University of Pretoria
Department of Economics Working Paper Series

Forecasting Real US House Price: Principal Components versus Bayesian Regressions

Rangan Gupta
University of Pretoria
Alain Kabundi
University of Johannesburg
Working Paper: 2009-07
February 2009

Department of Economics
University of Pretoria
0002, Pretoria
South Africa
Tel: +27 12 420 2413
Fax: +27 12 362 5207

FORECASTING REAL US HOUSE PRICE: PRINCIPAL COMPONENTS VERSUS BAYESIAN REGRESSIONS

Rangan Gupta[*] and Alain Kabundi[#]

Abstract

This paper analyzes the ability of principal component regressions and Bayesian regression methods under Gaussian and double-exponential prior in forecasting the real house price of the United States (US), based on a monthly dataset of 112 macroeconomic variables. Using an in-sample period of 1992:01 to 2000:12, Bayesian regressions are used to forecast real US house prices at the twelve-months-ahead forecast horizon over the out-of-sample period of 2001:01 to 2004:10. In terms of the Mean Square Forecast Errors (MSFEs), our results indicate that a principal component regression with only one factor is best-suited for forecasting the real US house price. Amongst the Bayesian models, the regression based on the double exponential prior outperforms the model with Gaussian assumptions.

Journal of Economic Literature Classification: C11, C13, C33, C53.
Keywords: Bayesian Regressions; Principal Components; Large-Cross Sections .

1. Introduction

This paper analyzes the ability of Bayesian regression methods under Gaussian and double-exponential prior in forecasting the real house price of the United States (US), based on a monthly dataset of 112 macroeconomic variables. Using an in-sample period of 1992:01 to 2000:12, Bayesian regressions are used to forecast real US house prices at the twelve-months-ahead forecast horizon over the out-of-sample period of 2001:01 to 2004:10. The forecast performance of the Bayesian regressions are then compared in terms of the Mean Square Forecast Errors (MSFEs) with the forecasts generated from the principal component regression, based on the same dataset of 112 variables. Our choice of the two Bayesian priors is motivated from the recent contribution by De Mol *et al.* (2008), and corresponds to the two interesting cases of variable aggregation and variable selection.[1]

With the methodologies in place, two questions arise immediately. First, why is forecasting real house price important? And second, why use large-scale models for this purpose? As far as the answer to the first question is concerned, the importance of predicting house price is motivated by a set of recent studies which conclude that asset prices help forecast both inflation and output (Forni *et al.*, 2003; Stock and Watson, 2003, Gupta and Das, 2008a,b and Das *et al.*, 2008a,b). Since a large amount of individual wealth is imbedded in houses, similar to other asset prices, house price movements are thus important in signaling inflation. Models that forecast real house price can give policy makers an idea about the direction of overall price level and, hence, economy-wide inflation in the future, and thus, can provide a better control for designing of appropriate

[*] To whom correspondence should be addressed. Associate Professor, University of Pretoria, Department of Economics, Pretoria, 0002, South Africa, Email: Rangan.Gupta@up.ac.za. Phone: +27 12 420 3460, Fax: +27 12 362 5207. We are grateful to De Mol et al. (2008) for making the replication files available publicly. A special thanks to Domenico Giannone for many helpful comments regarding the implementation of the codes.
[#] Senior Lecturer, University of Johannesburg, Department of Economics, Johannesburg, 2006, South Africa, Email: akabundi@uj.ac.za. Phone: +27 11 559 2061, Fax: +27 11 559 3039.
[1] See Section 2 for further details.

policies. In addition, given that movements in the housing market are likely to play an important role in the business cycle (Iacoviello and Neri, 2008), not only because housing investment is a very volatile component of demand (Bernanke and Gertler, 1995), but also because changes in house prices tends to have important wealth effects on consumption (International Monetary Fund, 2000) and investment (Topel and Rosen, 1988), and hence, the importance of forecasting house price is vital. The housing sector thus plays a significant role in acting as a leading indicator of the real sector of the economy, and predicting it correctly cannot be overemphasized, especially in the light of the recent credit crunch in the U.S. that started with the burst of the housing price bubble which, in turn, transmitted to the real sector of the economy driving it towards an imminent recession.

The rationale for using large-scale models to forecast real house price emanates from the fact that a large number of economic variables help in predicting real housing price (Cho, 1996; Abraham and Hendershott, 1996; Johnes and Hyclak, 1999; and Rapach and Strauss, 2007, 2008). For instance, income, interest rates, construction costs, labor market variables, stock prices, industrial production, consumer confidence index – which are amongst the 112 monthly series used by the models – act as potential predictors.

To realize the contribution of this study, it is important to place this paper in the context of current research that focusing on forecasting in the housing market. In this regard, few studies are worth mentioning: Rapach and Strauss (2007) used an autoregressive distributed lag (ARDL) model framework, containing 25 determinants, to forecast real housing price growth for the individual states of the Federal Reserve's Eighth District. Given the difficulty in determining *apriori* particular variables that are most important for forecasting real housing price growth, the authors also use various methods to combine the individual ARDL model forecasts, which result in better forecast of real housing price growth. Rapach and Strauss (2008) do the same for 20 largest US states based on ARDL models containing large number of potential predictors, including state, regional and national level variables. Once again, the authors reach similar conclusions as far as the importance of combining forecasts are concerned. On the other hand, Gupta and Das (2008b), look into forecasting the recent downturn in real house price growth rates for the twenty largest states of the US economy. In this paper, the authors use Spatial BVARs, based merely on real house price growth rates, to predict their downturn over the period of 2007:01 to 2008:01. They find that, though the models are quite well-equipped in predicting the recent downturn, they underestimate the decline in the real house price growth rates by quite a margin. They attribute this underprediction of the models to the lack of any information on fundamentals in the estimation process.

Given that in practice, forecasters and policymakers often use information from many series than the ones included in smaller models, like the ones used by Rapach and Strauss (2007, 2008), who also indicate the importance of combining forecast from alternative models, the role of a large-scale models cannot be ignored. In addition, one cannot condone the fact that the main problem of small models, as seen from the studies by Rapach and Strauss (2007, 2008), is in the decision regarding the choice of the correct potential predictors to be included. Due to this reason, Vargas-Silva (2008) and Gupta and Kabundi (2009a,b) uses Factor Augmented Vector Autoregression (FAVAR) models containing large number of macroeconomic variables in analyzing the impact of monetary policy shocks on the housing sector of the United States and South Africa. To the best of our knowledge, this is the first attempt to look into the ability of Bayesian and principal component regressions in forecasting real house price in the US.

In such a backdrop, our paper can thus be viewed as an extension of the abovementioned studies, in the sense that we use large-scale models that allow for the role of a wide possible set of fundamentals to affect the housing sector. The remainder of the paper is organized as follows: Section 2 lays out the basics of the alternative models. In Section 3 we discuss the data and evaluate the forecasting performances of the various models, and finally, Section 5 concludes.

2. The Models[2]

Consider the $(n \times 1)$ vector of covariance-stationary processes $Z_t = (z_{1t}, ..., z_{nt})'$. It will be assumed that they all have a mean of zero and a variance of unity. We are interested in forecasting linear transformations of some element(s) of Z_t based on all the variables as possible predictors. Formally, we are interested in estimating the linear projection:

$$y_{t+h/t} = proj\{y_{t+h} / \Omega_t\}$$

where $\Omega_t = span\{Z_{t-p}, p = 0,1,2,...\}$ is a potentially large time t information set and $y_{t+h} = z^h_{i,t+h} = f_h(L) z_{i,t+h}$ is a filtered version of z_{it}, for a specific i.

Traditionally, time series models approximate the projection using only a finite number, p, of lags of Z_t. In particular, we generally consider the following regression:

$$y_{t+h} = Z_t' \beta_0 + ... + Z_{t-p}' \beta_p + u_{t+h} = X_t' \beta + u_{t+h}$$

where $\beta = (\beta_0', ..., \beta_p')$ and $X_t = (Z_t', ..., Z_{t-p}')'$.

Given a sample of size of T, we will denote by $X = (X_{p+1}, ..., X_{T-h})'$ the $(T-h-p) \times n(p+1)$ matrix of observations for the predictors and by $y = (y_{p+1+h,...,} y_T)'$ the $(T-h-p) \times 1$ matrix of the observations on the dependent variable. The regression coefficients are generally estimated by Ordinary Least Squares (OLS), $\hat{\beta}^{LS} = (X'X)^{-1} X'y$, and the forecast is given by $\hat{y}^{LS}_{T+h/T} = X_T' \hat{\beta}^{LS}$. Naturally, when the size of the information set, n, is large, such projection involves the estimation of a large number of parameters. This leads to loss of degrees of freedom and large out-of-sample forecast errors. Besides, OLS is not feasible when the number of regressors is larger than the sample size, i.e., $n(p+1) > T$.To solve this problem of curse of dimensionality, the method that has been considered in the literature is to compute the forecast as a projection on the first few principal components (Stock and Watson, 2002a, b; Forni et al., 2005; Giannone et al. 2004).

Consider the spectral decomposition of the sample covariance matrix of the regressors:

$$S_x V = VD \qquad (1)$$

where $D = diag(d_1, ..., d_{n(p+1)})$ is a diagonal matrix, with the diagonal elements constituted of the eigenvalues of $S_x = \frac{1}{T-h-p} X'X$ in decreasing order of magnitude and

[2] This section relies heavily on the discussion available in De Mol et al. (2008), and, also retains their symbolic representations.

$V = \left(v_1, ..., v_{n(p+1)}\right)$ is the $n(p+1) \times n(p+1)$ matrix whose columns are the corresponding eigenvectors[3]. Given this, the normalized principal components (PC) are defined as:

$$\hat{f}_{it} = \frac{1}{\sqrt{d_i}} v_i' X_t \qquad (2)$$

for $i = 1, ..., N$, where N is the number of non zero eigenvalues[4].

If there is limited cross-correlation among the specific components of the data and, if most of the interactions amongst the variables in the information set emerge due to few common factors, the information contained in the large date set can be captured by few aggregates. While, the part not explained by the common factors can be predicted by means of traditional forecasting methods. In such instances, few principal components, $\hat{F}_t = \left(\hat{f}_{1t}, ..., \hat{f}_{rt}\right)$ with $r \ll n\left(p+1\right)$, are likely to provide a good approximation of the underlying factors.

Assuming for the sake of simplicity, that no lags of the dependent variable are required as additional regressors, the principal component forecast is defined as:

$$y^{PC}_{t+h/t} = proj\left\{y_{t+h} / \Omega^F_t\right\} \approx proj\left\{y_{t+h} / \Omega_t\right\} \qquad (3)$$

where $\Omega^F_t = span\left\{\hat{F}_t, \hat{F}_{t-1}, ...,\right\}$ is a parsimonious representation of the information set. Given the parsimonious approximation, the projection is now feasible, since it requires the estimation of a limited number of parameters. Under assumptions defining an approximate factor structure,[5] once common factors have been estimated via principal components, the projection is computed by OLS by treating the estimated factors as observable variables.

On the other hand, the Bayesian approach imposes limits on the length of β through priors and estimate the parameters as the posterior mode. Hence, here the parameters are used to compute the forecasts. As in De Mol *et al.* (2008), we also consider two alternative prior specifications, namely, Gaussian and double exponential priors.
Under the Gaussian prior, $u_t \sim i.i.d.N\left(0, \sigma^2_u\right)$ and $\beta \sim N\left(\beta_0, \Phi_0\right)$,and assuming for simplicity, that $\beta_0 = 0$, we have:

$$\hat{\beta}^{bay} = \left(X'X + \sigma^2_u \Phi_0^{-1}\right)^{-1} X'y .$$

The forecast then is computed as:

$$\hat{y}^{bay}_{T+h/T} = X'_T \hat{\beta}^{bay}$$

---

[3] The eigenvalues and eigenvectors are typically computed on $\frac{1}{T-p}\sum_{t=p+1}^{T} X_t X_t'$ (see Stock and Watson, 2002a). We follow De Mol *et al.* (2008) in computing them on $\frac{1}{T-h-p} X'X = \frac{1}{T-p-h}\sum_{t=p+1}^{T-h} X_t X_t'$ for comparability with other estimators considered in the paper.

[4] Note that $N \leq \min\left\{n(p+1), T-h-p\right\}$.

[5] See Section 3 for further details.

When the parameters are independently and identically distributed, i.e., $\Phi_0 = \sigma_\beta^2 I$, the estimates are equivalent to those produced by penalized Ridge regression with parameter $v = \frac{\sigma_u^2}{\sigma_\beta^2}$ [6]. Formally[7]:

$$\hat{\beta}^{bay} = \arg\min_{\beta} \left\{ \left\| y - X\beta \right\|^2 + v \left\| \beta \right\|^2 \right\}.$$

OLS, principal components regression and Gaussian Bayesian regression tends to weight all variables.[8] An alternative to this is to select variables. Under Bayesian regression, one can use a double exponential prior to do so, which, when uses a zero mean i.i.d. prior, is equivalent Lasso regression (least absolute shrinkage and selection operator). In this particular case, the method can also be seen as a penalized regression with a penalty on the coefficients involving the L_1 norm instead of the L_2 norm. Specifically:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \left\| y - X\beta \right\|^2 + v \sum_{i=1}^{n} \left| \beta_i \right| \right\} \qquad (4)$$

where $v = \frac{1}{\tau}$ where τ is the scale parameter of the prior density[9].

In comparison with the Gaussian density, the double-exponential puts more mass near zero and in the tails, which, in turn tends to produce coefficient estimates that are either large or zero. As a result, one often favors the recovery of a few large coefficients instead of many fairly small ones. Moreover, the double-exponential prior favors *sparse* regression coefficients (sparse mode), since it favors truly zero values instead of small ones.

In the case with non orthogonal regressors, the Lasso solution enforces sparsity on the variables rather than on the principal components, which implies a regression on few observables rather than on few linear combinations of the variables. Unfortunately, in the general case, the maximizer of the posterior distribution has no analytical form and has to be computed based on numerical methods. Following De Mol et al. (2008), we use the Least Angle Regression (LARS) algorithm developed recently by Efron et al. (2004) for this purpose.

The next section will consider the empirical performance of the three methods discussed in an out-of-sample forecast exercise based on a large panel of time series.

3. Data and Results:

The data set employed for the out of sample forecasting analysis is the same as the 111 major macroeconomic variables used by Boivin *et al.* (2008). With this data set ending at 2005:10, the endpoint of our sample is automatically determined. The data set contains a broad range of macroeconomic variables, such as industrial production, income,

[6] Though, homogenous variance and zero mean are too simplistic of assumptions, they are justified by the fact that the variables in the panel are standardized and demeaned. Note, this transformation is obvious to allow for comparison with principal components.

[7] $\|.\|$ denotes the L^2 matrix norm, i.e. for every matrix A, $\|A\| = \sqrt{\lambda \max(A'A)}$. For vectors it corresponds to the Euclidean norm.

[8] See De Mol et al. (2008) for further details.

[9] Recall that the variance of the prior density is proportional to $2\tau^2$.

employment and unemployment, housing starts, inventories and orders, stock prices, exchange rates, interest rates, money aggregates, consumer prices, producer prices, earnings, and consumption expenditure. As far as the US house price is concerned, the nominal house price figures were obtained from the Office of Federal Housing Enterprise Oversight (OFEO), and were converted to their real counterpart by dividing them with the personal consumption expenditure deflator. So, in total we have a balanced panel of 112 monthly series for the period running from 1991:01 to 2005:10. A full description of the dataset has been provided in the appendix of the paper.

Series are transformed to induce stationarity. In general, following, De Mol et al. (2008), all real variables, such as employment, industrial production, sales and the real US house price, we take monthly growth rate. While for series that are already expressed in rates, such as the unemployment rate, capacity utilization, interest rate and some surveys, we take first differences. Finally, fro nominal prices and wages, we take the first differences of their annual rates.

Defining *HP* as the monthly real US house price, the relevant variable that we forecast is: $z^{h}_{HP,t+h} = (hp_{t+h} - hp_{t}) = z_{HP,t+h} + ...z_{HP,t+1}$, where $hp_{t} = 100 \times log(HP_{t})$. The forecasts for the *log(HP)* is then recovered as: $hp^{F}_{T+h|T} = z^{h}_{HP,T+h|T} + hp_{T}$. The accuracy of the forecasts is evaluated using the mean-square forecast error (MSFE), given by: $MSFE^{h}_{hp} = \frac{1}{T_1 - T_0 - h + 1} \sum_{T=T_0}^{T_1-h} (hp^{F}_{T+h|T} - hp_{T+h})^2$.

The sample has a monthly frequency ranges from 1991:01 to 2005:10, with the starting point of the sample determined by the availability of monthly US house price. The out-of-sample period is 2001:01 to 2004:10, with data between 1992:01 and 2000:12 serving as the in-sample for the analysis, i.e., T_0 =2000:12. The last available time point is T_1 =2005:10. We consider rolling estimates with a window of 9 years. In other words, parameters are estimated at each time *T* using the most recent 9 years of data.[10]. All the procedures have been applied to standardized data, and, hence, mean and variance have been re-attributed to the forecasts accordingly. Following De Mol et al. (2008), the results for *h* = 12, under the principal components regression, and the Bayesian regressions under the Gaussian and double-exponential priors have been reported in Tables 1 through 3, respectively. We compare across the three models, and can draw the following conclusions, based on the MSFE relative to the random walk, and the variance of the forecasts relative to the variance of the actual data for real US house price:

(i) Principal Component Regression: Let us start with the principal component regression, where the results have been reported for the choice of *r* = 1, 3, 5, 10, 25, 50 and 75. Note when *r* = 0, we have the random walk model with drift on the log of HP, while, when *r* = *n*, we have the OLS model. As in De Mol et al. (2008), we only report results for *p* = 0, since this is the case for which the theory has been developed and is typically what is considered in standard macroeconomic applications. Results in Table 1 show principal components improve a lot over the random walk model, especially for *r* =1 and 10. While, for r = 3 it is nearly as good as the random walk model. But beyond *r* = 10, i.e., the advantage is lost, due to a possible loss in parsimony. Moreover, beyond *r*

[10] The choice of 9 years as the rolling-sample ensures that or out-of-sample horizon starts at 2001:01, but at the same time, this also allows us to use the maximum amount of data available for the in-sample analysis.

equal to 10 and beyond, the variance of the forecasts become larger than the series itself. As pointed out by De Mol et al. (2008), this can be explained by the large uncertainty of the regression coefficients when we have a large number of explanatory aggregates. Overall, a principal component model with one regressor is best suited in forecasting real US house price relative to the random walk model, not only because it produces the minimum MSFE relative to the random walk model , but also because it results in lower variance for the forecasts relative to the original series;

(ii) Bayesian (Ridge) Regression with Gaussian Prior: For comparability with the principal component regression, we focus on the case $p = 0$ also for the Bayesian regression, which implies that we do not consider any lags of the regressor. For the Bayesian regression under Gaussian prior, we run the regression using the first estimation period 1991 to 2000 for a grid of priors. Following De Mol et al. (2008), we then choose the priors which causes the in-sample fit to explain a given fraction 1- κ of the variance of the real US house price. We report the results for the different values of κ and υ, the latter kept fixed for the whole out-of-sample horizon. Note $\kappa = 0$ corresponds to a case where the prior is quite uninformative and would be very close to the OLS model, while, $\kappa = 1$ implies the random walk case. Based on results reported in Table 2, the ridge regression performs better than the random walk model for all values of κ beyond 0.1, but especially, well for values the same between 0.3 and 0.5, which, in turn, are associated with shrinkage parameters between thrice and ten times the cross-sectional dimension, *n*. However, the minimum MSFE of the Bayesian regression under the Gaussian prior relative to the MSFE of the random walk model is more than twice of the minimum obtained under the principal component regression with $r = 1$. However, the forecasts produced by the Ridge regressions are generally smoother than the principal component forecasts. Moreover, the principal component and the Ridge forecasts, as seen from the last line of Table 2, are highly correlated. Though, it is not the case that the correlation is maximal for priors giving the best forecasts, indicative of the fact that there does not exists a common explanation for the performance of the two methods;
(iii) Bayesian Regression with Double Exponential Prior: Finally, we consider the case of double-exponential priors. As in De Mol et al. (2008), instead of fixing the values of the parameter υ, a prior is selected that delivers a given number, say *k*, of non-zero coefficients at each estimation step in the out-of-sample period. We look at the cases of k = 1, 3, 5, 10, 25, 50, and 75 non-zero coefficients. Results reported in Table 3, show that good forecasts relative to the random walk model are obtained with predictors between 1 to 5, with the best being for the case of $k = 3$, which though is about 1.7 times more than the minimum obtained under the principal component regression. As far as correlation with principal component forecast is concerned for $k = 3$, the value is second-highest. Variance of the forecasts relative to the original data increases as the number of predictors increases, but, never exceeds the latter. Note the four variables selected for $k \approx 3$ at the beginning and at the end of the out-of-sample period have been reported in the last column of the Table A.2 describing the data in the appendix A. Three of the four variables selected relate to the housing market, namely housing start in the north-east, total new private housing authorized and mobile homes, with the former two being picked up both at the beginning and end of the forecast evaluation period, and the third one only appearing at the end of the out-of-sample horizon. The fourth variable, namely, the spread between the 10-year Treasury bonds yield and the Federal funds rate, is picked up at the beginning of the forecast evaluation period. Overall, these results tend to suggest the importance of the leading indicators related to the housing market, besides the long-term interest rate spread, as major determinants of the real US house price.

[INSERT TABLES 1 THROUGH 3]

5. Conclusions

This paper analyzes the ability of principal component regressions and Bayesian regression methods under Gaussian and double-exponential prior in forecasting the real house price of the United States (US), based on a monthly dataset of 112 macroeconomic variables. Using an in-sample period of 1992:01 to 2000:12, the alternative regressions are used to forecast real US house prices at the twelve-months-ahead forecast horizon over the out-of-sample period of 2001:01 to 2004:10. In summary, based on the 12-months-ahead forecast over the out-of-sample horizon of 2001:01 to 2004:10 and the MSFE relative to the random walk model, we can conclude that the principal component model with only one factor is best suited in forecasting the real US house price relative to the Bayesian regressions based on Gaussian and double-exponential priors. Within the two-types of Bayesian regressions, the Lasso forecasts with three non-zero coefficients tends to outperform the best-performing ridge-regression forecasts obtained under a shrinkage parameter of nearly six times the size of the cross-section.

Recent works by Banbura et al. (2008) and Gupta and Kabundi (2008a,b) have indicated that large-scale Bayesian Vector Autoregressions (LBVARs) tends to outperform Factor-Augmented VARs (FAVARs) in forecasting key macroeconomic variables. In such a backdrop, future research would be aimed at analyzing the ability LBVARs in forecasting house prices.

References

Abraham, J.M., & Hendershott, P.H. (1996). Bubbles in Metropolitan Housing Markets. Journal of Housing Research, 7(2), 191–207.

Bernanke, B., & Gertler, M. (1995). Inside the Black Box: the Credit Channel of Monetary Transmission. Journal of Economic Perspectives, 9(4), 27–48.

Boivin, J., Giannoni, M., & Mihov, I. (2008). Sticky Prices and Monetary Policy: Evidence from Disaggregated U.S. Data. Forthcoming American Economic Review.

Banbura, M., Giannone, D. & Reichlin, L. (2008). Large Bayesian VARs. Forthcoming Journal of Applied Econometrics.

Cho, M. (1996). House Price Dynamics: A Survey of Theoretical and Empirical Issues. Journal of Housing Research, 7(2), 145–172.

Das, S., Gupta, R., & Kabundi, A. (2008a). Is a DFM Well-Suited for Forecasting Regional House Price Inflation?" Working Paper No. 85, Economic Research Southern Africa.

Das, S., Gupta, R., & Kabundi, A. (2008b). Could We Have Forecasted the Recent Downturn in the South African Housing Market? Working Paper No. 200831, Department of Economics, University of Pretoria.

De Mol, C., Giannone, D. & Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian regression a valid alternative to principal components?, Journal of Econometrics, 146(2),318-328.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. Annals of Statistics, 32(2), 407–499.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2005). The Generalized Dynamic Factor Model, One Sided Estimation and Forecasting. Journal of the American Statistical Association, 100(471), 830–840.
Forni M., Hallin, M., Lippi, M., Reichlin, L. (2003). Do financial variables help forecasting inflation and real activity in the euro area*?* Journal of Monetary Economics,
Giannone, D., Reichlin, L., & Sala, L. (2004). Monetary Policy in Real Time, in NBER Macroeconomics Annual, ed. by M. Gertler, and K. Rogoff, pp. 161–200. MIT Press.
Gupta, R., & Das, S. (2008a). Spatial Bayesian Methods for Forecasting House Prices in Six Metropolitan Areas of South Africa. South African Journal of Economics*,* 76(2), 298-313.
Gupta, R., & Das, S. (2008b). Predicting Downturns in the US Housing Market. Forthcoming Journal of Real Estate Economics and Finance.
Gupta, R., & Kabundi, A. (2008a). Forecasting Macroeconomic Variables using Large Datasets: Dynamic Factor Model vs Large-Scale BVARs. Working Paper No. 200816, Department of Economics, University of Pretoria.
Gupta, R., & Kabundi, A. (2008b). Forecasting Macroeconomic Variables in a Small Open Economy: A Comparison between Small- and Large-Scale Models. Working Paper No. 200830, Department of Economics, University of Pretoria.
Iacoviello, M., & Neri, S. (2008). Housing Market Spillovers: Evidence from an Estimated DSGE Model. Working Paper No. 659, Boston College Department of Economics.
International Monetary Fund. World Economic Outlook: Asset Prices and the Business Cycle, 2000.
Johnes, G., & Hyclak, T. (1999). House Prices and Regional Labor Markets. Annals of Regional Science, 33(1), 33–49.
Rapach, D.E., & Strauss. J. K. (2008). Differences in Housing Price Forecast ability Across U.S. States. Forthcoming International Journal of Forecasting.
Rapach, D.E., & Strauss, J.K. (2007). Forecasting Real Housing Price Growth in the Eighth District States. Federal Reserve Bank of St. Louis. Regional Economic Development, 3(2), 33–42.
Stock, J.H., & Watson, M.W. (2003). Forecasting Output and Inflation: The Role of Asset Prices. Journal of Economic Literature, 41(3), 788-829.
Stock, J. H., & Watson, M. W. (2002a). Forecasting Using Principal Components from a Large Number of Predictors," Journal of the American Statistical Association, 97, 147–162.
Stock, J.H., & Watson, M.W. (2002b). Macroeconomic Forecasting Using Diffusion Indexes. Journalof Business and Economics Statistics, 20, 147–162.
Topel, R. H., & Rosen, S. (1988). Housing Investment in the United States. Journal of Political Economy, 96(4), 718–740.
Vargas-Silva, C. (2008).The Effect of Monetary Policy on Housing: A Factor Augmented Approach. Applied Economics Letters, 15(10), 749-752.

Table 1. Principal Component Forecasts

	Real US House Price (2001:01-2004:10)						
	Number of Principal Components						
	1	3	5	10	25	50	75
MSFE	**0.382**	0.9927	1.1137	0.5024	1.2403	1.0304	1.2592
Variance*	0.5323	0.5014	0.7336	1.0685	1.0865	1.0832	1.1328

MSFE are relative to Random Walk forecast. *The variance of the forecast relative to the variance of the series.

Table 2: Bayesian Forecasts with Gaussian Prior

	Real US House Price (2001:01-2004:10)								
	In-Sample Residual Variance								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ν	35	146	336	629	1066	1735	2855	5091	11790
MSFE (12-steps)	1.0189	0.8348	0.7905	**0.7893**	0.8072	0.8351	0.8692	0.9082	0.9517
Variance*	0.6827	0.5823	0.5027	0.4468	0.4064	0.3755	0.3502	0.3282	0.3085
Correlation with PC forecasts (r=1)	0.7284	0.8127	0.8614	0.8935	0.9147	0.9285	0.9374	0.9426	0.945

MSFE are relative to Random Walk forecast. *The variance of the forecast relative to the variance of the series.

Table 3: Lasso Forecasts

	Real US House Price (2001:01-2004:10)						
	Number of Non-Zero Coefficients						
	1	3	5	10	25	50	75
MSFE(12-Steps)	0.7367	**0.6557**	0.8048	0.9316	1.1529	1.4734	1.748
Variance*	0.4337	0.5981	0.6345	0.6838	0.7836	0.7894	0.6541
Correlation with PC forecasts (r=1)	0.8932	0.8432	0.7842	0.7367	0.6745	0.6008	0.5552

MSFE are relative to Random Walk forecast. *The variance of the forecast relative to the variance of the series.

APPENDIX

TABLE A1: Data Transformation

	DEFINITION	TRANSFORMATION
1	$x_{it} = z_{it}$	No transformation.
2	$x_{it} = \Delta z_{it}$	Monthly Difference
4	$x_{it} = \ln z_{it}$	Log
5	$x_{it} = \Delta \ln z_{it} \times 100$	Monthly Growth Rate
6	$x_{it} = \Delta \ln \frac{z_{it}}{z_{it-12}} \times 100$	Monthly difference of yearly growth rates

TABLE A2: Data Description

Code	Description	Transf.	HP
a0m052	Personal income (AR, bil. chain 2000 $)	5	
A0M051	Personal income less transfer payments (AR, bil. chain 2000 $)	5	
IPS10	INDUSTRIAL PRODUCTION INDEX - TOTAL INDEX	5	
IPS11	INDUSTRIAL PRODUCTION INDEX - PRODUCTS, TOTAL	5	
IPS299	INDUSTRIAL PRODUCTION INDEX - FINAL PRODUCTS	5	
IPS12	INDUSTRIAL PRODUCTION INDEX - CONSUMER GOODS	5	
IPS13	INDUSTRIAL PRODUCTION INDEX - DURABLE CONSUMER GOODS	5	
IPS18	INDUSTRIAL PRODUCTION INDEX - NONDURABLE CONSUMER GOODS	5	
IPS25	INDUSTRIAL PRODUCTION INDEX - BUSINESS EQUIPMENT	5	
IPS32	INDUSTRIAL PRODUCTION INDEX - MATERIALS	5	
IPS34	INDUSTRIAL PRODUCTION INDEX - DURABLE GOODS MATERIALS	5	
IPS38	INDUSTRIAL PRODUCTION INDEX - NONDURABLE GOODS MATERIALS	5	
IPS43	INDUSTRIAL PRODUCTION INDEX - MANUFACTURING (SIC)	5	
IPS67e	INDUSTRIAL PRODUCTION INDEX - MINING NAICS=21	5	
IPS68e	INDUSTRIAL PRODUCTION INDEX - ELECTRIC AND GAS UTILITIES	5	
IPS307	INDUSTRIAL PRODUCTION INDEX - RESIDENTIAL UTILITIES	5	
IPS316	INDUSTRIAL PRODUCTION INDEX - BASIC METALS	5	
PMP	NAPM PRODUCTION INDEX (PERCENT)	1	
LHEL	INDEX OF HELP-WANTED ADVERTISING IN NEWSPAPERS (1967=100;SA)	2	
LHELX	EMPLOYMENT: RATIO; HELP-WANTED ADS:NO. UNEMPLOYED CLF	2	
LHEM	CIVILIAN LABOR FORCE: EMPLOYED, TOTAL (THOUS.,SA)	5	
LHNAG	CIVILIAN LABOR FORCE: EMPLOYED, NONAGRIC.INDUSTRIES (THOUS.,SA)	5	
LHUR	UNEMPLOYMENT RATE: ALL WORKERS, 16 YEARS & OVER (%,SA)	2	
LHU680	UNEMPLOY.BY DURATION: AVERAGE(MEAN)DURATION IN WEEKS (SA)	2	
LHU5	UNEMPLOY.BY DURATION: PERSONS UNEMPL.LESS THAN 5 WKS (THOUS.,SA)	5	
LHU14	UNEMPLOY.BY DURATION: PERSONS UNEMPL.5 TO 14 WKS (THOUS.,SA)	5	
LHU15	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 WKS + (THOUS.,SA)	5	
LHU26	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 TO 26 WKS (THOUS.,SA)	5	

BLS_P-service EMP	Private Service-providing Employment - Seasonally Adjusted - CES0800000001	5	
BLS_LPNAG	Total Nonfarm Employment - Seasonally Adjusted - CES0000000001	5	
CES002	EMPLOYEES ON NONFARM PAYROLLS - TOTAL PRIVATE	5	
CES003	EMPLOYEES ON NONFARM PAYROLLS - GOODS-PRODUCING	5	
CES006	EMPLOYEES ON NONFARM PAYROLLS - MINING	5	
CES011	EMPLOYEES ON NONFARM PAYROLLS - CONSTRUCTION	5	
CES015	EMPLOYEES ON NONFARM PAYROLLS - MANUFACTURING	5	
CES017	EMPLOYEES ON NONFARM PAYROLLS - DURABLE GOODS	5	
CES033	EMPLOYEES ON NONFARM PAYROLLS - NONDURABLE GOODS	5	
CES046	EMPLOYEES ON NONFARM PAYROLLS - SERVICE-PROVIDING	5	
CES048	EMPLOYEES ON NONFARM PAYROLLS - TRADE, TRANSPORTATION, AND UTILITIES	5	
CES049	EMPLOYEES ON NONFARM PAYROLLS - WHOLESALE TRADE	5	
CES053	EMPLOYEES ON NONFARM PAYROLLS - RETAIL TRADE	5	
CES088	EMPLOYEES ON NONFARM PAYROLLS - FINANCIAL ACTIVITIES	5	
CES140	EMPLOYEES ON NONFARM PAYROLLS - GOVERNMENT	5	
CES151	AVERAGE WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFAR	1	
CES155	AVERAGE WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFAR	2	
BLS_LEHCC	Construction Average Hourly Earnings of Production Workers - Seasonally Adjusted - CES2000000006	5	
BLS_LEHM	Manufacturing Average Hourly Earnings of Production Workers - Seasonally Adjusted - CES3000000006	5	
PMEMP	NAPM EMPLOYMENT INDEX (PERCENT)	1	
HSFR	HOUSING STARTS:NONFARM(1947-58);TOTAL FARM&NONFARM(1959-)(THOUS.,SA	4	
HSNE	HOUSING STARTS:NORTHEAST (THOUS.U.)S.A.	4	**I-II**
HSMW	HOUSING STARTS:MIDWEST(THOUS.U.)S.A.	4	
HSSOU	HOUSING STARTS:SOUTH (THOUS.U.)S.A.	4	
HSWST	HOUSING STARTS:WEST (THOUS.U.)S.A.	4	
HSBR	HOUSING AUTHORIZED: TOTAL NEW PRIV HOUSING UNITS (THOUS.,SAAR)	4	**I-II**
HMOB	MOBILE HOMES: MANUFACTURERS' SHIPMENTS (THOUS.OF UNITS,SAAR)	4	**II**
RHPUS	Real US House Price (SA)	5	
PMI	PURCHASING MANAGERS' INDEX (SA)	1	
PMNO	NAPM NEW ORDERS INDEX (PERCENT)	1	
PMDEL	NAPM VENDOR DELIVERIES INDEX (PERCENT)	1	

PMNV	NAPM INVENTORIES INDEX (PERCENT)	1	
A0M008	Mfrs' new orders, consumer goods and materials (bil. chain 1982 $)	5	
A0M027	Mfrs' new orders, nondefense capital goods (mil. chain 1982 $)	5	
FM1	MONEY STOCK: M1(CURR,TRAV.CKS,DEM DEP,OTHER CK'ABLE DEP)(BIL$,SA)	6	
FM2	MONEY STOCK:M2(M1+O'NITE RPS,EURO$,G/P&B/D MMMFS&SAV&SM TIME DEP(BIL$,	6	
FM3	MONEY STOCK: M3(M2+LG TIME DEP,TERM RP'S&INST ONLY MMMFS)(BIL$,SA)	6	
FM2DQ	MONEY SUPPLY - M2 IN 1996 DOLLARS (BCI)	5	
FMFBA	MONETARY BASE, ADJ FOR RESERVE REQUIREMENT CHANGES(MIL$,SA)	6	
FMRRA	DEPOSITORY INST RESERVES:TOTAL,ADJ FOR RESERVE REQ CHGS(MIL$,SA)	6	
FMRNBA	DEPOSITORY INST RESERVES:NONBORROWED,ADJ RES REQ CHGS(MIL$,SA)	6	
FCLNQ	COMMERCIAL & INDUSTRIAL LOANS OUSTANDING IN 1996 DOLLARS (BCI)	6	
FCLBMC	WKLY RP LG COM'L BANKS:NET CHANGE COM'L & INDUS LOANS(BIL$,SAAR)	1	
CCINRV	CONSUMER CREDIT OUTSTANDING - NONREVOLVING(G19)	6	
FSPCOM	S&P'S COMMON STOCK PRICE INDEX: COMPOSITE (1941-43=10)	5	
FSPIN	S&P'S COMMON STOCK PRICE INDEX: INDUSTRIALS (1941-43=10)	5	
FSDXP	S&P'S COMPOSITE COMMON STOCK: DIVIDEND YIELD (% PER ANNUM)	2	
FSPXE	S&P'S COMPOSITE COMMON STOCK: PRICE-EARNINGS RATIO (%,NSA)	5	
FSDJ	COMMON STOCK PRICES: DOW JONES INDUSTRIAL AVERAGE	5	
PSCCOM	SPOT MARKET PRICE INDEX:BLS & CRB: ALL COMMODITIES(1967=100)	5	
FYFF	INTEREST RATE: FEDERAL FUNDS (EFFECTIVE) (% PER ANNUM,NSA)	2	
FYGM3	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,3-MO.(% PER ANN,NSA)	2	
FYGM6	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,6-MO.(% PER ANN,NSA)	2	
FYGT1	INTEREST RATE: U.S.TREASURY CONST MATURITIES,1-YR.(% PER ANN,NSA)	2	
FYGT5	INTEREST RATE: U.S.TREASURY CONST MATURITIES,5-YR.(% PER ANN,NSA)	2	
FYGT10	INTEREST RATE: U.S.TREASURY CONST MATURITIES,10-YR.(% PER ANN,NSA)	2	
FYAAAC	BOND YIELD: MOODY'S AAA CORPORATE (% PER ANNUM)	2	
FYBAAC	BOND YIELD: MOODY'S BAA CORPORATE (% PER ANNUM)	2	
sfygm3	fygm3-fyff	1	
sFYGM6	fygm6-fyff	1	
sFYGT1	fygt1-fyff	1	
sFYGT5	fygt5-fyff	1	

sFYGT10	fygt10-fyff	1	**I**
sFYAAAC	fyaaac-fyff	1	
sFYBAAC	fybaac-fyff	1	
EXRSW	FOREIGN EXCHANGE RATE: SWITZERLAND (SWISS FRANC PER U.S.$)	5	
EXRJAN	FOREIGN EXCHANGE RATE: JAPAN (YEN PER U.S.$)	5	
EXRUK	FOREIGN EXCHANGE RATE: UNITED KINGDOM (CENTS PER POUND)	5	
EXRCAN	FOREIGN EXCHANGE RATE: CANADA (CANADIAN $ PER U.S.$)	5	
PWFSA	PRODUCER PRICE INDEX: FINISHED GOODS (82=100,SA)	6	
PWFCSA	PRODUCER PRICE INDEX:FINISHED CONSUMER GOODS (82=100,SA)	6	
PWIMSA	PRODUCER PRICE INDEX:INTERMED MAT.SUPPLIES & COMPONENTS(82=100,SA)	6	
PWCMSA	PRODUCER PRICE INDEX:CRUDE MATERIALS (82=100,SA)	6	
PMCP	NAPM COMMODITY PRICES INDEX (PERCENT)	1	
PUNEW	CPI-U: ALL ITEMS (82-84=100,SA)	6	
PU83	CPI-U: APPAREL & UPKEEP (82-84=100,SA)	6	
PU84	CPI-U: TRANSPORTATION (82-84=100,SA)	6	
PU85	CPI-U: MEDICAL CARE (82-84=100,SA)	6	
PUC	CPI-U: COMMODITIES (82-84=100,SA)	6	
PUCD	CPI-U: DURABLES (82-84=100,SA)	6	
PUXF	CPI-U: ALL ITEMS LESS FOOD (82-84=100,SA)	6	
PUXHS	CPI-U: ALL ITEMS LESS SHELTER (82-84=100,SA)	6	
PUXM	CPI-U: ALL ITEMS LESS MIDICAL CARE (82-84=100,SA)	6	
HHSNTN	U. OF MICH. INDEX OF CONSUMER EXPECTATIONS(BCD-83)	2	

Note: I and II indicate the variables selected at the beginning of 2001:01 and/or at the end of 2004:10, respectively, by the Lasso regression.