School of Economics Working Paper
2016-07

University of East Anglia

SCHOOL OF
ECONOMICS

Model Selection with Factors and Variables

by Jack Fosten*

*School of Economics, University of East Anglia

Abstract

This paper provides consistent information criteria for the selection of forecasting models which use a subset of both the idiosyncratic and common factor components of a big dataset. This hybrid model approach has been explored by recent empirical studies to relax the strictness of pure factor-augmented model approximations, but no formal model selection procedures have been developed. The main difference to previous factor-augmented model selection procedures is that we must account for estimation error in the idiosyncratic component as well as the factors. Our first contribution shows that this combined estimation error vanishes at a slower rate than in the case of pure factor-augmented models in circumstances in which N is of larger order than √T, where N and T are the cross-section and time series dimensions respectively. Under these circumstances we show that existing factor-augmented model selection criteria are inconsistent, and the standard BIC is inconsistent regardless of the relationship between N and T. Our main contribution solves this issue by proposing new information criteria which account for the additional source of estimation error, whose properties are explored through a Monte Carlo simulation study. We conclude with an empirical application to long-horizon exchange rate forecasting using a recently proposed model with country-specific idiosyncratic components from a panel of global exchange rates.

JEL classification codes
C13, C22, C38, C52, C53

Keywords
Forecasting, Factor model, model selection, information criteria, idiosyncratic

School of Economics
University of East Anglia
Norwich Research Park
Norwich NR4 7TJ
United Kingdom
www.uea.ac.uk/economics

Model Selection with Factors and Variables

Jack Fosten[*]
University of East Anglia, UK

March 14, 2016

Abstract

This paper provides consistent information criteria for the selection of forecasting models which use a subset of both the idiosyncratic and common factor components of a big dataset. This hybrid model approach has been explored by recent empirical studies to relax the strictness of pure factor-augmented model approximations, but no formal model selection procedures have been developed. The main difference to previous factor-augmented model selection procedures is that we must account for estimation error in the idiosyncratic component as well as the factors. Our first contribution shows that this combined estimation error vanishes at a slower rate than in the case of pure factor-augmented models in circumstances in which N is of larger order than $\sqrt{T}$, where N and T are the cross-section and time series dimensions respectively. Under these circumstances we show that existing factor-augmented model selection criteria are inconsistent, and the standard BIC is inconsistent regardless of the relationship between N and T. Our main contribution solves this issue by proposing new information criteria which account for the additional source of estimation error, whose properties are explored through a Monte Carlo simulation study. We conclude with an empirical application to long-horizon exchange rate forecasting using a recently proposed model with country-specific idiosyncratic components from a panel of global exchange rates.

JEL Classification: C13, C22, C38, C52, C53

Keywords: Forecasting, Factor model, model selection, information criteria, idiosyncratic

1 Introduction

This paper provides consistent model selection criteria in predictive regressions involving both the common factors and the idiosyncratic components from a big dataset. The modelling environment differs from the standard two-stage "diffusion index" approach of Stock and Watson (2002a,b), which uses only the estimated factors as forecast model regressors and discards all remaining information which is idiosyncratic to each variable. We argue that the pure factor approach may be an excessive approximation in cases where a target variable has a strong relationship with a particular

[*] Correspondence to: School of Economics, University of East Anglia, Norwich, NR4 7TJ, UK. E-mail address: j.fosten@uea.ac.uk. I would like to thank Valentina Corradi, Federico Martellosio, Vasco Gabriel, Jean-Yves Pitarikis and Barbara Rossi for very important comments regarding this paper. For additional thoughts and comments, I would also like to thank participants of the 2015 Winter Meeting of the Econometric Society, IGIER-Università Bocconi and the 24th Symposium of the Society for Nonlinear Dynamics and Econometrics, University of Alabama, Tuscaloosa.

set of variables in the dataset. As an example, we might specify a forecasting model for inflation by exploiting the idiosyncratic variation of a small number unemployment variables as predicted by a Phillips curve model, and also use the factors in order to pick up the 'big data' effect. Hybrid models of this form have been considered as early as Stock and Watson (1999) and can provide benefits in terms of forecast accuracy by performing significant data reduction and limiting the effect of the so-called "curse of dimensionality". Recent empirical studies such as Castle et al. (2013), Luciani (2014) and Engel et al. (2015) have also used this type of hybrid model for macroeconomic forecasting. In the financial econometrics literature, models involving the idiosyncratic component also arise in studies such as Brownlees et al. (2015) which analyse networks in asset returns by using 'de-factored' log-prices obtained by subtracting the common factor component of each asset return.

The objective of this paper is to provide information criteria for model selection in models involving the estimated idiosyncratic component, which has not been addressed in the existing literature to the best of our knowledge. There has, however, been significant progress in research into model selection criteria in pure factor models and pure factor-augmented models. For selecting the number of factors present in a panel of observed variables, the seminal paper of Bai and Ng (2002) showed how to modify standard information criteria when both the cross-section (N) and time series dimension (T) of the dataset grow to infinity. Subsequently, Amengual and Watson (2007), Hallin and Liška (2007) and Onatski (2010) have proposed different methods to tackle the same problem. In choosing the relevant number of factors to use in time-series forecasting regressions, Bai and Ng (2009) proposed a boosting approach to determine the number of autoregressive lags and factors which enter the forecasting model. Other approaches include Groen and Kapetanios (2013), who propose modified Bayesian and Hannan-Quinn type information criteria and Djogbenou (2016) who uses cross-validation. However, none of these approaches are able to deal with the case where both factors and idiosyncratic components are estimated and used in the forecasting model. This yields new challenges which we must address.

The main issue which is new to this paper relative to the literature on pure factor-augmented models, is that we additionally use estimates of the idiosyncratic components as forecast model regressors. This requires new results showing that estimation error in the idiosyncratic component vanishes asymptotically in time series regressions, as both N and T grow to infinity. In the pure factor-augmented case, results regarding factor estimation error are now well known; see Stock and Watson (2002a,b), Bai and Ng (2002, 2006) and Gonçalves and Perron (2014). The first contribution of our paper is to provide the analogue of these results for the estimated idiosyncratic component. We show that the Principal Components estimation error from the idiosyncratic errors vanishes at a rate $\min\left\{\sqrt{T}, N\right\}$, whereas factor estimation error vanishes at a rate $\min\{T, N\}$. Given that we do not require any restrictions on the relationship between N and T for model selection,[1] then in all cases except when T grows faster than N^2, overall estimation error in the regression vanishes

[1] For example, the condition that $\sqrt{T}/N \to 0$ is required by papers such as Bai and Ng (2006), who use it to show that the distribution of the estimated factor-augmented coefficients are unaffected by the presence of factor estimation error. In this paper, we may even employ the assumption of Gonçalves and Perron (2014) that $\sqrt{T}/N \to c$.

at a slower rate than in the pure factor-augmented case. Since it is quite natural for N to be of similar order to T in macroeconomic forecasting, we require new results for model selection in this case where the idiosyncratic regressors are generated as well as the factors.

Our main contribution is to propose new information criteria for model selection which take into account this additional source of estimation error. We specify a general class of information criteria with a penalty function $g(N,T)$ which depends both on N and T. Our main result is a theorem on selection consistency, which shows that consistency only obtains for information criteria with a penalty satisfying the condition $\min\left\{\sqrt{T},N\right\}g(N,T)\to\infty$. This condition is a new finding in the literature and carries several important implications. Firstly, it suggests that we must modify otherwise standard information criteria in order to get consistent model selection. We propose a range of criteria whose properties are equivalent asymptotically but vary in finite samples. Secondly, we find that our result implies that the standard Bayesian Information Criterion (BIC), commonly used in time series applications, is inconsistent for any relative rate of increase between N and T. Thirdly, we find that even recent model selection procedures for the pure factor-augmented case are inconsistent in cases where N is large relative to $\sqrt{T}$. The inconsistency of existing criteria in our set-up is driven by the presence of the estimated idiosyncratic components. We conduct a set of Monte Carlo experiments which demonstrate the improvements of our methods relative to related criteria such as those of Groen and Kapetanios (2013).

We apply our model selection criteria to the challenging but important empirical problem of long-horizon exchange rate forecasting. The recent paper of Engel et al. (2015) proposes a factor-based approach to exchange rate forecasting. They suggest to use the idiosyncratic component from a global dataset of countries' exchange rates as the 'fundamental' in a regression model for a particular country's exchange rate. Therefore their approach precisely matches the modelling framework under which our methods apply. We extend their model to allow for cross-country exchange rate spillovers, modelled by the idiosyncratic components, and use our new information criteria to select between these spillover effects. Our results, applied to a range of OECD countries, show that it is very difficult to out-perform a naïve no-change model; a result mirrored in the majority of existing empirical evidence following the seminal work of Meese and Rogoff (1983b,a). However, we also find that our model selection criteria select a non-trivial number of idiosyncratic effects both in-sample and out-of-sample. This is in contrast to the standard BIC which is not consistent and severely overfits the model by selecting the maximum possible number of variables in many cases.

The rest of the paper is organized as follows. Section 2 provides the forecasting set-up and motivates the use of idiosyncratic components alongside the factors. Section 3 provides results on estimation error in the idiosyncratic components, proposes a class of information criteria for these models and establishes the new conditions required for selection consistency. Section 4 proposes specific functional forms for the penalty function of the information criteria which satisfy the conditions required for consistency, and compares them to existing information criteria. Section 5 provides a Monte Carlo analysis. Section 6 presents the empirical application to exchange rate forecasting and Section 7 concludes the paper.

2 Set-up and Motivation

The broad interest of this study is in predictive regressions for a target variable y_{t+h} when a large set of N candidate predictor variables, X_t, are available. As a statistical device to reduce the dimensionality of the problem, we assume as in Stock and Watson (2002a,b), that the predictors permit the common factor structure:

$$X_t = \Lambda F_t + u_t \tag{1}$$

where F_t is an unobserved $r \times 1$ vector of common factors, Λ is an $N \times r$ matrix of factor loadings and u_t an $N \times 1$ vector of idiosyncratic errors. The unknown factors and loadings can be estimated by methods such as principal components (PCA) as in Stock and Watson (2002a,b) and Bai and Ng (2002, 2006).[2]

The general predictive regression we consider has the following form:

$$y_{t+h} = \beta^{0\prime} F_t^0 + \alpha^{0\prime} u_t^0 + \varepsilon_{t+h} \tag{2}$$

where F_t^0 is an $r^0 \times 1$ subset of the factors, F_t, and u_t^0 is a finite $m^0 \times 1$ subset of the idiosyncratic error vector u_t. The selection of these components is the main aim of this study. The idea is that when $r^0 + m^0 << N$, significant data reduction can improve predictions of y_{t+h} by reducing the excess variability caused by parameter uncertainty.[3]

Forecasting models involving idiosyncratic components, where $\alpha^0 \neq 0$ in Equation (2), are a relatively recent development in the literature, with an increasing number of applications. Luciani (2014) suggests that the use of additional idiosyncratic terms may be useful in forecasting due to the effect of "pervasive shocks" which affect multiple variables. Castle et al. (2013) choose to look at hybrid models involving factors and variables, and use the *Autometrics* routine to select between alternative formulations. Engel et al. (2015) directly specify long-horizon exchange rate forecasting models as a function of the idiosyncratic component from a factor model of international exchange rates. Studies of financial asset networks such as that of Brownlees et al. (2015), mentioned above, also use a similar structure to Equation (2), although they are not explicitly focussed on the use of these equations in forecasting.

On the other hand, the "pure" factor-augmented approach, where $\alpha^0 = 0$ and only the factors are used in forecasting, has been used extensively in the applied forecasting literature. When $\alpha^0 = 0$ and F_t^0 is equal to the full factor vector F_t with no factors omitted, Equation (2) corresponds exactly to the factor-augmented, or "diffusion index" model of Stock and Watson (2002a,b) and Bai and Ng (2006). This model has been widely used in empirical studies; see Stock and Watson (2011)

[2] Other methods, such the maximum likelihood approach of Doz et al. (2011, 2012), could also be used to estimate the factors but we focus on PCA in this paper.

[3] We could also specify Equation (2) to include a set of other 'must-have' non-factor regressors, W_t, as in Bai and Ng (2006) but we omit these here for clarity.

for an overview. When $\alpha^0 = 0$ but F_t^0 is a smaller subset of F_t, only some factors are used in forecasting a given target variable. This type of model was motivated by Boivin and Ng (2006) in the context of real versus nominal macroeconomic factors. As mentioned above, model selection techniques have been proposed for the pure factor-augmented approach by Bai and Ng (2009) and Groen and Kapetanios (2013), but not for models involving estimated idiosyncratic components.

In addition to the empirical motivation for pursuing the model in Equation (2), we can also offer a simple analytical motivation for using this model. The specification of models involving the idiosyncratic terms has been somewhat overlooked in the literature for the two-stage approach of Stock and Watson (2002a,b). We firstly note that it is most common for predictive economic models for y_{t+h} to be be derived between the observed X_t variables. We can then use the factor model from Equation (1) as a way to approximate X_t in obtaining a low-dimensional model. This seems preferable than specifying economic models directly as a function of the factors, F_t, as the factors themselves do not have any direct economic interpretation.

We consider the simple linear data generating process (DGP-X) for y_{t+h} as a function of X_t, which we partition into X_t^0 and X_t^1 which are of dimension $m^0 \times 1$ and $\left(N - m^0\right) \times 1$ respectively, where m^0 is a finite integer as specified above.

(DGP-X)
$$y_{t+h} = \alpha^{0\prime} X_t^0 + \alpha^{1\prime} X_t^1 + e_{t+h} \tag{3}$$

The reason we split X_t into these two components is that it allows the small set of variables X^0 to have a 'large' impact on y_{t+h} through α^0 whereas each element in the high-dimensional vector X^1 has a 'small' impact individually, even though the aggregation of their impacts is non-negligible. This can be related to the above example where the target variable y_{t+h} is inflation, which is predicted strongly by a small set of unemployment-type series, X_t^0, and is lesser affected by the remainder of the big dataset, X_t^1.

Now using the factor model, which we can correspondingly partition into $X_t^0 = \Lambda^0 F_t + u_t^0$ and $X_t^1 = \Lambda^1 F_t + u_t^1$, we can re-write DGP-X in Equation (3) into the alternative formulation:

(DGP-F)
$$y_{t+h} = \beta' F_t + \alpha^{0\prime} u_t^0 + \alpha^{1\prime} u_t^1 + e_{t+h} \tag{4}$$

where $\beta' = \left(\alpha^{0\prime}\Lambda^0 + \alpha^{1\prime}\Lambda^1\right)$. In the pure factor-augmented case, the methods of Stock and Watson (2002a,b) have an overall regression error term $v_{t+h} = \alpha^{0\prime} u_t^0 + \alpha^{1\prime} u_t^1 + e_{t+h}$ and do not explicitly include u_t^0 in the model. Note that both DGP-X and DGP-F are high dimensional as $N \to \infty$ as they have a number of variables of order N. As noted by Bai and Ng (2009), it is unwise to form predictions using models involving all variables as the mean squared prediction error can be seen to increase in N due to the effect of parameter estimation.

However, it may be preferable to obtain forecasts using small-scale models involving the predictors F_t and/or u_t^0, which are both finite dimensional. We therefore use Equation (2) in making h-step ahead forecasts. Assuming for simplicity that F_t and u_t^0 are known, for the pure factor model case we have $\mathrm{E}\left(y_{T+h} | F_T\right) = \widehat{\beta}' F_T$, whereas for the model including the idiosyncratic components we have $\mathrm{E}\left(y_{T+h}|F_T, u_T^0\right) = \widehat{\beta}' F_T + \widehat{\alpha}' u_T^0$. It can be shown that the difference in mean squared forecast error (MSFE) between the two models is:

$$MSFE\left(y_{T+h}|F_T\right) - MSFE\left(y_{T+h}|F_T, u_T^0\right) = \\ \mathrm{E}\left[\alpha^{0\prime} u_T^0 u_T^{0\prime} \alpha^0\right] - \mathrm{E}\left[u_T^{0\prime}\left(\widehat{\alpha}^0 - \alpha^0\right)\left(\widehat{\alpha}^0 - \alpha^0\right)' u_T^0\right] \tag{5}$$

which can be interpreted as the trade-off between omitting u_t^0 and incurring an unavoidable loss in population, or including u_t^0 in the model and incurring additional parameter estimation error. To give a simple analytical example of this expression, we note as in Bai and Ng (2009) that $T\left(\widehat{\alpha}^0 - \alpha^0\right)\left(\widehat{\alpha}^0 - \alpha^0\right)'$ is $\chi^2_{m^0}$ with expectation m^0 and, if we assume that the idiosyncratic errors are homoskedastic and cross-sectionally uncorrelated with variance σ_u^2 and if the regression error e_{t+h} has variance σ_e^2, then this expression is approximately equal to:

$$\sigma_u^2 \alpha^{0\prime} \alpha^0 - \sigma_e^2 \frac{m^0}{T}$$

Therefore, in cases where α^0 is 'large', for example if $\alpha_i^0 = O(1)$ for $i = 1, ..., m^0$, then we expect that using the model with idiosyncratic components dominates the pure factor-augmented model in terms of $MSFE$. On the other hand, if α^0 is 'small' or local-to-zero such as $\alpha_i^0 = O\left(N^{-1/2}\right)$ or even $\alpha_i^0 = 0$, then there may be no gains in using the additional idiosyncratic components.

Motivated by the empirical and analytical reasons for using the model in Equation (2), the rest of this paper provides results on the optimal selection of the factors and idiosyncratic error terms which enter these types of regression. To the best of our knowledge this is the first paper to look at model selection in this context.

3 Model Estimation and Selection

3.1 Model Estimation

This study differs from previous work as we additionally have to estimate u_t^0 in Equation (2) as well as F_t^0. However, to the best of our knowledge there are no formal results in the literature on time series regression involving estimated idiosyncratic errors. Using Principal Components Analysis (PCA) estimation as in Stock and Watson (2002a,b) gives the following feasible analogue to the model in Equation (2):

$$y_{t+h} = \beta^{0\prime} \widehat{F}_t^0 + \alpha^{0\prime} \widehat{u}_t^{0\prime} + \varepsilon_{t+h} \tag{6}$$

where $\widehat{F}_t^0 \subseteq \widehat{F}_t$, and the $T \times r$ matrix $\widehat{F}$ consists of the r eigenvectors which correspond to the r largest eigenvalues of the $T \times T$ matrix XX', under the identifying normalization $\widehat{F}'\widehat{F}/T = I_r$. This yields the factor loading estimate $\widehat{\Lambda} = X'\widehat{F}/T$. Using both the factor estimates and the factor loadings estimates, along with Equation (1) yields an estimator for each variable i in $\widehat{u}_t^0 \subseteq \widehat{u}_t$ which is equal to:

$$\widehat{u}_{it} = X_{it} - \widehat{\lambda}_i' \widehat{F}_t \tag{7}$$

where $\widehat{\lambda}_i$ corresponds to the ith row of the estimated loadings matrix $\widehat{\Lambda}$. This gives the idiosyncratic part of each variable which is orthogonal to the factors. In papers such as Brownlees et al. (2015), $\widehat{u}_{it}$ represents the 'de-factored' part of the financial asset return X_{it}.

The first result we require is to analyse the estimation error $(\widehat{u}_{it} - u_{it})$ resulting from the time series regression in Equation (6). When OLS is used to estimate β^0 and α^0, this requires us to show that $(\widehat{u}_{it} - u_{it})$ has an asymptotically negligible covariance with the components of y_{t+h}, namely F_t^0, u_t^0 and ε_{t+h}. This is in a similar way to Bai and Ng (2006) and Gonçalves and Perron (2014) who show results regarding the negligibility of factor estimation error $\left(\widehat{F}_t - HF_t\right)$ in time series regression.

The crucial difference in our study is to note that, since $\widehat{u}_{it}$ involves the estimated common component $\widehat{\lambda}_i'\widehat{F}_t$, the estimation error term is:

$$\widehat{u}_{it} - u_{it} = \lambda_i' F_t - \widehat{\lambda}_i' \widehat{F}_t$$

which includes terms in both factor estimation error and factor loading estimation error. We require the following assumptions to analyse the asymptotic properties of this estimation error term.

Assumption 1: *The factors satisfy* $E\|F_t\|^4 \leq M$ *and* $\frac{1}{T}\sum_{t=1}^{T} F_t F_t' \underset{p}{\rightarrow} \Sigma_F > 0$ *as* $T \rightarrow \infty$.

Assumption 2: *The factor loadings are either deterministic such that* $\|\lambda_i\| \leq M$, *or stochastic such that* $E\|\lambda_i\|^4 \leq M$, *and they satisfy* $\frac{1}{N}\Lambda'\Lambda \underset{p}{\rightarrow} \Sigma_\Lambda > 0$ *as* $N \rightarrow \infty$.

Assumption 3: *The idiosyncratic errors are such that: (i)* $E(u_{it}) = 0$ *and* $E|u_{it}|^8 \leq M$ *for any* i *and* t, *(ii)* $E(u_{it}u_{js}) = \sigma_{ij,ts}$ *with* $|\sigma_{ij,ts}| < \bar{\sigma}_{ij}$ *for all* t *and* s, *and* $|\sigma_{ij,ts}| < \bar{\tau}_{ts}$ *for any* i *and* j, *and* $\frac{1}{N}\sum_{i,j=1}^{N}\bar{\sigma}_{ij} \leq M$, $\frac{1}{T}\sum_{s,t=1}^{T}\bar{\tau}_{ij} \leq M$ *and* $\frac{1}{NT}\sum_{i,j,t,s=1}\bar{\sigma}_{ij,ts} \leq M$, *(iii)* $E|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(u_{is}u_{it} - E(u_{is}u_{it}))|^4 \leq M$ *for all* t *and* s.

Assumption 4: *The variables* $\{\lambda_i\}$, $\{F_t\}$ *and* $\{u_{it}\}$ *are three mutually independent groups, although dependence within each group is allowed.*

Assumption 5: *The regression errors satisfy: (i)* $E(\varepsilon_{t+h}|\mathcal{F}_t) = 0$ *for* $h > 0$ *where* $\mathcal{F}_t = \sigma(y_t, F_t, u_t, y_{t-1}, F_{t-1}, u_{t-1}, ...)$ *and* $E\left(\varepsilon_{t+h}^4\right) \leq M$, *(ii)* $E|\frac{1}{\sqrt{TN}}\sum_{t=1}^{T}\sum_{i=1}^{N}\varepsilon_{t+h}(u_{is}u_{it} - E(u_{is}u_{it}))|^2 \leq M$ *for each* s *and for* $h > 0$, *(iii)* $E\|\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{i=1}^{N}\lambda_i u_{it}\varepsilon_{t+h}\|^2 \leq M$ *with* $E(\lambda_i u_{it}\varepsilon_{t+h}) = 0$ *for any* i *and* t, *and (iv)* $\frac{1}{\sqrt{T}}Z_t\varepsilon_{t+h} \underset{d}{\rightarrow} N(0, \Omega)$ *where* $Z_t = \left[F_t^0, u_t^0\right]$.

Assumptions 1-3 are standard in the literature of factor-augmented forecasting models for ensuring the existence of r factors, and allowing for heteroskedasticity and limited dependence in the idiosyncratic errors. These assumptions coincide with those of Bai and Ng (2006), with the exception of their Assumption C4 which they require to derive the asymptotic distribution of the factors, which is not required in this paper. Assumption 4 requires mutual independence of λ_i, F_t and u_{it} as in Bai and Ng (2006), though can be readily replaced with weaker assumptions along the lines of Gonçalves and Perron (2014). Finally, Assumption 5 follows Cheng and Hansen (2015) in placing standard moment restriction on the factor-augmented model, in this case modified to include the idiosyncratic components.

Under these assumptions, we can show the following Lemma concerning the covariance of $(\widehat{u}_{it} - u_{it})$ with the factors:

Lemma 1 *Let Assumptions 1-5 hold and let the factors and factor loadings be estimated by Principal Components. Then as $N, T \to \infty$,*

$$\frac{1}{T}\sum_{t=1}^{T} F_t \left(\widehat{u}_{it} - u_{it}\right) = O_p\left(\max\left\{\frac{1}{N}, \frac{1}{\sqrt{T}}\right\}\right)$$

The proof of Lemma 1 is shown in the Appendix along with two additional Lemmas.[4] The key difference of this result relative to existing results regarding the factor estimation error term in the case of pure factor-augmented regressions of Bai and Ng (2006), is that the estimation error term here vanishes at a rate $\min\left\{\sqrt{T}, N\right\}$ and not $\min\left\{T, N\right\}$. The intuition behind this result is that Bai (2003) shows that estimated factor loadings $\widehat{\lambda}_i$ are consistent at rate $\min\left\{\sqrt{T}, N\right\}$. Since the estimated idiosyncratic error terms in Equation (7) depend on $\widehat{\lambda}_i$, our result shows that the asymptotic rate for $\widehat{\lambda}_i$ carries over to time series regressions involving $\widehat{u}_{it}$.

The main implication of this result is that there are non-trivial cases regarding the relationship between N and T for which the overall estimation error in regressions involving the idiosyncratic component vanishes more slowly when compared to the pure factor-augmented case of Bai and Ng (2006). Since the rate given in Lemma 1 is $\min\left\{\sqrt{T}, N\right\}$, then in cases where $\sqrt{T}$ is of smaller order than N, in other words $T^{1/2} = N^{1-\varepsilon}$ for some $\varepsilon > 0$, then estimation error vanishes at a slower rate than the pure factor-augmented rate, $\min\left\{T, N\right\}$. This can affect many of the kinds of datasets used in applied economic forecasting exercises. For example, in the simple case where $T = N$ as $N, T \to \infty$, then estimation error vanishes more slowly than in the pure factor-augmented case. The situation where $T \approx N$ is relatively standard in many macroeconomic datasets based on monthly or quarterly data, such as that of Stock and Watson (2002a,b). In fact, some studies such as Eickmeier and Ng (2011) use a vast amount of variables from international databases, with a small number of quarterly observations. In which case, it may even be reasonable to assume N is of even larger order than T, meaning that the difference in estimation error relative to the pure factor-augmented case is expected to be yet bigger.[5]

Given that model selection techniques depend on the rate of consistency of the generated regressors, and since we have shown that the estimation of the idiosyncratic component may give a

[4] The covariance of the estimation error with u_{it} and ε_{t+h} are shown in the Appendix to vanish at the faster rate $\min\left\{N, T\right\}$, and are therefore not repeated in the text.

[5] This is in contrast to high-frequency financial data, where we may expect that N is of very small order relative to T. In this case, then the estimation error will be of the same order when the idiosyncratic errors are also generated and used as regressors.

different consistency rate to the pure factor-augmented case, then we require new results in order to obtain consistent model selection. In the next section we propose a general class of information criteria and show how the result of Lemma 1 is used to provide consistent model selection.

3.2 Model Selection

In this section we propose a class of information criteria for an empirical researcher to select between alternative specifications of Equation (2). We will refer to a model specification i which uses the variables $\left(\widehat{F}_t^i, \widehat{u}_t^i\right)$ containing r_i estimated factors and m_i estimated idiosyncratic errors. This combination yields the regression:

$$y_{t+h} = \beta^{i\prime}\widehat{F}_t^i + \alpha^{i\prime}\widehat{u}_t^i + \varepsilon_{t+h}^i \tag{8}$$

We wish to allow for full flexibility of model selection by searching over all possible combinations of factors and variables. In the case of the factors, since r is typically small and can be consistently estimated by information criteria of Bai and Ng (2002), we can search over every combination of the factors. On the other hand, in selecting variables from u_t , 2^N combinations may not be practical in terms of computational feasibility and, as in standard model selection procedures, we must first reduce the size of this model space over which we search using information criteria. In other words, we must choose a subset of $m^{max} << N$ variables to search over.

The choice of this candidate set is a decision for the individual researcher. For example, we could be guided by economic theory as in the above example of Stock and Watson (1999) who look at inflation models involving both factors and employment-type series. Therefore we could perform model selection over all r factors, and the subset of m^{max} idiosyncratic components relating to the employment and unemployment series in the dataset. A more general way of generating a candidate subset would be to use a device such as Forward Stepwise (FS) regression and search only over the first m^{max} variables from this procedure. Forward Stepwise methods and other sequential methods are surveyed extensively in the chapter of Ng (2013). The principle is to start out with an empty model and add one variable at a time in a way which maximises the fit of the regression. We do not claim to have an optimality result for this type of procedure in this paper, and acknowledge that this has certain shortcomings, as mentioned by Ng (2013). However, since we use this method only as a way to generate a candidate search set we expect it to perform relatively well, and this expectation is confirmed by Monte Carlo simulation. Future work may address this issue using penalized regression techniques such as LASSO, but this is outside the scope of the current paper.

The class of information criteria we use has a penalty function which depends on both the sample size T and the number of variables N. The difference in this paper with respect to other studies is that the information criterion is a function of estimated idiosyncratic errors as well as factors, which will impact the requirements on the penalty function for consistent model selection.

For the model i, the criterion depends on the number of variables, r_i and m_i, and the estimated sum of squared residuals in model (8):

$$IC\left(\widehat{F}^i, \widehat{u}^i\right) = \ln\left(V\left(\widehat{F}^i, \widehat{u}^i\right)\right) + (r_i + m_i)\, g(N, T)$$

where:

$$V\left(\widehat{F}^i, \widehat{u}^i\right) = \frac{1}{T}\sum_{t=1}^{T}\left(y_{t+h} - \widehat{\beta}^{i\prime}\widehat{F}_t^i - \widehat{\alpha}^{i\prime}\widehat{u}_t^i\right)^2$$

Selection consistency occurs when the probability limits of the generated regressors span the same space as the true factors and idiosyncratic errors F_t^0 and u_t^0. For the factors, it is well known that the Principal Components estimates converge to a particular rotation of the true factors $H^0 F_t^0$, as shown by Bai and Ng (2002).[6] On the other hand, the estimates of the idiosyncratic errors $\widehat{u}_t^i$ are consistent for the true u_t^i without rotation.

The following Theorem shows the conditions on $g(N, T)$ required for consistency of selection:

Theorem: *Let Assumptions 1-5 hold and let the factors and factor loadings be estimated by Principal Components. For two models i and j, if model i corresponds to the true model such that the probability limit of $\left(\widehat{F}^i, \widehat{u}^i\right)$ is $\left(H^0 F_t^0, u_t^0\right)$ for all t, and for model j one or both of $\widehat{F}_t^j$ and $\widehat{u}_t^j$ has different probability limit, then:*

$$\lim_{N,T\to\infty} \Pr\left(IC\left(\widehat{F}^j, \widehat{u}^j\right) < IC\left(\widehat{F}^i, \widehat{u}^i\right)\right) = 0$$

as long as (i) $g(N, T) \to 0$ and (ii) $\min\left\{\sqrt{T}, N\right\} g(N, T) \to \infty$ as $N, T \to \infty$.

The key difference of this result relative to previous information criteria is that Condition (ii) requires that $\min\left\{\sqrt{T}, N\right\} g(N, T) \to \infty$ rather than $\min\{T, N\}\, g(N, T) \to \infty$, which was the condition required in pure factor model studies such as Bai and Ng (2002) and Groen and Kapetanios (2013).[7] This comes as a direct consequence of Lemma 1, and the rate at which the penalty function vanishes to zero reflects the rate in Lemma 1 in incorporating the additional source of estimation error in $\widehat{u}^i$.

The main implication of this result is that many information criteria used in previous studies do not meet the requirements in Condition (ii) for consistent model selection, and should therefore not be used in specifying models which take this form. In particular, even model selection criteria which are modified to allow for factor estimation error are inconsistent as they only take the $\min\{T, N\}$ consistency rate of the factors into account. In papers such as Groen and Kapetanios (2013), who provide information criteria for pure factor-augmented models, their penalty function does not satisfy Condition (ii) of this paper which leads to inconsistent model selection.

[6] We could use alternative methods such as the Maximum Likelihood approach of Doz et al. (2011, 2012). As in Groen and Kapetanios (2013), this will change Conditions (i) and (ii) in the Theorem below, depending on the rate at which factor and idiosyncratic estimation error vanishes in those methods.

[7] Condition (i) is standard for information criteria in ensuring that inflation in $V(.)$ due to incorrect model specification is always larger than the penalty $g(N, T)$. We therefore do not discuss this condition further.

In the next section we make this point clear by first proposing a new set of information criteria which satisfy Conditions (i) and (ii). We then look at existing information criteria and show that they do not meet this condition and are consequently inconsistent.

4 Information Criteria

In suggesting different functional forms for the penalty function $g(N,T)$, we propose several different information criteria based both on a Mallows-type form and a Hannan-Quinn form. The penalty functions we propose make use of the fact that $\max\left\{\frac{1}{\sqrt{T}},\frac{1}{N}\right\}\approx\frac{\sqrt{T}+N}{\sqrt{T}N}$. Since we ensure that all of the criteria satisfy both Conditions (i) and (ii) of the above Theorem, they are all consistent and therefore equivalent asymptotically. However, their performance may differ in finite samples. These finite sample properties will be assessed later through Monte Carlo simulations.

The new information criteria are as follows:

$$IC_1\left(\widehat{F}^i,\widehat{u}^i\right)=\ln\left(V\left(\widehat{F}^i,\widehat{u}^i\right)\right)+(r_i+m_i)\ln\left(\frac{\sqrt{T}N}{\sqrt{T}+N}\right)\left(\frac{\sqrt{T}+N}{\sqrt{T}N}\right)$$

$$IC_2\left(\widehat{F}^i,\widehat{u}^i\right)=\ln\left(V\left(\widehat{F}^i,\widehat{u}^i\right)\right)+(r_i+m_i)\ln\left(\min\left\{\sqrt{T},N\right\}\right)\left(\frac{\sqrt{T}+N}{\sqrt{T}N}\right)$$

$$IC_3\left(\widehat{F}^i,\widehat{u}^i\right)=\ln\left(V\left(\widehat{F}^i,\widehat{u}^i\right)\right)+(r_i+m_i)\frac{\ln\left(\min\left\{\sqrt{T},N\right\}\right)}{\min\left\{\sqrt{T},N\right\}}$$

$$HQ_1\left(\widehat{F}^i,\widehat{u}^i\right)=\ln\left(V\left(\widehat{F}^i,\widehat{u}^i\right)\right)+2(r_i+m_i)\ln\ln\left(\frac{\sqrt{T}N}{\sqrt{T}+N}\right)\left(\frac{\sqrt{T}+N}{\sqrt{T}N}\right)$$

$$HQ_2\left(\widehat{F}^i,\widehat{u}^i\right)=\ln\left(V\left(\widehat{F}^i,\widehat{u}^i\right)\right)+2(r_i+m_i)\ln\ln\left(\min\left\{\sqrt{T},N\right\}\right)\left(\frac{\sqrt{T}+N}{\sqrt{T}N}\right)$$

$$HQ_3\left(\widehat{F}^i,\widehat{u}^i\right)=\ln\left(V\left(\widehat{F}^i,\widehat{u}^i\right)\right)+2(r_i+m_i)\frac{\ln\ln\left(\min\left\{\sqrt{T},N\right\}\right)}{\min\left\{\sqrt{T},N\right\}}$$

These criteria clearly satisfy both Conditions (i) and (ii) of the Theorem above. In comparing these criteria to previous literature, we firstly note that the IC_1, IC_2 and IC_3 criteria are similar in nature to those in Bai and Ng (2002), though since their criteria were for selecting the number of factors in the whole panel of variables, we do not compare our results to theirs. The Hannan-Quinn criteria HQ_1, HQ_2 and HQ_3 are similar in spirit to those in Groen and Kapetanios (2013). As such, we will now discuss the differences of our criteria to those, with particular reference to the conditions for selection consistency shown in the Theorem of the previous section.

The $BICM$ and $HQICM$ criteria of Groen and Kapetanios (2013)[8] were proposed for the case of pure factor-augmented model selection, with no estimated idiosyncratic components. Their

[8] In their paper they scale the criteria by T whereas in this paper we follow more closely the specifications of Bai and Ng (2002).

information criteria are as follows:

$$BICM\left(\widehat{F}^i,\widehat{u}^i\right)=\ln\left(V\left(\widehat{F}^i,\widehat{u}^i\right)\right)+\left(r_i+m_i\right)\ln\left(T\right)\left(\frac{T+N}{TN}\right)$$
$$HQICM\left(\widehat{F}^i,\widehat{u}^i\right)=\ln\left(V\left(\widehat{F}^i,\widehat{u}^i\right)\right)+2\left(r_i+m_i\right)\ln\ln\left(T\right)\left(\frac{T+N}{TN}\right)$$

While both of these criteria have $g\left(N,T\right)\to 0$ and therefore pass our Condition (i), they both fail Condition (ii) in cases where $\sqrt{T}$ is of smaller order than N, which includes the simple example given above where $N=T$ as $N,T\to\infty$. This means that in finite samples, panels where N is large relative to T, we expect our proposed criteria to provide significant improvements over both of the methods proposed by Groen and Kapetanios (2013).

It is also useful to compare our criteria to the standard AIC and BIC criteria which are commonly used in time series applications:

$$AIC\left(\widehat{F}^i,\widehat{u}^i\right)=\ln\left(V\left(\widehat{F}^i,\widehat{u}^i\right)\right)+\left(r_i+m_i\right)\frac{2}{T}$$
$$BIC\left(\widehat{F}^i,\widehat{u}^i\right)=\ln\left(V\left(\widehat{F}^i,\widehat{u}^i\right)\right)+\left(r_i+m_i\right)\frac{\ln\left(T\right)}{T}$$

Both the AIC and the BIC also fulfil Condition (i) as $g\left(N,T\right)\to 0$ but they both fail Condition (ii) for all configurations of N and T and are therefore inconsistent. This result is somewhat alarming as the vast majority of empirical forecasting studies tend to use the BIC for model selection. In our simulations, we therefore expect there to be overparameterization in all Monte Carlo specifications of N and T for these criteria. This result is unusual relative to the pure factor-augmented approach of Groen and Kapetanios (2013) as in that set-up the BIC is still consistent in cases where $N<<T$. In our case, the presence of the estimated idiosyncratic components drives inconsistency in the BIC, and we must use the modified the penalty functions proposed above.

5 Monte Carlo

In this section we provide Monte Carlo simulation evidence to support the Theorem above regarding selection consistency of the proposed information criteria. We use a similar Monte Carlo approach to those of Bai and Ng (2009) and Groen and Kapetanios (2013), modified to the case where a small number of idiosyncratic components are important predictors for the variable y_t.

5.1 Data Generating Process

Here we specify the Data Generating Processes (DGPs) for the factor model and the forecasting model. We follow the structure proposed in Section 2, where the high-dimensional DGP-X in Equation (3) is used for y_{t+h} along with the low-dimensional approximating model in Equation (2), and the factor model in Equation (1) is used for X_t. The $N\times 1$ vector of variables X_t is split into the sub-vectors X_t^0 and X_t^1 of dimension $m^0 \times 1$ and $\left(N - m^0\right) \times 1$ respectively. These two vectors of variables affect y_t differently through the 'large' coefficients α^0 and the 'small' coefficients α^1 in the following linear data generating process:

$$y_t = \alpha^{0\prime} X_t^0 + \alpha^{1\prime} X_t^1 + \sqrt{\theta}\varepsilon_t \tag{9}$$

$$X_t = \frac{1}{\sqrt{r}} \Lambda F_t + u_t \tag{10}$$

The main difference in our study is that we allow for cases in which both F_t and the first m^0 idiosyncratic components, u_t^0, have non-negligible explanatory power for y_t, as in Equation (2). This is ensured by making α^0 significantly larger than α^1, reflecting the motivation that some variables may have stronger predictive power for y_t, while the combination of the remaining variables has predictive power only through the factors. We do this by letting $\alpha^0 = \mathbf{1}_{m^0 \times 1}$ and $\alpha^1 = \mathbf{1}_{(N-m^0)\times 1}/\sqrt{N - m^0}$. This means that when $N = 50$ and $m^0 = 1$ the coefficient on the first variable is 7 times the size of the remaining coefficients.

The regression errors ε_t are drawn such that $\varepsilon_t \sim i.i.d.N\left(0, 1\right)$. The idiosyncratic errors are also drawn from a normal distribution, but the variance differs between the first m^0 variables and the remaining variables, with $u_{it} \sim i.i.d.N\left(0, K\right)$ for $i \leq m^0$ and $u_{it} \sim i.i.d.N\left(0, 1\right)$ for $i > m^0$. Finally, the factor loadings are drawn with a non-zero mean such that $\Lambda \sim i.i.d.N\left(1, 1\right)$ and the factors are $F_t \sim i.i.d.N\left(0, 1\right)$. The rescaling by $1/\sqrt{r}$ in Equation (10) ensures that the variance of X_{it} is K for $i \leq m^0$ and 1 for $i > m^0$.

We then choose the parameters K and θ to fix the signal to noise ratio in Equation (9), and also to equate the signal to noise ratio between F_t and u_t^0. In order to do this, note that we can combine Equations (9) and (10) to give the alternative expression as in DGP-F of Equation (4):

$$y_t = \underbrace{\frac{1}{\sqrt{r}} \left(\alpha^{0\prime}\Lambda^0 + \alpha^{1\prime}\Lambda^1\right) F_t + \alpha^{0\prime} u_t^0}_{\text{Explained}} + \underbrace{\left(\alpha^{1\prime} u_t^1 + \sqrt{\theta}\varepsilon_t\right)}_{\text{Unexplained}}$$

The aim of model selection is to include the first m^0 idiosyncratic components in the model, therefore only the remaining errors u_t^1 enter the unexplained part of the regression. This is unlike in Groen and Kapetanios (2013) where the unexplained part of the pure factor-augmented model contains all N idiosyncratic errors. Given the distributions of Λ, F, u and ε, the overall regression R^2 is:

$$\begin{aligned} R^2 &= 1 - \frac{\alpha^{1\prime}\alpha^1 + \theta}{\alpha^{0\prime}\alpha^0 + \alpha^{1\prime}\alpha^1 + K\alpha^{0\prime}\alpha^0 + \alpha^{1\prime}\alpha^1 + \theta} \\ &= 1 - \frac{1 + \theta}{m^0 + 1 + Km^0 + 1 + \theta} \end{aligned}$$

We therefore set $K = \left(m^0 + 1\right)/m^0$ to equate the variation of y_t explained by F_t and u_t^0 and we set θ so that $R^2 = 1/2$ which requires that $\theta = 1 + 2m^0$.

For the sample sizes, we consider a fixed set of values for the time series dimension corresponding to $T = 50, 100, 200, 400$. For the cross section dimension we consider different asymptotic rules corresponding to $N = c_1T$ and $N = c_2\sqrt{T}$. These two particular asymptotic set-ups are chosen because, in the case where $N = c_1T$, as mentioned in Section 4, we expect consistent model selection only for the new information criteria proposed in this paper, although in finite samples the pure factor augmented model criteria of Groen and Kapetanios (2013) may perform reasonably when N is small. When $N = c_2\sqrt{T}$, we expect inconsistency of the AIC and the BIC, though the criteria of Groen and Kapetanios (2013) should be consistent along with those suggested in this paper.

We therefore select 2 different levels of c_1 and c_2 which fix the level of N equal to 20 and 50 when $T = 50$. We call these the "Small N" and "Large N" scenarios:

Table 1: Scenarios for Sample Sizes of T and N

	"Small N"		"Large N"	
Scenario:	1	2	3	4
T	$N = 0.4T$	$N = 2.83\sqrt{T}$	$N = T$	$N = 7.07\sqrt{T}$
50	20	20	50	50
100	40	28	100	71
200	80	40	200	100
400	160	57	400	141

Notes: The coefficients multiplying T and $\sqrt{T}$ in the expressions for N are set so that $N = 20, 50$ when $T = 20$ in the "Small" and "Large" scenarios. Note that $2.83 \approx 20/\sqrt{50}$ and $7.07 \approx 50/\sqrt{50}$.

These combinations of N and T are typical of many of the types of dataset commonly used in applied forecasting studies. Further results for other combinations of sample size are available upon request.

We are interested in two aspects of the results: that the correct number of variables are selected, and that the identity of their selection is correct, both for factors and idiosyncratic errors. From the results in the Theorem above we expect that the AIC and BIC criteria overestimate r^0 and m^0 for all combinations of T and N and that the $MBIC$ and $HQICM$ of Groen and Kapetanios (2013) overestimate in cases where $T << N$.

In order to assess the number of selected variables, we will use the average number of selected idiosyncratic components and factors, $\widehat{m}^0$ and $\widehat{r}$, over $B = 1000$ Monte Carlo replications. To assess variable selection, we use a mean squared deviation (MSD) statistic. If we denote $\widehat{S}^F$ and $\widehat{S}^u$ as the $r \times 1$ and $N \times 1$ binary selection vectors of 1's and 0's which correspond to the minimization of a given criterion, and S_0^F and S_0^u are the true inclusion vectors according to the data generating process in Equation (9) then the MSD statistics for F and u are:

$$MSD^F = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{S}_b^F - S_0^F\right)'\left(\widehat{S}_b^F - S_0^F\right) \tag{11}$$

$$MSD^u = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{S}_b^u - S_0^u\right)'\left(\widehat{S}_b^u - S_0^u\right) \tag{12}$$

where b indexes the Monte Carlo replication. Values of MSD equal to zero therefore represent perfect model selection, whereas large values of MSD represent poor model selection.

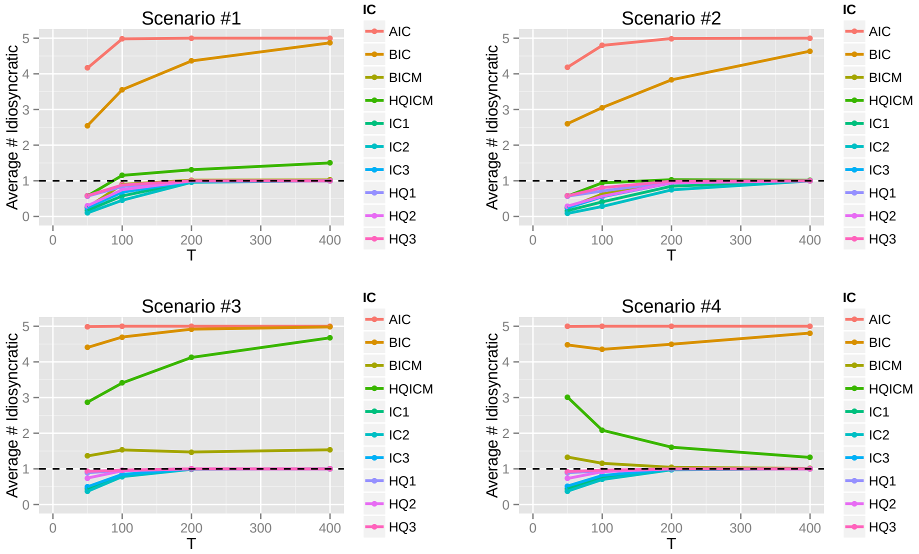
In the following section we document the results for the scenario where $r = 1$ and we let $m^0 = 1, 2$. Results for other specifications are available upon request. We search over a maximum possible set of $r^{max} = 5$ factors and $m^{max} = 5$ idiosyncratic errors. The largest model we consider in the search procedure therefore contains 10 variables, and we also search over potentially redundant factors. The candidate set of variables in $\widehat{u}_t$ is selected using Forward Stepwise regression, as mentioned above.

5.2 Results

Figures 1 to 4 show the number of chosen idiosyncratic errors, $\widehat{m}^0$, and factors, $\widehat{r}$, averaged across the 1,000 Monte Carlo replications for the 2 specifications of $r = 1$ and $m^0 = 1, 2$. Figures 5 to 8 in the Appendix also show the corresponding Mean Squared Deviation statistics MSD^u and MSD^F for the same configurations.

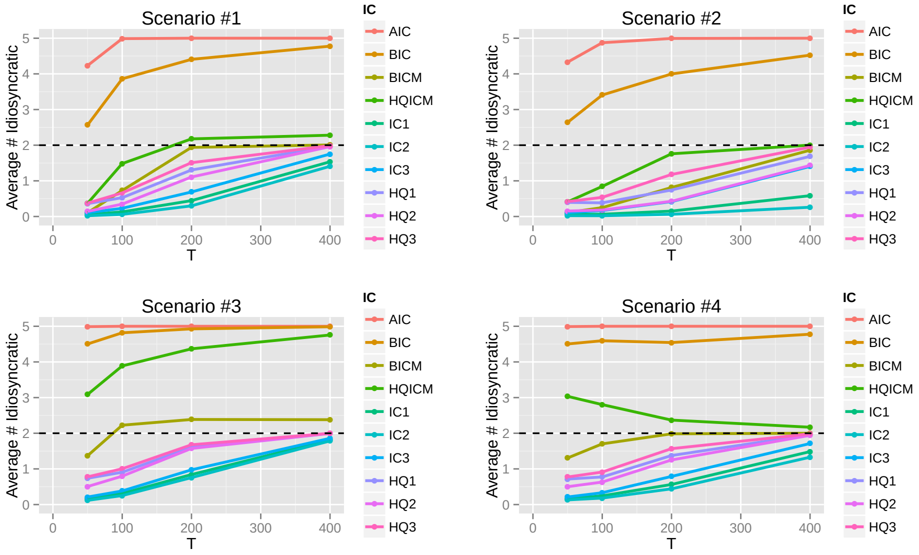
The results clearly demonstrate the selection consistency of the newly proposed information criteria IC_1, IC_2, IC_3, HQ_1, HQ_2 and HQ_3. There are a few key features of the results to highlight

Figure 1: Average number of selected idiosyncratic components ($\widehat{m}^0$) for different information criteria over 1,000 Monte Carlo replications when the true $r = 1$ and $m^0 = 1$



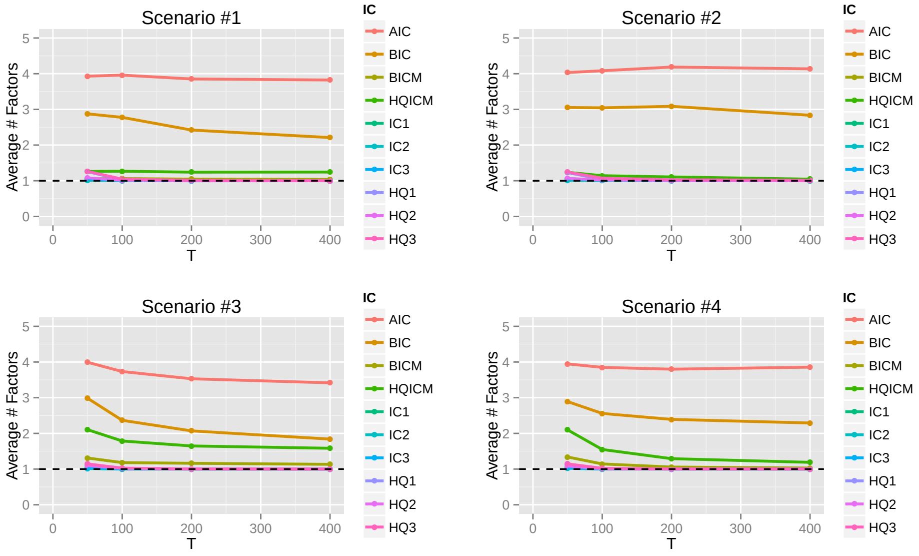
Notes: The horizontal dashed line represents the true number of idiosyncratic components $m^0 = 1$. Scenarios 1-4 are described in Section 5.1 and each of the information criteria are described in Section 4.

Figure 2: Average number of selected idiosyncratic components ($\widehat{m}^0$) for different information criteria over 1,000 Monte Carlo replications when the true $r = 1$ and $m^0 = 2$



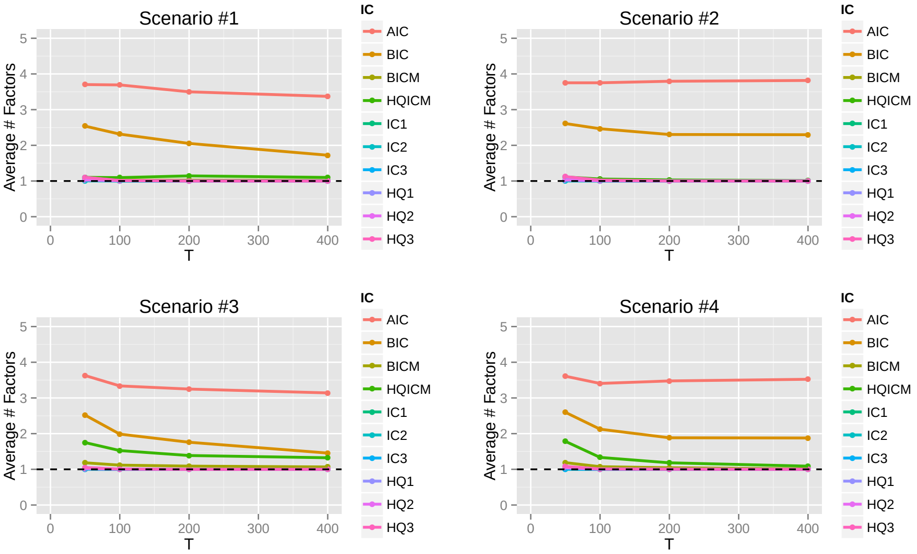
Notes: The horizontal dashed line represents the true number of idiosyncratic components $m^0 = 2$. Scenarios 1-4 are described in Section 5.1 and each of the information criteria are described in Section 4.

Figure 3: Average number of selected factors ($\widehat{r}$) for different information criteria over 1,000 Monte Carlo replications when the true $r = 1$ and $m^0 = 1$



Notes: The horizontal dashed line represents the true number of factors $r = 1$. Scenarios 1-4 are described in Section 5.1 and each of the information criteria are described in Section 4.

Figure 4: Average number of selected factors $(\widehat{r})$ for different information criteria over 1,000 Monte Carlo replications when the true $r = 1$ and $m^0 = 2$



Notes: The horizontal dashed line represents the true number of factors $r = 1$. Scenarios 1-4 are described in Section 5.1 and each of the information criteria are described in Section 4.

in particular. Firstly, consistency occurs over all different combinations of N and T considered in Scenarios 1 through 4. This is unlike the cases of the AIC and BIC which display inconsistent model selection in all configurations, and the $BICM$ and $HQICM$ criteria of Groen and Kapetanios (2013), where overfitting occurs in the cases where N is large relative to T. Secondly, while the new criteria behave the same way asymptotically, it appears from these results that the Hannan-Quinn type criteria perform better than IC_1, IC_2 and IC_3, with HQ_3 having the best finite sample performance with MSD figures closest to 0 in Figures 5 to 8. Thirdly, all six new criteria have particularly strong selective power with respect to the factors, even for very low T and N. This is in contrast to the $BICM$ and $HQICM$ criteria when N is large. Finally, we see that as the number of idiosyncratic terms m^0 increases from 1 to 2, selection worsens and the criteria become more conservative. This is another fairly standard result, although the HQ_3 criteria still appears to perform well when the panel dimensions are reasonable.

Turning to the other information criteria, as mentioned above, we can see that both the AIC and BIC in Figures 1 to 4 overfit the model for all values of N and T, both for the selection of factors and variables. In many cases the number of selected factors and variables is close to the maximum number considered in the search procedure which is 5. This result is unsurprising for the AIC which is known to be inconsistent in all standard set-ups. The result for the BIC, as mentioned before, is somewhat unusual as in the pure-factor augmented set-up of Bai and Ng (2009) and Groen and Kapetanios (2013), where the search takes place only over factor estimates, the BIC is consistent for $N << T$. In our set-up the BIC is inconsistent in *all* cases; a result which is due to the search over additional estimates of the idiosyncratic error terms.

Finally, the results for the $BICM$ and $HQICM$ criteria of Groen and Kapetanios (2013) illustrate the property shown in the previous section, that these criteria for pure factor-augmented models are only consistent for cases where N is small relative to $\sqrt{T}$. This can be seen from Scenarios 1 and 2 in Figures 1 to 4 which correspond to the "Small N" scenario, in which the $BICM$ and $HQICM$ perform reasonably. However, in the cases of "Large N" the inconsistency of these criteria under Scenarios 3 is clear, indicating that care must be used if applying these criteria when estimated idiosyncratic components are used. Even in Scenario 4, where we expect $BICM$ and $HQICM$ to give consistent model selection, their finite sample performance tends to be worse than the new criteria IC_1 to HQ_3.

6 Empirical Application: Long-Horizon Exchange Rate Modelling

In this section we provide an empirical application of these model selection procedures in a very challenging predictive environment: long-horizon exchange rate modelling. We extend a recent approach of Engel et al. (2015) who suggest to predict exchange rate growth using the idiosyncratic component from a factor model of countries' exchange rates. This model provides a natural application for evaluating our new consistent model selection criteria.

6.1 Background

Since the seminal work of Meese and Rogoff (1983b,a), there has been growing interest in the predictive ability of log exchange rates over long forecast horizons. There have been many subsequent empirical studies looking to relate the growth of exchange rates at different horizons to macroeconomic 'fundamentals'. These approaches are comprehensively surveyed in Rossi (2013). The standard empirical approach in the literature tends to follow that of Mark (1995) and Kilian (1999) in specifying regression models such as:

$$s_{i,t+h} - s_{it} = \mu + \beta \left(s_{it} - f_{it}\right) + \varepsilon_{i,t+h} \tag{13}$$

where s_{it} is the logarithm of the exchange rate of country i, usually relative to the U.S. dollar (\$), and $(s_{it} - f_{it})$ is the deviation from an equilibrium relationship between s_{it} and the fundamentals. Typically, the type of fundamentals considered are variables such as trade balances, inflation, income and money supply, based on varying macroeconomic models of exchange rates.

Recently, the paper of Engel et al. (2015) moved away from the standard set of fundamentals by suggesting to use the common factors from a vector of N countries' exchange rates $s_t = [s_{1t,...,}s_{Nt}]'$ relative to the U.S. dollar (\$). They propose to use the estimated common component of country i as a fundamental. In other words they use a factor model of the form of Equation (1) above for log exchange rates:

$$s_t = \Lambda F_t + u_t$$

Having estimated Λ and F_t, for a given country i they use the estimated idiosyncratic error $\widehat{u}_{it} = s_{it} - \widehat{\lambda}_i' \widehat{F}_t$ and substitute this directly into the forecasting Equation (13) in place of $(s_{it} - f_{it})$. Therefore their model is exactly the form of model which is the focus of this paper. Our proposed information criteria may therefore be used to select between alternative specifications.

In this application we choose to go one step further than Engel et al. (2015) and use the full set of (cross-country) idiosyncratic errors to allow for potential exchange rate spillover effects from other economies, all relative to the common factor in the panel of exchange rates. Empirical studies into exchange rate spillover in both mean and variance has been considered by papers such as Hong (2001) and others. These spillover effects, if present, will be selected using the information criteria proposed in this paper. We therefore combine the literatures of long-horizon exchange rate forecasting, factor models and exchange rate spillovers.

Specifically, we will augment the model in Equation (13) and use regressions of the form:

$$s_{i,t+h} - s_{it} = \mu + \beta \left(s_{it} - f_{it}\right) + \alpha_i \widehat{u}_{it} + \alpha_j \widehat{u}_{jt} + \varepsilon_{i,t+h} \tag{14}$$

where $\widehat{u}_{it}$ is the domestic idiosyncratic error as in Engel et al. (2015), whereas $\widehat{u}_{jt}$, for $j \neq i$, is the exchange rate spillover effect new to this paper.[9] Therefore, when $\alpha_j = 0$, Equation (14) corresponds exactly to the specification of Engel et al. (2015). For the macroeconomic fundamentals, we will use the PPP model since this outperforms other models such as Taylor-rules in Engel et al. (2015). In other words for the fundamental f_{it} in Equation (14) we will use $\pi_{it} - \pi_t^*$, the long-run inflation differential between country i and the United States.

6.2 Data

To form the factor dataset, we choose the same countries as used in Engel, Mark and West (2015). We have monthly data for the monthly average closing exchange rate of 18 OECD countries plus the Eurozone relative to the U.S. Dollar over the time period August 1988 to May 2015. The data we use for the inflation rate differential is the CPI, which is available at a monthly level for all countries except for Australia. All data is extracted from the Haver Analytics databases USECON, G10 and EMERGE.[10] We split the sample at the end of 1998 and will perform the analysis for the pre- and post-Euro sub samples. The countries are listed in Table 2.

This means that for the pre-Euro subsample the dataset is of size $T = 197$, $N = 17$ and for the post-Euro subsample we have $T = 137$, $N = 10$. We will look at the monthly forecast horizons $h = 1, 3, 6, 9, 12, 18, 24$, which is similar to other studies such as McCracken and Sapp (2005) and Engel et al. (2015) who use quarterly data and a horizon of 2 or 3 years. We estimate the factors and idiosyncratic components by Principal Components, and let the number of factors be $r = 2$ as

[9] Equation (14) is written for only a one-country spillover though we will consider up to m^{max} spillover countries.

[10] Data accessed 18th June 2015.

Table 2: List of countries in the dataset, split by pre-Euro and post-Euro subsamples with cut-off date December 1998.

Pre-Euro	Post-Euro
Australia (AUS)	Australia (AUS)
Austria (AUT)	
Belgium (BEL)	
Canada (CAN)	Canada (CAN)
Denmark (DNK)	Denmark (DNK)
	Europe (EUR)
Finland (FIN)	
France (FRA)	
Germany (DEU)	
Italy (ITA)	
Japan (JPN)	Japan (JPN)
Korea (KOR)	Korea (KOR)
Netherlands (NLD)	
Norway (NOR)	Norway (NOR)
Spain (ESP)	
Sweden (SWE)	Sweden (SWE)
Switzerland (CHE)	Switzerland (CHE)
United Kingdom (GBR)	United Kingdom (GBR)

in the main results of Engel et al. (2015).[11]

6.3 Full Sample Model Selection Results

We first present the model selection results using all available data in both the pre- and post-Euro subsamples. Our interest is whether the new model selection criteria select any spillover idiosyncratic components $\widehat{u}_{jt}$ in model (14) over and above the domestic idiosyncratic component $\widehat{u}_{it}$, and the PPP fundamental. There are several reasons why we may expect very few additional variables to be chosen. Firstly, the sample size is particularly small in the cross section dimension N, for which we know our selection criteria can be slightly conservative. Secondly, we know from previous evidence that additional variables tend to have weak predictive power over and above the "no-change" benchmark, which may mean that few additional variables will be selected.

Tables 3 and 4 present the model selection results for the spillover effects in the pre- and post-Euro subsamples respectively. These Tables show the identity of the spillover countries j for a given country i and in parentheses the number of spillovers chosen. We present only the results for the $h = 1$ and $h = 12$ horizons for brevity. The upper panel of each table shows the results for the new HQ_3 criterion and the lower panel presents those for the standard BIC by way of comparison. As in the Monte Carlo section, we search over a maximum possible number of $m^{max} = 5$ idiosyncratic error spillover components.

[11] The results for other values of r are available upon request.

Table 3: In-Sample Model Selection Results: 1988-1998 pre-Euro era.

	Selection Criterion: HQ_3	
	$h = 1$	$h = 12$
	Countries ($\widehat{m}$)	Countries ($\widehat{m}$)
AUS	- (0)	KOR, FIN, BEL, ESP (4)
AUT	- (0)	DNK, BEL (2)
BEL	- (0)	DNK (1)
CAN	- (0)	DNK, GBR, FIN, ITA, SWE (5)
DNK	- (0)	BEL (1)
FIN	- (0)	KOR (1)
FRA	- (0)	DNK, BEL (2)
DEU	- (0)	DNK, BEL (2)
ITA	- (0)	KOR, FIN, CAN (3)
JPN	- (0)	ITA, CAN (2)
KOR	- (0)	ESP, BEL, CHE (3)
NLD	- (0)	DNK, BEL (2)
NOR	- (0)	KOR, DNK (2)
ESP	- (0)	KOR, FIN (2)
SWE	- (0)	FRA, GBR, FIN (3)
CHE	- (0)	KOR (1)
GBR	- (0)	FIN (1)

	Selection Criterion: BIC	
	$h = 1$	$h = 12$
	Countries ($\widehat{m}$)	Countries ($\widehat{m}$)
AUS	GBR, CHE (2)	KOR, FIN, ITA, BEL, ESP (5)
AUT	BEL, DNK (2)	KOR, DNK, BEL (3)
BEL	DNK (1)	KOR, DNK (2)
CAN	SWE, GBR (2)	DNK, GBR, FIN, ITA, SWE (5)
DNK	BEL, FIN (2)	FIN, KOR, BEL (3)
FIN	NOR (1)	KOR, SWE, ITA, CHE (4)
FRA	BEL (1)	DNK, BEL, KOR (3)
DEU	NOR (1)	KOR, DNK, BEL (3)
ITA	NOR (1)	KOR, FIN, CAN (3)
JPN	NOR (1)	ITA, AUS, ESP (3)
KOR	NOR, BEL, CHE (3)	ESP, BEL, CHE, DNK, ITA (5)
NLD	FIN, BEL, DNK (3)	DNK, BEL, KOR (3)
NOR	- (0)	KOR, DNK, FIN, CHE (4)
ESP	NOR (1)	KOR, FIN, ITA, SWE, CAN (5)
SWE	NOR, FIN (2)	FRA, GBR, FIN, ITA (4)
CHE	NOR (1)	KOR, DNK, BEL, FIN (4)
GBR	- (0)	FIN, JPN, KOR (3)

Notes: For each country, the column "Countries ($\widehat{m}$)" displays the identity of selected spillover countries, and the number of these selected countries in parentheses. Model selection is performed using the HQ_3 criterion (upper panel) and BIC criterion (lower panel). The models are assuming $r = 2$ factors and results are displayed only for the horizons $h = 1$ and $h = 12$. Spillovers are chosen over and above the PPP and domestic idiosyncratic fundamentals as described in the text.

Table 4: In-Sample Model Selection Results: 1999-2015 post-Euro era.

	Selection Criterion: HQ_3	
	$h = 1$	$h = 12$
	Countries ($\widehat{m}$)	Countries ($\widehat{m}$)
AUS	- (0)	- (0)
CAN	- (0)	SWE (1)
DNK	- (0)	SWE (1)
EUR	- (0)	SWE (1)
JPN	- (0)	CHE (1)
KOR	- (0)	- (0)
NOR	- (0)	SWE (1)
SWE	- (0)	- (0)
CHE	- (0)	SWE (1)
GBR	- (0)	- (0)
	Selection Criterion: BIC	
	$h = 1$	$h = 12$
	Countries ($\widehat{m}$)	Countries ($\widehat{m}$)
AUS	- (0)	SWE, JPN, EUR, DNK (4)
CAN	SWE (1)	SWE, DNK, EUR (3)
DNK	NOR (1)	NOR, CAN, CHE, GBR (4)
EUR	NOR (1)	NOR, CAN, CHE, GBR (4)
JPN	CHE (1)	CHE, DNK, EUR, SWE (4)
KOR	- (0)	CHE, SWE, NOR, CAN, AUS (5)
NOR	- (0)	SWE (1)
SWE	NOR (1)	NOR, AUS, KOR (3)
CHE	CAN (1)	SWE, JPN, EUR (3)
GBR	- (0)	CHE, SWE (2)

Notes: For each country, the column "Countries ($\widehat{m}$)" displays the identity of selected spillover countries, and the number of these selected countries in parentheses. Model selection is performed using the HQ_3 criterion (upper panel) and BIC criterion (lower panel). The models are run assuming $r = 2$ factors and results are displayed only for the horizons $h = 1$ and $h = 12$. Spillovers are chosen over and above the PPP and domestic idiosyncratic fundamentals as described in the text.

There are several key features to highlight from these results. The most surprising result is that, particularly in the pre-Euro era which includes the major European economies, the consistent HQ_3 criterion displays a non-trivial selection of spillover effects for the $h = 12$ horizon. In most cases there are 2 or 3 spillover effects chosen. This is contrary to the intuition that the model selection criterion would be over-conservative, both because of the small sample size, and because of the harsh predictive environment. In many cases, the selected spillover effects also have some reasonable interpretation, even though there is no theoretical model of spillovers in place. For example, we see linkages between the major European economies such as France, Belgium and

Germany. However, due to the small panel we consider, the results should also be treated with caution. For example, the countries Australia, Canada and Korea are particularly isolated within our sample and, as such, the model selection criterion appears to deliver spillovers which may not be deemed sensible. This gives motivation for a more extensive study involving a larger panel of global economies.

In the post-Euro era, however, much fewer spillover effects are chosen at the $h = 12$ horizon by the consistent HQ_3 criterion. This may be in some part due to the small number of countries in this subsample, $N = 10$. This is a number smaller than we considered in the Monte Carlo simulations, and is likely to have quite conservative model selection. The other main features of the results are that, at the short $h = 1$, there are no spillover effects chosen by the HQ_3 criterion. This is in line with the assertion of Mark (1995), and many others, that predictive ability of exchange rates at short horizons is likely to be low.

Finally, we see in the lower panels of Tables 3 and 4 that model selection using the standard BIC gives rise to much larger models. This is what was expected from the results in the previous sections, and confirms that it is unwise to rely on standard selection criteria such as the BIC in empirical studies using the idiosyncratic components estimated from a factor model.

6.4 Pseudo Out-of-Sample Results

In order to assess how the model in Equation (14) performs out of sample, when spillovers are selected by information criteria such as our new HQ_3, we perform a pseudo out-of-sample forecasting experiment. The evidence on out-of-sample predictive ability of long-horizon exchange rate models is very mixed in the literature. The assertion of Meese and Rogoff (1983b,a) was that no model could outperform the naïve no-change benchmark model. Since their work, there has been a great many papers attempting to overturn this result, for example Mark (1995). The paper of McCracken and Sapp (2005) also suggested that, using a range of different models and a range of predictive ability test statistics, there is some positive evidence of beating the no-change forecasting model. Nevertheless, we are not too hopeful of very positive evidence from an out-of-sample perspective. However, it will be useful to see how parsimonious our model selection procedures are in a pseudo out-of-sample setting.

For the pre- and post-Euro subsamples, we denote the total sample size as $T + h$ and split the sample into an 'in-sample' and 'out-of-sample' portion $T = R + P - 1$, having lagged the regressors h periods for the direct forecasting scheme. We proceed to form P out-of-sample forecasts by rolling through the sample with a rolling window length R, starting with the first R observations of the subsample. In every rolling window, we first estimate the factor model and the idiosyncratic errors using Principal Components. We then perform model selection using the HQ_3 criterion and estimate the parameters of the model by OLS before making the h-step ahead forecast.

This pseudo out-of-sample procedure yields a string of P forecast errors for the model in Equation (14) which we can write for country i as:

$$\widehat{\varepsilon}_{i,t+h} = (s_{i,t+h} - s_{it}) - \left(\widehat{\mu}_t + \widehat{\beta}_t (s_{it} - f_{it}) + \widehat{\alpha}_{it}\widehat{u}_{it}^{(t)} + \widehat{\alpha}_{jt}\widehat{u}_{jt}^{(t)}\right) \tag{15}$$

for all $t = R, ..., T$, where the estimated parameters $\widehat{\mu}_t, \widehat{\beta}_t, \widehat{\alpha}_{it}, \widehat{\alpha}_{jt}$ are the OLS estimators using the rolling window of data from $t - R + 1, ..., t$, and the idiosyncratic errors $\widehat{u}_{it}^{(t)}$ and spillovers $\widehat{u}_{jt}^{(t)}$ are superscripted by t as they have also been estimated by Principal Components using the same rolling window of the data.

The competitor model is the no-change forecast, which gives rise to the P forecast errors:

$$\widehat{\varepsilon}_{i,t+h}^{NC} = (s_{i,t+h} - s_{it}) - 0 \tag{16}$$

for all $t = R, ..., T$. For a country i, we will compare the two sets of forecasts in the expressions (15) and (16) using the Relative Mean Squared Forecast Error (RMSFE) statistic:

$$RMSFE_i = \frac{MSFE_i}{MSFE_i^{NC}} = \frac{\frac{1}{P}\sum_{t=R}^{T} \widehat{\varepsilon}_{i,t+h}^{2}}{\frac{1}{P}\sum_{t=R}^{T}\left(\widehat{\varepsilon}_{i,t+h}^{NC}\right)^2} \tag{17}$$

Therefore, a value greater than 1 implies that the no-change benchmark outperforms the exchange rate model, whereas a value less than 1 implies that the exchange rate model improves over the no-change benchmark. In the results, rather than present individual statistics by country and model, we instead look at the median RMSFE statistic across countries, in a similar way to Engel et al. (2015).[12]

Since we select the models using the HQ_3 criterion in each rolling window, it is possible that the model changes in every period, and it is therefore inappropriate to use Diebold-Mariano type approaches to unconditional predictive ability. We will instead use the approach of Giacomini and White (2006) for conditional predictive ability testing which allows the comparison of 'forecasting methods' and not 'forecasting models.'

Table 5 displays the results for three versions of the model in Equation (14) relative to the no-change benchmark. The first version uses only the domestic idiosyncratic component and omits the PPP and spillover terms by setting $\beta = 0$ and $\alpha_j = 0$. the second version omits the spillover effects by setting only $\alpha_j = 0$ and the final version is the full unrestricted model.

The results in Table 5 to a large extent confirm the results of other studies: that it is very difficult to beat the no-change model in out-of-sample prediction. The median RMSFE is above 1 in most cases, except for at the largest forecast horizons. This is consistent with the findings of Mark (1995), that predictive ability should increase with horizon. On the other hand, purely from a qualitative point of view, we can see that the addition of spillover effects above the PPP model reduces the median RMSFE in almost all cases. The statistic #(RMSFE<1) also indicates that this result holds not just for the median country, but it also increases the number of countries for

[12] Results for the individual countries are available on request from the author.

Table 5: Pseudo Out-of-Sample Forecasting Results.

		Pre-Euro						
		$h = 1$	$h = 3$	$h = 6$	$h = 9$	$h = 12$	$h = 18$	$h = 24$
$\widehat{u}_{it}$ only	Median RMSFE	1.07	1.23	1.32	1.41	1.37	1.15	1.15
	# (RMSFE<1)	1	0	0	0	1	1	1
$\widehat{u}_{it}$ + PPP	Median RMSFE	1.08	1.27	1.49	1.44	1.49	1.08	0.85
	# (RMSFE<1)	2	1	0	2	2	6	10
$\widehat{u}_{it}$ + PPP + $\widehat{u}_{jt}$	Median RMSFE	1.07	1.12	1.07	1.21	1.35	1.02	0.77
	# (RMSFE<1)	3	4*	4	4*	3	8	11
	Median $\widehat{m}$	0.35	1.55	1.47	1.40	1.23	1.28	1.71

		Post-Euro						
		$h = 1$	$h = 3$	$h = 6$	$h = 9$	$h = 12$	$h = 18$	$h = 24$
$\widehat{u}_{it}$ only	Median RMSFE	1.04	1.08	1.10	1.13	1.13	1.21	1.24
	# (RMSFE<1)	2	0	0	0	1	2	2
$\widehat{u}_{it}$ + PPP	Median RMSFE	1.06	1.10	1.13	1.16	1.21	1.33	1.46
	# (RMSFE<1)	1	0	0	1	0	2	4
$\widehat{u}_{it}$ + PPP + $\widehat{u}_{jt}$	Median RMSFE	1.05	1.08	1.15	1.16	1.01	1.30	1.26
	# (RMSFE<1)	1	0	2*	2	4	4	4
	Median $\widehat{m}$	0.00	0.13	0.61	0.65	0.64	0.98	1.08

Notes: Results based on a number of factors $r = 2$. The Median RMSFE statistic is the median of the countries' relative mean squared forecast error statistic from Equation (17). The median is taken over $N = 17$ countries in the Pre-Euro panel and $N = 10$ countries for post-Euro. The # (RMSFE<1) statistic counts the number of countries for which the RMSFE is less than 1, i.e. the model has lower MSFE than the no-change benchmark. The three entries with * are cases where a single country had a statistically significant improvement over the no-change model using the test of Giacomini and White (2006) at a 10% nominal size.

which the RMSFE is below 1. This result holds particularly in the pre-Euro area where for the $h = 24$ horizon, 11 out of 17 countries have a RMSFE less than 1. However, the evidence from the Giacomini and White (2006) test all but eliminates this result, as only a handful of results are statistically significantly in favour of the exchange rate model over and above the no-change benchmark.

Purely from a model selection point of view, which is the main purpose of this paper, we can see that the HQ_3 criterion selects a non-trivial amount of spillover effects on average. The Median $\widehat{m}$ statistic takes the median of all countries' number of selected spillover effects, averaged over all pseudo out-of-sample observations. From this we can see that the number of chosen idiosyncratic components tends to increase over the forecast horizon, indicating that the explanatory power of these variables tends to be for longer-term prediction. The model selection, again, becomes more conservative over the post-Euro period which may be due to the small sample size in N, which results in between 0 and 1 idiosyncratic components being selected on average.

7 Conclusion

In this paper we have proposed information criteria for performing model selection in regressions involving both estimated factors and idiosyncratic components. We show that existing information criteria, even those which account for factor estimation error, are inconsistent when we additionally use estimated idiosyncratic components in the regression. The first main contribution of the paper presents results regarding the estimation error of the idiosyncratic component when used in time series regression. We show that, under Principal Components estimation, this error vanishes at a rate $\min\left\{\sqrt{T}, N\right\}$, whereas factor estimation error vanishes at a rate $\min\{T, N\}$. This result implies that, in cases where N is of larger order than $\sqrt{T}$, which we consider to cover the majority of important macroeconomic datasets, then existing model selection criteria are inconsistent, even those which allow for factor estimation in pure factor-augmented models such as Groen and Kapetanios (2013). Also, we find that the standard BIC is inconsistent regardless of the relative rate of increase between N and T, which means that this criterion should not be used in specifying models involving factors and idiosyncratic components.

We therefore propose a class of information criteria with a penalty function which takes the estimation error in the idiosyncratic component into account. We show that these new criteria perform well in Monte Carlo simulations, relative to the existing information criteria which severely overfit the models in some or all configurations of N and T. We illustrate these model selection methods with an empirical application to forecasting exchange rates, extending the recent model of Engel et al. (2015) to allow for exchange rate spillover effects. We find that our methods select a non-zero amount of spillover effects, even in a challenging predictive environment when the models do not perform much better than a no-change benchmark. Future work can look at penalized LASSO-type regression for selection of these models. This was proposed without formal justification by the empirical study of Luciani (2014). Formal results using penalized regressions would provide a useful alternative to the information criteria proposed in this paper.

8 Appendix A: Proofs of Lemmas and Theorem

The proof of the Theorem in the text makes use of the following Lemmas on estimation error in the idiosyncratic components. Lemma 1 was discussed in the text, and Lemmas 2 and 3 are also required. The proof of each of these Lemmas makes use of the following identity:

$$\begin{aligned}\widehat{u}_{it} - u_{it} &= \left(X_{it} - \widehat{\lambda}_i'\widehat{F}_t\right) - \left(X_{it} - \lambda_i'F_t\right)\\ &= \lambda_i'H^{-1}HF_t - \widehat{\lambda}_i'\widehat{F}_t\\ &= \lambda_i'H^{-1}\left(HF_t - \widehat{F}_t\right) + \left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)'\widehat{F}_t\end{aligned}$$

$$
\begin{aligned}
&= \lambda_i' H^{-1}\left(HF_t - \widehat{F}_t\right) + \left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)' HF_t \\
&+ \left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)'\left(HF_t - \widehat{F}_t\right)
\end{aligned} \tag{18}
$$

where H is the rotation matrix described in Bai (2003) and Bai and Ng (2006).

Lemma 1 *Let Assumptions 1-5 hold and let the factors and idiosyncratic errors be estimated by Principal Components. Then as* $N, T \to \infty$,

$$
\frac{1}{T}\sum_{t=1}^{T} F_t\left(\widehat{u}_{it} - u_{it}\right) = O_p\left(\max\left\{\frac{1}{N}, \frac{1}{\sqrt{T}}\right\}\right) \tag{19}
$$

Proof of Lemma 1

We can use Equation (18) to write for any i:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} F_t\left(\widehat{u}_{it} - u_{it}\right) &= \frac{1}{T}\sum_{t=1}^{T} F_t\lambda_i' H^{-1}\left(HF_t - \widehat{F}_t\right) + \frac{1}{T}\sum_{t=1}^{T} F_t\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)' HF_t \\
&+ \frac{1}{T}\sum_{t=1}^{T} F_t\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)'\left(HF_t - \widehat{F}_t\right) \\
&= \left(\frac{1}{T}\sum_{t=1}^{T} F_t\left(HF_t - \widehat{F}_t\right)'\right) H'^{-1}\lambda_i + \left(\frac{1}{T}\sum_{t=1}^{T} F_tF_t'\right) H'\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right) \\
&+ \left(\frac{1}{T}\sum_{t=1}^{T} F_t\left(HF_t - \widehat{F}_t\right)'\right)\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)
\end{aligned}
$$

By Assumptions 1-5, the first term is $O_p\left(\max\left\{\frac{1}{N}, \frac{1}{T}\right\}\right)$ using Lemma A.1 of Bai and Ng (2006) and as $H'^{-1}\lambda_i$ is $O_p(1)$. We also know that the final term is of smaller order since $\left(\widehat{\lambda}_i - H'^{-1}\lambda_i\right)$ is $o_p(1)$ by Bai (2003) Theorem 2. However, for the middle term:

$$
\begin{aligned}
\left(\frac{1}{T}\sum_{t=1}^{T} F_tF_t'\right) H'\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right) &= O_p(1) \times \left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right) \\
&= O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\max\left\{\frac{1}{N}, \frac{1}{T}\right\}\right)
\end{aligned}
$$

Since Bai (2003) Theorem 2 shows that for the Principal Components estimator of $\widehat{\lambda}_i$ we have:

$$
\widehat{\lambda}_i - H'^{-1}\lambda_i = H\frac{1}{T}\sum_{t=1}^{T} F_t u_{it} + O_p\left(\max\left\{\frac{1}{N}, \frac{1}{T}\right\}\right)
$$

Now since we do not place any restriction on the relative rate of increase of T and N, $O_p\left(\frac{1}{\sqrt{T}}\right) +$

$O_p\left(\max\left\{\frac{1}{N},\frac{1}{T}\right\}\right) = O_p\left(\max\left\{\frac{1}{N},\frac{1}{\sqrt{T}}\right\}\right)$ and therefore:

$$\frac{1}{T}\sum_{t=1}^{T} F_t\left(\widehat{u}_{it} - u_{it}\right) = O_p\left(\max\left\{\frac{1}{N},\frac{1}{\sqrt{T}}\right\}\right)$$

as required.

Lemma 2 *Let Assumptions 1-5 hold and let the factors and idiosyncratic errors be estimated by Principal Components. Then as $N, T \to \infty$,*

$$\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{u}_{it} - u_{it}\right)\varepsilon_{t+h} = O_p\left(\max\left\{\frac{1}{N},\frac{1}{T}\right\}\right) \tag{20}$$

Proof of Lemma 2

We can use Equation (18) to write for any i that:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{u}_{it} - u_{it}\right)\varepsilon_{t+h} &= \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{t+h}\lambda_i' H^{-1}\left(HF_t - \widehat{F}_t\right) + \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{t+h}\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)' HF_t \\ &+ \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{t+h}\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)'\left(HF_t - \widehat{F}_t\right) \\ &= \lambda_i' H^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\left(HF_t - \widehat{F}_t\right)\varepsilon_{t+h}\right) + \left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)' H\left(\frac{1}{T}\sum_{t=1}^{T}F_t\varepsilon_{t+h}\right) \\ &+ \left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)'\left(\frac{1}{T}\sum_{t=1}^{T}\left(HF_t - \widehat{F}_t\right)\varepsilon_{t+h}\right)\end{aligned}$$

By Assumptions 1-5, the first part is $O_p\left(\max\left\{\frac{1}{N},\frac{1}{T}\right\}\right)$ by Lemma A.1 of Bai and Ng (2006) and as $\lambda_i' H^{-1} = O_p(1)$. The last part is of smaller order because $\widehat{\lambda}_i - H'^{-1}\lambda_i = o_p(1)$ by Bai (2003) Theorem 2. For the middle part,

$$\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)' H\left(\frac{1}{T}\sum_{t=1}^{T}F_t\varepsilon_{t+h}\right) = \left[O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\max\left\{\frac{1}{N},\frac{1}{T}\right\}\right)\right] \times O_p\left(\frac{1}{\sqrt{T}}\right)$$

since $\mathrm{E}[F_t\varepsilon_{t+h}] = 0$ implies that $\left(\frac{1}{T}\sum_{t=1}^{T}F_t\varepsilon_{t+h}\right) = O_p\left(\frac{1}{\sqrt{T}}\right)$ and as $\left(\widehat{\lambda}_i - H'^{-1}\lambda_i\right) = O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\max\left\{\frac{1}{N},\frac{1}{T}\right\}\right)$ by Bai (2003) Theorem 2. This part is therefore $O_p\left(\max\left\{\frac{1}{\sqrt{T}N},\frac{1}{T}\right\}\right)$. Combining these three results shows that

$$\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{u}_{it} - u_{it}\right)\varepsilon_{t+h} = O_p\left(\max\left\{\frac{1}{N},\frac{1}{T}\right\}\right)$$

as required.

Lemma 3 *Let Assumptions 1-5 hold and let the factors and idiosyncratic errors be estimated by Principal Components. Then as* $N, T \to \infty$,

$$\frac{1}{T}\sum_{t=1}^{T} u_{it}\left(\widehat{u}_{it} - u_{it}\right) = O_p\left(\max\left\{\frac{1}{N}, \frac{1}{T}\right\}\right) \tag{21}$$

Proof of Lemma 3

In a similar way to Lemmas 1 and 2, we can write:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} u_{it}\left(\widehat{u}_{it} - u_{it}\right) &= \frac{1}{T}\sum_{t=1}^{T} u_{it}\lambda_i' H^{-1}\left(HF_t - \widehat{F}_t\right) + \frac{1}{T}\sum_{t=1}^{T} u_{it}\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)' HF_t \\
&+ \frac{1}{T}\sum_{t=1}^{T} u_{it}\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)'\left(HF_t - \widehat{F}_t\right) \\
&= \left(\frac{1}{T}\sum_{t=1}^{T} u_{it}\left(HF_t - \widehat{F}_t\right)'\right) H'^{-1}\lambda_i + \left(\frac{1}{T}\sum_{t=1}^{T} F_t' u_{it}\right) H'\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right) \\
&+ \left(\frac{1}{T}\sum_{t=1}^{T} u_{it}\left(HF_t - \widehat{F}_t\right)'\right)\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right)
\end{aligned}$$

Now in this case, the first term is $O_p\left(\max\left\{\frac{1}{N}, \frac{1}{T}\right\}\right)$ by Assumptions 1-5 and Bai (2003) Lemma B.1, the final part is of smaller order as $\widehat{\lambda}_i - H'^{-1}\lambda_i$ is $o_p(1)$ and for the middle part we have:

$$\left(\frac{1}{T}\sum_{t=1}^{T} F_t' u_{it}\right) H'\left(H'^{-1}\lambda_i - \widehat{\lambda}_i\right) = O_p\left(\frac{1}{\sqrt{T}}\right) \times \left[O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\max\left\{\frac{1}{N}, \frac{1}{T}\right\}\right)\right]$$

since $\mathrm{E}[F_t' u_{it}] = 0$ implies that $\left(\frac{1}{T}\sum_{t=1}^{T} F_t' u_{it}\right) = O_p\left(\frac{1}{\sqrt{T}}\right)$ and $\left(\widehat{\lambda}_i - H'^{-1}\lambda_i\right) = O_p\left(\frac{1}{\sqrt{T}}\right) + O_p\left(\max\left\{\frac{1}{N}, \frac{1}{T}\right\}\right)$ from Bai (2003) Theorem 2. This term is therefore $O_p\left(\max\left\{\frac{1}{\sqrt{T}N}, \frac{1}{T}\right\}\right)$. Combining the three results gives:

$$\frac{1}{T}\sum_{t=1}^{T} u_{it}\left(\widehat{u}_{it} - u_{it}\right) = O_p\left(\max\left\{\frac{1}{N}, \frac{1}{T}\right\}\right)$$

as required.

Proof of Theorem

We have two regression specifications i and j:

$$y_{t+h} = \beta^{i\prime}\widehat{F}_t^i + \alpha^{i\prime}\widehat{u}_t^i + \varepsilon_{t+h}^i$$

with $\widehat{F}_t^i$ and β^i of dimension $r_i \times 1$ and α^i and $\widehat{u}_t^i$ of dimension $m_i \times 1$; and:

$$y_{t+h} = \beta^{j\prime} \widehat{F}_t^j + \alpha^{j\prime} \widehat{u}_t^j + \varepsilon_{t+h}^j$$

with $\widehat{F}_t^j$ and β^j of dimension $r_j \times 1$ and α^j and $\widehat{u}_t^j$ of dimension $m_j \times 1$. For simplicity we will write these compactly as:

$$y_{t+h} = \theta^{i\prime} \widehat{Z}_t^i + \varepsilon_{t+h}^i$$

and

$$y_{t+h} = \theta^{j\prime} \widehat{Z}_t^j + \varepsilon_{t+h}^j$$

Where $\widehat{Z}_t^i = \left[\widehat{F}_t^{i\prime}, \widehat{u}_t^{i\prime}\right]^\prime$ and $\widehat{Z}_t^j = \left[\widehat{F}_t^{j\prime}, \widehat{u}_t^{j\prime}\right]^\prime$. Now in what follows we will relate the generated regressors $\widehat{F}_t^i$ and $\widehat{u}_t^i$ to their probability limits $H^i F_t^i$ and u_t^i where H^i is the relevant submatrix of H described in Bai and Ng (2006) and u_t^i is not rotated as the common component is identified without rotation. This gives the probability limit vectors $Z_t^i = \left[F_t^{i\prime} H^{i\prime}, u_t^{i\prime}\right]^\prime$ and $Z_t^j = \left[F_t^{j\prime} H^{j\prime}, u_t^{j\prime}\right]^\prime$.

Now we can rewrite the sum of square error functions $V(.)$ both for the estimated factor and idiosyncratic components, and for their probability limits as:

$$\begin{aligned}
V\left(\widehat{F}^i, \widehat{u}^i\right) &= V\left(\widehat{Z}^i\right) = \frac{1}{T} y^\prime \widehat{M}^i y \\
V\left(\widehat{F}^j, \widehat{u}^j\right) &= V\left(\widehat{Z}^j\right) = \frac{1}{T} y^\prime \widehat{M}^j y \\
V\left(F^i H^{i\prime}, u^i\right) &= V\left(Z^i\right) = \frac{1}{T} y^\prime M^i y \\
V\left(F^j H^{j\prime}, u^j\right) &= V\left(Z^j\right) = \frac{1}{T} y^\prime M^j y
\end{aligned}$$

where $\widehat{M}^i = I - \widehat{Z}^i \left(\widehat{Z}^{i\prime} \widehat{Z}^i\right)^{-1} \widehat{Z}^{i\prime}$, $M^i = I - Z^i \left(Z^{i\prime} Z^i\right)^{-1} Z^{i\prime}$ and similarly for $\widehat{M}^j$ and M^j. Therefore we can rewrite the statement in Theorem 1 as:

$$\lim_{N,T\to\infty} \Pr\left(\ln\left[\frac{\frac{1}{T} y^\prime \widehat{M}^j y}{\frac{1}{T} y^\prime \widehat{M}^i y}\right] < \left(r_i + m_i - r_j - m_j\right) g\left(N, T\right)\right) = 0$$

which we can manipulate in order to relate the estimated regressors back to the true (rotated) factors and idiosyncratic errors as follows:

$$\lim_{N,T\to\infty} \Pr\left(\ln\left[\frac{\frac{1}{T} y^\prime M^j y}{\frac{1}{T} y^\prime M^i y}\right] + \ln\left[\frac{\frac{1}{T} y^\prime \widehat{M}^j y}{\frac{1}{T} y^\prime M^j y}\right] - \ln\left[\frac{\frac{1}{T} y^\prime \widehat{M}^i y}{\frac{1}{T} y^\prime M^i y}\right] < \left(r_i + m_i - r_j - m_j\right) g\left(N, T\right)\right) = 0 \tag{22}$$

The second and third terms on the left of this expression are both estimation error terms involving the estimated factors and idiosyncratic components. We first show the convergence rate of these two terms as T and N grow large, using Lemmas 1-3. This proof deviates from those in Bai and Ng (2006) and Groen and Kapetanios (2013) as the matrices $\widehat{M}^i$ and $\widehat{M}^j$ do not just contain estimated factors, they additionally contain estimation error due to the idiosyncratic errors.

Consider the first of these for the model specification i. (That for j will follow the same argument).

$$\begin{aligned}\frac{1}{T}y'\widehat{M^i}y - \frac{1}{T}y'M^iy &= \frac{1}{T}y'\left(\widehat{M^i} - M^i\right)y \\ &= \frac{1}{T}y'\left(P^i - \widehat{P}^i\right)y\end{aligned}$$

where $\widehat{P}^i$ and P^i are projection matrices $\widehat{P}^i = \widehat{Z}^i\left(\widehat{Z}^{i\prime}\widehat{Z}^i\right)^{-1}\widehat{Z}^{i\prime}$ and $P^i = Z^i\left(Z^{i\prime}Z^i\right)^{-1}Z^{i\prime}$. Now make the following expansion:

$$\begin{aligned}\frac{1}{T}y'\left(P^i - \widehat{P}^i\right)y &= \frac{1}{T}y'\left(Z^i\left(Z^{i\prime}Z^i\right)^{-1}Z^{i\prime} - \widehat{Z}^i\left(\widehat{Z}^{i\prime}\widehat{Z}^i\right)^{-1}\widehat{Z}^{i\prime}\right)y \\ &= \frac{1}{T}y'\left(Z^i\left(Z^{i\prime}Z^i\right)^{-1}Z^{i\prime}\right. \\ &\quad \left. - \left(\widehat{Z}^i - Z^i + Z^i\right)\left(\widehat{Z}^{i\prime}\widehat{Z}^i\right)^{-1}\left(\widehat{Z}^i - Z^i + Z^i\right)'\right)y \\ &= \frac{1}{T}y'Z^i\left[\left(Z^{i\prime}Z^i\right)^{-1} - \left(\widehat{Z}^{i\prime}\widehat{Z}^i\right)^{-1}\right]Z^{i\prime}y \\ &\quad - \frac{1}{T}y'\left(\widehat{Z}^i - Z^i\right)\left(\widehat{Z}^{i\prime}\widehat{Z}^i\right)^{-1}Z^{i\prime}y \\ &\quad - \frac{1}{T}y'Z^i\left(\widehat{Z}^{i\prime}\widehat{Z}^i\right)^{-1}\left(\widehat{Z}^i - Z^i\right)'y \\ &\quad - \frac{1}{T}y'\left(\widehat{Z}^i - Z^i\right)\left(\widehat{Z}^{i\prime}\widehat{Z}^i\right)^{-1}\left(\widehat{Z}^i - Z^i\right)'y \\ &= I + II + III + IV\end{aligned}$$

Consider part II:

$$\begin{aligned}II &= \frac{1}{T}y'\left(\widehat{Z}^i - Z^i\right)\left(\widehat{Z}^{i\prime}\widehat{Z}^i\right)^{-1}Z^{i\prime}y \\ &= \frac{1}{T}y'\left(\widehat{Z}^i - Z^i\right)\left(\frac{1}{T}\widehat{Z}^{i\prime}\widehat{Z}^i\right)^{-1}\frac{1}{T}Z^{i\prime}y\end{aligned}$$

The product of the second and last part of this expression $\left(\frac{1}{T}\widehat{Z}^{i\prime}\widehat{Z}^i\right)^{-1}\frac{1}{T}Z^{i\prime}y$ gives a column vector of dimension $(r_i + m_i) \times 1$ which is $O_p(1)$. The first part is a $1 \times (r_i + m_i)$ row vector. Now re-write this part in terms of the estimates $\widehat{F}_t^i$ and $\widehat{u}_t^i$ using $\widehat{Z}_t^i = \left[\widehat{F}_t^{i\prime}, \widehat{u}_t^{i\prime}\right]'$, or in matrix form $\widehat{Z}^i = \left[\widehat{F}^i, \widehat{u}^i\right]$, and substituting in the true model for y from Equation (2) we have:

$$\begin{aligned}\frac{1}{T}y'\left(\widehat{Z}^i - Z^i\right) &= \frac{1}{T}y'\left[\left(\widehat{F}^i - F^iH^{i\prime}\right), \left(\widehat{u}^i - u^i\right)\right] \\ &= \frac{1}{T}\left(F^0\beta + u^0\alpha + \varepsilon\right)'\left[\left(\widehat{F}^i - F^iH^{i\prime}\right), \left(\widehat{u}^i - u^i\right)\right]\end{aligned}$$

Now the first r_i elements of this row vector corresponding to factor estimation error are results which have already been shown in the literature, namely:

$$\begin{aligned}\beta'\frac{1}{T}F^{0\prime}\left(\widehat{F}^i - F^iH^i\right) &= \beta'\frac{1}{T}\sum_{t=1}^{T}F_t^0\left(\widehat{F}_t^i - H^iF_t^i\right)' \\ &= O_p\left(\max\left\{\frac{1}{N},\frac{1}{T}\right\}\right)\end{aligned}$$

and

$$\begin{aligned}\frac{1}{T}\varepsilon'\left(\widehat{F}^i - F^iH^i\right) &= \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{t+h}\left(\widehat{F}_t^i - H^iF_t^i\right)' \\ &= O_p\left(\max\left\{\frac{1}{N},\frac{1}{T}\right\}\right)\end{aligned}$$

both by Bai and Ng (2006) Lemmas A.1 (ii) and A.1 (vi), and

$$\begin{aligned}\alpha'\frac{1}{T}u^{0\prime}\left(\widehat{F}^i - F^iH^i\right) &= \alpha'\frac{1}{T}\sum_{t=1}^{T}u_t^0\left(\widehat{F}_t^i - H^iF_t^i\right)' \\ &= O_p\left(\max\left\{\frac{1}{N},\frac{1}{T}\right\}\right)\end{aligned}$$

by Bai (2003) Lemma B.1 since the dimension of u_t^0 is m^0 which is finite. However, the remaining three terms are new to this paper and are analysed in the Lemmas above, namely:

$$\begin{aligned}\beta'\frac{1}{T}F^{0\prime}\left(\widehat{u}^i - u^i\right) &= \beta'\frac{1}{T}\sum_{t=1}^{T}F_t^0\left(\widehat{u}_t^i - u_t^i\right)' \\ &= O_p\left(\max\left\{\frac{1}{N},\frac{1}{\sqrt{T}}\right\}\right)\end{aligned}$$

and

$$\begin{aligned}\frac{1}{T}\varepsilon'\left(\widehat{u}^i - u^i\right) &= \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{t+h}\left(\widehat{u}_t^i - u_t^i\right)' \\ &= O_p\left(\max\left\{\frac{1}{N},\frac{1}{T}\right\}\right)\end{aligned}$$

and finally:

$$\begin{aligned}\alpha'\frac{1}{T}u^{0\prime}\left(\widehat{u}^i - u^i\right) &= \alpha'\frac{1}{T}\sum_{t=1}^{T}u_t^0\left(\widehat{u}_t^i - u_t^i\right)' \\ &= O_p\left(\max\left\{\frac{1}{N},\frac{1}{T}\right\}\right)\end{aligned}$$

The same results are used in showing the other parts I, III and IV and as such we do not repeat them here. Combining all of these results yields:

$$\frac{1}{T}y'\widehat{M}^i y - \frac{1}{T}y'M^i y = O_p\left(\max\left\{\frac{1}{N}, \frac{1}{\sqrt{T}}\right\}\right)$$

This differs to that of Groen and Kapetanios (2013), due to the consistency rate in Lemma A. This result implies:

$$\frac{\frac{1}{T}y'\widehat{M}^i y}{\frac{1}{T}y'M^i y} = 1 + O_p\left(\max\left\{\frac{1}{N}, \frac{1}{\sqrt{T}}\right\}\right)$$

which in turn implies:

$$\ln\left[\frac{\frac{1}{T}y'\widehat{M}^i y}{\frac{1}{T}y'M^i y}\right] = O_p\left(\max\left\{\frac{1}{N}, \frac{1}{\sqrt{T}}\right\}\right)$$

And the same result holds for model j. Therefore we can rewrite the expression in Equation (22) to be:

$$\lim_{N,T\to\infty} \Pr\left(\ln\left[\frac{\frac{1}{T}y'M^j y}{\frac{1}{T}y'M^i y}\right] + O_p\left(\max\left\{\frac{1}{N}, \frac{1}{\sqrt{T}}\right\}\right) < (r_i + m_i - r_j - m_j)\, g\,(N, T)\right) = 0 \qquad (23)$$

To show Theorem 1, we assume that model i is correct and that the probability limits of $\widehat{F}_t^i$ and $\widehat{u}_t^i$ are $H^0 F_t^0$ and u_t^0 for all t, which because we impose orthogonality on the factors, means that model i contains the true number of variables with $r_i = r^0$ and $m_i = m^0$. This means that $M^i F^0 = 0$ and $M^i u^0 = 0$ so as in Groen and Kapetanios (2013), the denominator of the first part in (23) becomes:

$$\begin{aligned}\frac{1}{T}y'M^i y &= \frac{1}{T}y'M^i y \\ &= \frac{1}{T}\varepsilon' M^i \varepsilon \\ &= \sigma_\varepsilon^2 + O_p\left(\frac{1}{T}\right)\end{aligned}$$

assuming homoskedasticity of ε for simplicity. To assess statement (23) we now take two exhaustive cases in which the candidate model j is incorrectly specified:

Case 1: *The probability limits of $\widehat{F}_t^j$ and $\widehat{u}_t^j$ are such that: (i) $M^j F^0 = 0$ but $M^j u^0 \neq 0$ (Model j has correct factor specification but not all relevant idiosyncratic errors are included), (ii) $M^j F^0 \neq 0$ but $M^j u^0 = 0$ (not all relevant factors are included, but all relevant idiosyncratic errors are included) or (iii) $M^j F^0 \neq 0$ and $M^j u^0 \neq 0$ (model missing relevant factors and relevant idiosyncratic errors).*

In any of these three cases (i)-(iii), the numerator in Equation (23) is:

$$\begin{aligned}\frac{1}{T}y'M^j y &= \frac{1}{T}\varepsilon' M^j \varepsilon + \frac{1}{T}\left(F^0\beta + u^0\alpha\right)' M^j \left(F^0\beta + u^0\alpha\right) \\ &= \sigma_\varepsilon^2 + \tau_1 + O_p\left(\frac{1}{T}\right)\end{aligned}$$

where $\tau_1 > 0$ and the form of τ_1 depends on whether we are in Case 1 (i), (ii) or (iii). Therefore:

$$\frac{1}{T}y'M^j y - \frac{1}{T}y'M^j y = \tau_1 + O_p\left(\frac{1}{T}\right)$$

which implies that

$$\ln\left[\frac{\frac{1}{T}y'M^j y}{\frac{1}{T}y'M^j y}\right] \geq \tau_2 > 0$$

for some known τ_2. Therefore using Equation (23), the statement in Theorem 1 will hold in Case 1 as long as we can show that:

$$\lim_{N,T\to\infty} \Pr\left(\tau_2 + O_p\left(\max\left\{\frac{1}{N}, \frac{1}{\sqrt{T}}\right\}\right) < (r_i + m_i - r_j - m_j)\, g(N,T)\right) = 0$$

and since $(r_i + m_i - r_j - m_j)$ is finite, this statement is true when $g(N,T) \to 0$. This is required by Condition (i) stated in the Theorem, which proves what was required in Case 1.

Turning to Case 2:

Case 2: *The probability limits of $\widehat{F}_t^j$ and $\widehat{u}_t^j$ are such that both $M^j F^0 = 0$ and $M^j u^0 = 0$, but more than the relevant variables are included with either (i) $r_j = r^0$ but $m_j > m^0$ (correct factor specification but too many idiosyncratic errors included), (ii) $r_j > r^0$ but $m_j = m^0$ (too many factors specified, but correct idiosyncratic error specification) or (iii) $r_j > r^0$ and $m_j > m^0$ (too many factors and idiosyncratic errors included).*

In this case, the numerator in (23) is:

$$\begin{aligned}\frac{1}{T}y'\widehat{M}^j y &= \frac{1}{T}\varepsilon' M^j \varepsilon + \frac{1}{T}\left(F^0\beta + u^0\alpha\right)' M^j \left(F^0\beta + u^0\alpha\right)\\ &= \sigma_\varepsilon^2 + O_p\left(\frac{1}{T}\right)\end{aligned}$$

Therefore:

$$\frac{1}{T}y'M^j y - \frac{1}{T}y'M^j y = O_p\left(\frac{1}{T}\right)$$

which implies that:

$$\ln\left[\frac{\frac{1}{T}y'M^j y}{\frac{1}{T}y'M^j y}\right] = O_p\left(\frac{1}{T}\right)$$

Therefore using (23), the statement in Theorem 1 will hold in Case 2 as long as we can show that:

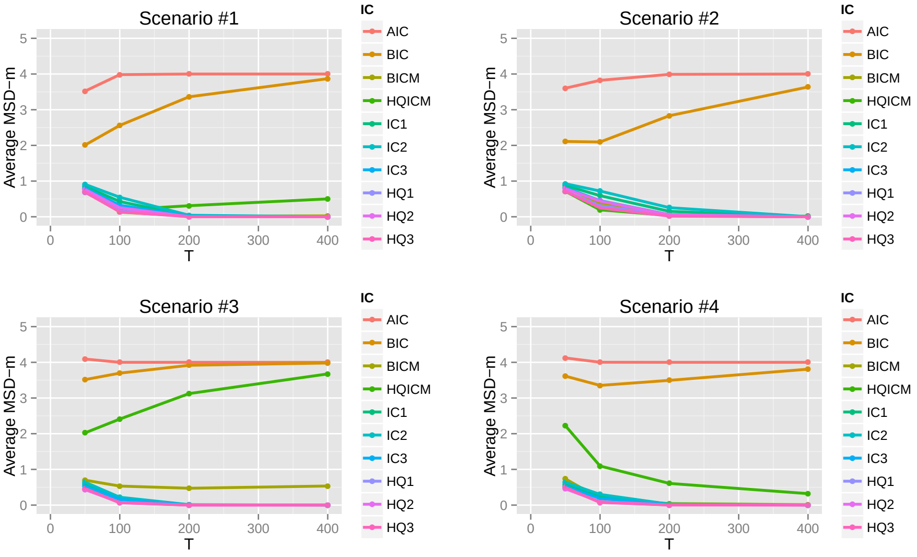
$$\lim_{N,T\to\infty} \Pr\left(O_p\left(\max\left\{\frac{1}{N}, \frac{1}{\sqrt{T}}\right\}\right) < (r_i + m_i - r_j - m_j)\, g(N,T)\right) = 0$$

Now since model i is the correct model with $r_i = r^0$ and $m_i = m^0$, each of Case 2 (i), (ii) and (iii) imply that $(r_i + m_i - r_j - m_j) < 0$. Therefore this statement holds when $g(N,T)\min\left\{\sqrt{T}, N\right\} \to \infty$ and the right hand side diverges to $-\infty$ at a quicker rate than the estimation error. Since this corresponds to Condition (ii) stated in the Theorem, this shows what was required in Case 2.

This completes the proof of the Theorem.

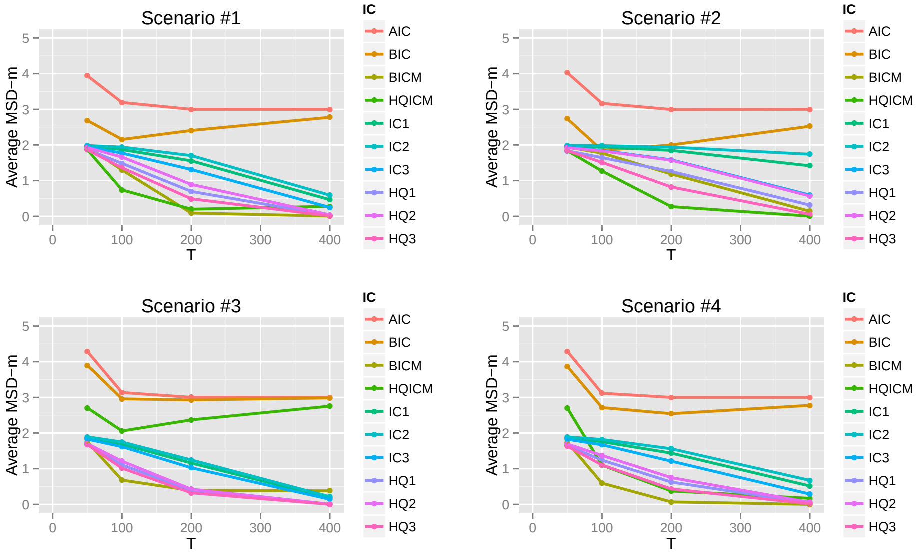
9 Appendix B: Monte Carlo MSD Results

Figure 5: MSD^u for different information criteria over 1,000 Monte Carlo replications when the true $r = 1$ and $m^0 = 1$



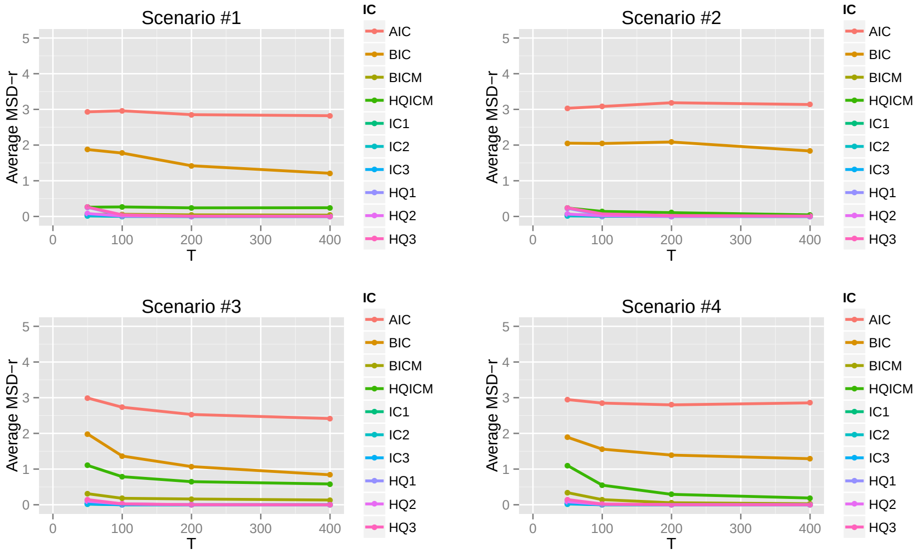
Notes: Scenarios 1-4 are described in Section 5.1 and each of the information criteria are described in Section 4.

Figure 6: MSD^u for different information criteria over 1,000 Monte Carlo replications when the true $r = 1$ and $m^0 = 2$



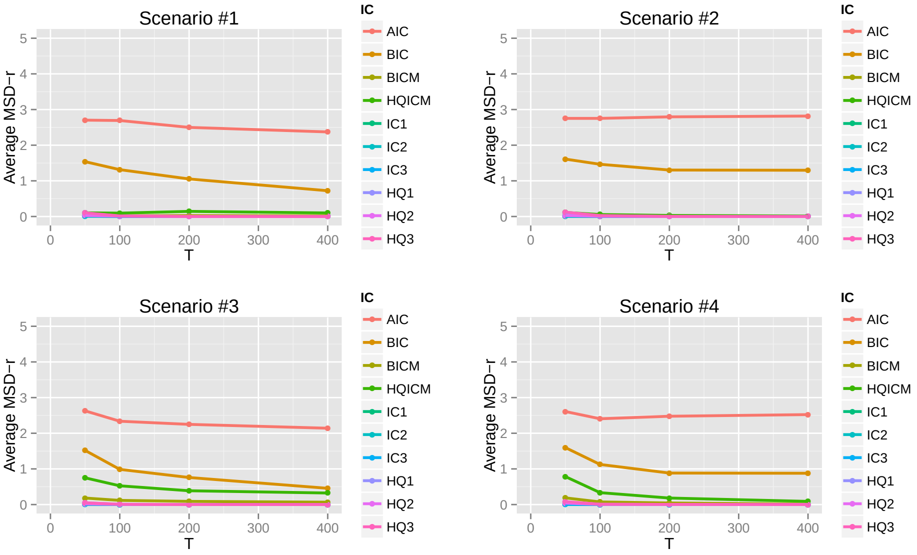
Notes: Scenarios 1-4 are described in Section 5.1 and each of the information criteria are described in Section 4.

Figure 7: MSD^r for different information criteria over 1,000 Monte Carlo replications when the true $r = 1$ and $m^0 = 1$



Notes: Scenarios 1-4 are described in Section 5.1 and each of the information criteria are described in Section 4.

Figure 8: MSD^r for different information criteria over 1,000 Monte Carlo replications when the true $r = 1$ and $m^0 = 2$



Notes: Scenarios 1-4 are described in Section 5.1 and each of the information criteria are described in Section 4.

References

Amengual, D. and M. W. Watson (2007). Consistent estimation of the number of dynamic factors in a large n and t panel. *Journal of Business and Economic Statistics 25*(1), 91–96.

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

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

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

Bai, J. and S. Ng (2009). Boosting diffusion indices. *Journal of Applied Econometrics 24*(4), 607–629.

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

Brownlees, C. T., E. Nualart, and Y. Sun (2015). Realized networks. *Available at SSRN 2506703*.

Castle, J. L., M. P. Clements, and D. F. Hendry (2013). Forecasting by factors, by variables, by both or neither? *Journal of Econometrics 177*(2), 305–319.

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

Djogbenou, A. (2016). Model Selection in Factor-Augmented Regressions with Estimated Factors. *Mimeo*.

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

Doz, C., D. Giannone, and L. Reichlin (2012). A quasi-maximum likelihood approach for large, approximate dynamic factor models. *The review of economics and statistics 94*(4), 1014–1024.

Eickmeier, S. and T. Ng (2011). Forecasting national activity using lots of international predictors: An application to new zealand. *International Journal of Forecasting 27*(2), 496–511.

Engel, C., K. West, and N. Mark (2015). Factor model forecasts of exchange rates. *Econometric Reviews 34*(1), 32–55.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*(6), 1545–1578.

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

Groen, J. J. J. and G. Kapetanios (2013). Model selection criteria for factor-augmented regressions. *Oxford Bulletin of Economics and Statistics 75*(1), 37–63.

Hallin, M. and R. Liška (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association 102*(478), 603–617.

Kilian, L. (1999). Exchange rates and monetary fundamentals: What do we learn from long-horizon regressions? *Journal of Applied Econometrics 14*(5), 491–510.

Luciani, M. (2014). Forecasting with approximate dynamic factor models: the role of non-pervasive shocks. *International Journal of Forecasting 30*(1), 20–29.

Mark, N. (1995). Exchange rates and fundamentals: evidence on long-horizon predictability. *The American Economic Review 85*(1), 201–218.

McCracken, M. W. and S. G. Sapp (2005). Evaluating the predictability of exchange rates using long-horizon regressions: Mind your p's and q's! *Journal of Money, Credit and Banking 37*(3), 473–494.

Meese, R. and K. Rogoff (1983a). The out-of-sample failure of empirical exchange rate models: Sampling error or misspecification? In J. A. Frenkel (Ed.), *Exchange Rates and International Macroeconomics*, pp. 67–112. Chicago: University of Chicago Press.

Meese, R. A. and K. Rogoff (1983b). Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics 14*(1), 3.

Ng, S. (2013). Variable selection in predictive regressions. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2. North-Holland.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics 92*(4), 1004–1016.

Rossi, B. (2013). Exchange rate predictability. *Journal of Economic Literature 51*(4), 1063–1119.

Stock, J. and M. Watson (2011). Dynamic factor models. In M. P. Clements and D. F. Hendry (Eds.), *The Oxford handbook of economic forecasting*, pp. 35–60. New York: Oxford University Press.

Stock, J. H. and M. W. Watson (1999). Forecasting inflation. *Journal of Monetary Economics 44*(2), 293–335.

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

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