House Price Forecasts with Factor Combinations

Charles Rahal†

This version: January 10, 2015
For the latest version please click here
Department of Economics Discussion Paper Series
University of Birmingham

Abstract

We take a computational approach to forecasting real and nominal house prices, comparing a large number of models varying by the choice of factors, ‘observable endogenous variables’ and the number of lags, in addition to classical and modern econometric models. We utilize various optimal model selection and model averaging techniques, comparing them against classical benchmarks. Using six original datasets with large cross-sectional dimensions, we include recent developments in the factor literature as part of our model set. Within a ‘pseudo real-time’ out of sample forecasting context, results show that model averaging across a large set of candidate factor models is able to consistently generate forecasts with favorable properties, but are unable to consistently generate the singularly lowest forecast error. Other results include forecast error increasing in horizon, error magnitude varying by country (the variance of the underlying series) and that errors are lower for nominal, as opposed to real indexes.

Keywords: House Prices, Forecasting, Factor Error Correction Models, FAVARs
JEL Classification: C53, C55, R30

---

1Charles Rahal, Department of Economics, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK., E-mail address: charlierahal at gmail dot com
1 Introduction

Considering the recently verified importance of housing markets on an institutional, international scale, it may be thought of as surprising that such little space in academic journals has been devoted to forecasting either real or nominal house price series. In this paper we combine various strands of the econometric, forecasting and real estate literature in order to show where forecasting gains over classical models can be generated. In particular, vector autoregressive (VAR) models have been a predominant feature of the forecasting literature since their inception during the 1980’s, and our approach overcomes the deficit of most classical empirical analyses which only focus upon a small number of variables. That deficit is that the ‘classical’ VAR model typically includes between three and ten variables, potentially creating an omitted variable bias, with serious consequences for forecasting. Even when considering an approach which overcomes this, then the choice of which combination of either variables or factors to combine is itself based upon subjective empirical methodologies.

While the integration of large datasets into the econometric profession has been growing due to increased computational power and improved algorithmic efficiency, such analysis is an ‘exception rather than the rule’ (Boivin and Ng, 2005, p.1), despite factor-based approaches appearing to consistently generate more accurate forecasts through their use of large cross-sections. However, when confronted with uncertainty over which model to use in a forecasting context, pooling over all forecasts may seem like a natural way to deal with the issue. In addition to this, an increasing amount of literature on model averaging is also shown to generate favorable results compared with a single econometric specification. We utilize such techniques to forecast a variable which has recently been shown to influence consumer spending, financial markets and the macroeconomy more generally. We do this across a range of markets, in contrast to previous studies primarily focused solely upon either the U.S. aggregate, or a regional U.S. housing market with a small model set (see Section 2). Unlike most other studies, our sample period includes the recent crisis period.

In particular, the methods used in this paper provide a way of incorporating a large number of available time series, reconciling the growing availability of data while keeping the framework itself empirically tractable. The restricted and unrestricted ‘large-dimensional approximate’ factor models which we implement permit the idiosyncratic errors to be weakly serially and cross-sectionally correlated with large cross-sectional (N) and temporal (T) dimensions. We model the co-variability of the series in our large datasets in terms of a small number of latent factors. We then estimate the relationship between these factors and our ‘observed’ variables. Of particular interest, we consider recent developments in the forecasting literature in the form of the Factor Error Correction Model (FECM/FECMc), as proposed in Banerjee and Marcellino (2008) and implemented in Banerjee et al. (2014), as well as the ‘two-step’ variety of unrestricted factor augmented vector autoregressive models (FAVAR) as used by Bernanke et al. (2005). Stationary I(0) factors are extracted as per Stock and Watson (2002a), and non-stationary I(1) factors are extracted using the technique of Bai (2004), as is common in the existing factor literature.

Particular inspiration for the model-averaging routines in this paper comes from Bai and Ng (2008), Bai and Ng (2009) and Cheng and Hansen (2012) who discuss the situation whereby the conventional ordering of factors by their importance in driving the co-variability of many predictors may not be consistent with the forecasting power
for a specific series of interest. As a result, we employ model specification techniques\textsuperscript{1} to determine which combination of factors (and which lag lengths) should be used in the models in addition to various weightings of forecasts from the set of candidate models.

Until recently, the effect of generated regressors on model selection and combination had not been previously investigated, in contrast to inference on the coefficients, where studies as early as Pagan (1984) show that generated regressors affect the sampling distribution. Importantly, Cheng and Hansen (2012) show, in a contrast to other results in the literature, that the generated regressor issue can be safely ignored when considering factor combination, without restrictions on $N$ and $T$. However, in this paper, we make no claims regarding the theoretical optimality of any of our model selection or averaging routines, with a sole desire to analyze empirical results in as robust a framework as possible.

House price forecasts and their corresponding accuracy over volatile periods, now more than ever considering the recent financial crisis, are of increasing importance. Given the dramatic increases in financial innovation, their accuracy is not only scrutinized by active market participants, but those involved in the securitization of the financial products underlying related mortgages, as well as central bankers and regulatory agencies monitoring economic stability and mortgage pricing and underwriting. Early literature such as Foster and Van Order (1985) and Crawford and Rosenblatt (1995) shows that the volatility of house prices is a major determinant of not only mortgage default, but also prepayment risk. Non-recourse loans, prevalent in some developed economies, facilitated the advent of situations where the optimal strategy for home-buyers was to exercise their option to default in the scenario of falling home values, further depressing the market (the frictionless option-theoretic model).

Given only the recent advent of factors and their applicability to modern macroeconometrics, and while there has been much research which considers factor models alone, there have been a relatively small number of studies which analyze the forecast performance of models which jointly consider both factors and variables. There are even fewer papers which provide empirical support for such a joint inclusion (Castle et al., 2013, p.17), motivating our study of the ‘small macroeconometric model’. This may appear surprising, considering the intuition that factor augmented models are able to overcome one of the primary criticisms of the classical VAR approach to macroeconomics, in that factors are able to conserve degrees of freedom while being able to span huge information sets typically available to practitioners.

In particular, early notable work along this dimension (Bernanke and Boivin (2003)) had considerable success with factors in improving the estimation of the policy reaction functions of the Federal Reserve. Such a modeling practice is entirely rationalizable given the assumption that central banks face information constraints similar to econometricians, who may be unable to directly observe various economic phenomena. In particular, the problem lies in the omission of key variables - a problem which has for example been used to explain anomalies such as the price puzzle. However, other macroeconomic applications such as Banerjee et al. (2005), which compares single indicators with factor models, do not find improvements from factors for euro aggregates, and, using similar methodologies, nor do Schumacher and Dreger (2004).

\textsuperscript{1}In particular: the IPC\textsubscript{2} for determining the appropriate number of factors, and then the SIC for determining the number of lags, selecting models out of our combination set, and then for approximate Bayesian Model Averaging (BMA).
In a recent comparative exercise, Stock and Watson (2012) present results which suggest that shrinkage methods offer ‘little improvements’ over dynamic factor models on average over all the series in question. In addition to this, as recognized in Bjørnland et al. (2009), a number of central banks have also been incorporating combination approaches to short-term forecasting problems. This includes, but is not limited to the Bank of England, the Riksbank, the Reserve Bank of New Zealand and the Bank of Canada which estimate multiple models and average across them. Typical model sets may include bi-variate and multi-variate VAR models, Bayesian vector autoregressions (BVARs), factor augmented models, indicator models and medium-large sized dynamic general equilibrium models.

Models which forecast real house price inflation are able to give an impression of the overall macroeconomy, and therefore aid in the formulation of policy. Economic theory dictates that house prices help to predict mortgage defaults, property taxes, and a range of consumption, investment and other household decisions. While the literature relevant to this paper is discussed in Section 2, it shows that house prices are strongly characterized by a persistent process subject to booms and busts. This further motivates our study of univariate and multivariate autoregressive processes which are not only augmented (parsimoniously) with factors, but also contain key variables relevant to ascertaining when a system is in disequilibrium. Figure 1 shows a time plot of both real and nominal house prices, and from this, we can consider the heterogeneous nature of our exercise: forecasting a range of different series during periods of increased volatility (2004:Q4-2013:Q3).

While there is existing research which considers the performance of either factor or regression models, very few consider both in the context of housing forecasts. However, combining a large number of variables condensed into factors as well as explicitly incorporating key observed variables decided upon by theory in a cointegrating (i.e. -FECM) context is an especially promising avenue for house price forecasting given the characteristics of housing markets. Following the discussion of the literature in Section 2, we then briefly discuss the large datasets used for factor extraction in Section 3.

---

**Figure 1: Cross-Country House Prices**

(a) Real Prices  
(b) Nominal Prices

*Source: Dallas Fed. Shaded regions represent U.S. recessionary periods (NBER).*
Then, in Section 4 we outline the methodology of the model set to be estimated, as well as providing an outline of our factor extraction, model selection and model averaging techniques. Following this, we present all of the results in Section 5, and finally, Section 6 concludes.

2 Literature Review

While the literature in the broader subject area is expansive, we attempt to give a short précis of the studies relevant to our approach. Early studies include Gau (1984) and Hamilton and Schwab (1985), which take a forecasting approach to determining market efficiency using data on commercial income generating markets in Vancouver and national residential markets in the United States respectively. However, these studies are typically constrained by small sample sizes and relatively outdated methodologies. The classic ‘bad forecast’ is that of Mankiw and Weil (1990), who forecast a real house price decline of 47% based upon the aging of the baby boom (‘baby bust’) generation and a falling demand for housing. While this appears to be more of an exercise in forecasting demographics, human capital and economic prosperity, the most direct critique comes from Hendershott (1991), who claims that their equation only fits data from the 1950s and 1960s, and that even between 1970-1987, the in-sample forecast error is unacceptably large. In a final example of the early literature (a paper which is subsequently explicitly mentioned as the motivation for several other academic works on the subject), Case and Shiller (1990) show how markets are forecastable using a repeat sales transaction index over a short time horizon due to serial dependence, citing an inability to make excess returns because of high transaction costs.

Patrick et al. (2000) compare ARIMA and spectral methods and their ability to forecast the turning points of US, UK and Australian securitized property markets. The authors find that while exponential models perform best during periods incorporating a steady trend, both ARIMA and spectral regression are capable of predicting turning points. In related work, Crawford and Fratantoni (2003) consider ARIMA, GARCH and regime switching models in five MSAs between 1979:Q1-2001:Q4, with out-of-sample forecasts generated over a range of different periods. The paper explicitly rationalizes the decision to consider only univariate methods due to the inherent difficulties in obtaining regional data. The authors conclude that while regime switching models perform best in-sample, with the ability to categorize boom and bust cycles, they only performed best (minimum RMSE) in three of the fifteen out-of-sample forecasts. However, the performance of the switching model in this paper is severely restricted by the fact that the authors only estimate it as a simple autoregressive specification due to concerns about the number of estimable parameters involved.

One particular paper of interest is that of Guirguis et al. (2004). The authors examine and compare six (VECM, AR, GARCH, Kalman Filter random walk, Kalman Filter autoregressive process, and exponential smoothing) models for the U.S housing market which are estimated between 1975-1985, with forecasts generated from 1985-1998. The comparison of the out-of-sample forecasts indicates that those generated by the Kalman Filter with an AR representation and the rolling GARCH models outperform all other specifications considered. However, the validity of the forecasting exercise is called into question through the claim that they ‘include the error correction term calculated from
the whole sample period extending from 1975:01 to 1997:2’ (p.45), granting the model more information than would have been available at the time of a hypothetical ex-ante out of sample forecast.

Turning to models which consider a small number of explanatory variables for house prices, Brown et al. (1997) apply the Kalman Filter to U.K. aggregate house prices over a boom period (1989:Q3-1992:Q2), having assumed that housing markets are structurally unstable. Such an approach allows two of the coefficients (nominal user cost of housing and expected gain on housing) to vary with time. Both static and dynamic forecasts are generated and compared against three fixed parameter models, which it outperforms, providing initial support for time varying parameter models in a house price forecasting context. In a similar study, Hall et al. (1997) use the Markov-switching regression proposed by Goldfeld and Quandt (1973) with an error-correction model (ECM) which includes four variables (real house prices, real income, owner occupied stock and mortgage rate). The system is a two state Markov process (stable and unstable) with transition probabilities dependent on how far out of equilibrium the system is. The model is successful in identifying the boom periods of UK housing of 1971-1974, 1977-1979 and 1986-1989, confirming parameter instability due to economic fluctuations and the degree of deviation from a long-run equilibrium.

In the main paper which informs our choice of observable variables, using only a standard small scale VECM estimation framework, Gattini and Hiebert (2010) include euro area house prices, housing investment, real disposable income per capita and a mixed maturity measure of the real interest rate. The variable choice is entirely rationalization given economic theory and various potential specifications of the data generating process for house prices. The authors find that ‘the reduced form model tracks closely turning points in house prices when examining out-of-sample one and two-step ahead forecasts’ (Gattini and Hiebert, 2010, p.47). While the paper presents results robust to specification tests, it only considers a single model class with one specification for one economic area and also only considers one and two step (quarters) ahead forecasts, encouraging further development in this area.

Rapach and Strauss (2007) consider state level forecasts with 25 potential predictors within an autoregressive distributed lag (ARDL) framework. The authors find that useful information is contained in the state level housing price/income ratio, unemployment, and national inflation rate. Various forecast combination methods are also considered with some degree of success. Rapach and Strauss (2009) follows in much the same trail, finding that across the twenty largest U.S. states, different housing markets can be better predicted with different information sets.

In the first of the ‘large dataset’ type house price forecast papers which we consider, Gupta (2013) uses 145 (monthly) fundamental variables within a dynamic factor and Bayesian shrinkage framework to forecast house price growth for the four census regions of the U.S. and the aggregate U.S. market as a whole. Using the standard Minnesota prior as well as additional priors which restrict the sum of the coefficients, the paper compares forecasts over an out-of-sample horizon of 1995:M1-2009:M3 based on an in-sample of 1968:M2-1994:M12 against a classical random walk and VAR model. The paper concludes that one of the factor models outperform in at least four of the five specifications. In the same vein, Gupta et al. (2009) implement several Bayesian and classical models for 20 U.S. states, comparing Bayesian shrinkage and FAVAR models (with spatial or causality priors), showing that factor-augmented models outperform.
Exploring the subject in even more detail, Das et al. (2009a) develops FAVAR and large BVAR models using a dataset of 126 monthly series for forecasting annualized real house price growth rates across nine census divisions of the United States. The results again indicate a preference for not only factor models, but show that large-scale models outperform their smaller counterparts - a result we are especially interested in. Considering potentially the most complete of all the studies on the subject, Gupta et al. (2011) consider the US real house price index (and its turning point in 2006:Q2) using not only large BVAR and FAVARs, but also factor augmented AR, Bayesian-shrinkage models and the DSGE model from Iacoviello and Neri (2008). The study concludes that while the small-scale Bayesian version of the model outperforms other models, it is only the DSGE model which is able to forecast the turning point with any accuracy.

Das et al. (2009b) consider forecasting the recent downturn in the South African housing market. Using 268 quarterly series to forecast small, medium and large-segment housing, the authors again compare large BVAR, classical and Bayesian versions of univariate and multi-variate (vector) autoregression models with a DFM model. Based on the RMSE, the result is that the large BVAR not only outperforms, but also manages to predict the downturn in the middle-segment. Turning to the metropolitan areas of South Africa, Gupta and Das (2008) estimate a spatial BVAR based on First-Order Spatial Contiguity and Random Walk Averaging priors. The results in this study show that the larger BVAR models outperform for all markets except that for large scale dwellings.

Bork and Møller (2012) use a common factor approach based on 122 quarterly time series for U.S. house prices. They find that a three-factor model can not only explain about 50% of one-quarter in-sample forecasting regressions but is also able to consistently outperform their AR benchmark at all horizons. The authors also document the substantial out-of-sample predictive power by comparing their model against AR benchmarks and computationally intensive forecast combination methods.

Moving away from the realm of house price forecasting, of particular interest to our work at hand is Banerjee and Marcellino (2006), who compare the augmented (with leading indicators) univariate approach of Cecchetti et al. (2000), the automated model selection procedure developed by Hendry and Krolzig (1999), and a dynamic factor model based upon the factor extraction technique of Stock and Watson (2002a). The paper implicitly compares the performance of models which include factors against those which use a combination of leading indicators against a single leading indicator which are then themselves compared against a standard autoregressive benchmark, finding that factor models are ‘on the whole disappointing’ for forecasting performance, in comparison to the result by Stock and Watson (2002b). Interestingly, Banerjee et al. (2005) uses a similar model, finding factor models perform differently depending on the underlying datasets, with results varying across country. A common finding is that gains from factor models tend to be larger for nominal series (see, for example, Boivin and Ng (2005)).

In particular, some authors such as Boivin and Ng (2006) argue that it may not be optimal to always use all available data when constructing principle components. However, Bernanke and Boivin (2003) find that the forecast performance of their model is improved when the dataset is enlarged from 78 to 215 variables. In addition to this, the standard approach calculates factors independent of the target variable. For example, Bai and Ng (2008) calculate factors (using the quadratic principal components method) from ‘targeted predictors’ (based on soft-thresholding). Their main finding is that for forecasting inflation over all horizons, there are gains from using fewer, but
more informative predictors, commenting that holding the set of predictors fixed may be unnecessarily restrictive.

One study of particular importance is that of Eickmeier and Ziegler (2008), who conduct a meta-study of 46 forecasting exercises involving dynamic factor models which are compared to a benchmark. The merit in this study comes from the fact that it is often difficult to draw conclusions across methods due to differences in data and implementation (e.g. different benchmarks). The authors find mixed results for factor model forecasts, with out-performance in some instances (e.g. across variables, or economic region) but not others. Studies such as this reinforce the importance of any examination of the topic being confined to a specific economic variable while examining it in a robust fashion across a range of different datasets.

Forecast combination methods have increasingly been used (since the seminal article of Bates and Granger (1969)) to improve forecasting performance in more conventional applications such as GDP and currency markets, but also with meteorological data, wilderness area use and the forecasting of football games. Clemen (1989) provides an extensive review of the early combination literature featuring over 200 items (‘an explosion in the number of articles on the combination of forecasts’), and a thorough over-view of this literature is far beyond the scope of this paper. Hendry and Clements (2004) provide various rationales for the success of forecast pooling (alternatively known as model averaging in a forecast context). Their arguments are based on the notion that if individual forecasts are based upon different subsets of information, or if they are differently biased, a combination may improve the accuracy, or offset the impact of structural breaks (through a reduction of the variance). Considering specific empirical successes from a policy making point of view, papers such as Kapetanios et al. (2008a) and Kapetanios et al. (2008b) indicate that while individual central bank models have difficulty in beating a simple benchmark AR model, combined forecasts frequently outperform. However, one commonly held notion is that of the ‘forecast combination puzzle’, whereby simple combinations ($\frac{1}{N}$) of point forecasts are repeatedly found to outperform, which we subsequently examine in Section 4 as one of our model weighting procedures.

3 Data

We utilize six new large datasets (largely dependent on availability), including data on the United Kingdom, the United States, Australia, Japan, Canada and Sweden. In order to keep our ‘observed’ endogenous variables as consistent as possible, we utilize real and nominal house prices from Mack and Martínez-García (2013). In addition to this, we take real gross domestic production and interest rates on mortgages from Oxford Economics, and real residential investment and all of our deflators from the OECD. For all datasets, $N > T$, ranging from $N = 162$ (Sweden) to $N = 324$ (US). All data used for factor extraction (which excludes our ‘observable’ variables) come from a range of national and international sources, such as central banks, official statistical agencies and other third party data providers, dependent upon availability. All series were collected from Datastream, and the optional appendix available with this paper contains the applicable Datastream identification codes.

There is presently no definitive guide in the literature as to what data should go into factor analysis. The macroeconomic panels are ‘put together’ by the researcher, and the
factors are inherently sample dependent by construction. However, the structure of our datasets follow the existing literature and are roughly categorized into groups such as monetary, banking, prices, foreign exchange rates, government, consumption, investment and other variables directly relevant to house price movements. While econometric theory might postulate that the size of the cross-sectional dimension positively influences the quality of the forecasts, data quality/reliability (i.e. source) remains a primary consideration.

As seen in the optional appendix which uses standard transformation codes (e.g. 5 - log first difference, 4 - logs, etc.), the data is automatically transformed where appropriate. In addition to this, the routine automatically seasonally adjusts where necessary (i.e. where economic theory dictates a seasonal component and the series is not adjusted at source), and also deflates where appropriate. We apply the identical transformations to all variables of the same type. While the use of ‘real-time’ data is increasingly accounted for in papers such as Clark and McCracken (2010) and Faust and Wright (2009), it remains the exception rather than the norm. While other research utilizes sources such as the ALFRED real-time database, international availability of real-time sources remains limited, and we prefer to instead utilize a ‘pseudo-real time’ forecasting exercise in order to exploit a cross-country dimension.

3.1 Number of Factors

The plots of all estimated factors can be seen in Figures 2 and 3. One essential modelling decision is regarding the number of factors to include in a model, or the maximum permitted within a determination routine such as that of Bai and Ng (2002). Our choices are similar to the strategies discussed in Remark 2 and Remark 3 of Chen et al. (2014): the maximum number of factors (both I(0) and I(1)) is imposed as a prior: based on the maximum number of stationary and non-stationary factors based on Banerjee et al. (2014), Figure 4, Figure 5 and trials of large values in the IPC$^2$ criteria of Bai and Ng (2002) discussed in Section 4.6.1. For three of our models, we then recursively estimate the number of factors using the IPC$^2$ itself.

4 Methodology

In this section we detail the range of different model classes considered, the model selection/model averaging techniques and the approach taken for the extraction of stationary and non-stationary factors. The exact approach taken to model estimation, specification and forecast generation is the focus of Section 4.6. Regarding the time dimension, $t = 1, \ldots, T$ denotes the initial ‘calibration’ period; that is, in all six of our datasets: 1980:Q1-2004:Q3. The recursive procedure assumes models estimated over $t = 1, \ldots, \{T + r - 1\}, \ldots, \{T + R - 1\}$ where $r = 1, \ldots, R$, and $R = 28$. Point forecasts for $H = \{1, 2, \ldots, 8\}$ are then recursively generated for the horizons $t = \{T + r - 1 + H\}$ to $t = \{T + R - 1 + H\}$ from 2004:Q4 onwards.

4.1 Factor Extraction

Recalling that not the true factors, their loadings nor idiosyncratic errors are directly observable, for stationary factors, we consider the static principle components method of
Stock and Watson (2002a) which is fast, easy to implement, and, given our sample size, performs at an equivalent level to dynamic principle components, with both assuming the series are composed of two mutually orthogonal (common and idiosyncratic) components. However, a range of research shows that the forecasts produced using various estimators of factors are highly collinear, producing similar $R^2$ values. Based upon their simulation results, Boivin and Ng (2005) claims: “With respect to static versus dynamic factor estimates . . . the estimator per se is not a choice of first-order importance” (p.132). Stock and Watson (2002a) show consistency of the factors if the errors are i.) stationary, ii.) have non-trivial loadings and iii.) have idiosyncratic errors with weak serial and cross-sectional correlation.

Factors are extracted recursively\(^3\), using the entire cross-section of variables ($N$) in our datasets excluding the four ‘observed’ endogenous variables (as is common in the literature; see, for example, Koop (2011)) for each economic area. For the I(0) routine, the variables are transformed into stationary variables which have zero mean and unit variance. The maximum number of common factors extracted is assumed to be $k_{\text{max}} \ll N$, with each factor($z$) drawn from each of the datasets, denoted by $F_{z,t}$ where $z=1,...,k_{\text{max}} \backslash _i k_{\text{max}}$. Specific factor loadings drive the cross-sectional co-movement of the variables. Formally: allow $X_i$ to be a $T \times 1$ vector of observations for variable $i$, allowing $F^0_0, \Lambda^0_t$ to denote the true common factors:

$$X_i \overset{\mathcal{F}}{=} F^0_0 \overset{\mathcal{r}}{\times \mathcal{T}} \Lambda^0_0 \overset{\mathcal{r} \times \mathcal{1}}{\times \mathcal{T}} + \xi_i \overset{\mathcal{T} \times \mathcal{1}}{\times \mathcal{T}} \tag{1}$$

where $X_i = (X_{i1}, X_{i2}, \ldots, X_{iT})'$, $F^0 = (F^0_1, F^0_2, \ldots, F^0_T)' \in \mathbb{R}^{T \times r}$ and $\xi_i = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{iT})$ and $r$ is the assumed ‘true’ number of factors. We take $k_{\text{max}} = 5$ for our I(0) routine, and $k_{\text{max}} = 4$ (i._$k_{\text{max}}$) for our I(1) routine (discussed below). For our ‘macro-panel’ of variables:

$$X \overset{\mathcal{F}}{=} F^0_0 \overset{\mathcal{T} \times \mathcal{r}}{\times \mathcal{T}} \Lambda^0_t \overset{\mathcal{r} \times \mathcal{N}}{\times \mathcal{T}} + \xi \overset{\mathcal{T} \times \mathcal{N}}{\times \mathcal{T}} \tag{2}$$

We estimate the common factors non-parametrically using asymptotic principal components, obtained by solving the optimization problem:

$$V(k) = \min_{\Lambda, F_{k_{\text{max}}}} (N \times T)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{i,t} - \Lambda_{k_{\text{max}}}^{k_{\text{max}}} F_{k_{\text{max}}})^2 \tag{3}$$

The number of factors chosen to solve the minimization problem (typically, in order to explain a certain proportion of the variance), $F_t = (F_{1,t}, \ldots, F_{k_{\text{max}},t})$ becomes a $T \times k_{\text{max}}$ matrix of estimated factors\(^4\) with corresponding loadings $\Lambda = (\Lambda_{1,t}, \ldots, \Lambda_{k_{\text{max}},t})$ matrix which is $k_{\text{max}} \times N$. At this point, note that a model with additional factors (e.g. $k_{\text{max}}+1$) can fit no worse than a model with less ($k_{\text{max}}$), but efficiency is lost as more loadings are estimated. To econometrically identify the model, we apply the standard normalization that $F^k_{k_{\text{max}} F^k_{k_{\text{max}}}} = I_{k_{\text{max}}}$.\(^5\)

---

\(^3\)We re-write and adapt the Gauss routine used by Banerjee et al. (2014) (available at the homepage of Igor Masten at www.econlab.si/igor.masten) for the I(0), cyclical and I(1) cases where $N > T$.

\(^4\)In our I(0) routine, the dimension shrinks to $T - 1$, to account for first differences taken to induce stationarity.
Figure 2: Stationary Factors
For the I(0) factors, the estimated factor matrix is \( \hat{F}^{I(0)} = \sqrt{T} \hat{P} \) where \( \hat{P} \) is a \( T \times k_{max} \) matrix of eigenvectors corresponding to the \( k_{max} \) largest eigenvalues of \( XX' \). The corresponding matrix of factor loadings is \( \hat{\Lambda}^{k_{max}} = \frac{\hat{F}^{I(0)} X}{T} \). Note that while this solution is not unique, it is a computationally less costly alternative when (as is our case) \( N > T \).

**Figure 4:** Percent of Variance Explained

As one may expect, the theory of principal components estimation requires modification when applied to integrated and cointegrated variables. For the case where \( N > T \), for the I(1) methodology proposed by Bai (2004), each column of our matrix of variables \( X \) is again mean zero with unit variance, but this time is not transformed to contain only stationary variables. That is, it contains both stationary and non-stationary variables as per the ‘raw’ datasets described in Section 3. While Banerjee et al. (2014) find no difference between a measure which excludes all stationary variables in their I(1) routine, we deem it favorable to preserve the rich cross sectional dimension, neglecting to discard the I(0) variables in I(1) factor extraction. The factors and loadings are extracted recursively as: \( \hat{F}^{I(1)} = T \hat{P} \) and \( \hat{\Lambda} = \frac{\hat{F}^{I(1)} X}{T} \) respectively.

The percent of variance explained by each of the I(0) and I(1) factors can be seen in Figure 4, showing an average cumulative variance explained across the \( k_{max} \) I(0) factors at around 35-60%, and for the \( i \_ k_{max} \) I(1) factors, upwards of 80%, completely in accordance with the literature. In addition to this, our choice of five I(0) factors and four I(1) factors appears entirely justifiable considering the Cattle Scree plots in Figure 5 which depict the

---

5 Other earlier research (Banerjee et al. (2005)) shows an irrelevance between modelling price series as I(1) or I(2).
regions in which adding additional factors ceases to explain a sufficient amount of variance (the ‘kink’). For the idiosyncratic cyclical factor, we extract one factor using the same method as the I(0) routine above, but instead of extracting eigenvectors from $XX^T$, we extract one cyclical factor from $\chi = (XX^T - F^{I(1)}I^{I(1)})$. These additional I(0) factors may be thought to represent ‘common cycles’ (denoting the $c$ in FECMc), as opposed to the ‘common trends’ represented by the I(1) factors. These factors are generated recursively to represent our ‘pseudo-real time’ forecasting exercise. While it would be possible to give interpretation to our factors as in papers such as Bork and Møller (2012) who regress each factor with each component series, plotting against series category, such an exercise provides minimal value due to the combinatoric nature of our forecasting regime.

4.2 Classical Models

In order to compare our factor based methodology with classical ‘observed’ models, we estimate three benchmarks. The first of which is a univariate model which is estimated based on the algorithm of Hyndman and Khandakar (2008). While not the same step-wise approach, it estimates models in the possible parameter space of an ARIMA($p,d,q$) of the form:

$$\Phi(B)(1 - B^d)Y_t = c + \theta(B)\epsilon_t$$

(4)

where $\{\epsilon_t\}$ is a white noise process with mean zero and variance $\sigma^2$, $B$ is a backshift operator and $\Phi$ and $\theta$ are polynomials of order $p$ and $q$. We discount any models where $\Phi$ and $\theta$ have no roots for $|z| < 1$. We search the model set where $p = 1, \ldots, 4$, $d = 0, 1, 2$ and $q = 1, \ldots, 4$, first running a KPSS test, and then estimating the ARMA($p,q$) model on the
suitably transformed data (at the 5% level). The resulting model is the one which satisfies
the above condition and minimizes the Schwarz Information Criterion (SIC) in the interest
of parsimony, and, like all of the following models, the specifications are determined and
estimated at each recursive step for each country. Our endogenous variable \(Y_t\) is the (log
of the) real or nominal house prices from Mack and Martínez-García (2013) as discussed
above. Naturally, we also estimate corresponding (symmetric) unrestricted (VAR) and
restricted (VECM) vector autoregressive models as per Eq. 5 and Eq. 6:

\[
\begin{bmatrix}
\Delta Y_t \\
\Delta Z_t \\
\end{bmatrix} = C + \Phi(L) \begin{bmatrix}
\Delta Y_{t-1} \\
\Delta Z_{t-1} \\
\end{bmatrix} + \begin{bmatrix}
\epsilon_{y,t} \\
\epsilon_{z,t} \\
\end{bmatrix} \tag{5}
\]

\[
\begin{bmatrix}
\Delta Y_t \\
\Delta Z_t \\
\end{bmatrix} = \alpha(\delta' \begin{bmatrix} Y_{t-1} \\ Z_{t-1} \end{bmatrix} + \rho_0) + \Phi(L) \begin{bmatrix}
\Delta Y_{t-1} \\
\Delta Z_{t-1} \\
\end{bmatrix} + \alpha \perp \gamma_0 + \begin{bmatrix}
\epsilon_{y,t} \\
\epsilon_{z,t} \\
\end{bmatrix} \tag{6}
\]

We denote by \(\Phi(L)\) a standard conformable lag polynomial of finite order, with the order
determined by an information criteria and the cointegrating rank (the size of \(\alpha\) and \(\beta\) is
\(r \times k\), where \(k\) is the number of endogenous variables and \(r\) is the rank) determined by the
Johansen Trace Test at the 5% level (as are the FECM/FECMc models in Section 4.3.1)
where \(Z_t\) is a vector of all ‘observable’ endogenous variables (real GDP, real residential
investment, mortgage rate) similar to Gattini and Hiebert (2010), discussed in Section
2. Our deterministic specification (\(C\) in Eq. 5, \(\rho_0\) and \(\gamma_0\) in Eq. 6) is based on the
assumption that the level data have linear trends but the cointegrating equations have
only intercepts.

### 4.3 Factor Augmented Models

Turning to factor models, we estimate our unrestricted systems as the ‘two-step’ variant
of Bernanke et al. (2005), where the factors are extracted by the methods of Stock
and Watson (2002a), prior to the estimation of the FAVAR. It is duly noted that it
would instead be possible to undertake the exercise as a ‘direct’ multi-step forecast
from regressing \(Y_{t+h}\) on \(Z_t, \hat{F}_t, Y_t\) and their lags. However, we instead use system
methodology in conjunction with the one-step ahead forecasting equation, iterating the
model forward. Our justification is based on Marcellino et al. (2006) who give particular
empirical validation over longer forecast horizons. In addition to this, it is claimed:
‘when there is more than one factor, vector-autoregressive forecasts of the factors should
be considered, not univariate autoregressive forecasts’ (Boivin and Ng, 2005, p.125).

We can model the joint dynamics of our FAVAR system as:

\[
\begin{bmatrix}
\Delta Y_t \\
\Delta Z_t \\
\hat{F}^{(0)}_t \\
\end{bmatrix} = C + \Phi(L) \begin{bmatrix}
\Delta Y_{t-1} \\
\Delta Z_{t-1} \\
\hat{F}^{(0)}_{t-1} \\
\end{bmatrix} + \begin{bmatrix}
\epsilon_{y,t} \\
\epsilon_{z,t} \\
\epsilon_{f,t} \\
\end{bmatrix} \tag{7}
\]

where the \(\hat{F}^{(0)}_t\) are extracted from suitably transformed (i.e. I(0)) data. If the relative
coefficients of \(\Phi\) are zero, there exists a direct mapping onto our earlier models (that is
to say that Eq. 5 is nested in Eq. 7), allowing us to analyze the contributions of \(\hat{F}_t\) (i.e.
the difference between forecast errors). Conversely, if the true system is indeed a FAVAR,
omission of factors will lead to biased coefficient estimates in the VAR framework.

---

\(^6\)This is claimed to produce ‘more plausible’ structural responses than the computationally intensive
Bayesian method based upon Gibbs sampling.
4.3.1 FECM: Factor Error Correction Models

While a full derivation of the model, from VAR to ECM to FECM, is provided in Banerjee and Marcellino (2008) (with Monte Carlo\empirical forecasting applications in Banerjee et al. (2014)) we merely show the final models to be estimated, following similar notation. In particular, from the Granger representation theorem, there exists an error correction of the form:

\[
\begin{bmatrix}
\Delta Y_t \\
\Delta Z_t \\
\Delta  \hat{F}_{t}^{(1)} \\
\Delta \hat{C}_{t}^{(0)}
\end{bmatrix} = \alpha \begin{bmatrix} \delta' \\
Y_{t-1} \\
Z_{t-1} \\
\hat{F}_{t-1}^{(1)} \\
\hat{C}_{t-1}^{(0)}
\end{bmatrix} + \rho_0 + \Phi(L) \begin{bmatrix} \Delta Y_{t-1} \\
\Delta Z_{t-1} \\
\Delta \hat{F}_{t-1}^{(1)} \\
\Delta \hat{C}_{t-1}^{(0)}
\end{bmatrix} + \alpha \perp \gamma_0 + \begin{bmatrix} \varepsilon_{y,t} \\
\varepsilon_{z,t} \\
\varepsilon_{f,t}
\end{bmatrix}
\]

whereby the I(1) factors (\hat{F}_{t}^{(1)}) are extracted from the dataset using the methodology of Bai (2004). When this model is augmented with additional lags of the differenced dependent variables, we have what is defined as an FECM by Banerjee and Marcellino (2008). From a practical point of view, one issue of particular importance regards the number of cointegrating vectors to estimate. Following Banerjee et al. (2014), we determine this empirically through the Johansen Trace Test (at the 5% level), although it would be possible to estimate all models with all viable r. As discussed in Section 4.1, and as per Banerjee et al. (2014), we also extract idiosyncratic cyclical components. We then include them in a subset of our FECM models, which, when augmented with additional lag lengths, form the FECMc:

\[
\begin{bmatrix}
\Delta Y_t \\
\Delta Z_t \\
\Delta  \hat{F}_{t}^{(1)} \\
\Delta  \hat{C}_{t}^{(0)}
\end{bmatrix} = \alpha \begin{bmatrix} \delta' \\
Y_{t-1} \\
Z_{t-1} \\
\hat{F}_{t-1}^{(1)} \\
\hat{C}_{t}^{(0)}
\end{bmatrix} + \rho_0 + \Phi(L) \begin{bmatrix} \Delta Y_{t-1} \\
\Delta Z_{t-1} \\
\Delta \hat{F}_{t-1}^{(1)} \\
\Delta \hat{C}_{t-1}^{(0)}
\end{bmatrix} + \alpha \perp \gamma_0 + \begin{bmatrix} \varepsilon_{y,t} \\
\varepsilon_{z,t} \\
\varepsilon_{f,t} \\
\varepsilon_{c,t}
\end{bmatrix}
\]

4.4 Candidate Models

Equations 4 to 9 are used to generate the first six of the twelve forecasts which we evaluate, as discussed in Section 4.6.1 and 4.6.2. The second set of six come from selecting (two), and averaging (across three different model sets with two different averaging methods) over the set of candidate models, to be discussed below. The largest model in our set of candidate models (for the FAVAR \( k_{max}=5 \) and the FECM \( i \_ k_{max}=4 \), augmented with one additional cyclical factor), is the one which includes all endogenous ‘observed’ variables and the maximum number of factors for both FAVAR and FECMc style models. Within this large model, we estimate all possible subsets of endogenous ‘observable variables’ with all possible subsets of factor combination, with \( Y_t \) always remaining as an endogenous variable. That is; any given model may include all factors or variables, a subset of factors or a subset of variables. Only four possible ‘observed’ variables are chosen (instead of a larger number, such as the ten variables used in Iacoviello and Neri (2008), as per Gupta et al. (2011)) in order to follow the advice of Cheng and Hansen (2012) so that the total number of models is practically and computationally feasible.

As in Cheng and Hansen (2012), our strategy is to treat the lag and factor structure as unknown, selecting and weighting across the model set. Given that there are three

7Our equations are similar in nature to equations nine and ten in Banerjee and Marcellino (2008) and seven and eight in Banerjee et al. (2014) respectively.
supplementary ‘observable’ variables and five possible factors (whereby we exclude a null
subset of factors which would render us equivalent to Eq. 5 or Eq. 6), we estimate
\(2^5 \times (2^5 - 1)\) models for both the FAVAR and FECMc, with \(k = \{1, 2, 3, 4, 5\}\), giving a
total of 992 factor models for each of the FAVAR and FECMc model classes, generated at
each of 28 recursive periods across six datasets, for a total of 665,952 \((2 \times 28 \times 1982 \times 6)\)
candidate factor models.8

4.5 Estimation Techniques

We initially estimate our AR, VAR and FAVAR models with ordinary least squares. For
the VECM and FECM models (using the same procedure for both), we use a two-step
method. The estimate of \(\hat{\delta}\) of the cointegrating matrix in Eq. 6 and Eq. 9 is computed
by maximum likelihood in the first step. This is then substituted in as the true \(\delta\) with all
other parameters estimated in the second stage of the least squares estimation (Lütkepohl,
2006, p.301) for \(\alpha\) and \(\Phi\). We determine the size of \(\alpha\) and \(\delta\) as per the Johansen Trace
Test (5%) recursively where applicable, and separately for each different combination of
observed variables and factors.

4.6 Forecasting Approach

The forecasting approach which we take is as follows. We utilize six datasets described in
Section 3. For each of these datasets, we attempt to create a ‘pseudo-real time’ forecasting
environment through a recursive procedure. We begin by calibrating our models up until
2004Q3 and then generate multi-step point forecasts which are evaluated at horizons of
1, 2, \ldots, 8 periods ahead \((h = 1, \ldots, 8)\). The models are then re-estimated one step ahead
(that is, for example where \(r = 2, 2004Q4: t = (T+r-1)\)) with forecasts being generated
from 2005Q1 (that is, a point forecast for \(T+r-1+h\) onwards recursively. This procedure
is repeated until 28 recursions \((r = 1, 2, \ldots, 28(R))\) are complete.

4.6.1 Optimal Model Selection Routine

In general, if we let \(w = ((w(1), \ldots, w(M)))'\) denote the weight vector, requiring that the
weights sum to one: \(\sum_i^M w_i = 1\) and that they are non-negative \((0 < w(i) < 1)\), model
selection is akin to a combination routine where \(w(i) = 1\). More generally, combination
involves selecting the weight vector \(w\) from \(H^M\), the unit simplex in \(\mathbb{R}^M\). For the \(p\) and
\(q\) of the ARIMA(p,d,q) model, and \(k\) of the VAR and VECM models, we utilize the SIC:

\[
\text{SIC}_i = \frac{-2 \times \ln(L_i)}{n} + \frac{p \times \ln(n)}{n} \quad (10)
\]

where \(p\) represents the number of estimated coefficients, \(n\) the number of observations
and \(L_i\) the log likelihood of model \(i\) in the equation for house prices (as opposed to the
system log likelihood).

Following this, we optimally select the number of I(0) and I(1) factors for an ‘optimal’
FAVAR, FECM and FECMc models respectively using the IPC\(_2\) of Bai and Ng (2002)
as per Banerjee et al. (2014), as, unlike the PC criterion, the IPC\(_2\) is not contingent.

8If any models cannot be estimated (for reasons of singularity), they are not candidates for model
selection and are subsequently assigned a weight of zero in the following weighting routines.
on $k_{max}$ or $i_{-}k_{max}$. The procedure is as follows, where we choose $k$ to minimize IPC$_2$ criteria recursively (although IPC$_1$ and IPC$_3$ criterion are also generated and determined a robustness check), and our maximum number of permitted factors is as per Section 4.1. First, we estimate the sum of squared residuals divided by $N \times T$ when $k$ factors are estimated:

$$V(k, \hat{F}_t) = \min_{\Lambda} \frac{1}{N \times T} \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{i,t} - \lambda_i^k \hat{F}_k)^2$$  \hspace{1cm} (11)$$

As per Bai and Ng (2002), denote $C^2_{N,T} = \min\{N, T\}$, and then the IPC$_2$ is defined for $k$ factors as:

$$IPC_2(k) = \ln(V(k, \hat{F}_k)) + (k \times \frac{(N + T)}{(N \times T)} \times \ln C^2_{N,T})$$  \hspace{1cm} (12)$$

where we choose the number of factors ($k$) which minimizes IPC$_2$. Having used the SIC to select the ARIMA specification, and the lags for each of the VAR and VECM models, we then use it to determine the number of lags in our IPC$_2$ determined FAVAR/FECM/FECMc models, as well as to ‘model select’ (determine the FAVAR and the FECMc) each of two models which minimize the SIC over the entire set of candidate models discussed in Section 4.4.

### 4.6.2 Model Weighting Algorithms

Following the selection/identification of our three classical models, our FAVAR/FECM/FECMc models and our two ‘combination’ models, we also utilize four forecasts created from weighting across the entire set of candidate models: three of which are approximate Bayesian Model Averaging weights, with the fourth a simple naïve weighted average. In forecast combinations generally, a large number of individual models are combined to produce a weighted forecast: $\hat{Y}_{t+h\mid t}$ from $i$ different forecasts ($\hat{Y}_{i_t+h\mid t}$) with weights from each model as $\hat{w}_i^t$. The combination is thus calculated as:

$$\hat{Y}_{t+h\mid t} = \sum_{i=1}^{M} \hat{w}_i^t \hat{Y}_{i_t+h\mid t}$$  \hspace{1cm} (13)$$

Having calculated the SIC$_i$ for each model($i$) from Eq. 10, we ‘weight in evidence’ in favor of the model $i$ being the best approximating model out of $M$ total models. We construct ‘exponential’ weights as is common in the literature such as Cheng and Hansen (2012) based on the original work on Bayesian Model Averaging of Min and Zellner (1992):

$$w_i = \frac{\exp(-\frac{1}{2}SIC_i)}{\sum_{m=1}^{M} \exp(-\frac{1}{2}SIC_i)}$$  \hspace{1cm} (14)$$

where $M = 1,984$. For the SIC case, assuming diffuse priors and equal model prior probabilities, Eq. 14 can be interpreted as approximate Bayesian Model Average (BMA) weights. Our four weighted forecasts are created from: 1.) an approximate BMA across all FAVAR models, 2.) An approximate BMA across all FECM/FECMc models, 3.) An approximate BMA across both FAVAR and FECM/FECMc models and 4.) a uniform weight across all estimated models. Weights allocated to each of the 1984 models, summed across recursions and countries, but split between real and nominal house prices, can be seen in Figure 6. Interesting points to note in Figure 6 are that for both real and
nominal indexes, it is the final element on the x-axis which is allocated least weight (Model 1984) - and this is the least parsimonious of all models: the full FECMc model (with all ‘observable’ variables, four I(1) factors, a cyclical factor and the maximum number of lags \(k = 4\)). The models allocated the most weight are FECMs with house prices and real residential investment and three factors, a cyclical factor and one lag (Model 1269) for nominal house prices, and two factors and two lags (Model 1277) for real house prices. A summary of all 12 forecasts generated can be seen in Section 5. The fact that models with fewer parameters are assigned higher weights than the fully parameterized model is no surprise given the well documented properties of the information criterion being used.

4.7 Analysis

Models are re-estimated and re-specified for 28 recursions (separately for each country) up until 2011q2, at each recursion generating multi-step forecasts, with the last forecast occurring at the final recursion for \(h=8\) in 2013:Q3. Most academic textbooks on the subject recommend the use of the MAPE (such as Hanke and Reitsch (1986)), and this is the metric we use in this paper. Specifically, we use the absolute percentage error (APE - Eq. 15), the mean absolute percentage error across recursions (MAPE\(_r\) - Eq. 16), or the mean absolute percentage error averaged across recursions and horizons (MAPE\(_{r,h}\) - Eq. 17). For \(h=1,\ldots,8(H)\) and \(r=1,\ldots,28 (R)\):

\[
APE_{T+r+h} = \left| \frac{\hat{y}_{T+r+h} - y_{T+r+h}}{y_{T+r+h}} \right|
\]

\[
MAPE^r_{T+R+h} = \frac{\sum_{T+R+h}^{T+R+h} |\hat{y}_{T+r+h} - y_{T+r+h}|}{R}
\]

\[
MAPE^{r,h}_{T+R+H} = \frac{\sum_{T+R+H}^{T+R+H} \sum_{T+R+h}^{T+R+h} |\hat{y}_{T+r+h} - y_{T+r+h}|}{R \times H}
\]
Which of Eq. 15-17 is used where in Section 5 is discussed below.

5 Results

In Tables 1 and 2, we recap the 12 forecasts generated which are described above and the results from which are plotted in Figures 7 - 13.

Table 1: Forecasts from Classic Models and Vanilla Factor Models

<table>
<thead>
<tr>
<th>1. ARIMA</th>
<th>2. VAR</th>
<th>3. VECM</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. FAVAR - IPC/SIC</td>
<td>5. FECM - IPC/SIC</td>
<td>6. FECMe - IPC/SIC</td>
</tr>
</tbody>
</table>

Table 2: Forecasts from Candidate Model Set

<table>
<thead>
<tr>
<th>7. FAVAR - SIC</th>
<th>8. FECM - SIC</th>
<th>9. FAVAR - BMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>10. FECM - BMA</td>
<td>11. All Models - Uniform</td>
<td>12. All Models - BMA</td>
</tr>
</tbody>
</table>

5.1 Forecast Error

Figures 7 and 8 (below) display the MAPE\(^r\) metric (Eq. 16) for each of the eight horizons (one to eight quarters ahead) averaged across each of the recursions, displayed across each country and then for each of real and nominal house prices.\(^9\)

9 The height of the column divided by 8 represents the MAPE\(^r,h\) metric.
Figure 8: Forecast Error: Nominal House Prices

Considering the performance of specific models, one consistent feature is that Model 9 - the approximate BMA weighting across all FAVAR models consistently performs worst under this metric. Classical VAR models consistently provide the lowest average forecast errors compared to their counterparts - with lower average RMSEs than both the VECM (11 of 12 times) and the ARIMA (10 of 12 times). Concerning which models are able to generate consistently low average forecast errors (other than the VAR) are the forecasts generated from both uniform and BMA weights across the entire candidate model sets (Models 11 and 12) respectively. Their equivalent performance is a function of the similar weights placed on each of the 1,984 models in the candidate model set by both the BMA and the uniform procedures. This is due to the fact that the SICs are not sufficiently different to generate large differences in $w_i$ when the denominator of Eq. 14 is large, and this is emphasized in Figure 6 above.

There is a significant difference in magnitude in forecast error between countries. As may have been expected, house price indexes with higher variance (e.g. the United Kingdom) typically display a larger average forecast error compared to lower variance indexes (e.g. Japan). This relationship is plotted in Figure 9, which shows the average forecast errors for each model for each country (referenced as per Tables 1 and 2) on the y-axis, with the x-axis displaying the variance of the underlying series which it is forecasting on the x-axis. Figure 9 displays forecast errors averaged over both recursion and horizon (Eq. 17), and allows further visualization of the large errors generated by Model 9, and the consistently low forecast errors of the classical VAR model (Model 3) and the model weights over the candidate set (Model 11 and Model 12).

As may be predicted by econometric theory, one recurring observation is that forecast error is increasing in the horizon - average forecast error is several times larger in

---

10In the order of highest to lowest variances for real house prices: Australia, United Kingdom, Sweden, Canada, Japan, US, with the orders reversed of the final two for nominal house prices.
Figure 9: Forecast Error and Series Variance

(a) Real Prices

(b) Nominal Prices

Figure 10: Forecast Error and Horizon
magnitude at the two year horizon than the one quarter horizon. Figure 10 uses exactly the same data (forecast error matrices) to visualize the results in a different way, plotting forecast error of each of the twelve models as the horizon increases on the x-axis for each country for real and nominal series (again averaged over recursions). The plots again provide good evidence for preferring the forecasts weighted across the candidate model sets, regardless of horizon, as well as emphasizing the poor performance of Model 9 (and Model 3-9 more generally).

5.2 Binary Measures of Outperformance

Figure 11: Beat All Other Models

The number of times that a point forecast from one model has the lowest APE (Eq. 15) out of the twelve, per recursion, per country is displayed in Figure 11. The most interesting point to note is the strong performance of Model 9 (FAVARs weighted using approximate BMA), in comparison to the above analysis. This forecast performs best 11.61% and 13.61% of the time in real and nominal forecasts respectively (against a prior uniform expectation of 8.33%), with the strongest performance coming in the first four horizons (i.e. 16.07% at the $h=3$ horizon for real, and 14.88% for nominal models). Again in contrast to the above, the weakest performance under this metric comes from the forecasts created through model averaging across the entire candidate set. This may be unsurprising given that by their very nature, weighted forecasts have lower variance, and are less likely to produce an outlying forecast which captures movements (such as turning points) in markets such as housing.

Turning to individual level model comparisons, summing across recursions, countries and real/nominal indices, calculations show that the APE is lower for classical VAR models than VECM models 62.61% of the time. In addition to this, FECMc models outperform FECM models 52.16% of the time, although the FECMc is only able to outperform FAVAR models 42.86% of the time. The FAVAR BMA outperforms the FAVAR SIC 36.64%, and the FECM BMA 48.69%, pointing to the superiority of the SIC selection technique across the set of candidate models over an approximate BMA method. When comparing this SIC selection technique across the set of candidate models to the
more conventional approach of augmenting a set of observable variables with factors (as in Bernanke et al. (2005) or Banerjee et al. (2014)), the SIC selection method performs best only 46.76% of the time for FAVARs, but 59.82% of the time for FECMs. It may be considered surprising that the SIC selection method does not perform similarly well for the FAVARs, given that Models 4-6 are nested, and are able to be selected from the set of candidate models. Finally, the BMA across the entire model set outperforms our uniform weighting scheme 46.71% of the time, lending support to the naïve forecast combination puzzle widely discussed in the literature (such as Smith and Wallis (2009)).

5.2.1 Beating the ARIMA Benchmark

One of the most important binary output metrics is the ability of our classical, factor, or candidate model set to outperform a standard univariate benchmark (Model 1). Again in this metric, the classical VAR model performs admirably, outperforming the univariate benchmark 60.57% and 58.63% for real and nominal indexes respectively. Considering forecasts from our factor models, the uniform and BMA weights across the entire candidate sets again perform similarly well, outperforming the ARIMA benchmark 55.21% and 55.069% for nominal indexes, and 57.29% and 57.22% for real indexes respectively.

Figure 12: Beating the ARIMA Benchmark

(a) Real Prices

(b) Nominal Prices

As shown in Table 3, further analysis indicates similarity with the findings of Banerjee et al. (2005), which uses a similar model, finding that results of a factor model against an AR counterpart depend on the underlying (country-level) datasets. Table 3 displays the percent of times the FAVAR model (Model 4) of Bernanke et al. (2005) and the FECMc model (Model 6) of Banerjee et al. (2014) can outperform the Hyndman and Khandakar (2008) algorithm:

5.2.2 Real Vs. Nominal House Price Forecasts

The percentage of times real house price forecasts are more accurate than their nominal equivalent is shown in Figure 13. The amount of times that forecasts for real house prices are more accurate is: 46.58%, 52.34%, 55.51%, 40.14%, 39.14%, 44.98%, for the

24
U.K., U.S., Australia, Japan, Canada and Sweden respectively (for an average across all models, horizons, recursions and countries of 46.44%). The observation is indicative of the fact that while real house prices may be less volatile (as indicated on the x-axis of Figure 9), it is more prone to unexpected turning points. The results are fairly homogenous across model class, although the two models which perform better for real models than nominal are the ‘vanilla’ FAVAR (Model 4) and the model selected FAVAR (Model 7).

5.2.3 Direction of Change

Figure 14 displays a simple ‘direction of change’ metric: if a forecast predicts the index will increase, and it increases, or if the forecast predicts a downturn for that period, and the index decreases, the metric is satisfied, irregardless of magnitude of change. The initial comment to make is the extremely large success rate of all twelve models for nominal indexes at the h=1 horizon. This is due to the fact that for our six nominal indexes, there are a very small amount of turning points (which our autoregressive models are not well suited to detect). The nature of the models in question is that they will create forecasts with strong autocorrelation in sign, and this metric provides a way for assessing their performance in this regard. Again, our final two combination models (Model 11 and Model 12) consistently perform well, frequently featuring as one of the best models under this criterion. In addition to this, Model 4 (the ‘vanilla’ FAVAR) performs admirably and again, Model 9 (BMA FAVAR) performs poorly. As may be expected, the univariate benchmark also performs poorly under this evaluation measure.
6 Conclusion

In this paper we develop a computationally intensive forecasting algorithm to forecast both real and nominal aggregate national house prices over a range of different developed countries using six original large datasets. Extracting a maximum of five stationary and four non-stationary and one stationary cyclical factor, we estimate all possible combinations of factors and key ‘observable’ variables in a small macroeconometric model for house prices in unrestricted (FAVAR) and restricted (FECM) models and compare the results of classical autoregressive models against a range of model selection and model weighting algorithms.

The results show that if our objective lies in minimizing a the total forecast error, then using either an approximate BMA or a naïve weight of every model in our candidate set (which contains every combination of ‘observable’ endogenous variables and factors) is generally able to perform well. If, however, rather than comparing forecast errors, we are interested in a model which generates the absolute lowest forecast error against all of its competitors, it is an approximate BMA weighting over all FAVAR models which, while unable to generate low forecast errors on average, is able to generate the absolute lowest out of all its counterparts most frequently. We are able to show the ability of our forecasts weighted across the total model set to consistently beat an automated univariate benchmark, and also how results under this metric vary across countries, as indicated by earlier literature. We are also able to provide other results, such as the forecast error increasing in the volatility of the underlying series, and the horizon. We are able to show results regarding direction of change and turning points, and that our forecasts are systematically more accurate for nominal, as opposed to real series.

While Model 11 and Model 12 provide reasonable evidence of the ability of factor based
models to perform well, there are two major reasons why models 4-10 do not perform better. The first of which is potentially due to the quality of the underlying data which is utilized for factor extraction. While datasets such as the one compiled from US aggregates are typical in the literature, few studies have focused on Sweden, Canada or Australia due to a shortage of the series necessary to generate factors which fully capture common unobserved forces. A further point to note is that while all of our models (excluding the uniform average) are consistently identified in some way using the SIC, this is only one of several potential methods of both specification and weighting, and others may provide results more favorable for factor based models. Recent literature such as Chen et al. (2014) has also focused on time variance of factor loadings (parameter shifts which are not local to zero). Further research would draw inspiration from earlier papers such as Banerjee et al. (2008) who claim that it is the instability of factor loadings which may cause factor-based forecasts to fail to outperform classical models.

One avenue for further research is to obtain and develop more ‘real-time’ ‘big datasets’ for forecasting purposes and then to estimate and compare a similar specification/combination methodologies with the ones presented in this paper. Another possibility would be to instead estimate the models as ‘direct’ rather than ‘iterated’ forecasts. Other interesting recent developments in the field could lead to applications of time varying parameter factor (such as TVP-FAVAR) models being applied to house price forecasting, with the potential for a further reduction in forecast error.

Acknowledgements

The author would like to thank Dr. William Pouliot, Professor Anindya Banerjee, Dr. Marco Barassi and Professor John Fender for their discussions. In addition to this, comments from conference and seminar participants at the Universities of Birmingham, West England and Nottingham and participants at the ESRC Festival of Science are also gratefully acknowledged.

References


