F) 1 \left\{ \sigma_{DM}^2(F) < \underline{s} \right\},$$

⁶Specifically, in the case of square error loss with $g(\cdot) = q(\cdot) = (\cdot)^2$, the terms are given by $\widehat{V}_{j,P} = \frac{4}{P} \sum_{t=1}^{R-1-l_P} \sum_{\tau=-l_P}^{l_P} \omega(\tau, l_P) \widehat{u}_{1,t+1} \widehat{u}_{1,t+1+\tau} Z_{j,t} Z'_{j,t+\tau}$, while $\widehat{H}_{j,P} = \frac{2}{P} \sum_{t=1}^{R-1} Z_{j,t} Z'_{j,t}$.

where the lower bound $\underline{s} > c > 0$ for some $c > 0$ can be set to regulate CPL^\dagger in cases where σ_{DM}^2 is very close to zero, e.g. $\underline{s} = 0.1$ in our simulations. Consequently, we construct the empirical counterpart as:

$$\widehat{CPL}^{\dagger\dagger} = \widehat{CPL}^\dagger 1\{\widehat{\sigma}_{DM,P}^2 \geq \underline{s}\} + \widehat{CPL}^\dagger \widehat{\sigma}_{DM,P}^2 1\{\widehat{\sigma}_{DM,P}^2 < \underline{s}\}. \quad (13)$$

The reason for doing so is that whenever $\widehat{\sigma}_{DM,P}^2 < \underline{s}$, we have that $\widehat{CPL}^{\dagger\dagger} < \widehat{CPL}^\dagger$ and $\widehat{CPL}^{\dagger\dagger}$ can be seen as a consistent estimator of the lower bound:

$$CPL^{\dagger\dagger} = -\phi\left(z_{\alpha/2} - \frac{1}{2}\left(z_{\alpha/2} - \sqrt{4 + z_{\alpha/2}^2}\right)\right)\left(z_{\alpha/2} - \sqrt{4 + z_{\alpha/2}^2}\right)$$

when $\sigma_{DM}^2(F) < \underline{s}$.

Finally, we can use this to obtain an estimate of $\eta_{P,R}$ from Equation (12):

$$\widehat{\eta}_{P,R} = \left(\frac{\widehat{CSD}^\dagger}{\widehat{CPL}^{\dagger\dagger}} R^{-1/2} \sqrt{2 \ln \ln P} \sqrt{2 \ln \ln T}\right)^{1/3},$$

where \widehat{CSD}^\dagger is obtained as described below Equation (12). From the proof of Theorem 2.1, it becomes apparent that CSD^\dagger is obtained as an upper bound to the contribution of the estimation error component.

With the data-driven regularisation parameter $\widehat{\eta}_{P,R}$ in mind, we can now define:

$$\widehat{DM}_P(\widehat{\eta}_{P,R}) = \frac{\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{2,t+1})) + \widehat{\eta}_{P,R} \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} e_t}{\sqrt{\widehat{\sigma}_{DM,P}^2 + \widehat{\eta}_{P,R}^2 \frac{1}{P} \sum_{t=R}^{T-1} e_t^2}}. \quad (14)$$

The following result establishes the asymptotic equivalence of the feasible DM statistic $\widehat{DM}_P(\widehat{\eta}_{P,R})$ and its counterpart $\widehat{DM}_P(\eta_{P,R})$.

Theorem 2.2: Let Assumptions A.1-A.7 hold. Then, if either $g(\cdot) = q(\cdot)$ and $0 \leq \pi < \infty$ or $g(\cdot) \neq q(\cdot)$ and $\pi = 0$,

$$\sup_{F \in \mathcal{F}} \left| \widehat{DM}_P(\widehat{\eta}_{P,R}) - \widehat{DM}_P(\eta_{P,R}) \right| = o_{P_F}(1).$$

Remark 2: For overlapping null DGPs the limiting distribution is driven by the added artificial randomness. When undertaking a test at 5%, for the same original sample, 95% of researchers do not reject the null of equal expected predictive ability, while 5% do.⁷ This is the price we pay to have a test which asymptotically controls the probability of a Type I error at the nominal level across all null DGPs. If we are willing to give up on this uniform control of the Type I error probability

⁷Of course, in practice for a given data set, this is only likely to be an issue in “knife-edge” cases where the decision lies in between reject and not reject (e.g., because the p -value is 0.048).

and have a conservative test in the overlapping null case, there is an alternative. One could draw K sets of P random draws from the standard normal distribution and compute $P^{-1/2} \sum_{t=1}^P e_t^{(k)}$ for each set of draws $k = 1, \dots, K$. Denote $\widehat{DM}_P^{(k)}(\widehat{\eta}_{P,R})$ the corresponding statistic based on random draws $e_t^{(k)}$ and let PV_k be the corresponding two sided p -value. We can then devise the rule to reject H_0 if $\max_{k=1, \dots, K} PV_k \leq \alpha$ and to not reject H_0 otherwise. Under strictly non-nested DGPs or when the variance converges to zero sufficiently slowly, artificial randomness does not matter asymptotically, so the probability that $PV_1 = PV_2 = \dots = PV_K$ converges to one with P , and the test has correct asymptotic size equal to α . On the other hand, in the overlapping null case when the variance collapses sufficiently fast, because of the independence of the random draws it holds that $\Pr_F(\max_{k=1, \dots, K} PV_k \leq \alpha) = \prod_{k=1}^K \Pr_F(PV_k \leq \alpha) = \alpha^K$. Thus, for $K \geq 3$, $\Pr(\max_{k=1, \dots, K} PV_k \leq \alpha)$ is very close to zero so that the test becomes very conservative quickly. The main drawback of this approach is therefore a loss in the control of the Type I error probability at level α over all null DGPs, and lower power against local (to overlapping) alternatives.

3 Superior Predictive Ability Test

3.1 Set-up and Limiting Distribution

In this section we outline an SPA test (White, 2000; Hansen, 2005) which can be used in the presence of possibly overlapping models. Specifically, we allow for all competing models to overlap with the benchmark. We take model 1 to be the benchmark, and models 2, ..., J to be the competitors. For simplicity, we will assume that $g(\cdot) = q(\cdot)$ and $0 \leq \pi < \infty$ in what follows, which arguably represents the more relevant case in practice.

For some $F \in \mathcal{F}$, the null hypothesis is given by:

$$H_0^{RC} : \mu_1 - \mu_k \leq 0 \text{ for } k = 2, \dots, J \tag{15}$$

$$H_A^{RC} : \mu_1 - \mu_k > 0 \text{ for at least one } k$$

where $\mu_j = E_F(g(u_{j,t+1}))$ for $j = 1, \dots, J$.

It is immediate to see that H_0^{RC} is a composite hypothesis, formed by the intersection of $J - 1$ null hypotheses each given by $H_{0,k} : \mu_1 - \mu_k \leq 0$, while H_A^{RC} is the union of the $J - 1$ complements of $H_{0,k}$, $k = 2, \dots, J$. To test H_0^{RC} vs. H_A^{RC} , we use a statistic proposed in Andrews and Soares (2010). This statistic satisfies Assumption 2 in Hansen (2005) and builds on the insight from the latter paper that statistics for multiple inequality comparisons that are studentized and that do not recenter all inequalities unless evidence suggests they are binding, exhibit superior power properties

in general. Thus, letting the sample analogue of $\mu_1 - \mu_k$ in Equation (15) be given by:

$$\widehat{m}_{k,P} = \frac{1}{P} \sum_{t=R}^{T-1} (g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{k,t+1})),$$

where $\widehat{u}_{1,t+1}$ and $\widehat{u}_{k,t+1}$ denote forecast errors based on the estimated models 1 and k , respectively, we use the following test statistic:

$$\widehat{S}_P(\widehat{\eta}_{P,R}) = \sum_{k=2}^J \left(\max \left\{ 0, \widehat{DM}_{k,P}(\widehat{\eta}_{P,R}) \right\} \right)^2. \quad (16)$$

where for each $k \in \{2, \dots, J\}$:

$$\widehat{DM}_{k,P}(\widehat{\eta}_{P,R}) = \frac{\sqrt{P}\widehat{m}_{k,P} + \widehat{\eta}_{k,P,R} \frac{1}{\sqrt{P}} \sum_{t=1}^P e_{k,t}}{\sqrt{\widehat{\sigma}_{k,P}^2 + \widehat{\eta}_{k,P,R}^2 \frac{1}{P} \sum_{t=1}^P e_{k,t}^2}} \quad (17)$$

Here, $e_{k,t} \stackrel{i.i.d.}{\sim} N(0, 1)$ with $E(e_{k,t}e_{k',t}) = 0$ for $k \neq k'$, as implied by Assumption A.7, and $\widehat{\sigma}_{k,P}^2$ is again a HAC estimator of $\lim_{P,R \rightarrow \infty} \text{var}_{F_P} \left(\sqrt{P}m_{k,P} \right)$ with $m_{k,P} = \frac{1}{P} \sum_{t=R}^{T-1} (g(u_{1,t+1}) - g(u_{k,t+1}))$, while F_P denotes an element of the sequence $\{F_P\}_{P=1}^{\infty}$ as defined below.

For each $k \in \{2, \dots, J\}$, the data-driven regularisation constant for the model comparison pair 1 and k , denoted $\widehat{\eta}_{k,P,R}$, is chosen under the least favourable case of the null hypothesis where the moment is binding. In particular, following derivations analogous to the previous section, but examining a sequence of one sided alternatives of the form δ_k/\sqrt{P} for some $\delta_k > 0$ instead of the previously used two-sided version, we obtain:

$$\widehat{\eta}_{k,P,R} = \left(\frac{\widehat{CSD}_k^\dagger}{\widehat{CPL}_k^{\dagger\dagger}} R^{-1/2} \sqrt{2 \ln \ln P} \sqrt{2 \ln \ln T} \right)^{1/3},$$

where:

$$\widehat{CPL}_k^{\dagger\dagger} = \widehat{CPL}_k^\dagger \mathbf{1} \{ \widehat{\sigma}_{k,P}^2 \geq \underline{s} \} + \widehat{CPL}_k^\dagger \widehat{\sigma}_{k,P}^2 \mathbf{1} \{ \widehat{\sigma}_{k,P}^2 < \underline{s} \},$$

with the optimal δ_k^\dagger in \widehat{CPL}_k^\dagger given by $\delta_k^\dagger = \frac{\widehat{\sigma}_k}{2} \left(-z_\alpha + \sqrt{4 + \frac{z_\alpha^2}{2}} \right)$ for some $\alpha \in (0, 1/2)$, see Remark 2.1 in the supplementary material for the derivation. Similarly:

$$\widehat{CSD}_k^\dagger = \phi(z_\alpha) \frac{\widehat{C}_{k,SD}}{\eta} R^{-1/2} \sqrt{2 \ln \ln P} \sqrt{2 \ln \ln T}$$

where, for each $k = 2, \dots, J$, $\widehat{C}_{k,SD}$ is constructed as in Subsection 2.2. Hereafter, define again

$\mathcal{F}_0^{RC} \equiv \{F \in \mathcal{F} : H_0^{RC} \text{ holds}\}$ and let for any sequence $\{F_P\}_{P=1}^\infty$ with $F_P \in \mathcal{F}_0^{RC}$ for all P :

$$\sigma_{k,P}^2 = \text{var}_{F_P} \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(u_{1,t+1}) - g(u_{k,t+1})) \right)$$

as in the pairwise case. Note that, in a similar way to before, for every k , \mathcal{F}_0^{RC} contains both DGPs with $\lim_{P,R \rightarrow \infty} \sigma_{k,P}^2 = \sigma_k^2 > 0$, as well as $\lim_{P,R \rightarrow \infty} P\sigma_{k,P}^2 = c$, $0 \leq c < \infty$, and $\lim_{P,R \rightarrow \infty} P\sigma_{k,P}^2 = \infty$.

More specifically, if $\lim_{P,R \rightarrow \infty} E_{F_P}(m_{k,P}) \equiv m_k < 0$, observe that since $\frac{1}{\sqrt{P}} \sum_{t=1}^P e_{k,t}$ converges weakly to a standard normal random variable, we have for P sufficiently large:

$$\Pr_F \left(\max \left\{ 0, \widehat{DM}_{k,P}(\widehat{\eta}_{P,R}) \right\} > 0 \right) \rightarrow 0,$$

so that model k does not contribute to the test statistic asymptotically. By contrast, for a given DGP under the alternative with $m_k > 0$, as P gets large,

$$\max \left\{ 0, \widehat{DM}_{k,P}(\widehat{\eta}_{P,R}) \right\} = \frac{\sqrt{P}\widehat{m}_{k,P} + \widehat{\eta}_{k,P,R} \frac{1}{\sqrt{P}} \sum_{t=1}^P e_{k,t}}{\sqrt{\widehat{\sigma}_{k,P}^2 + \widehat{\eta}_{k,P,R}^2 \frac{1}{P} \sum_{t=1}^P e_{k,t}^2}} = \frac{\sqrt{P}\widehat{m}_{k,P}}{\sqrt{\widehat{\sigma}_{k,P}^2}} (1 + o_{\Pr_F}(1)) = O_{\Pr_F}(\sqrt{P}).$$

By contrast, as in the pairwise case, if $m_k = 0$, the term in Equation (17) converges weakly to a standard normal random variable. In fact, in the non-overlapping case with $\sigma_k^2 > 0$, this limiting distribution is driven by $\sqrt{P}\widehat{m}_{k,P}/\widehat{\sigma}_{k,P}$, while in the overlapping case with $\sigma_k^2 = 0$ it is driven by either the first or second term in Equation (17) depending on whether $\sigma_{k,P}^2$ approaches zero at a faster or slower rate than $\eta_{k,P,R}$ (as the statistics with $\widehat{\eta}_{k,P,R}$ and $\eta_{k,P,R}$ are asymptotically equivalent).

In the multiple comparison case, we need to take into account also the correlation among different moment conditions. For example, if both model k and k' do not overlap with model 1, then in general the associated empirical moment conditions are correlated. On the other hand, if model k and/or k' overlap with the benchmark, the corresponding moment conditions are asymptotically independent since the added random noise is independent across moment comparison pairs.

Finally, when $m_k = 0$, both $\sqrt{P}\widehat{m}_{k,P}$ and $\sqrt{P}\bar{e}_k$ are close to zero with high probability, where $\bar{e}_k = \frac{1}{P} \sum_{t=1}^P e_{k,t}$. If this occurs, we may have that $\sqrt{P}\widehat{m}_{k,P} > 0$ and $\sqrt{P}\widehat{m}_{k,P} + \sqrt{P}\bar{e}_k < 0$, or the other way round, with positive probability. However, this is not a problem asymptotically, provided that the bootstrap statistic properly mimics such switching behaviour.

Hereafter, for some sequence $\{F_P\}_{P=1}^\infty$ with $F_P \in \mathcal{F}_0^{RC}$ for all P , let:

$$\varsigma_k = \lim_{P,R \rightarrow \infty} \frac{\sqrt{P}E_{F_P}(m_{k,P})}{\sqrt{\sigma_{k,P}^2 + \eta_{k,P}^2}}, \quad (18)$$

so that if $\lim_{P,R \rightarrow \infty} \sigma_{k,P}^2 > 0$, then $\varsigma_k = 0$ if $\lim_{P,R \rightarrow \infty} \sqrt{P}E_{F_P}(m_{k,P}) = 0$ and $\varsigma_k = -\infty$ if $\lim_{P,R \rightarrow \infty} \sqrt{P}E_{F_P}(m_{k,P}) = -\infty$.

Now, suppose the case of $E_{F_P}(m_{k,P}) = -cP^{-(1/2+\varepsilon)}$ for some $c > 0$ and $0 < \varepsilon < 1/2$, which is consistent with both overlapping and non-overlapping DGPs. If $\lim_{P,R \rightarrow \infty} (\sigma_{k,P}^2/\eta_{k,P}^2) = 0$ and $P^{-\varepsilon}/\eta_{k,P} \rightarrow \infty$ (i.e. $\eta_{k,P}$ goes to zero faster than $P^{-\varepsilon}$ and slower than $\sigma_{k,P}$), then we have that $\lim_{P,R \rightarrow \infty} \sqrt{P}E_{F_P}(m_{k,P}) = 0$ but $\lim_{P,R \rightarrow \infty} \frac{\sqrt{P}E_{F_P}(m_{k,P})}{\sqrt{\sigma_{k,P}^2 + \eta_{k,P}^2}} = -\infty$. Thus, for the same value of $E_{F_P}(m_{k,P})$, ς_k can be either zero or minus infinity, depending on the degree to which model k overlaps with the benchmark. Finally, define Ω as $(J-1) \times (J-1)$ matrix with ii -th element $\omega_{ii} = 1$, and for $i \neq j$, ij -th element given by:

$$\omega_{ij} = \text{acov} \left(\widehat{DM}_{i,P}(\widehat{\eta}_{P,R}), \widehat{DM}_{j,P}(\widehat{\eta}_{P,R}) \right), \quad (19)$$

so that $\omega_{ij} = 0$ if model i and/or model j overlap with the benchmark, and (generally) different from zero if both model i and j do not overlap with the benchmark.

Letting $\mathcal{F}_A^{RC} \equiv \{F \in \mathcal{F} : H_A^{RC} \text{ holds}\}$ in analogy to \mathcal{F}_0^{RC} , we have the following pointwise result regarding the limiting distribution.

Theorem 3.1: Let Assumptions A.1-A.7 hold. Then,

(i) Under H_0^{RC} , for any given $F \in \mathcal{F}_0^{RC}$

$$\widehat{S}_P(\widehat{\eta}_{P,R}) \xrightarrow{d} \sum_{k=2}^J \left(\max \{0, (\Omega^{1/2}Z)_k + \varsigma_k\} \right)^2$$

where $(\Omega^{1/2}Z)_k$ is the k -th element of $\Omega^{1/2}Z$, with Z being a $J-1$ dimensional standard normal, where Ω is defined in Equation (19) and ς_k is defined in Equation (18).

(ii) Under H_A^{RC} for any given $F \in \mathcal{F}_A^{RC}$, there exists an $\varepsilon > 0$ such that:

$$\lim_{P,R \rightarrow \infty} \Pr_F \left(P^{-1} \widehat{S}_P(\widehat{\eta}_{P,R}) > \varepsilon \right) = 1$$

The critical values of the limiting distribution in (i) above can be obtained by the bootstrap, which in turn ensures that the contribution of the added randomness to the selected moment conditions is the same for the statistic and its bootstrap counterpart. Note that the statement in Theorem 3.1 holds pointwise for any DGP F in the null set, while we want to have critical values which are valid uniformly over \mathcal{F}_0^{RC} . In the next subsection we therefore devise a bootstrap procedure which we show to deliver critical values that control the rejection rate at level α uniformly over \mathcal{F}_0^{RC} .

3.2 Bootstrap for SPA

To generate critical values, we will rely on the moving block bootstrap procedure of Künsch (1989). Since PEE is asymptotically negligible in our set-up, we do not need to re-estimate the parameters

using resampled observations, and so can directly resample the forecast error series. This is true regardless of the estimation scheme we use. We draw b_P blocks of length l_P such that $l_P b_P = P$ from $\widehat{\mathbf{u}}_{t+1} = (\widehat{u}_{1,t+1}, \dots, \widehat{u}_{J,t+1})$ for $t = R, \dots, R + P - 1$ to obtain bootstrap resamples $\widehat{\mathbf{u}}_{t+1}^* = (\widehat{u}_{1,t+1}^*, \dots, \widehat{u}_{J,t+1}^*)$. In what follows, with a slight abuse of notation, we set the length of block equal to the lag truncation parameter used for the HAC covariance matrix estimation, l_P . We then also take P independent draws from $\mathbf{e}_t = (e_{2,t}, \dots, e_{J,t})$ for $t = 1, \dots, P$ to generate $\mathbf{e}_t^* = (e_{2,t}^*, \dots, e_{J,t}^*)$. Note that we require a resample of the latter since we will adopt the Generalized Moment Selection (GMS) procedure of Andrews and Soares (2010), which involves trimming slack empirical moment inequalities. Specifically, letting $\bar{e}_k = \frac{1}{P} \sum_{t=1}^P e_{k,t}$, $\bar{e}_k^* = \frac{1}{P} \sum_{t=1}^P e_{k,t}^*$, and $1\{\cdot\}$ denote the indicator function, we use the following bootstrap statistic:

$$\widehat{S}_P^*(\widehat{\eta}_{P,R}) = \sum_{k=2}^J \left(\max \left\{ 0, \left(\widehat{DM}_{k,P}^*(\widehat{\eta}_{k,P,R}) \right) \times 1 \left\{ \widehat{DM}_{k,P}(\widehat{\eta}_{k,P,R}) \geq -2\kappa_{k,P} \right\} \right\} \right)^2, \quad (20)$$

where, for $k = 2, \dots, J$, in analogy to before:

$$\widehat{DM}_{k,P}^*(\widehat{\eta}_{k,P,R}) = \frac{\sqrt{P} (\widehat{m}_{k,P}^* + \widehat{\eta}_{k,P,R} \bar{e}_k^*) - \sqrt{P} (\widehat{m}_{k,P} + \widehat{\eta}_{k,P,R} \bar{e}_k)}{\sqrt{\widehat{\sigma}_{k,P}^{*2} + \widehat{\eta}_{k,P,R}^2 \bar{e}_k^{*2}}}$$

with $\widehat{m}_{k,P}^* = \frac{1}{P} \sum_{t=R}^{T-1} (g(\widehat{u}_{1,t+1}^*) - g(\widehat{u}_{k,t+1}^*))$ and $\widehat{\sigma}_{k,P}^{*2}$ is the variance estimator for the moving block bootstrap from Gonçalves and White (2004) using the resampled observations $(g(\widehat{u}_{1,t+1}^*) - g(\widehat{u}_{k,t+1}^*))$. Note that the term $\sqrt{P} (\widehat{m}_{k,P} + \widehat{\eta}_{k,P,R} \bar{e}_k)$ in Equation (20) is the standard recentering term in the bootstrap statistic, while the indicator function trims empirical moment inequalities that are found to be “too slack”, in other words where $\widehat{DM}_{k,P}(\widehat{\eta}_{k,P,R})$ is smaller than $-2\kappa_{k,P}$ with $\frac{\sqrt{2 \ln \ln P}}{\kappa_{k,P}} \rightarrow c$ for some $0 \leq c < \infty$ and $\frac{\kappa_{k,P}}{\sqrt{P}} \rightarrow 0$. The critical values are therefore only based on moment inequalities which are found to be empirically binding to avoid overly conservative inference. Hansen (2005) introduced the idea of a trimming rule using the law of the iterated logarithm, while Andrews and Soares (2010) later proposed other (rate) choices. In addition, note that $\widehat{\eta}_{k,P,R}$ is used throughout the bootstrap statistic. This is because we resample directly from the forecast errors, while resampling also $\widehat{\eta}_{k,P,R}$ would require to resample the original data.⁸

The logic underlying the statistic in Equation (20) is simple. If $m_k < 0$, with probability approaching one, we have that $\widehat{m}_{k,P} < 0$ and $1 \left\{ \widehat{DM}_{k,P}(\widehat{\eta}_{k,P,R}) \geq -2\kappa_{k,P} \right\} = 0$ for sufficiently large P , which means that model k contributes to neither the statistic nor to its bootstrap analogue. On the other hand, if $m_k > 0$, which can occur only in the non-overlapping case, then the event $1 \left\{ \widehat{DM}_{k,P}(\widehat{\eta}_{k,P,R}) \geq -2\kappa_{k,P} \right\} = 1$ occurs with probability approaching one, and $\widehat{DM}_{k,P}^*(\widehat{\eta}_{k,P,R})$

⁸In fact, since under the conditions in Theorem 3.2 we would have that $\Pr_F^* \left(|\widehat{CSD}^{\dagger*} - \widehat{CSD}^{\dagger}| > \delta \right) = o_{\Pr_F}(1)$ for any $\delta > 0$, where $\Pr_F^*(\cdot)$ denotes the bootstrap probability measure, we expect little difference in terms of the finite sample results.

weakly converges to a standard normal random variable conditional on the sample, with probability converging to one, while $\sqrt{P}\widehat{m}_{k,P}$ diverges in probability to infinity. Finally, when $m_k = 0$, regardless of whether model 1 and k overlap or not, the event $1\left\{\widehat{DM}_{k,P}(\widehat{\eta}_{k,P,R}) \geq -2\kappa_{k,P}\right\} = 1$ occurs with probability approaching one, and so $\widehat{DM}_{k,P}(\widehat{\eta}_{k,P,R})$ as well as $\widehat{DM}_{k,P}^*(\widehat{\eta}_{k,P,R})$ have the same limiting distribution conditional on the sample, with probability converging to one.

In what follows, let $c_{B,P,R,1-\alpha}^*$ denote the $1 - \alpha$ percentile of the empirical distribution of $\widehat{S}_P^*(\widehat{\eta}_{P,R})$, based on B bootstrap replications. We now have the following result:

Theorem 3.2: Let Assumptions A.1-A.7 hold.⁹ Also, as $P \rightarrow \infty$ and for every $k \in \{2, \dots, J\}$, $\frac{\sqrt{2 \ln \ln P}}{\kappa_{k,P}} \rightarrow c$ for some $0 \leq c < \infty$ and $\frac{\kappa_{k,P}}{\sqrt{P}} \rightarrow 0$ then for any $0 < \alpha < 1/2$:

(i) under H_0^{RC} ,

$$\lim_{B,R,P \rightarrow \infty} \sup_{F \in \mathcal{F}_0^{RC}} \Pr_F \left(\widehat{S}_P(\widehat{\eta}_{P,R}) \geq c_{B,P,R,1-\alpha}^* \right) \leq \alpha$$

and if for some $k = 2, \dots, J$, $m_k = 0$, then:

$$\lim_{B,R,P \rightarrow \infty} \sup_{F \in \mathcal{F}_0^{RC}} \Pr_F \left(\widehat{S}_P(\widehat{\eta}_{P,R}) \geq c_{B,P,R,1-\alpha}^* \right) = \alpha$$

(ii) under H_A^{RC} , for any $F \in \mathcal{F}_A^{RC}$

$$\lim_{P,R \rightarrow \infty} \Pr_F \left(\widehat{S}_P(\widehat{\eta}_{P,R}) \geq c_{B,P,R,1-\alpha}^* \right) = 1.$$

This Theorem demonstrates that the bootstrap critical values are asymptotically valid, uniformly over all DGPs under the null. Furthermore, if at least one competitor is as good as the benchmark, i.e. if $m_k = 0$ for some k , then we have a test with asymptotic rejection rate equal to α . Finally, under any alternative DGP, the null is rejected with probability approaching one.

Remark 3: In contrast to the SPA statistic used in this paper, the SPA statistic originally proposed by Hansen (2005) is $\widehat{H}_P = \max \left\{ \max_{k=2, \dots, J} \frac{\sqrt{P}\widehat{m}_{k,P}}{\widehat{\sigma}_{k,P}}, 0 \right\}$, and once we add the artificial randomness component it becomes $\widehat{H}_P(\widehat{\eta}_{P,R}) = \max \left\{ \max_{k=2, \dots, J} \frac{\sqrt{P}\widehat{m}_{k,P} + \sqrt{P}\widehat{\eta}_{k,P,R}\bar{e}_k}{\sqrt{\widehat{\sigma}_{k,P}^2 + \widehat{\eta}_{k,P,R}^2\bar{e}_k^2}}, 0 \right\}$. One could then derive the limiting distribution of $\widehat{H}_P(\widehat{\eta}_{P,R})$ using similar arguments to ours, following Andrews and Soares (2010).¹⁰ Another difference is that we draw blocks from $(g(\widehat{u}_{1,t+1}), \dots, g(\widehat{u}_{J,t+1}))$ whereas Hansen (2005) draws blocks from $\left(\frac{g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{2,t+1})}{\widehat{\sigma}_{2,P}}, \dots, \frac{g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{J,t+1})}{\widehat{\sigma}_{J,P}} \right)$ for the bootstrap. As a consequence,

⁹Note that A.6(i) ensures that $l_P/\sqrt{P} \rightarrow 0$ and so b_P grows at a rate faster than \sqrt{P} .

¹⁰Note that a difference between Andrews and Soares (2010) and Hansen (2005) is that the latter derives the limiting distribution under a fixed null DGP, while the former derive the limiting behaviour under drifting null sequences, thus providing uniform inference results.

our bootstrap statistic in Equation (20) is scaled by the variance estimators based on resampled observations, while Hansen (2005) scales by the variance estimator based on the original observations.

4 Monte Carlo Simulations

4.1 Set-up

In this section we explore the finite sample properties of our modified Diebold-Mariano test for pairwise model comparisons as well as the SPA test for multiple model comparisons. We assume a simple linear autoregressive DGP and make use of different misspecified linear forecasting models which will allow us to verify the performance of the tests in the overlapping and non-nested cases under the null, as well as under the alternative. The design we use is in the spirit of the simulation study of Clark and McCracken (2014) who also investigate overlapping models.

The target variable y_{t+1} is a linear function of an autoregressive term as well as a lag of five predictor variables $\mathbf{x}_t = (x_{1,t}, x_{2,t}, x_{3,t}, x_{4,t}, x_{5,t})'$:

$$y_{t+1} = 0.5y_t + \theta' \mathbf{x}_t + u_{t+1} \quad (21)$$

where $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)'$ is a parameter vector whose values are specified below. The variables $x_{j,t}$ for $j = 1, \dots, 5$ each follow simple autoregressive processes:

$$x_{j,t+1} = 0.5x_{j,t} + e_{j,t+1} \quad (22)$$

such that we have a six-variable VAR(1) system with zero restrictions. The innovation terms are drawn from independent Gaussian distributions meaning that there is no correlation among $x_{j,t+1}$:

$$\begin{pmatrix} u_{t+1} \\ e_{1,t+1} \\ \vdots \\ e_{5,t+1} \end{pmatrix} \sim \text{i.i.d. } N \left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 8 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & 1 \end{pmatrix} \right)$$

however we will check the robustness of the results when the $x_{j,t+1}$ variables are correlated.

In the pairwise comparison case we consider the following two forecasting models where the first model makes use only of the variable $x_{1,t}$ and the second model makes use only of $x_{2,t}$:

$$y_{t+1} = a_0 + a_1x_{1,t} + a_2x_{1,t-1} + \zeta_{1,t+1} \quad (23)$$

$$y_{t+1} = b_0 + b_1x_{2,t} + b_2x_{2,t-1} + \zeta_{2,t+1} \quad (24)$$

where both models crucially omit the autoregressive term present in the true DGP, allowing us to

assess performance in a situation of dynamic misspecification where the forecast errors are serially dependent.

With this set-up we can generate all cases of interest by varying the parameter vector θ in Equation (21) in the following three ways:

- DGP1:** $\theta = (0, 0, 0, 0, 0)'$ (Size Overlapping)
- DGP2:** $\theta = (0.5, 0.5, 0.5, 0.5, 0.5)'$ (Size Non-Nested)
- DGP3:** $\theta = (0, 2, 0, 0, 0)'$ (Power)

The first parameter value choice (DGP1) ensures we are in the overlapping case under the null hypothesis where neither forecasting model has any predictive ability. The second choice (DGP2) gives the non-nested case under the null where both of the models have equal predictive ability since variables $x_{1,t}$ and $x_{2,t}$ have equal non-zero weight in the DGP. The third choice (DGP3) allows us to observe the power properties of the test as only $x_{2,t}$ contributes to the DGP in Equation (21) and so model 2 is superior to model 1 even though both are dynamically mis-specified. Under these three versions of the DGP parameters, the R^2 in the true DGP in Equation (21) will be different. As an example, under DGP1 in the overlapping case, the R^2 is 25% whereas the R^2 in the actual forecasting models is zero as they both omit the autoregressive term and use only irrelevant predictors. In a similar way for DGP2 and DGP3, the forecast model R^2 is much lower than in the true DGP due to the omission of regressors from the DGP.

For the multiple model comparison we will add some additional models similar to those described in Equations (23) and (24). Specifically, we will now make use of the five variables $x_{j,t+1}$ for $i = 1, \dots, 5$. We will consider the one-lag ($x_{j,t}$), two-lag ($x_{j,t}, x_{j,t-1}$) and three-lag ($x_{j,t}, x_{j,t-1}, x_{j,t-2}$) versions of these models for each individual predictor which gives a total of $J = 15$ models. This number of models aligns with our empirical application below. With the three DGP parameter values described above, this set-up will respectively assess the SPA test performance in the overlapping case, the empirical size in the least favourable case with non-nested models, as well as the power of the test.

The remaining details of the simulation study are as follows. The sample sizes we use are all combinations of $R = 100, 200, 300$ and $P = 100, 200, 300$ for the in-sample and out-of-sample windows, which gives rise to sample split ratio P/R between $1/3$ and 3 . In the supplementary material we also provide some further sample sizes up to $R = P = 1000$ and for a wider sample split ratio of $1/5$ and 5 . The models are estimated using OLS and model-implied conditional mean forecasts are obtained using the recursive estimation scheme which is commonly applied in practice. The models will be evaluated using the squared error loss function for $g(\cdot)$. The significance level $\alpha = 0.05$ will be used for the tests. We perform $M = 2500$ simulation replications and for simulating the bootstrap SPA test we will make use of the Warp Speed bootstrap of Giacomini et al. (2013) making a single bootstrap draw per Monte Carlo replication with a block length $l_P = 4$ and a moment selection rule of $\kappa_{k,P} = \ln \ln P$. As recommended in the related paper of Hsu and Shi (2017), the data are standardised to have mean zero and unit variance after they have been drawn from the distributions

described above, while \underline{s} from CPL^\dagger is set to 0.1.

4.2 Results

We first present the results of the pairwise comparison using the two models described in Equations (23) and (24) above. Alongside the results of our test $\widehat{DM}_P(\widehat{\eta}_{P,R})$, we will also compute the rejection rates for the unadjusted DM test. For this standard DM test we will display two variants: the version with the sample variance estimator and the version with the HAC estimator, denoted \widehat{DM}_P^{S0} and \widehat{DM}_P^S respectively. Table 1 reports the empirical size results in the overlapping and non-nested cases, followed by the power.

Starting with the overlapping case, the standard DM test \widehat{DM}_P^{S0} tends to over-reject as it does not account for serial dependence in the forecast errors, whereas the \widehat{DM}_P^S test severely under-rejects in the degenerate case when P is large relative to R , as expected from the theory. For instance, empirical rejection rates are as poor as 1.4% in the case where $R = 100$ and $P = 300$. Our proposed test, on the other hand, is seen to have much more robust size properties in the overlapping case with the rejection rate much closer to the nominal level for all sample sizes. These rejection rates are roughly around 4-6% in all cases. In the non-nested case, as expected we see that the standard DM test with sample variance estimate is wildly over-rejecting with rejection rates of 17-18% in some cases. The DM test with HAC variance estimate also over-rejects with rejection rates typically around 7-9%. This is in contrast with our test, which provides up to a 2% improvement in the rejection rate across all sample sizes and gives rejection rates around 6% as the sample size grows. In Tables S2 and S3 of the supplementary material, we also show results for higher and lower degrees of dynamic mis-specification (with autoregressive parameter 0.9 and 0.1 respectively). These results show that our test performs even better with lower persistence, and it greatly moderates the extreme over-rejection of the unadjusted DM test in situations with high persistence. We also find that our test performs well with correlated $x_{j,t+1}$ variables, as seen in Table S4 in the supplementary material.

Regarding the power of the test, we see the rejection rate rise swiftly to unity for the standard DM tests. Our test has slightly lower power which is to be expected, both as it is not over-rejecting in the non-nested case and also as we expect some power loss given the perturbation we use. Nevertheless, the power loss is minimal as we only see power as low as 77% in the case where $P = 100$ and the rejection rate rises to unity as the sample size grows to a level which is consistent with the sample sizes used in most empirical applications. We also explore the local power of the test which shows that, especially in cases such as $P = 100$, $R = 300$, our test shows slightly better local power properties than the standard DM test when being close to the overlapping case (see Figure S1 in the supplementary material).

We next present the results of the multiple model comparison test. As a reference for our SPA test, we will also compute the equivalent of the SPA test $\widehat{S}_P(\widehat{\eta}_{P,R})$ in Equation (16) when the perturbation is not used, in other words the SPA test based on the standard DM test with HAC standard errors.

Table 1: Rejection Rate - Pairwise Comparison

	R	P	\widehat{DM}_P^{S0}	\widehat{DM}_P^S	$\widehat{DM}_P(\widehat{\eta}_{P,R})$
Size Overlapping	100	100	0.055	0.034	0.044
	100	200	0.036	0.016	0.040
	100	300	0.033	0.014	0.039
	200	100	0.083	0.050	0.048
	200	200	0.051	0.029	0.045
	200	300	0.050	0.024	0.044
	300	100	0.096	0.054	0.055
	300	200	0.082	0.042	0.052
	300	300	0.068	0.030	0.051
Size Non-Nested	100	100	0.161	0.082	0.070
	100	200	0.160	0.067	0.058
	100	300	0.154	0.064	0.059
	200	100	0.172	0.078	0.073
	200	200	0.180	0.079	0.067
	200	300	0.176	0.082	0.062
	300	100	0.172	0.089	0.068
	300	200	0.171	0.072	0.070
	300	300	0.180	0.079	0.060
Power	100	100	0.982	0.948	0.775
	100	200	1.000	1.000	0.950
	100	300	1.000	1.000	0.988
	200	100	0.980	0.938	0.798
	200	200	1.000	1.000	0.972
	200	300	1.000	1.000	0.994
	300	100	0.972	0.926	0.801
	300	200	1.000	0.999	0.975
	300	300	1.000	1.000	0.998

Notes: The nominal level is 5%. With forecasting models given by Equations (23) and (24), the overlapping case is DGP1 where $\theta = (0, 0, 0, 0, 0)'$ in Equation (21), the non-nested case is DGP2 where $\theta = (0.5, 0.5, 0.5, 0.5, 0.5)'$ and the power case is DGP3 where $\theta = (0, 2, 0, 0, 0)'$.

We denote this \widehat{S}_P^{DM} . For further comparison, we will also compute the SPA test of Hansen (2005) which we denote \widehat{H}_P .

The results are displayed in Table 2. Here we see very promising results from our tests, in a similar way to the pairwise model comparison case. When we have overlapping models, the rejection rate of our test becomes close to the nominal level as the sample size grows, improving around 2-3% on the non-adjusted \widehat{S}_P^{DM} test which has a rejection rate below 3% in all cases. The non-nested results in the least favourable case again show that our test has very good size properties, better even than the pairwise comparison test above. On the other hand, the unadjusted \widehat{H}_P test is seen to over-reject in

almost all cases by up to 3-4 percentage points, whereas the \widehat{S}_P^{DM} has a low rejection rate. In terms of the power, our test has comparable properties to the other two tests, with power only below 90% when the sample size $P = 100$ is used. The local power results are also an improvement relative to non-adjusted \widehat{S}_P^{DM} test when we are close to the overlapping case (see Figure S2 in the supplementary material).

Table 2: Rejection Rate - Multiple Model Comparison

	R	P	\widehat{S}_P^{DM}	\widehat{H}_P	$\widehat{S}_P(\widehat{\eta}_{P,R})$
Size Overlapping	100	100	0.015	0.056	0.032
	100	200	0.012	0.039	0.032
	100	300	0.014	0.035	0.026
	200	100	0.028	0.070	0.046
	200	200	0.023	0.047	0.035
	200	300	0.025	0.039	0.046
	300	100	0.019	0.060	0.048
	300	200	0.029	0.050	0.036
	300	300	0.025	0.043	0.046
Size Non-Nested	100	100	0.026	0.091	0.047
	100	200	0.036	0.064	0.050
	100	300	0.041	0.048	0.044
	200	100	0.026	0.092	0.041
	200	200	0.043	0.069	0.040
	200	300	0.049	0.075	0.056
	300	100	0.028	0.077	0.043
	300	200	0.047	0.078	0.057
	300	300	0.053	0.072	0.059
Power	100	100	0.482	0.884	0.660
	100	200	0.966	0.994	0.928
	100	300	0.998	1.000	0.994
	200	100	0.523	0.854	0.670
	200	200	0.968	0.994	0.967
	200	300	0.999	1.000	0.996
	300	100	0.464	0.840	0.700
	300	200	0.966	0.994	0.974
	300	300	0.999	1.000	0.999

Notes: The nominal level is 5%. There are $J = 15$ different models, with Equation (23) being the benchmark. The remaining models are all one-, two- and three-lag models based on one of five predictor variables as described above. As in Table 1, the overlapping case is DGP1 where $\theta = (0, 0, 0, 0, 0)'$ in Equation (21), the non-nested case is DGP2 where $\theta = (0.5, 0.5, 0.5, 0.5, 0.5)'$ and the power case is DGP3 where $\theta = (0, 2, 0, 0, 0)'$.

Overall, our simulation results show that our proposed tests work very well relative to the un-

adjusted versions of the same tests. In particular, we are able to significantly improve the severe under-rejection which occurs in the overlapping case, while also mitigating the over-rejection which occurs in the non-nested case. This demonstrates that our tests control size across a range of different sample sizes. Moreover, we do not find any substantial issue with power loss, with our tests providing power approaching unity for reasonable sample sizes, both in the pairwise and multiple comparison tests. In the supplementary material we also provide additional simulation results for the cases where the evaluation loss function is changed to Linex, and where both estimation and evaluation loss functions are the check loss function. These results echo those presented here and demonstrate that our test can be applied in a wide range of empirical settings.

5 Empirical Application

In this section we apply our test to analyse the predictive ability of various different models for predicting excess bond returns using both financial and macroeconomic variables. The analysis of bond risk premia continues to receive significant attention in the recent literature on financial economics. It is now well established that bond returns of various maturities are predictable in excess of the short rate. There have been a variety of different financial predictors used in previous studies, with the most widely-used being: forward spreads (Fama and Bliss, 1987), yield spreads (Campbell and Shiller, 1991) and a linear combination of forward rates (Cochrane and Piazzesi, 2005). In more recent years, it has also been suggested that macroeconomic factors contain predictive content for bond risk premia, over and above the information contained in financial predictors. The idea of using macro factors dates back to Ludvigson and Ng (2009) and has been expanded in many ways by recent studies including: Coroneo et al. (2016), Ghysels et al. (2018), Gargano et al. (2019), Andreasen et al. (2019), Massacci (2019), Bianchi et al. (2021) and Andreasen et al. (2021).

Since there is a lack of consensus about which specific financial and/or macroeconomic factors to use in forecasting excess bond returns, it is important to explore the out-of-sample predictive ability of different models which use different combinations of predictors. Our study aims to re-examine and build on existing research by combining the variables used across studies to assess which combination of the prevailing financial and macro predictors perform best in forecasting exercises. As we outline in the coming section, this gives a very natural setting to apply our tests as the candidate models can be non-nested, overlapping or nested.

5.1 Data and Set-up

We have monthly U.S. data from 1959:M1 through to 2021:M12. We use data on zero-coupon bond prices taken from the Fama-Bliss database at the Center for Research in Security Prices.¹¹ This gives

¹¹See: <https://wrds-www.wharton.upenn.edu/pages/get-data/center-research-security-prices-crsp/> [Last accessed: 03/11/22]

price data for n -year bonds with maturity n ranging from one to five years. The macroeconomic data we use are taken from the FRED-MD database of McCracken and Ng (2016).¹² This dataset contains a harmonised series of 127 U.S. variables, mostly macroeconomic series with some financial series, which is regularly updated and is now widely used in empirical “big data” forecasting. These variables cover groups including output and income; labour market; money and credit; housing and so on. The full list of variables and their groupings, as well as the data transformations for each series, can be found on the Federal Reserve Bank of St. Louis website.

The dependent variable for the study is the one-year-ahead log excess return of the n -year bond, denoted $rx_{t+12}^{(n)}$. Letting $p_t^{(n)}$ be the log price of the n -year bond and $y_t^{(n)} = -\frac{1}{n}p_t^{(n)}$ be the log yield, the excess return is then calculated as the spread between the one-year log holding return (from buying an n -year bond in time t and selling it on as an $(n - 1)$ -year bond in period $t + 12$) and the one-year yield, in other words $rx_{t+12}^{(n)} = p_{t+12}^{(n-1)} - p_t^{(n)} - y_t^{(1)}$. We will analyse all available maturities $n = 2, \dots, 5$.

We will generate out-of-sample forecasts from various models in a recursive fashion, with the first year-ahead forecast being made for 1990:M1 (using data up to 1989:M1) and the last forecast made at the end of the sample in 2021:M12. This splits the total sample of $T = 744$ months, after lagging the explanatory variables, into in-sample and out-of-sample evaluation windows of $R = 360$ and $P = 384$ respectively.

We consider three different types of model for $rx_{t+12}^{(n)}$ based on different combinations of predictors, which are summarised in Table 3. The first type of model is the traditional approach which expresses $rx_{t+12}^{(n)}$ as a linear function of some other bond market variable, $Z_t^{(n)}$, which may or may not depend on the maturity n :

$$rx_{t+12}^{(n)} = \beta_0^{(n)} + \beta_1^{(n)} Z_t^{(n)} + \varepsilon_{t+12}^{(n)}. \quad (25)$$

The benchmark model we use in our study will be Equation (25) where, following Ludvigson and Ng (2009), for $Z_t^{(n)}$ we will use the single forward factor of Cochrane and Piazzesi (2005), denoted CP_t , which is a linear combination of the one-year yield y_t^1 and the four forward rates, $f_t^{(n)}$, for maturities $n = 2, \dots, 5$.^{13,14} In other words, in the benchmark case we set $Z_t^{(n)} = CP_t$, noting that the main finding of Cochrane and Piazzesi (2005) was that the $Z_t^{(n)}$ factor did not in fact depend on the maturity. We will also use another two candidate predictors for $Z_t^{(n)}$ which *do* depend on the maturity, namely the n -year forward rate spread, $fs_t^{(n)} = f_t^{(n)} - y_t^{(1)}$, as in Fama and Bliss (1987), and the yield spread $s_t^{(n)} = y_t^{(n)} - y_t^1$ as in Campbell and Shiller (1991). These three variables are given code names CP, FB and CS, as detailed in Table 3, to allow them to be easily identified within competing models.

¹²See: <https://research.stlouisfed.org/econ/mccracken/fred-databases/> [Last accessed: 09/11/22]

¹³In the standard way, the one year forward rate for loans from period $t + n - 1$ to $t + n$ uses the zero-coupon yield curve for maturities $n - 1$ and n , specifically: $f_t^n = p_t^{(n-1)} - p_t^{(n)}$.

¹⁴In computing the CP_t variable, which takes a weighted average of forward rates based on the regression coefficients of the average excess return on all four forward rates and the one-year yield $y_t^{(1)}$, we obtain the weights only once in the first out-of-sample estimation window to avoid look-ahead bias.

Table 3: Summary of Competing Models

Model	Code	Fwd Rate Factor (CP) (1 variable)	Fwd Spreads (FB) (1 variable)	Yield Spreads (CS) (1 variable)	FRED-MD (LN) (4 variables)	Full FRED-MD (8 variables)	Single FRED-MD (1 variable)
Benchmark	1	CP	✓				
Bond Market Data Only	2	FB	✓				
	3	CS		✓			
Macro Data Only	4	LN			✓		
	5	FRED8				✓	
	6	FRED1					✓
	7	CP, LN	✓		✓		
	8	CP, FRED8	✓			✓	
	9	CP, FRED1	✓				✓
Bond Market and Macro Data	10	FB, LN	✓		✓		
	11	FB, FRED8	✓			✓	
	12	FB, FRED1	✓				✓
	13	CS, LN		✓	✓		
	15	CS, FRED8		✓		✓	
	16	CS, FRED1		✓			✓

The second type of model is based only on macroeconomic predictors:

$$rx_{t+12}^{(n)} = \alpha_0^{(n)} + \alpha^{(n)'} F_t + \varepsilon_{t+12}^{(n)} \quad (26)$$

where F_t is an $r \times 1$ vector of latent factors which are assumed to underpin the large set of N macroeconomic variables X_t with the following factor model representation:

$$X_t = \Lambda F_t + v_t$$

where Λ is an $N \times r$ matrix of factor loadings and v_t is an $N \times 1$ vector of idiosyncratic disturbances. As in Gonçalves et al. (2017), Ludvigson and Ng (2009), Massacci (2019) and Bianchi et al. (2021), we estimate the factors by principal components analysis (PCA). Since we do this in a recursive fashion as in Gonçalves et al. (2017), this results in factor estimates $\widehat{F}_{j,t}$ for all observations $j = 1, \dots, t$ in recursive windows $t = R, \dots, T - 12$. The last factor estimate in each window is used to obtain the forecasts in the feasible equivalent to Equation (26):

$$\widehat{rx}_{t+12}^{(n)} = \widehat{\alpha}_{0,t}^{(n)} + \widehat{\alpha}_t^{(n)'} \widehat{F}_{t,t} \quad (27)$$

We will use three different versions of the factors in our results. The first set is the same subset used by Ludvigson and Ng (2009), specifically $\widehat{F}_t^{LN} = [\widehat{F}_{1t}, \widehat{F}_{3t}, \widehat{F}_{4t}, \widehat{F}_{8t}]'$, except here we exclude the cubic term they use in their specifications, \widehat{F}_{1t}^3 . This is because the higher-order polynomial produces erratic out-of-sample predictions when the dataset is extended beyond the end of their original sample. The second set of factors we consider uses all of the first $r = 8$ factors from the FRED-MD database. The use of eight factors is for comparability with Ludvigson and Ng (2009) and Bianchi et al. (2021).¹⁵ Finally, we also use a specification with only the first factor from the FRED-MD database, \widehat{F}_{1t} . Since the first factor is the strongest and often referred to as the “real factor,” this seems like a natural specification to explore, as well as resulting in a more parsimonious forecasting model than the other sets of factors. It also matches more closely with studies like Coroneo et al. (2016) who use a much smaller set of factors estimated from a smaller database. These three sets of factors are given code names LN, FRED8 and FRED1 respectively, as in Table 3.

The third and final type of model combines Equations (25) and (27) to use both financial and macroeconomic predictors in making the forecasts:

$$\widehat{rx}_{t+12}^{(n)} = \widehat{\alpha}_{0,t}^{(n)} + \widehat{\alpha}_t^{(n)'} \widehat{F}_{t,t} + \widehat{\beta}_t^{(n)'} Z_t^{(n)} \quad (28)$$

which will allow us to assess whether the macroeconomic factors have additional out-of-sample pre-

¹⁵We note that the FRED-MD database we use does not match exactly with the original database of Ludvigson and Ng (2009), which used the 132 variables from Stock and Watson (2002), as some variables have since been discontinued. The FRED-MD database was created by McCracken and Ng (2016) to match the Stock and Watson database as closely as possible, so this should be the most comparable dataset we can use in our study.

dictive ability relative to the models which only include bond market data. Therefore, as detailed in Table 3, we have a total of 15 models including the benchmark: three models with only bond market variables, three models with only macroeconomic factors, and all nine combinations of both.¹⁶ The benchmark is non-nested and potentially overlapping with respect to many of these models (2-6 and 10-16), and is nested within others (7-9). This gives us a natural setting to use our test.

While the theoretical results developed in Sections 2 and 3 are not explicitly written for the case where estimated factors are used in generating forecasts, we can easily accommodate this set-up in our tests. To do this, we can rely on the results of Gonçalves et al. (2017) who show the conditions required for the forecasts made using recursively estimated factors $\widehat{F}_{t,t}$ to behave asymptotically like the predictions made using the true factors F_t . This, in turn, provides a justification for treating the DM-type statistics in our paper as if they were based on the true factors, even though these have been recursively estimated by PCA. The assumptions laid out in A1-A6 of Gonçalves et al. (2017) are rather standard assumptions from the factor model literature, modified slightly for the recursive estimation scheme. We do not repeat the assumptions here. The main thing to note is on the asymptotic rates for T and N , the time series dimension and cross-sectional dimension of the data matrix, say X . Gonçalves et al. (2017) show that factor estimation error does not contribute to predictive ability tests when $\sqrt{T}/N \rightarrow 0$. Their results mirror the earlier work on factor-augmented models of Bai and Ng (2006) who note that requiring $\sqrt{T}/N \rightarrow 0$ is a relatively weak assumption which provides a good approximation for the majority of datasets such as ours. In fact, under squared error loss and by an application of Theorem 4.1 as well as Lemma 4.1 in Gonçalves et al. (2017), it is possible to show that a similar conclusion holds true also for our set-up provided that $\sqrt{P}/\min\{N, R\} \rightarrow 0$.

The models described above are estimated using least squares and evaluated using the squared error loss function. All data are standardised to have zero mean and unit variance. For the bootstrap implementation we use $B = 2500$ draws and a block length of $l_P = 4$, and will compute the critical values at a significance level of 5%. For the replicability of the random perturbation, the random number seed is set to 1000 before commencing the study (R version 4.3.1), though we discuss below how the results are unaffected by the random number seed chosen.

5.2 Results

The results of the out-of-sample experiment are displayed in Table 4 which displays the MSFE for each of the models for excess returns at maturities two through five. The model numbers and their specifications can be found in Table 3 above. The lowest MSFE is highlighted in bold in each column, along with the test statistic and 5% critical value for our test $\widehat{S}_P(\widehat{\eta}_{P,R})$ along with the unadjusted version, $\widehat{S}_P^D M$, and the Hansen (2005) SPA test, \widehat{H}_P .

¹⁶In fact, for maturity $n = 2$ there are some redundant models as the 2-year forward rate spread is a scale multiple of the 2-year yield spread when yields are computed using continuous discounting. This is not the case for maturities larger than $n = 2$.

Table 4: MSFE and SPA Test Results

Model	Code	MSFE			
		$rx_t^{(2)}$	$rx_t^{(3)}$	$rx_t^{(4)}$	$rx_t^{(5)}$
1	CP	0.831	0.927	1.001	1.028
2	FB	0.642	0.675	0.682	0.677
3	CS	0.642	0.678	0.680	0.677
4	LN	0.663	0.686	0.701	0.705
5	FRED8	0.743	0.747	0.754	0.740
6	FRED1	0.586	0.635	0.673	0.698
7	CP.LN	0.627	0.697	0.754	0.776
8	CP.FRED8	0.610	0.675	0.718	0.727
9	CP.FRED1	0.699	0.803	0.891	0.931
10	FB.LN	0.678	0.719	0.738	0.706
11	FB.FRED8	0.732	0.794	0.803	0.755
12	FB.FRED1	0.600	0.649	0.666	0.663
13	CS.LN	0.678	0.705	0.711	0.708
14	CS.FRED8	0.732	0.779	0.786	0.770
15	CS.FRED1	0.600	0.642	0.656	0.660

Test Statistics					
\widehat{S}_P^{DM}	64.340	79.019	86.427	91.895	
CV 5%	33.560	34.076	36.688	35.948	
\widehat{H}_P	3.312	3.329	3.422	3.295	
CV 5%	2.127	2.132	2.149	2.158	
$\widehat{S}_P(\widehat{\eta}_{P,R})$	33.420	41.926	46.811	48.780	
CV 5%	22.182	23.804	24.608	24.581	

Notes: The model numbers and codes are detailed in Table 3. Figures in bold denote the lowest MSFE per column.

There are several interesting results to take away from the MSFE numbers, before turning to the test results. The first finding is that using macroeconomic data seems to deliver the lowest MSFE across all maturities. Secondly, we see that the best results seem to come from only using the first factor from the FRED-MD database, with either model 6 (FRED1) or 15 (CS.FRED1) coming top across all maturities. This result suggests that, while using macroeconomic data can reduce the MSFE, it may not be the case that all information is useful and there could be some worsening of forecast model performance from using too many factors. Finally, while macroeconomic data seem to generally be useful in all cases, we do observe some variation in the best model across maturities. For the two- and three-year maturities, model FRED1 with only the single factor has the lowest MSFE. On the other hand, for the four- and five-year maturities, model CS.FRED1 with yield spreads and the single macro factor has lowest RMSFE. This indicates that some maturity-specific modelling may be appropriate, which mirrors the broader conclusions of Bianchi et al. (2021).

We now turn to the results of the SPA test for the null hypothesis described in Equation (15) above. Relative to the benchmark of the Cochrane and Piazzesi (2005) model, across all four maturities we are able to strongly reject the null hypothesis at the 5% significance level using all tests.¹⁷ This indicates statistical evidence to confirm that macroeconomic factors are important predictors of excess bond returns relative to this benchmark model. In the case of $n = 2$ and $n = 3$ the superior model uses only the single FRED-MD factor, as written above, whereas for $n = 4$ and $n = 5$ this factor is in the best model alongside the yield spread variable. In each of the maturities, the best model improves by 30% or more against the benchmark in terms of MSFE. We also checked the sensitivity of our test with respect to the random number seed used in obtaining the random perturbation and found that the test conclusions were unaffected by this change.¹⁸ Finally the simulation evidence from Section 4 suggests that we may be rather confident that the test has been carried out at the nominal level.

6 Conclusion

This paper proposes new tests for equal predictive ability between two or more competing models. Asymptotically, the tests control the Type I error probability at nominal level uniformly over strictly non-nested and overlapping DGPs thus avoiding the need for pre-testing to determine whether the variance of the test is significantly different from zero or not. Our methods are simple to implement and involve the addition of a random perturbation to standard DM and SPA type tests that vanishes asymptotically if the models are strictly non-nested or if the variance collapses to zero sufficiently slowly. Our tests are valid in a wide set of scenarios, accommodating dynamic misspecification, general loss functions and different out-of-sample estimation schemes. The main conclusion from our theoretical and simulation results is that our tests exhibit well controlled asymptotic size, and may be able to discern models that are close to overlapping in finite samples. This improves over existing DM-type tests which have poor size properties, becoming degenerate in the overlapping case and tending to over-reject in the non-nested case.

We apply our test to predicting U.S. excess bond returns using combinations of financial and macroeconomic factors employed in the existing literature. Our analysis concludes that there is statistical evidence in favour of using macroeconomic information over and above bond market data in predicting year-ahead excess bond returns across various maturities. We also conclude that the predictive ability of the models we consider can differ across different bond maturities.

¹⁷We are also able to reject at the 1% level though here we only present the 5% critical values as we calculate our test statistic using the data-driven constant obtained to balance the size-power trade-off at the nominal level of 5%.

¹⁸For instance, for variable $rx_t^{(2)}$ the test statistic and 5% critical value using random number seed 1000 are 64.34 and 33.56, whereas with a random number seed of 9999 these become 64.34 and 30.87.

7 Appendix

This Appendix presents some auxiliary Lemmas as well as the proofs of Theorems 1.1-1.3, 3.1-3.2. The proofs of the former as well as the proofs of Theorem 2.1 and 2.2 can be found in the supplementary material. Moreover, for the subsequent auxiliary Lemmas, we require the following definitions for $f \in \{g, q\}$:

$$\nabla_{\theta_j}^{(1)} f(y_{t+1}, Z_{j,t}; \theta_j) \equiv \nabla^{(1)} f(y_{t+1} - m(Z_{j,t}; \theta_j)) \nabla_{\theta_j}^{(1)} m(Z_{j,t}; \theta_j) \quad (29)$$

and:

$$\begin{aligned} \nabla_{\theta_j}^{(2)} f(y_{t+1}, Z_{j,t}; \theta_j) &\equiv \nabla^{(1)} f(y_{t+1} - m(Z_{j,t}; \theta_j)) \nabla_{\theta_j}^{(2)} m(Z_{j,t}; \theta_j) \\ &\quad + \nabla^{(2)} f(y_{t+1} - m(Z_{j,t}; \theta_j)) \nabla_{\theta_j}^{(1)} m(Z_{j,t}; \theta_j)^2. \end{aligned} \quad (30)$$

Lemma A1. Let Assumptions A.1-A.6 hold with $g(\cdot) = q(\cdot)$ and $0 \leq \pi < \infty$. Moreover, for $j = 1, 2$, recall the definitions of $H_{j,F}^{-1}$ as the inverse of $H_{j,F}$ defined in Equation (7) and of $V_{j,F}$ defined in Equation (6) of Assumption A.4(ii). Then, for any given $F \in \mathcal{F}_0$:

$$\begin{aligned} &\left| \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(\hat{u}_{1,t+1}) - g(\hat{u}_{2,t+1})) - \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(u_{1,t+1}) - g(u_{2,t+1})) \right| \\ &\leq 2\sqrt{\frac{T}{R}} \max \{ \text{tr}(H_{1,F}^{-1}V_{1,F}), \text{tr}(H_{2,F}^{-1}V_{2,F}) \} R^{-1/2} \sqrt{2 \ln \ln P} \sqrt{2 \ln \ln T} \end{aligned}$$

almost surely.

Lemma A2. A.1-A.6 hold with $g(\cdot) \neq q(\cdot)$ and $\pi = 0$. Moreover, for $j = 1, 2$, recall the definitions of $H_{j,F}^{-1}$ as the inverse of $H_{j,F}$ defined in Equation (7) and of $V_{j,F}$ defined in Equation (6) of Assumption A.4(ii), while:

$$G_{j,F} \equiv \lim_{P,R \rightarrow \infty} \mathbb{E}_F \left(\frac{1}{P} \sum_{t=R}^{T-1} \nabla_{\theta_j}^{(1)} g(y_{t+1}, Z_{j,t}; \theta_j^\dagger) \right) \quad (31)$$

with $\nabla_{\theta_j}^{(1)} g(y_{t+1}, Z_{j,t}; \theta_j^\dagger)$ as defined in Equation (29). Then, for any given $F \in \mathcal{F}_0$:

$$\begin{aligned} &\left| \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(\hat{u}_{1,t+1}) - g(\hat{u}_{2,t+1})) - \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(u_{1,t+1}) - g(u_{2,t+1})) \right| \\ &\leq 2\frac{\sqrt{P}}{\sqrt{R}} \sqrt{2 \ln \ln T} \times \max \{ |G'_{1,F} H_{1,F}^{-1} V_{1,F}|, |G'_{2,F} H_{2,F}^{-1} V_{2,F}| \} \end{aligned}$$

almost surely (F).

Proof of Theorem 1.1:

(i) Without loss of generality, we consider the recursive estimation scheme and consider sequence $\{F_P\}_{P=1}^\infty$ such that $F_P \in \mathcal{F}_0$ for every P . Now, when $g(\cdot) = q(\cdot)$, from Lemma A1 we have that for any given $F_P \in \mathcal{F}_0$:

$$\begin{aligned} \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(\widehat{u}_{1,t+1}) - g(\widehat{u}_{2,t+1})) &= \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(u_{1,t+1}) - g(u_{2,t+1})) \\ &\quad + O_{\text{Pr}_{F_P}} \left(R^{-1/2} \sqrt{\ln \ln T} \sqrt{\ln \ln P} \right) \end{aligned}$$

and so:

$$\begin{aligned} &\widehat{DM}_P(\eta_{P,R}) \\ &= \frac{\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(u_{1,t+1}) - g(u_{2,t+1})) + O_{\text{Pr}_{F_P}} \left(R^{-1/2} \sqrt{\ln \ln T} \sqrt{\ln \ln P} \right) + \eta_{P,R} \frac{1}{\sqrt{P}} \sum_{s=1}^P e_s}{\sqrt{\widehat{\sigma}_{DM,P}^2 + \eta_{P,R} \frac{1}{P} \sum_{s=1}^P e_s^2}}. \end{aligned}$$

On the other hand, in the case of $g(\cdot) \neq q(\cdot)$ with $\pi = 0$, the $O_{\text{Pr}_{F_P}} \left(R^{-1/2} \sqrt{\ln \ln T} \sqrt{\ln \ln P} \right)$ term in the above expressions can be replaced by $O_{\text{Pr}_{F_P}} \left((P/R)^{-1/2} \sqrt{\ln \ln T} \right) = o_{\text{Pr}_{F_P}}(1)$, which follows from Lemma A2 and the fact that $\pi = 0$. When F_P approaches the strictly non-nested case, i.e. $\sigma_{DM,P}^2 \rightarrow \sigma_{DM}^2 > 0$ as $P \rightarrow \infty$, note that A.1, A.4 and A.6(i) ensure that Assumptions B and C in Andrews (1991) are satisfied, and so by his Theorem 1 we have that $\widehat{\sigma}_{DM,P}^2 - \sigma_{DM}^2 = O_{\text{Pr}_{F_P}} \left(\sqrt{\frac{l_P}{P}} \right)$. On the other hand, since, $\frac{1}{\sqrt{P}} \sum_{s=1}^P e_s = O_{\text{Pr}_{F_P}}(1)$, $\frac{1}{P} \sum_{s=1}^P e_s^2 = 1 + o_{a.s.}(1)$, and $\eta_{P,R} \rightarrow 0$ as $P, R \rightarrow \infty$, we have that for a given F_P :

$$\widehat{DM}_P(\eta_{P,R}) = \frac{\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(u_{1,t+1}) - g(u_{2,t+1})) + o_{\text{Pr}_{F_P}}(1)}{\sqrt{\sigma_{DM}^2 + o_{\text{Pr}_{F_P}}(1)}} \xrightarrow{d} N(0, 1)$$

Next, recall that in the overlapping case we have that:

$$\sigma_{DM,P}^2 = \text{var}_{F_P} \left(\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(u_{1,t+1}) - g(u_{2,t+1})) \right) \rightarrow 0.$$

as $P, R \rightarrow \infty$. Hereafter, we let $\sigma_{DM,P}^2 = c \cdot d_P (1 + o(1))$ for some generic, constant c with $0 < c < \infty$, and some deterministic sequence $d_P \rightarrow 0$ as $P \rightarrow \infty$ such that either $Pd_P \rightarrow C < \infty$ or $Pd_P \rightarrow \infty$, where C is again a generic constant satisfying $0 \leq C < \infty$. In either case, note that

$$d_P^{-1/2} \frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(u_{1,t+1}) - g(u_{2,t+1}))$$

converges weakly to a standard normal random variable, under F_P . Also, note that:

$$\widehat{\sigma}_{DM,P}^2 = \sigma_{DM,P}^2 + O_{Pr_{F_P}} \left(d_P \sqrt{\frac{l_P}{P}} \right) + O_{Pr_{F_P}} \left(\frac{l_P}{P} \right), \quad (32)$$

where $O_{Pr_{F_P}} \left(\frac{l_P}{P} \right)$ denotes the rate at which the estimation error component of the variance estimator approaches zero. Given Lemma A1, we have that:¹⁹

$$\begin{aligned} & \widehat{DM}_P(\eta_{P,R}) \\ = & \frac{\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(u_{1,t+1}) - g(u_{2,t+1})) + O_{Pr_{F_P}} \left(R^{-1/2} \sqrt{\ln \ln T} \sqrt{\ln \ln P} \right) + \eta_{P,R} \frac{1}{\sqrt{P}} \sum_{s=1}^P e_s}{\sqrt{\sigma_{DM,P}^2 + O_{Pr_{F_P}} \left(\max \left\{ \frac{l_P}{P}, d_P \sqrt{\frac{l_P}{P}} \right\} \right) + \eta_{P,R}^2 (1 + o_{a.s.}(1))}}. \end{aligned}$$

Now, if $d_P^{1/2}/\eta_{P,R} \rightarrow 0$, since by Assumption A.6(ii) it holds that $(Pl_P^{-1})^{1/4} \eta_{P,R} \rightarrow \infty$,

$$\widehat{DM}_P(\eta_{P,R}) = \frac{\eta_{P,R} \frac{1}{\sqrt{P}} \sum_{s=1}^P e_s \left(1 + o_{Pr_{F_P}}(1) \right)}{\sqrt{cd_P + O_{Pr_{F_P}} \left(\max \left\{ \frac{l_P}{P}, d_P \sqrt{\frac{l_P}{P}} \right\} \right) + \eta_{P,R}^2 (1 + o_{a.s.}(1))}} \xrightarrow{d} N(0, 1).$$

If instead $d_P^{1/2}/\eta_{P,R} \rightarrow \infty$,

$$\widehat{DM}_P(\eta_{P,R}) = \frac{\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(u_{1,t+1}) - g(u_{2,t+1})) + o_{Pr_{F_P}}(1)}{\sqrt{cd_P + O_{Pr_{F_P}} \left(\max \left\{ \frac{l_P}{P}, d_P \sqrt{\frac{l_P}{P}} \right\} \right) (1 + o(1))}} \xrightarrow{d} N(0, 1).$$

Finally, if $d_P^{1/2}/\eta_{P,R} \rightarrow C$, $0 < C < \infty$, then both components matter and:

$$\begin{aligned} & \widehat{DM}_P(\eta_{P,R}) \\ = & \frac{\frac{1}{\sqrt{P}} \sum_{t=R}^{T-1} (g(u_{1,t+1}) - g(u_{2,t+1})) + \eta_{P,R} \frac{1}{\sqrt{P}} \sum_{s=1}^P e_s}{\sqrt{cd_P + O_{Pr_{F_P}} \left(\max \left\{ \frac{l_P}{P}, d_P \sqrt{\frac{l_P}{P}} \right\} \right) + \eta_{P,R}^2 (1 + o_{a.s.}(1))}} + o_{Pr_{F_P}}(1) \\ & \xrightarrow{d} N(0, 1). \end{aligned}$$

Hence, regardless of the speed at which $d_P^{1/2}$ and $\eta_{P,R}$ approach zero, we have weak convergence to a standard normal.

¹⁹The same follows from Lemma A2 with the $O_{Pr_{F_P}} \left(R^{-1/2} \sqrt{\ln \ln T} \sqrt{\ln \ln P} \right)$ term again replaced by $O_{Pr_{F_P}} \left((R/P)^{-1/2} \sqrt{\ln \ln T} \right)$.

(ii) Under H_A , for any $\{F_P\}_{P=1}^\infty$ with $F_P \in \mathcal{F}_A$ for all P , we have that:

$$\widehat{DM}_P(\eta_{P,R}) = \frac{\sqrt{P}E_{F_P}(g(u_{1,t+1}) - g(u_{2,t+1}))}{\sqrt{\sigma_{DM}^2 + o_{\text{Pr}_{F_P}}(1)}} + O_{\text{Pr}_{F_P}}(1).$$

Given A.4 and the fact that $E_{F_P}(g(u_{1,t+1}) - g(u_{2,t+1})) \neq 0$, the result of the statement follows. ■

Proof of Theorem 1.2: Both parts are immediate from the proof of Theorem 1.1. ■

Proof of Theorem 1.3: Note that the result of Theorem 1.1(i) holds for any sequence $\{F_P\}_{P=1}^\infty$ with $F_P \in \mathcal{F}_0$ for every P . The result then follows from the same argument as in the proof of Theorem 11.4.5 in Lehmann and Romano (2005). ■

Proof of Theorem 3.1: Recall that:

$$\varsigma_k = \lim_{P,R \rightarrow \infty} \frac{\sqrt{P}E_{F_P}(m_{k,P})}{\sqrt{\sigma_{k,P}^2 + \eta_{k,P}^2}},$$

and recall that $m_{k,P}$ be defined as $\widehat{m}_{k,P}$, but with $\widehat{u}_{1,t+1}, \widehat{u}_{k,t+1}$ replaced by $u_{1,t+1}, u_{k,t+1}$. Given Theorem 2.2, by a similar argument as that used in the proof of Theorem 1.1,

$$\begin{aligned} & \frac{\sqrt{P}\widehat{m}_{k,P} + \widehat{\eta}_{k,P,R} \frac{1}{\sqrt{P}} \sum_{t=1}^P e_{k,t}}{\sqrt{\widehat{\sigma}_{k,P}^2 + \widehat{\eta}_{k,P,R}^2 \frac{1}{P} \sum_{t=1}^P e_{k,t}^2}} \\ &= \frac{\sqrt{P}\widehat{m}_{k,P} + \eta_{k,P,R} \frac{1}{\sqrt{P}} \sum_{t=1}^P e_{k,t}}{\sqrt{\widehat{\sigma}_{k,P}^2 + \eta_{k,P,R}^2 \frac{1}{P} \sum_{t=1}^P e_{k,t}^2}} (1 + o_{\text{Pr}_F}(1)) \\ &= \frac{\sqrt{P}m_{k,P} + \eta_{k,P,R} \frac{1}{\sqrt{P}} \sum_{t=1}^P e_{k,t}}{\sqrt{\sigma_{k,P}^2 + \eta_{k,P,R}^2 \frac{1}{P} \sum_{t=1}^P e_{k,t}^2}} (1 + o_{\text{Pr}_F}(1)) \end{aligned}$$

where $\lim_{P,R \rightarrow \infty} \sigma_{k,P}^2 = 0$ if models 1 and k overlap, while $\lim_{P,R \rightarrow \infty} \sigma_{k,P}^2 > 0$ if models 1 and k do not overlap. Hereafter, let:

$$\widehat{S}_{k,P}(\widehat{\eta}_{k,P,R}) = \frac{\sqrt{P}\widehat{m}_{k,P} + \widehat{\eta}_{k,P,R} \frac{1}{\sqrt{P}} \sum_{t=1}^P e_{k,t}}{\sqrt{\widehat{\sigma}_{k,P}^2 + \widehat{\eta}_{k,P,R}^2 \frac{1}{P} \sum_{t=1}^P e_{k,t}^2}}.$$

Then, for a given $F \in \mathcal{F}_0$:

$$\begin{pmatrix} \widehat{S}_{2,P}(\widehat{\eta}_{P,R}) - \varsigma_2 \\ \vdots \\ \widehat{S}_{J/2,P}(\widehat{\eta}_{P,R}) - \varsigma_{J/2} \\ \widehat{S}_{J/2+1,P}(\widehat{\eta}_{P,R}) - \varsigma_{J/2+1} \\ \vdots \\ \widehat{S}_{J,P}(\widehat{\eta}_{P,R}) - \varsigma_J \end{pmatrix} \xrightarrow{d} N(0, \Omega), \quad (33)$$

where for $i, j = 2, \dots, J$, the matrix Ω with elements ω_{ij} is defined as in the statement of the theorem. The result follows by a similar argument as in Andrews and Soares (2010). The limiting distribution in Equation (33) holds for any given $F \in \mathcal{F}_0^{RC}$ and for all k , where $\varsigma_k = 0$ or $\varsigma_k = -\infty$, depending on whether $\lim_{P,R \rightarrow \infty} \frac{\sqrt{P}E_F(m_{k,P})}{\sqrt{\sigma_{k,P}^2 + \eta_{k,P}^2}} = 0$ or $-\infty$. ■

Proof of Theorem 3.2:

(i) Hereafter, let $\Pr_F^* = \Pr_{1,F}^* \times \Pr_{2,F}^*$, where $\Pr_{1,F}^*$ is the probability measure governing the resampling of the data with distribution F , and $\Pr_{2,F}^*$ is the probability measure governing the resample of the added randomness $e_{k,t}$ (and so in fact $\Pr_{2,F}^* \equiv \Pr_2^*$ under any F). Also, in what follows we write E_F^* and var_F^* to denote the \Pr_F^* expectation and variance operator conditional on the original sample draws with distribution F . Then, recalling that $\Pr_{1,F}^*$ and $\Pr_{2,F}^*$ are independent of each other, note that $E_F^*(\widehat{m}_{k,p}^* + \bar{e}_k^*) = E_{1,F}^*(\widehat{m}_{k,p}^*) + E_2^*(\bar{e}_k^*)$ and $\text{var}_F^*(\widehat{m}_{k,p}^* + \bar{e}_k^*) = \text{var}_{1,F}^*(\widehat{m}_{k,p}^*) + \text{var}_2^*(\bar{e}_k^*)$. Given Lemma A1, for all $k = 2, \dots, J$:

$$E_{1,F}^*(\widehat{m}_{k,p}^*) = m_{k,P} + O_{\Pr_F} \left(\frac{l_P}{P} \right) \text{ and } E_2^*(\bar{e}_k^*) = \bar{e}_k = O_{\Pr_F} \left(\frac{1}{\sqrt{P}} \right)$$

and by the same argument as in the proof of Theorem 3 in Corradi and Swanson (2006),

$$\text{var}_{1,\Pr_F}^* \left(\sqrt{P}\widehat{m}_{k,p}^* \right) = \widehat{\sigma}_{k,P}^2 + O_{\Pr_F} \left(\frac{l_P}{\sqrt{P}} \right) \quad (34)$$

and:

$$\text{cov}_{1,F}^* \left(\sqrt{P}\widehat{m}_{k,p}^* \sqrt{P}\widehat{m}_{k',p}^* \right) = \widehat{\text{cov}} \left(\sqrt{P}m_{k,P} \sqrt{P}m_{k',P} \right) + O_{\Pr_F} \left(\frac{l_P}{\sqrt{P}} \right),$$

while:

$$\text{var}_2^* \left(\sqrt{P}\bar{e}_k^* \right) = \frac{1}{P} \sum_{t=1}^P e_{t,k}^2.$$

Using the comparison of the benchmark and model 2 as example, note that by Theorem 2.3:

$$\begin{aligned} \widehat{DM}_{2,P}^*(\widehat{\eta}_{2,P,R}) &= \frac{\sqrt{P}(\widehat{m}_{2,P}^* + \widehat{\eta}_{2,P,R}\bar{e}_2^*) - \sqrt{P}(\widehat{m}_{2,P} + \widehat{\eta}_{2,P,R}\bar{e}_2)}{\sqrt{\widehat{\sigma}_{2,P}^{*2} + \widehat{\eta}_{2,P,R}^2\bar{e}_2^{*2}}} \\ &= \frac{\sqrt{P}(\widehat{m}_{2,P}^* + \eta_{2,P,R}\bar{e}_2^*) - \sqrt{P}(\widehat{m}_{2,P} + \eta_{2,P,R}\bar{e}_2)}{\sqrt{\widehat{\sigma}_{2,P}^{*2} + \eta_{2,P,R}^2\bar{e}_2^{*2}}} + o_{\Pr_F}(1). \end{aligned}$$

Then, conditional on the sample and for all the samples but a set of probability measure approaching zero,

$$\begin{aligned} &\begin{pmatrix} \frac{\sqrt{P}(\widehat{m}_{2,P}^* + \eta_{2,P,R}\bar{e}_2^*) - \sqrt{P}(\widehat{m}_{2,P} + \eta_{2,P,R}\bar{e}_2)}{\sqrt{\widehat{\sigma}_{2,P}^{*2} + \eta_{2,P,R}^2\bar{e}_2^{*2}}} \\ \vdots \\ \frac{\sqrt{P}(\widehat{m}_{J,P}^* + \eta_{J,P,R}\bar{e}_J^*) - \sqrt{P}(\widehat{m}_{J,P} + \eta_{J,P,R}\bar{e}_J)}{\sqrt{\widehat{\sigma}_{J,P}^{*2} + \eta_{J,P,R}^2\bar{e}_J^{*2}}} \end{pmatrix} \\ &\xrightarrow{d^*} N(0, \Omega), \end{aligned} \tag{35}$$

with probability converging to one, where d^* denotes convergence in distribution according to \Pr_F^* conditional on the sample. Note that the limiting distribution on the RHS of Equation (35) is the same as that of $\left(\frac{\sqrt{P}(\widehat{m}_{2,P} + \eta_{2,P,R}\bar{e}_2)}{\sqrt{\widehat{\sigma}_{2,P}^2 + \eta_{2,P,R}^2\bar{e}_2^2}} - \varsigma_2, \dots, \frac{\sqrt{P}(\widehat{m}_{J,P} + \eta_{J,P,R}\bar{e}_J)}{\sqrt{\widehat{\sigma}_{J,P}^2 + \eta_{J,P,R}^2\bar{e}_J^2}} - \varsigma_J\right)'$ where $\varsigma_k = 0$, and so corresponds to the limiting distribution for the least favourable case under the null with all moment inequalities binding.

Now, let $c_{B,P,R,1-\alpha}^*$ denote the $1 - \alpha$ critical value of $\widehat{S}_P^*(\widehat{\eta}_{P,R})$, based on B bootstrap replications. Also let $c_{1-\alpha, \pi^*}$ be the $(1 - \alpha)$ critical value of $\sum_{k=2}^J (\max\{0, (\Omega^{1/2}Z)_k + \pi_k^*\})^2$, where $\pi_k^* = 0$ if $\pi_k = 0$ and $\pi_k^* = -\infty$ if $\pi_k < 0$ with $\pi = (\pi_2, \dots, \pi_J) \in \mathbb{R}_{-\infty}^{(J-1)}$, with $\mathbb{R}_- = \{x \in \mathbb{R}, x \leq 0\}$ and $\mathbb{R}_{-\infty} = \mathbb{R}_- \cup \{-\infty\}$. Similarly, let $h = (h_2, \dots, h_J) \in \mathbb{R}_{-\infty}^{(J-1)}$. Finally, let $\{\gamma_P\}_{P=1}^\infty$ be a sequence with each $\gamma_P \in \Gamma$ such that $\sqrt{P}\gamma_P \rightarrow h$, and for $\kappa_P \rightarrow \infty$, $\kappa_P/\sqrt{P} \rightarrow 0$, $\kappa_P^{-1}\sqrt{P}\gamma_P \rightarrow \pi$, $\pi = (\pi_2, \dots, \pi_J)$ as $P \rightarrow \infty$.²⁰ Then, given A.1-A.7, as $P, B \rightarrow \infty$, it follows that $\limsup_{P,R \rightarrow \infty} \Pr_{\gamma_P} \left(\widehat{S}_P(\widehat{\eta}_{P,R}) > c_{B,P,R,1-\alpha}^*\right) \leq \alpha$ from Lemma 2(a) and (b) in the supplement of Andrews and Soares (2010). In fact, there are only two differences with respect to the latter: the first one consists of the fact that we use bootstrap critical values rather than plug-in based ones, say $\widehat{c}_{P,1-\alpha}$. However, since $\widehat{c}_{P,1-\alpha} - c_{B,P,R,1-\alpha}^* = o_{\Pr_{\gamma_P}}(1)$ as $B, P \rightarrow \infty$, this does not alter the argument of the proofs. The second difference lies in the definition of Ω and ς_P , which depend also on the regularisation parameter $\eta_{P,R}$. Again, however, note that neither of these terms plays a role in the proof of their Lemma 2. Finally, if for some $k = 2, \dots, J$, it holds that $m_k = 0$, then Assumption 7 in Andrews and Soares (2010) is satisfied, and so it also follows that $\limsup_{P,R \rightarrow \infty} \Pr_{\gamma_P} \left(\widehat{S}_P(\widehat{\eta}_{P,R}) > c_{B,P,R,1-\alpha}^*\right) = \alpha$ (cf. Lemma 3 in Andrews and Soares, 2010). The statement then follows from the subsequence arguments of the proof of Theorem

²⁰As in Andrews and Soares (2010), we use Γ and \mathcal{F}_0^{RC} and the corresponding operators interchangeably in what follows.

1 in Andrews and Soares (2010, supplement).

(ii) Without loss of generality, suppose that the null is violated for the first K models with $K + K' = J$, and $K > 0$. Specifically, for $k \in \{1, \dots, K\}$ we have that $\lim_{P,R \rightarrow \infty} m_{k,P}/\sqrt{P} > 0$ and so $m_{k,P} = O_{\text{Pr}_F}(\sqrt{P})$. Thus, for sufficiently large P we have that:

$$\begin{aligned} \widehat{S}_P(\widehat{\eta}_{P,R}) &= \sum_{k=2}^K \left(\frac{\sqrt{P}\widehat{m}_{k,P} + \widehat{\eta}_{k,P,R} \frac{1}{\sqrt{P}} \sum_{t=1}^P e_{k,t}}{\sqrt{\widehat{\sigma}_{k,P}^2 + \widehat{\eta}_{k,P,R}^2 \frac{1}{P} \sum_{t=1}^P e_{k,t}^2}} \right)^2 \\ &\quad + \sum_{k=K+1}^J \left(\max \left\{ 0, \frac{\sqrt{P}\widehat{m}_{k,P} + \widehat{\eta}_{k,P,R} \frac{1}{\sqrt{P}} \sum_{t=1}^P e_{k,t}}{\sqrt{\widehat{\sigma}_{k,P}^2 + \widehat{\eta}_{k,P,R}^2 \frac{1}{P} \sum_{t=1}^P e_{k,t}^2}} \right\} \right)^2 + o_{\text{Pr}_F}(1) \\ &= \sum_{k=2}^K \left(\frac{\sqrt{P}\widehat{m}_{k,P} + \widehat{\eta}_{k,P,R} \frac{1}{\sqrt{P}} \sum_{t=1}^P e_{k,t}}{\sqrt{\widehat{\sigma}_{k,P}^2 + \widehat{\eta}_{k,P,R}^2 \frac{1}{P} \sum_{t=1}^P e_{k,t}^2}} - \frac{\sqrt{P}m_{k,P}}{\sigma_{k,P}} \right)^2 + \frac{Pm_{k,P}^2}{\sigma_{k,P}^2} \\ &\quad - 2 \frac{\sqrt{P}m_{k,P}}{\sigma_{k,P}} \sum_{k=2}^K \left(\frac{\sqrt{P}\widehat{m}_{k,P} + \widehat{\eta}_{k,P,R} \frac{1}{\sqrt{P}} \sum_{t=1}^P e_{k,t}}{\sqrt{\widehat{\sigma}_{k,P}^2 + \widehat{\eta}_{k,P,R}^2 \frac{1}{P} \sum_{t=1}^P e_{k,t}^2}} - \frac{\sqrt{P}m_{k,P}}{\sigma_{k,P}} \right) + o_{\text{Pr}_F}(1) \end{aligned}$$

which diverges to plus infinity with probability approaching one. On the other hand, $\widehat{S}_P^*(\widehat{\eta}_{P,R})$ converges in distribution conditional on the sample (under any given $F \in \mathcal{F}_A^{RC}$), and for all samples but for a set of probability measure approaching zero. The statement then follows. ■

References

- Adrian, T., N. Boyarchenko, and D. Giannone (2019). Vulnerable growth. *American Economic Review* 109(4), 1263–1289.
- Amburgey, A. and M. McCracken (2023). On the Real-Time Predictive Content of Financial Conditions Indices for Growth. *Journal of Applied Econometrics* 38(2), 137–163.
- Andreasen, M., K. Jørgensen, and A. Meldrum (2019). Bond Risk Premiums at the Zero Lower Bound. Technical Report 2019-40, FEDS Working Paper.
- Andreasen, M. M., T. Engsted, S. V. Møller, and M. Sander (2021). The yield spread and bond return predictability in expansions and recessions. *The Review of Financial Studies* 34(6), 2773–2812.
- Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59(3), 817–858.
- Andrews, D. W. K. and G. Soares (2010). Inference for parameters defined by moment inequalities generalized moment selection. *Econometrica* 78(1), 119–157.
- Bai, J. and S. Ng (2006). Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions. *Econometrica* 74(4), 1133–1150.
- Bianchi, D., M. Büchner, and A. Tamoni (2021). Bond risk premiums with machine learning. *The Review of Financial Studies* 34(2), 1046–1089.

- Campbell, J. Y. and R. J. Shiller (1991). Yield spreads and interest rate movements: A bird's eye view. *The Review of Economic Studies* 58(3), 495–514.
- Clark, T. E. and M. W. McCracken (2001). Tests of Equal Forecast Accuracy and Encompassing for Nested Models. *Journal of Econometrics* 105(1), 85–110.
- Clark, T. E. and M. W. McCracken (2014). Tests of equal forecast accuracy for overlapping models. *Journal of Applied Econometrics* 29(3), 415–430.
- Cochrane, J. H. and M. Piazzesi (2005). Bond risk premia. *American Economic Review* 95(1), 138–160.
- Coroneo, L., D. Giannone, and M. Modugno (2016). Unspanned macroeconomic factors in the yield curve. *Journal of Business and Economic Statistics* 34(3), 472–485.
- Coroneo, L. and F. Iacone (2020). Comparing predictive accuracy in small samples using fixed-smoothing asymptotics. *Journal of Applied Econometrics* 35, 391–409.
- Coroneo, L., F. Iacone, A. Paccagnini, and P. Santos Monteiro (2023). Testing the predictive accuracy of covid-19 forecasts. *International Journal of Forecasting* 39(2), 606–622.
- Corradi, V. (1999). Deciding between I(0) and I(1) via Flil-Based Bounds. *Econometric Theory* 5(15), 643–663.
- Corradi, V. and N. Swanson (2006). Bootstrap conditional distribution tests in the presence of dynamic misspecification. *Journal of Econometrics* 133, 779–806.
- Corradi, V. and N. Swanson (2007). Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes. *International Economic Review* 48(1), 67–109.
- Corradi, V. and N. R. Swanson (2002). A consistent test for nonlinear out of sample predictive accuracy. *Journal of Econometrics* 110(2), 353–381.
- Diebold, F. X. and R. S. Mariano (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics* 13(3), 253–263.
- Fama, E. F. and R. R. Bliss (1987). The information in long-maturity forward rates. *The American Economic Review* 77(4), 680–692.
- Gargano, A., D. Pettenuzzo, and A. Timmermann (2019). Bond return predictability: Economic value and links to the macroeconomy. *Management Science* 65(2), 508–540.
- Ghysels, E., C. Horan, and E. Moench (2018). Forecasting through the rearview mirror: Data revisions and bond return predictability. *The Review of Financial Studies* 31(2), 678–714.
- Giacomini, R., D. N. Politis, and H. White (2013). A Warp-Speed Method for Conducting Monte Carlo Experiments Involving Bootstrap Estimators. *Econometric Theory* 29(3), 567–589.
- Giacomini, R. and H. White (2006). Tests of Conditional Predictive Ability. *Econometrica* 74(6), 1545–1578.
- Gonçalves, S., M. W. McCracken, and B. Perron (2017). Tests of Equal Accuracy for Nested Models with Estimated Factors. *Journal of Econometrics* 198(2), 231–252.
- Goncalves, S. and H. White (2004). Maximum likelihood and the bootstrap for nonlinear dynamic models. *Journal of Econometrics* 119, 199–219.
- Hansen, P. and E.-I. Dumitrescu (2022). How should parameter estimation be tailored to the objective? *Journal of Econometrics* 230, 535–558.
- Hansen, P. and A. Timmermann (2015). Equivalence Between Out-of-Sample Forecast Comparisons and Wald Statistics. *Econometrica* 83(6), 2485–2505.
- Hansen, P. R. (2005). A Test for Superior Predictive Ability. *Journal of Business and Economic Statistics* 23(4), 365–380.

- Hsu, Y.-C. and X. Shi (2017). Model-selection tests for conditional moment restriction models. *The Econometrics Journal* 20(1), 52–85.
- Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* 17(3), 1217–1241.
- Lehmann, E. and J. Romano (2005). *Testing Statistical Hypotheses*. Springer, New York.
- Ludvigson, S. C. and S. Ng (2009). Macro factors in bond risk premia. *The Review of Financial Studies* 22(12), 5027–5067.
- Massacci, D. (2019). Unstable Diffusion Indexes: With an Application to Bond Risk Premia. *Oxford Bulletin of Economics and Statistics* 81(6), 1376–1400.
- McCracken, M. (2000). Robust out-of-sample inference. *Journal of Econometrics* 99, 195–223.
- McCracken, M. and S. Ng (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics* 34(4), 574–589.
- McCracken, M. W. (2007). Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics* 140(2), 719–752.
- Newey, W. K. and K. D. West (1987). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica (1986-1998)* 55(3), 703–708.
- Perlman, M. and L. Wu (1999). The emperor’s new tests. *Statistical Science* 14(4), 355–381.
- Pitarakis, J.-Y. (2023). A Novel Approach to Predictive Accuracy Testing in Nested Environments. *Econometric Theory (forthcoming)* 0, 1–44.
- Schennach, S. M. and D. Wilhelm (2017). A Simple Parametric Model Selection Test. *Journal of the American Statistical Association* 112(520), 1663–1674.
- Shi, X. (2015). A nondegenerate Vuong test. *Quantitative Economics* 6, 85–121.
- Stock, J. H. and M. W. Watson (2002). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business and Economic Statistics* 20(2), 147–162.
- Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–333.
- West, K. D. (1996). Asymptotic Inference about Predictive Ability. *Econometrica* 64(5), 1067–1084.
- White, H. (2000). A reality check for data snooping. *Econometrica* 68(5), 1097–1126.
- White, H. and I. Domowitz (1984). Nonlinear regression with dependent observations. *Econometrica* 52(1), 143–162.
- Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association* 81(394), 446–451.
- Zhu, Y. and A. Timmermann (2020). Can two forecasts have the same conditional expected accuracy? arxiv:2006.03238v1, arXiv working paper.