Prediction Using Several Macroeconomic Models*

Gianni Amisano† and John Geweke‡

April 24, 2012

Abstract

Prediction of macroeconomic aggregates is one of the primary functions of macroeconometric models, including dynamic factor models, dynamic stochastic general equilibrium models, and vector autoregressions. This study establishes methods that improve the predictions of these models, using a representative model from each class and a canonical 7-variable postwar US data set. It focuses on prediction over the period 1966 through 2011. It measures the quality of prediction by the probability densities assigned to the actual values of these variables, one quarter ahead, by the predictive distributions of the models in real time. Two steps lead to substantial improvement. The first is to use full Bayesian predictive distributions rather than substitute a "plug-in" posterior mode for parameters. Across models and quarters, this leads to a mean improvement in probability of 50.4%. The second is to use an equally-weighted pool of predictive densities from the three models, which leads to a mean improvement in probability of 41.9% over the full Bayesian predictive distributions of the individual models. This improvement is much better than that afforded by Bayesian model averaging. The study uses several analytical tools, including pooling, analysis of predictive variance, and probability integral transform tests, to understand and interpret the improvements.

Keywords: Analysis of variance, Bayesian model averaging, dynamic factor model, dynamic stochastic general equilibrium model, prediction pools, probability integral transform test, vector autoregression model

*A report on some aspects of this work using a different data set appears as Geweke and Amisano (2012a).

†European Central Bank, Frankfurt, gianni.amisano@ecb.int.The views expressed do not represent those of the ECB. Much of Amisano's work was carried out while on secondment (ECB "External Working Experience" program) at UTS. Amisano thanks UTS for providing warm hospitality, a very stimulating working environment and excellent computing facilities.

‡University of Technology Sydney (Australia), Erasmus University (The Netherlands) and University of Colorado (USA), John.Geweke@uts.edu.au. Geweke acknowledges partial financial support from Australian Research Council grant DP110104372.

1 Introduction

Normative decision-making theory, based on expected utility, requires that the decision-maker have a coherent probability distribution over relevant unknown magnitudes. Often these include future events, and often there are alternative approaches that the decision maker, and the decision-maker's staff, can take in approaching this task. This requires that the approaches somehow be brought together in order to arrive at a single distribution. While this is a rarefied depiction of actual decision-making, in the real world substantial time, effort and resources are devoted to distilling the logical implications of disparate views of uncertainty that emerge from even the most disinterested and skillful modelling of the course of future events.

This work examines some of the ways in which this task can be addressed in the particular case of macroeconometric models that are designed for the purpose of assigning probabilities to the future course of principal economic aggregates. Central banks, in particular, routinely utilize these models, and increasingly focus on probability distributions for future events as opposed to point forecasts. In many cases research departments maintain and improve several alternative models and bring forward their different predictions to a fairly advanced point in support of monetary policy. Reconciling differences among these models will not soon, if ever, be reduced to one formal procedure. However, there are well-defined steps in this direction.

This work takes up some alternative approaches to formulating predictive distributions from macroeconomic models, and reconciling their implications with full allowance for the fact that none of the models corresponds – or even comes close – to reality. In doing this it brings together a number of analytical tools, with some refinements beyond those in the literature, and uses them to sort through different approaches, understand their differences, and make practical recommendations for prediction in central banks.

The study focuses on three macroeconometric models: a dynamic factor model (DFM), a dynamic stochastic general equilibrium (DSGE) model, and a vector autoregression (VAR) model. We have several reasons for taking this approach. First, these are representatives of the three major families of macroeconometric models used for prediction in central banks. Second, these families differ in the ways in which they attempt to use general equilibrium theory as a source of information in formulating the model and conducting statistical inference. Third, the models take different approaches in the marshalling of prior information that is required if useful predictions about complex phenomena are to be constructed from relatively sparse data. Finally, our recent methodological work on some aspects of this problem (Geweke and Amisano, 2011) suggests that model combination is most fruitful when the models at hand are dissimilar.

Section 2 provides summary detail of the three models. We concentrate exclusively on the same canonical US data for seven macroeconomic aggregates used by Smets and Wouters (2007), incorporating subsequent revisions of their data and extending it through the last quarter of 2011. The paper studies predictive performance over 184 quarters, 1966 through 2011, breaking this up into three periods of interest described in

Section 2. The work here is all based on probability distributions over a single-period (one quarter) horizon. Posterior distributions or modes are constructed for each sample, the first ending in 1965:4 and used for prediction of the seven aggregates for 1966:1, and the last ending in 2011:3 and used for prediction of the seven aggregates for 2011:4. In that sense the analysis here is "out of sample," mimicking what a real econometrician would do in real time. However this study does not attempt to limit attention to the most recent data revisions available each quarter, nor does it grapple with the question of which vintage of revised data should be used in assessing the predictions. The most recent revisions available on February 16, 2012, are used for all purposes through the paper.

Section 3 summarizes the four principal analytical tools used in our construction and interpretation of predictive distributions from several models and alternative approaches to inference. Two are competing approaches to model combination, Bayesian model averaging and linear prediction pools. We use analysis of predictive variance (Geweke and Amisano, 2012b) to understand the gains from using prediction pools and to interpret the superiority of full Bayesian predictive distributions to predictive distributions based on point estimates that emerges in this work. The probability integral transform tests adduce evidence that all of the models studied here are grossly unrealistic in particular dimensions. We regard this fact both as fundamental in guiding approaches to model combination for purposes of prediction and in explaining the relatively poor performance of Bayesian model averaging in the empirical work.

The balance of the paper is devoted to the formulation, evaluation and understanding of the predictive distributions that emerge in the three models and in different combinations of these distributions. Section 4 studies two leading methods for prediction, one based on substituting the posterior mode for the parameter vector and the other using the full Bayesian predictive distribution. In general – though not entirely without exception – the latter performs much better than the former. It shows that these differences can be traced to quarters that turn out to be "outliers," realizations that have relatively low probability as assessed by any of the models. It shows that the effect arises because parameter uncertainty is an important source of variance in full predictive distributions that is ignored when only posterior modes are used. Finally, it shows that the effect is smallest in the DSGE model (the one with the fewest parameters, in which predictive variance due to parameter uncertainty is also smallest) and strongest in the VAR model (the one with the most parameters, in which predictive variance due to parameter uncertainty is also largest).

Section 5 takes up the combination of models for purposes of prediction. A theorem of McConway (1981) implies that for prediction of a multivariate random vector, only linear combinations of predictive densities will provide coherent results. This restriction renders the analysis tractable while still leaving open a number of interesting possibilities. The simple average of predictive distributions turns out to be very effective and imposes essentially no demands on the econometrician beyond those required to evaluate the predictive performance of the different models in the first place. An alternative is an optimal linear pool with weights updated at the end of each quarter to combine the

model predictive densities for the next quarter. This turns out to fall somewhat short of the equally-weighted pool. This comparison is specific to the three models used and the data employed here, and may well be attributed to the fact that we use three models that have all held their own in the marketplace of macroeconomic prediction. Bayesian model averaging falls well short of either pool for the entire period studied. The explanation is rooted in the fact that Bayesian model averaging conditions on one of the models being fully correctly specified (though which one is not known *a priori*), a condition that is manifesting unrealistic here.

The paper concludes with a short quantitative recapitulation of the results.

2 Models and data

The models that we study are a representative selection of forecasting models used in macroeconomic policy environments. Some models are typically specified in order to incorporate features directly drawn from economic theory. Among these models, dynamic stochastic general equilibrium models have been widely used in many central banks to produce forecasts, historical decompositions aiming at assessing the relevance of different kinds of macroeconomic shocks, and counterfactual analyses. Smets and Wouters (2003, 2007) have shown that these models can be successfully estimated with satisfactory fit and forecasting properties. Del Negro and Schorfheide (2012) provides interesting discussion of how DSGE models have fared in the recent past and how external information can be brought to bear to improve them.

On the other hand, policy-oriented macroeconomic forecasting is often based on time series models that are more agnostic with respect to general equilibrium theory. Among these models, the most widely used are the vector autoregression models introduced by Sims (1980). These models are characterized by a parameter space of high dimension and tpically employ prior distributions with carefully chosen hyperparameters.

It is often said that economic policy requires very large amounts of economic information being taken into consideration in order to guide decision making (Bernanke and Boivin, 2003). Dynamic factor models (Geweke, 1977; Sargent and Sims, 1977; Stock and Watson, 2002a; Forni, Hallin, Lippi and Reichlin, 2005) are well suited to this task. In these models the joint behavior of a large number of economic time series is jointly modelled by specifying that the series are driven by a small set of persistent common factors and by idiosyncratic shocks.

All these considerations lead us to include in our analysis three specific models representative of these classes, referred to subsequently as DFM, DSGE and VAR, in order to provide a compact, yet representative, basis to span the model set commonly used in an economic policy environment. Going forward, in circumstances where the ordering of the models is arbitrary, we refer to these models in alphabetical order of their acronyms.

2.1 Three models

The observable time series of interest in each of the three models are the log growth rates of real consumption, investment, income (GDP), and wages; the logarithm of a per capita weekly hours worked index, inflaton as measured by the growth rate of the GDP deflator, and the nominal Federal Funds rate. In the DFM and DSGE models the first series appear in exactly this form. We examine two variants of VAR models: in the VARD (VAR-differences) model the series also appear in this form; in the VARL (VAR-levels) the first four series appear as log-levels rather than growth rates.

2.1.1 The DFM

In a dynamic factor model a set of time series is driven by a typically small set of common factors and by idiosyncratic shocks. When the number of series being jointly considered is high, usually non-parametric estimation is used. In this regard, Stock and Watson (2002b) and Forni, Hallin, Lippi and Reichlin (2005) show how to use static and dynamic principal components methods to obtain consistent estimates of the space spanned by common factors.

In this study, we use a very small dynamic factor model, with a set of $n = 12$ variables that includes the 7 series common to all three models. The set of additional variables is chosen to consider, in a highly stylized way, some economic phenomena which are neglected by the information set used to estimate the DSGE and VAR models. These additional variables are stock returns, the term structure slope, the risk premium, the unemployment rate and the rate of change of money.

With such a compact information set, inference can be entirely parametric and we employ the model of Stock and Watson (2005) as follows for the $n \times 1$ vector of time series y_t,

$$y_t = \Gamma f_t + v_t, \;\; f_t = c + \sum_{j=1}^{p} A_j f_{t-j} + \eta_t, \;\; v_t = \sum_{j=1}^{q} B_j v_{t-j} + \varepsilon_t,$$

where f_t is a $k \times 1$ vector of latent factors and v_t is an $n \times 1$ vector of mutually independent indiosyncratic shocks.

The parameters are subject to several restrictions. For identification of factors, $\gamma_{ij} = 0 \;(j > i)$ and therefore there are $r = nk - k\,(k-1)\,/2$ free parameters of the matrix Γ that can be collected in an $r \times 1$ vector γ. Because the indiosyncratic shocks are independent $B_j = diag\,(b_j) \;(j = 1, \ldots, q)$.

The shocks are Gaussian and independent:

$$\begin{pmatrix} \eta_t \\ \varepsilon_t \end{pmatrix} \overset{iid}{\sim} N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} I_k & 0 \\ 0 & H^{-1} \end{bmatrix} \right\}.$$

The prior distribution has four independent components:

$$\begin{gathered}
\gamma \sim N\left(0, 10 \cdot I_r\right); \\
c \sim N\left(0, 10 \cdot I_n\right) \\
vec\left(A_j\right) \overset{iid}{\sim} N\left(0, I_k\right) \quad (j = 1, \ldots, p); \\
b_j \overset{iid}{\sim} N\left(0, I_n\right) \quad (j = 1, \ldots, q); \\
0.4 \cdot h_i \overset{iid}{\sim} \chi^2\left(4\right) \quad (i = 1, \ldots, n).
\end{gathered}$$

After preliminary investigations to settle on a satisfactory specification, we chose $k = 3$ common factors, VAR dynamics of order $p = 2$, and idiosyncratic shock dynamics of order $q = 2$. The total number of parameters in the model is $r + n + pk^2 + nq + n = 99$.

The linear-Gaussian state space nature of the model easily lends itself to Gibbs sampling with data augmentation. The conditional distributions of $(a_1, \ldots, a_p)$ and $(b_1, \ldots, b_q)$ require a Metropolis-Hastings within Gibbs approach, while the conditional distributions of γ and $(h_1, \ldots, h_n)$ are known analytically; the simulation of the latent common factors is straightforward. For each of the subsamples being analyzed, we run 65,000 MCMC iterations, after 15,000 burn-in draws to ensure convergence of the algorithm. The MCMC output was thinned to 10,000 recorded draws to save on storage space. We also computed posterior modal values for the parameters by using the Gibbs-based simulated annealing algorithm proposed by Doucet, Godsill and Robert (2002).

2.1.2 The DSGE model

The dynamic stochastic general equilibrium model we use in this study is exactly the model described in Smets and Wouters (2007), which details and discusses the model's specification. To briefly summarize the model's main characteristics, we just recall that the model has two sectors: an intermediate goods and a final good sector. The first sector produces differentiated goods in a monopolistically competitive setting, while the second sector is perfectly competitive. Both wages and prices are sticky and the infinitely-lived representative consumer has utility characterized by consumption habits. Monetary policy is specified as a Taylor rule and real variables are all characterized by a common deterministic growth rate which is estimated with all the other parameters of the model. The model has seven shocks: an aggregate total factor productivity shock, an investment specific shock, a risk premium shock (which acts as proxy for financial disturbances), wage and price mark-up shocks and fiscal and monetary policy shocks.

The model has 39 free parameters which we endow with exactly the same prior structure as in Smets and Wouters (2007). We carry out Bayesian and posterior mode estimation by using the linearized solution and, as customary in the applied DSGE literature (An and Schorfheide, 2007), we use a random walk Metropolis-Hastings algorithm with candidate tailored using the Hessian of the log posterior computed at its mode. For each posterior, the mode is found using numerical optimization. The results obtained

are based on 40,000 MCMC draws for each subsample, after 12,000 burn in iterations. For each of the subsamples we thin the draws to 10,000.

2.1.3 The VAR model

In the VAR the conditional distribution of the series takes the form of a normal multivariate regression model in which the covariates consist of an intercept term and the first four lagged values of each series. Thus the model has 29 coefficients in each of 7 equations, which together with the conditional variance matrix makes a total of 231 parameters.

For both the VARD and VARL variants of this model we utilize the "Minnesota prior" distribution (Litterman, 1986) in which the coefficients are Gaussian and independent and their variances are functions of a smally number of hyperparameters. Denoting $a_{ij,h}$ the coefficient on the h^{th}lag of variable j (Y_{jt-h}) in the equation where the dependent variable is Y_{it}, and c_i the intercept in the same equation, prior variances are specified as follows:

$$\begin{aligned} \operatorname{var}(a_{ij,h}) &= \begin{cases} \left(\pi_1 \times \pi_3^{-h} \times \frac{\sigma_{ii}}{\sigma_{jj}}\right)^2 & (i=j) \\ \left(\pi_1 \times \pi_2 \times \pi_3^{-h} \times \frac{\sigma_{ii}}{\sigma_{jj}}\right)^2 & (i \neq j) \end{cases} \\ \operatorname{var}(c_i) &= (\pi_1 \times \pi_4)^2, \end{aligned}$$

with $\pi_1 = .2$, $\pi_2 = .9$, $\pi_3 = 1, \pi_4 = 1$. In this prior distribution $E(a_{ij,h}) = 0$ in for all coefficients, except that $E(a_{ii,1}) = 1$ if variable i appears in levels. Hence, in the VARL variant all coefficients pertaining to own first have mean 1, while in the VARD variant only hours, inflation and the short term interest rate have own first lag coefficient with prior variance equal to one and the other four variables have first own lag coefficient prior mean equal to zero.

As regard the shocks precision matrix H, we specified a Wishart distribution with parameters $\underline{\nu} = 9$ and $\underline{S}$ set to match the OLS estimate of H based on the first subsample.

The posterior simulation of the model is based on a straightforward two block Gibbs algorithm. For each of the subsamples analyzed, we obtained 65,000 draws, the first 15,000 of which were discarded and the remaining 50,000 were thinned to produce 10,000 retained draws.

The posterior mode estimates were obtained by analytically marginalizing the joint posterior distribution with respect to the coefficients $a_{ij,h}$ and c_i. The resulting marginal posterior of H was numerically optimized yielding a modal value for H. With the modal value of H in hand, it is immediate to find the modal values of the coefficients $a_{ij,h}$ and c_i, because their conditional posterior distribution is Gaussian with known moments.

Table 1: Data sources

Series Mnemonics	Definition	Source
AAA	Moody's Seasoned Aaa Corporate Bond Yield	FRED
BBB	Moody's Seasoned Baa Corporate Bond Yield	FRED
CE16OV	Civilian Employment	FRED
FEDFUNDS	Effective Federal Funds Rate	FRED
FPI	Fixed Private Investment	FRED
GDPC96	Real Gross Domestic Product, 3 Decimal	FRED
GDPDEF	Gross Domestic Product: Implicit Price Deflator	FRED
GS10	10-Year Treasury Constant Maturity Rate	FRED
LNS10000000	Civilian noninstutitionalize population	BLS
M2SL	M2 stock	FRED(*)
PCEC	Personal Consumption Expenditures	FRED
PRS85006023	Nonfarm Business Sector: Average Weekly Hours	BLS
PRS85006103	Nonfarm Business Hourly Compensation	BLS
SP500C	SP500 composite index	FRED
UNRATE	Civilian unemployment rate	FRED
(*) Prior to 1959 the M2 series comes from Balke and Gordon (1986)		

2.2 Data and periods of interest

The series we use for the DSGE and VAR models are exactly those described in Smets and Wouters (2007), extended through 2011 and incorporating the latest available revisions. The DFM utilizes five additional series: stock returns, defined as the log differences of quarterly averages of S&P 500 Composite index; the aggregate civilian unemployment rate; the term premium, measured as the difference between the yields on ten year and three month US Treasury bond; the risk premium, measured as the difference between the Risk premium: BAA and AAA corporate bond spread; and the growth rate in the money supply M2.

The relevant series were all obtained from the Federal Reserve Bank of St. Louis Data Repository (FRED [1]) and from the Bureau Labor of Statistics website[2], in their version available on February, 16th 2012, when the series were downloaded. Series mnemonics and data sources are summarized in Table (1)

The time series plots in Figure 1 convey well known features of the seven series of interest. In particular, real variables seem to be affected by severe negative shocks in contractions, while the behavior of nominal variables is dominated by secular forces the build-up to the great inflation of the 1970s and the early 1980s, and by the subsequent drastic disinflation. All series show large abrupt movements at the outset of the Great Financial Crisis, from 2007:4 onwards.

[1] http://research.stlouisfed.org/fred2/.

[2] http://www.bls.gov/.

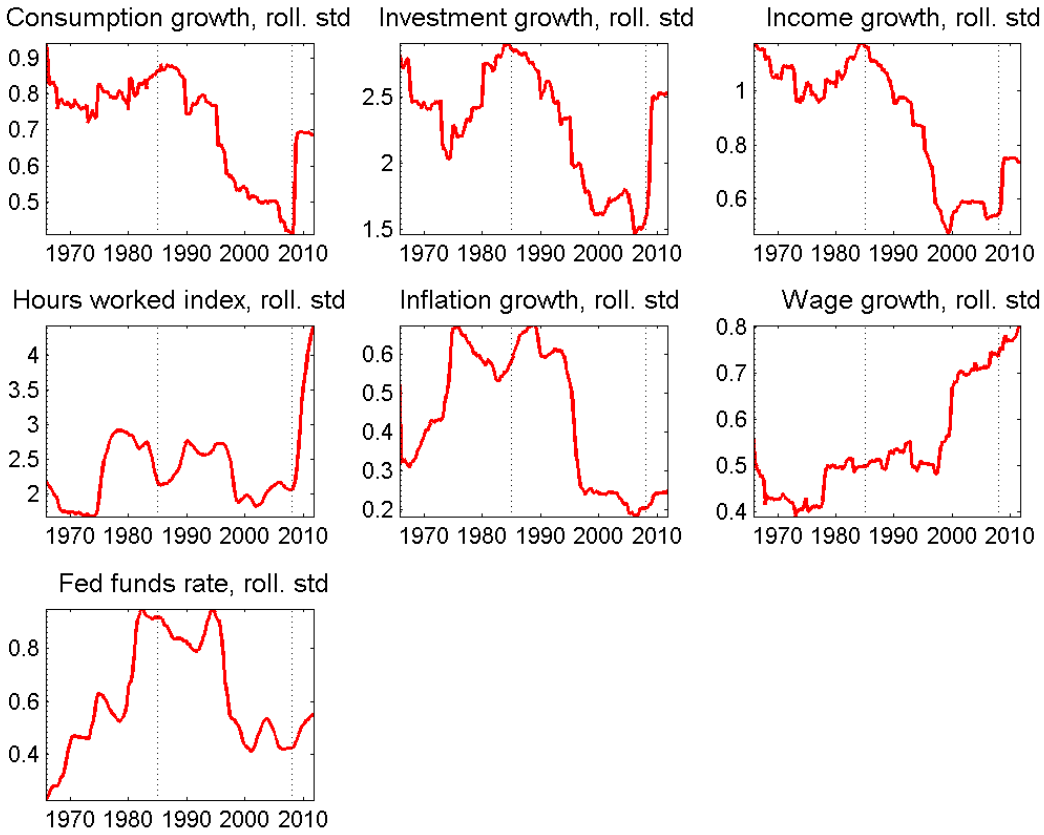

Figure 1: Series being predicted

Another well known feature of these series is the fact that they exhibit non-constant variance. This phenomenon is known as the great moderation (McDonnell and Perez-Quiros, 2000; Stock and Watson, 2004). Figure 2 provides standard deviations computed on rolling windows of 60 quarters and it shows that the movements in volatility have been gradual and substantial. With the exception of hours worked and real wage growth, all other series are affected by a gradual decrease of volatility taking place from the mid 1980s onwards which was sharply reversed with the onset of the GFC.

This phenomenon can be alternatively described by computing standard deviations based on sub-samples, as we do in Table 2.

We distinguish the whole period used in the analysis into four different subsamples

1. Initial: 1951:1-1965:4, used to initialize the estimation of each of the models and not used for forecasting evaluation;

2. Pre moderation: 1966:1 - 1984:4, characterized by higher volatility;

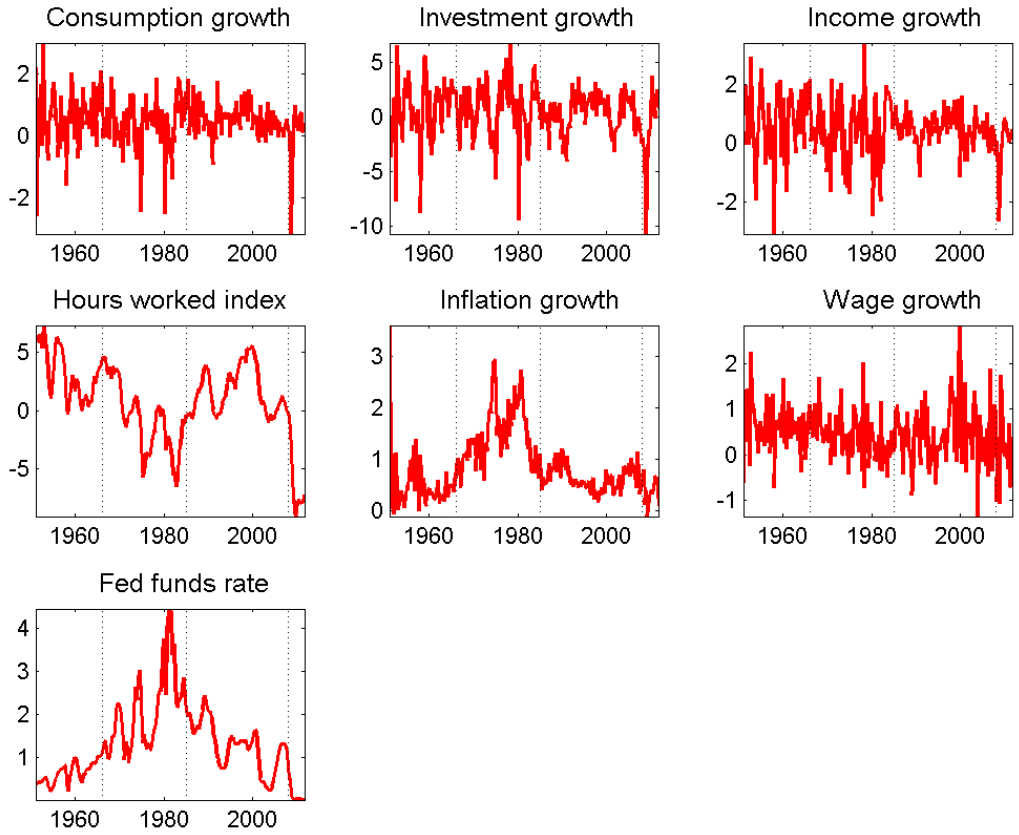

Figure 2: Standard deviations of the series being predicted. Computations based on rollinw window of 60 observations

3. Great moderation: 1985:1-2007:4, characterized by smaller volatility;

4. Post moderation: 2008:1-2011:4, characterized by a return to higher volatility.

Throughout the paper t indexes quarters, with $t = 1$ being the first quarter predicted, 1966:1, and $t = T = 184$ being the last quarter predicted, 2011:4. We denote the 7×1 vector of random variables of interest Y_t, and their realized values (the data) y_t. Each model specifies conditional densities $p\left(Y_t \mid Y_{1:t-1}, \theta_i, A_i\right)$ and a prior density $p\left(\theta_i \mid A_i\right)$, where θ_i is the parameter vector in model A_i. In the interest of non over-burdening the reader with pedantry, this abuses the notation somewhat: The conditioning data set always goes back to 1951; and, in the DFM, the conditioning data includes the histories of the five additional series from 1951 through quarter $t-1$.

In particular, note that all series volatilities dropped during the great moderation and for many series volatility returned to pre moderation levels (or higher) post moderation.

Table 2: Standard deviations of observed data computed on different subsamples

		Period		
Series	Initial	Pre moderation	Great moderation	Post moderation
Consumption growth	0.94	0.82	0.51	1.04
Investment growth	2.82	2.72	1.63	3.88
Income growth	1.16	1.09	0.53	0.97
Hours worked index	2.18	3.06	1.95	3.10
Inflation	0.52	0.58	0.24	0.26
Wage growth	0.56	0.49	0.67	0.75
Fed funds rate	0.23	0.89	0.54	0.24

Notice also that the federal funds rate was characterized by very low volatility in the years before 1970 and that the most recent years have reduced its volatility to the pre-1970s levels. This is a consequence of the extremely expansive policy being pursued by the Fed as a consequence of the Great Financial Crisis.

3 Analytical methods

This work uses an eclectic methodology in order to understand the performance of macroeconometric models in prediction and to examine how an econometrician might best use the models at her disposal to form predictive distributions. With the exception of one technique discussed in Section 3.4, all of these methods appear in the literature. The treatment here is short, intended to establish notation and indicate precisely the tools used in Sections 4 and 5.

3.1 Bayesian model averaging

Bayesian model averaging is implied by the $\mathcal{M}$-closed perspective of fully subjective Bayesian inference (Bernardo and Smith, 1993, Section 6.1.2). This approach conditions on a set of models $A_1, \ldots, A_n$, each with a parameter vector $\theta_i \in \Theta_i$, a prior density $p(\theta_i \mid A_i)$, and a specification of conditional densities $p(Y_t \mid Y_{1:t-1}, \theta_i, A_i)$ for a common set of observable vectors $Y_{1:T} = \{Y_1, \ldots, Y_T\}$. (Upper case Y denotes random vectors and lower case y the corresponding realizations.) Model prior probabilities $p(A_i)$, with $\sum_{i=1}^{n} p(A_i) = 1$, place models, parameter vectors and observables in a common probability space.

The laws of probability then imply the sequence of one-step-ahead predictive densities

$$p(Y_t \mid y_{1:t-1}) = \sum_{i=1}^{n} p(A_i \mid y_{1:t-1}) \cdot p(Y_t \mid y_{1:t-1}, A_i).$$

The model predictive densities

$$p\left(Y_t \mid y_{1:t-1}, A_i\right) = \int_{\Theta_i} p\left(Y_t \mid y_{1:t-1}, \theta_i, A_i\right) p\left(\theta_i \mid y_{1:t-1}, A_i\right) d\theta_i \;\; (t = 1, \ldots, T) \tag{1}$$

are accessed as described in Section 2.1. The conditional probabilities $p\left(A_i \mid y_{1:t-1}\right)$, also known as Bayesian model averaging weights, are

$$p\left(A_i \mid y_{1:t-1}\right) \propto p\left(A_i\right) p\left(y_{1:t-1} \mid A_i\right) = p\left(A_i\right) \prod_{s=1}^{t-1} p\left(y_s \mid y_{1:s-1}, A_i\right). \tag{2}$$

The marginal likelihoods $p\left(y_s \mid y_{1:s-1}, A_i\right)$ are evaluated as described in Section 2.1.

Given weak regularity conditions, including the existence of a true data generating process $p\left(Y_t \mid Y_{1:t-1}, D\right)$,

$$t^{-1} \sum_{s=1}^{t} \log p\left(Y_s \mid Y_{1:s-1}, A_i\right) \xrightarrow{a.s.} LS_i$$

as detailed in Geweke and Amisano (2011). So long as $j = \arg\max_i \left(LS_i\right)$ is unique $p\left(A_j \mid Y_{1:t-1}\right) \xrightarrow{a.s.} 1$, and $\left(A_i \mid Y_{1:t-1}\right) \xrightarrow{a.s.} 0$ for $i \neq j$. If in fact $A_k = D$ for some $k \in \{1, \ldots .n\}$ then $j = k$.

3.2 Pooling

Suppose that $R_1, \ldots, R_n$ are prediction rules, each specifying a sequence of predictive densities $p\left(Y_t; y_{1:t-1}, R_i\right)$ $(t = 1, \ldots, T)$. A prediction rule R_i could coincide with a sequence of model predictive densities (1), but it might also be any sequence of legitimate predictive densities that depends only on information actually available at $t-1$. For example it could be the sequence

$$p\left(Y_t \mid y_{1:t-1}, \widehat{\theta}_i\left(t-1\right), A_i\right) \;\; (t = 1, \ldots, T) \tag{3}$$

where

$$\widehat{\theta}_i\left(t-1\right) = \arg\max_{\theta_i} p\left(\theta_i \mid A_i\right) p\left(y_{1:t-1} \mid \theta_i, A_i\right), \tag{4}$$

the posterior mode.

A linear pool of n prediction rules $R_1, \ldots, R_n$ is the sequence of predictive densities

$$p\left(Y_t; y_{1:t-1}, \mathbf{w}_{t-1}, R_1, \ldots, R_n\right) = \sum_{i=1}^{n} w_{t-1,i} p\left(Y_t; y_{1:t-1}, R_i\right) \;\; (t = 1, \ldots, T) \tag{5}$$

where $\mathbf{w}_{t-1}$ is a point in the n-dimensional unit simplex, i.e. $w_{t-1,i} \geq 0$ $(i = 1, \ldots, n)$ and $\sum_{i=1}^{n} w_{t-1,i} = 1$. The subscript $t-1$ indicates the requirement that $\mathbf{w}_{t-1}$ also depends only on information actually available at $t-1$. Arguably the simplest pool is the one

that assigns equal weights to prediction rules, $w_{t-1,i} = n^{-1}$ $(t = 1, \ldots, T; i = 1, \ldots, n)$. We refer to each of these pools subsequently as an equally weighted pool (EWP).

Any prediction rule R formed at time t can be evaluated using the log scoring criterion

$$\sum_{s=1}^{t} \log p\left(y_s; y_{1:s-1}, R\right). \tag{6}$$

There are several compelling arguments for this rule, summarized in Geweke and Amisano (2011). Note that if $p\left(y_t \mid y_{1:t-1}, R\right) = p\left(y_t \mid y_{1:t-1}, A_i\right)$ $(t = 1, \ldots, T)$, the sequence of predictive likelihoods for model A_i, then the criterion (6) is the log marginal likelihood $\log p\left(y_{1:t} \mid A_i\right)$. An optimal prediction pool selects $\mathbf{w}_{t-1}^*$ to maximize this criterion:

$$\mathbf{w}_{t-1}^* = \arg\max_{\mathbf{w}_{t-1}} \sum_{s=1}^{t-1} \log \left[\sum_{i=1}^{n} w_{t-1,i} p\left(y_s; y_{1:s-1}, R_i\right)\right].$$

subject to the constraint that $\mathbf{w}_{t-1}$ be in the n-dimensional unit simplex. This is a simple convex programming problem.

This process generates a sequence of weight vectors and corresponding pools, each of which we refer to subsequently as a real-time optimal pool (RTOP). Because $\mathbf{w}_{t-1}$ and $p\left(y_{t-1}; y_{1:t-2}, R_i\right)$ involve only information actually available at the end of period $t-1$, this mimics a procedure that could have been carried out by an econometrician in real time. In order to summarize the behavior of pools over various time intervals, we shall sometimes refer to static optimal pools of the form

$$\mathbf{w}_{r:t}^* = \arg\max_{\mathbf{w}} \sum_{s=r}^{t} \log \left[\sum_{i=1}^{n} w_i p\left(y_s; y_{1:s-1}, R_i\right)\right] \tag{7}$$

for particular choices of s and t. Note that $\mathbf{w}_{1:t-1}^* = \mathbf{w}_{t-1}^*$. The log score of a static optimal pool cannot be less (and is generally greater) than the log score of the corresponding equally weighted pool. The log score of a RTOP can be less than that of the corresponding equally weighted pool.

3.3 Analysis of predictive variance

The variance implicit in a Bayesian predictive distribution can be decomposed into several sources as described in Geweke and Amisano (2012b). The approach proves useful here in understanding the relative performance of two popular approaches to forming predictive densities in macroeconometric models, the fully Bayesian predictive distribution with the sequence of densities (1) and the posterior mode or "plug in" approach (3). Due to the particular characteristics of predictive distributions in the three models, discussed in Section 2.1, the technical steps differ from those in Geweke and Amisano (2012b) and are somewhat simpler.

Consider first the predictive distributions for a single model A_i. By the law of total probability

$$\begin{aligned}\text{var}\left(Y_t \mid y_{1:t-1}, A_i\right) &= \text{E}_{\theta_i}\left\{\left[\text{var}\left(Y_t \mid y_{1:t-1}, \theta_i\right), A_i\right] \mid y_{1:t-1}, A_i\right\} \\ &\quad + \text{var}_{\theta_i}\left\{\left[\text{E}\left(Y_t \mid y_{1:t-1}, \theta_i\right), A_i\right] \mid y_{1:t-1}, A_i\right\}. \end{aligned} \tag{8}$$

Following Geweke and Amisano (2012b), the first term on the right side of (8) is the intrinsic variance of the predictive distribution, so named because it is the variance in the predictive distribution that would exist even if θ_i were known, averaged over the distribution of θ_i specified in the relevant posterior distribution. The second term on the right side of (8) is the extrinsic variance of the predictive distribution, so named because it is the variance in the conditional mean that arises from the fact that θ_i is not degenerate in the relevant posterior distribution.

As detailed in Section 2.1, the distribution

$$Y_t \mid \left(y_{1:t-1}, \theta_i, A_i\right) \sim N\left[\mu\left(y_{1:t-1}, \theta_i\right), V\left(y_{1:t-1}, \theta_i\right)\right]$$

in all of the macroeconometric models considered in this work. The vector $\mu\left(y_{1:t-1}, \theta_i\right)$ and matrix $V\left(y_{1:t-1}, \theta_i\right)$ have closed form expressions that are easy to evaluate. Corresponding to the vectors $\theta_{t-1,i}^{(m)}$ $(m = 1, \ldots, M)$ from the posterior simulator for model A_i and the sample $y_{1:t-1}$, let

$$\mu_{t-1,i}^{(m)} = \mu\left(y_{1:t-1}, \theta_{t-1,i}^{(m)}\right), \quad V_{t-1,i}^{(m)} = V\left(y_{1:t-1}, \theta_{t-1,i}^{(m)}\right)$$

$(t = 1, \ldots, T-1; i = 1, \ldots, n; m = 1, \ldots, M)$, and

$$\mu_{t-1,i} = M^{-1} \sum_{m=1}^{M} \mu_{t-1,i}^{(m)} \quad (t = 1, \ldots, T-1; i = 1, \ldots, n).$$

Then the relevant numerical approximations in (8) are

$$\text{E}_{\theta_i}\left\{\left[\text{var}\left(Y_t \mid y_{1:t-1}, \theta_i\right)\right] \mid y_{1:t-1}, A_i\right\} \approxeq M^{-1} \sum_{m=1}^{M} V_{t-1,i}^{(m)} = VI_{t-1,i} \tag{9}$$

for intrinsic variance and

$$\begin{aligned}&\text{var}_{\theta_i}\left\{\left[\text{E}\left(Y_t \mid y_{1:t-1}, \theta_i\right)\right] \mid y_{1:t-1}, A_i\right\} \\ \approxeq\ &(M-1)^{-1} \sum_{m-1}^{M} \left[\mu_{t-1,i}^{(m)} - \mu_{t-1,i}\right] \left[\mu_{t-1,i}^{(m)} - \mu_{t-1,i}\right]' = VE_{t-1,i} \end{aligned} \tag{10}$$

for extrinsic variance $(t = 1, \ldots, T-1; i = 1, \ldots, n)$. The approximation of total variance is the sum of (9) and (10). Then the fraction of the variance that is extrinsic may

be computed in the obvious way for each component of Y_t, using the diagonal elements of $VI_{t-1,i}$ and$VE_{ti-1,i}$.

For any pool with weight vector $\mathbf{w}$ this analysis can be extended to remove the conditioning on model A_i. From Geweke and Amisano (2012b, Proposition 4) the governing law of total probability is

$$\begin{aligned}\mathrm{var}\left(Y_t \mid y_{1:t-1}\right) &= \mathrm{E}_{A_i,\theta_i}\left\{\left[\mathrm{var}\left(Y_t \mid y_{1:t-1},\theta_i\right)\right] \mid y_{1:t-1},A_i\right\} \\ &\quad +\mathrm{E}_{A_i}\left\langle \mathrm{var}_{\theta_i}\left\{\left[\mathrm{E}\left(Y_t \mid y_{1:t-1},\theta_i\right)\right] \mid y_{1:t-1},A_i\right\} \mid A_i\right\rangle \\ &\quad +\mathrm{var}_{A_i}\left[\mathrm{E}\left(Y_{t+1} \mid A_i\right)\right]. \end{aligned} \tag{11}$$

The first term on the right side of (11) is the intrinsic variance, and its numerical approximation is $\sum_{i=1}^{n} w_i V_{t-1,i}^{int}$. The second term is the within-model extrinsic variance, and its numerical approximation is $\sum_{i=1}^{n} w_i V_{t-1,i}^{ext}$. The last term is the between-model extrinsic variance, and its numerical approximation is $\sum_{i=1}^{n} w_i \left[\mu_{t-1}^i - \mu_{t-1}\right]\left[\mu_{t-1}^i - \mu_{t-1}\right]'$ where $\mu_{t-1} = \sum_{i=1}^{n} w_i \mu_{t-1}^i$.

In the applications in this work the original posterior simulation samples of size 10,000 are thinned to simulation samples of size $M = 1,000$, which becomes the relevant simulation sample for the approximations just described.

3.4 Model evaluation with probability integral transforms

A data generating process D for a vector time series Y_t implies conditional cumulative distribution functions for any element Y_{jt} of Y_t,

$$F_j\left(x; Y_{1:t-1}, D\right) = P\left(Y_{jt} \leq x \mid Y_{1:t-1}, D\right).$$

Rosenblatt (1952) showed that the sequence $\pi_{jt} = F_j\left(Y_{jt}; Y_{1:t-1}, D\right)$ is independent, each π_{jt} uniformly distributed on the unit interval. Smith (1985) noted that the sequence $z_{jt} = \Phi^{-1}\left(\pi_{jt}\right)$, where Φ is the cumulative distribution function of the standard normal distribution, is i.i.d. $N\left(0,1\right)$; see also Berkowitz (2001).

These properties are the foundations of probability integral transform (PIT) tests of correct model specification. For the stated distributions of $\{\pi_{jt}\}$ and $\{z_{jt}\}$ to be literally true a model specification would have to be dogmatic and correct for θ_i. The usual criterion of correct specification of an econometric model is weaker: that for some value of θ_i, the distribution of $Y_{1:T}$ coincides with D. The actual size of test statistics for the properties of $\{\pi_{jt}\}$ or $\{z_{jt}\}$ is likely to be larger than the nominal size when the model is correctly specified up to unknown parameters. More relevant is the degree to which different models depart from the ideal of the PIT, and the particular ways in which this happens for different models; Geweke and Amisano (2010) illustrates this use of PIT tests.

The i.i.d. normal distribution of $\{z_{jt}\}$ under the hypothesis of correct specification is analytically more tractable than that of $\{\pi_{it}\}$ and the tests here proceed from $\{z_{it}\}$.

These series are by-products of the computation of variance decompositions described in the previous section. Corresponding to each parameter vector $\theta_{t-1,i}^{(m)}$ compute

$$\pi_{jt,i}^{(m)} = \Phi^{-1}\left[\left(y_{jt} - \mu_{j,t-1,i}^{(m)}\right) \cdot \left(v_{jj,t-1,i}^{(m)}\right)^{-1/2}\right] (m = 1, \ldots, M)$$

and then $M^{-1}\sum_{m=1}^{M} \pi_{j,t-1,i}^{(m)} \cong \pi_{jt,i}$and $\Phi\left(\pi_{jt,i}\right) \cong z_{jti}$. For each model A_i and each constituent time series j the ideal of correct model specification implies

$$z_{jti}\left(t = 1, \ldots, T\right) \overset{iid}{\sim} N\left(0, 1\right).$$

This hypothesis can be tested in a great many ways, each with its own power against alternatives. This work uses PIT tests developed in Geweke and Amisano (2012c). That paper provides further detail and derives the properties of the tests, which are simply stated here. Moment PIT tests are based on the distribution of a $Q \times 1$ vector m_{ji} of raw moments, each element of the form $T^{-1}\sum z_{jt,i}^{q}$ where q is a particular positive integer unique to that element. The asymptotic (in T) distribution is itself Gaussian with known parameters implied by the moments of the univariate standard normal distribution, which leads to a single test statistic with an asymptotic (in T) $\chi^2\left(Q\right)$ distribution. Autocorrelation PIT tests are based on the distribution of an $L \times 1$ vector r_{ji} of cross products, each element of the form $\left(T-\ell\right)^{-1}\sum_{t=1}^{T-\ell} z_{jti} \cdot z_{j,t-\ell,i}$ where ℓ is a particular positive integer unique to that element. This vector is also asymptotically normal with known parameters and leads to a test statistic with an asymptotic $\chi^2\left(L\right)$ distribution. The sum of the two test statistics has an asymptotic $\chi^2\left(Q+L\right)$ distribution.

Under the hypothesis of correct model specification the exact distribution of the test statistics depends only on T, and it is easy to access this distribution by simulating $z_t \overset{iid}{\sim} N\left(0,1\right)$ $\left(t = 1, \ldots, T\right)$. The work here uses 10^5 simulations, which reliably establishes p-values of PIT test statistics in the first three decimal places; moreover, except for very small p-values it turns out that the asymptotic approximations are quite good.

4 Model comparison and evaluation

This work concentrates on models designed for prediction, and specifically for the purpose of assigning probabilities to future events. This section addresses some details of this task using models individually, before taking up the matter of model combination in Section 5. It employs the log scoring rule for model comparison and, using this criterion, shows that full Bayesian inference is decisively superior to a "plug in" rule that substitutes the posterior mode $\widehat{\theta}_i$ for the parameter vector θ_i (Section 4.1). It uses similar methods to contrast levels and first-difference formulations of VAR models (Section 4.2).

Such model comparison exercises do not address the calibration of models – the degree to which subjective probability distributions for events *ex ante* are consistent with observed frequencies *ex post*. PIT tests of the models (Section 4.3) show that predictive probabilities and realized frequencies are inconsistent in varying degrees, depending

mainly on the events in question and to a lesser degree on the particular model. This finding sets the stage for taking up model combination methods that do not invoke the assumption that one of the models is true in Section 5.

4.1 Bayesian predictive distributions and prediction using posterior modes

A formal Bayesian approach with a single model A_i uses the predictive distribution (1) of Y_t conditional on $y_{1:t-1}$. Given the the output $\theta_i^{(m)} \sim p(\theta_i \mid y_{1:t-1}, A_i)$ $(m = 1, \ldots, M)$ of a posterior simulator, this can always be done by means of subsequent simulations

$$Y_t^{(m,s)} \sim p\left(Y_t \mid y_{1:t-1}, \theta_i^{(m)}\right) \quad (m = 1, \ldots, M; s = 1, \ldots, S). \tag{12}$$

Since the latter density is that of a multivariate normal distribution in the cases of the models studied in this work, methods like those discussed in Sections 3.3 and 3.4 can often be used to avoid the supplementary simulations (12). We refer to this approach subsequently in this section as "full Bayes" (FB).

A common alternative approach is to find the posterior mode $\widehat{\theta}_i(t-1)$ (4) and then replace θ_i with $\widehat{\theta}_i(t-1)$ in the conditional predictive density $p(Y_t \mid y_{1:t-1}, \theta_i)$ yielding (3). The same substitution in (12) can be used to access the resulting distribution of Y_t; again, in the case of the models used in this work, the subsequent simulation can be avoided. We refer to this approach subsequently in this section as "posterior mode" (PM). Whereas FB fully accounts for uncertainty about the parameter vector in θ_i, PM ignores it completely.

We emphasize that both approaches are fully out-of-sample procedures and can therefore be implemented in real time.

Table 3 compares these approaches using the log scoring rule. The entries in the third column of the table are $\sum p(y_t \mid y_{1:t-1}, A_i)$ and those in the fourth column are $\sum p\left(y_t \mid y_{1:t-1}, \widehat{\theta}_i(t-1), A_i\right)$, the range of summation being indicated by the first column and the model A_i by the second column in each case. The fifth column provides the difference in these log scores. Each row of the table also provides the weight on the full Bayes prediction rule (sixth column) and the posterior mode prediction rule (seventh column) in a static optimal pool of the two models. The right-most column indicates the log score of the optimal pool.

For the entire period full Bayes prediction clearly outperforms posterior mode (fifth column). The effect is smallest for the DSGE model, with successive increases for the DFM, VARD and VARL models. The same rankings occur in the pre moderation and post moderation periods, though the effects are substantially greater before than after the great moderation. The rankings do not characterize the great moderation, where differences are much smaller even though the great moderation period is slightly longer than the pre-moderation period. The optimal pools are consistent with these comparisons.

Table 3: Comparison of full Bayesian and posterior mode predictive distributions

		Log scores			Pool weights		Pool
Period	Model	FB	PM	FB-PM	FB	PM	Log score
Entire	DFM	-1083.86	-1135.10	51.24	0.816	0.184	-1082.40
Entire	DSGE	-1097.03	-1128.23	31.20	1.000	0.000	-1097.03
Entire	VARL	-1146.87	-1306.41	159.54	0.955	0.045	-1146.78
Entire	VARD	-1122.43	-1265.46	143.03	1.000	0.000	-1122.43
Pre moderation	DFM	-540.66	-581.09	40.44	0.805	0.195	-539.58
Pre moderation	DSGE	-559.74	-593.05	33.31	1.000	0.000	-559.74
Pre moderation	VARL	-599.67	-753.80	154.135	1.000	0.000	-599.67
Pre moderation	VARD	-609.05	-752.78	143.73	1.000	0.000	-609.05
Great moderation	DFM	-437.97	-443.57	5.60	0.750	0.250	-437.25
Great moderation	DSGE	-436.55	-433.88	-2.66	0.110	0.890	-433.73
Great moderation	VARL	-430.71	-424.40	-6.301	0.130	0.870	-418.38
Great moderation	VARD	-410.87	-402.98	-7.89	0.133	0.867	-399.36
Post moderation	DFM	-105.23	-110.44	5.20	1.000	0.000	-105.23
Post moderation	DSGE	-100.75	-101.30	0.44	1.000	0.000	-100.75
Post moderation	VARL	-116.49	-128.21	11.71	1.000	0.000	-116.49
Post moderation	VARD	-102.50	-109.69	7.19	1.000	0.000	-102.50

Table 4: Extrinsic fraction of total predictive variance

Series	DFM	DSGE	VARL	VARD
Consumption growth	0.0522	0.0257	0.1289	0.0847
Investment growth	0.0557	0.0240	0.1343	0.0866
Income growth	0.0627	0.0200	0.1247	0.0839
Hours worked index	0.0549	0.0202	0.1324	0.0942
Inflation	0.0480	0.0355	0.1354	0.0979
Wage growth	0.0583	0.0311	0.1415	0.0933
Fed funds rate	0.0435	0.0193	0.1386	0.1038

One would expect the advantage of full Bayes prediction to be greater in those models in which parameter uncertainty in the posterior distribution is greater. To make this somewhat more precise, one might expect the advantage to increase along with the number of parameters. That is the case here: the DSGE has 39 parameters, the DFM 99 parameters, amd the VARD and VARL models 233 parameters. A more specific characterization can be given in terms of the decomposition of predictive variance described in Section 3.3: the advantage of FB over PM prediction should be greater to the extent that extrinsic predictive variance is relatively more important in a model.

This is in fact confirmed by the variance components of the predictive distributions, approximated as described in Section 3.3. Table 4 provides the fraction of predictive variance that is extrinsic, averaged over the $T = 184$ predictive distributions. Without exception across the seven series, the ordering is the same as that of the difference in log scores between FB and PM shown in Table 3 for the entire time period. More detailed consideration of the variance decomposition, not presented here, reinforces this interpretation: with all the models, the extrinsic fraction of predictive variance is lower in the great moderation period than it is either pre moderation or post moderation.

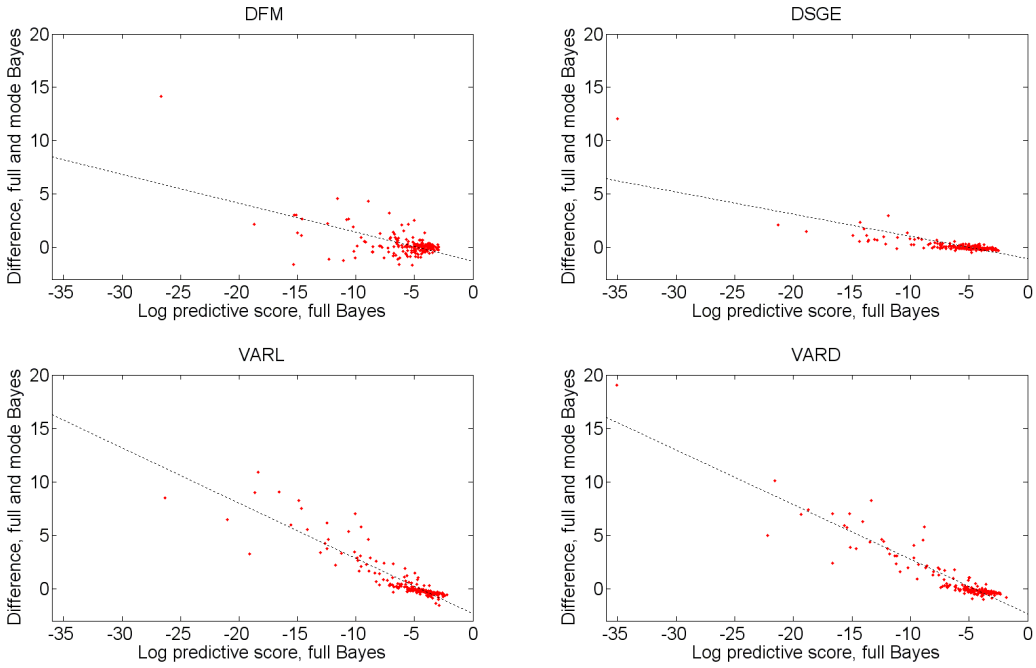

Figure 3: Each panel has one point for each of 184 quarters. The dotted line shows the least squares fit.

The full Bayes predictive distributions and posterior mode predictive distributions have no systematic differences in location, but the former are systematically more dispersed than the latter. The intrinsic component of the predictive distribution is

Gaussian. The extrinsic component is not, but appears to have broadly similar decay with increasing distance from the mode. The arithmetic of the Gaussian distribution then implies that as realizations y_t are farther from the center of the predictive distribution, the corresponding log score under the posterior mode predictive distribution decays more swiftly than it does under the full Bayes predictive distribution.

Figure 3 illustrates this effect in the four models. It is exhibited most sharply in the DSGE model and least sharply in the DFM model, but is clearly present in all four. Consistent with the evidence in Tables 3 and 4, the trade-off is strongest for the VAR models and weakest for the DSGE model. Of course, other effects can be at work as well: for instance, recent behavior (y_{t-s}, s small) that is historically atypical should magnify the effect of extrinsic uncertainty. However there is little evidence that such effects are of any quantitative importance. The main feature working to the disadvantage of predictive distributions based on posterior modes is their inability to account for regular but more extreme realizations of Y_t relative to full Bayes predictive distributions.

4.2 VAR models

To this point we have systematically considered two variants of VAR models, VARL and VARD. There are important differences in these models, chiefly with respect to long-run dynamics: VARL permits stationarity, random walk (with drift) and explosive behavior, whereas VARD imposes a unit root (with drift) thus precluding stationarity but permitting explosive behavior. These differences account for the greater number of parameters in the VARL variants. One might conjecture that the greater importance of extrinsic variance in the VARL predictive distributions originates in the variability in long-run dynamics. This is reinforced by the logical implausibility of both stationary and explosive behavior, though it seems unlikely that this would matter much in one-quarter-ahead predictive distributions.

Because of the potential simplicity afforded in going forward with three models rather than four in the model combination work in Section 5, we compare the VARL and VARD variants using the log predictive scoring criterion. Table 5 compares VARL and VARD in the same way that Table 3 compared FB and PM variants of the models.

For the entire period VARD performs substantially better than VARL. A formal Bayes factor would place the odds in favor of VARD over VARL at over $10^{10} : 1$, but this conclusion assumes that one or the other of the two models coincides with the data generating process. The static optimal pool for the entire period strongly contradicts this assumption, achieving an improvement of almost 20 points over VARD. This appears to be driven mainly by the pre moderation and great moderation periods. Comparisons based on PM are less interesting due to their inferiority relative to FB, but the implications for comparison of VARD and VARL are similar.

Going forward we exclude VARL from further analysis, and in particular undertake the model combination work in Section 5 using the three models DFM, DSGE and VARD. We have also undertaken these exercises using all four models. Model pools achieve higher log scores, but the improvements are slight compared with the dramatic

Table 5: Comparison of VARL and VARD predictive distributions

		Log scores			Pool weights		Pool
Period	Method	VARL	VARD	VARL-VARD	VARL	VARD	Log score
Entire	FB	-1146.87	-1122.43	-24.45	0.380	0.620	-1102.74
Entire	PM	-1306.41	-1265.46	-40.95	0.379	0.621	-1229.98
Pre moderation	FB	-599.67	-609.05	9.39	0.596	0.404	-584.60
Pre moderation	PM	-753.80	-752.78	-1.02	0.497	0.503	-717.17
Great moderation	FB	-430.71	-410.87	-19.84	0.124	0.876	-409.77
Great moderation	PM	-424.40	-402.98	-21.42	0.216	0.784	-399.34
Post moderation	FB	-116.49	-102.50	-13.99	0.000	1.000	-102.50
Post moderation	PM	-128.21	-109.69	-18.52	0.000	1.000	-109.69

improvement in the three-model pools over any of the individual models. If VARL substitutes for VARD in these exercises, then the three-model pools have modestly poorer performance. These results are consistent with the view that the predictive distributions of the VARL and VARD models are much closer to each other than they are to the predictive distributions of any of the other models.

4.3 Model performance

PIT tests of specification clearly indicate that each model is misspecified. The variant of the portmanteau test (Section 3.4 and Geweke and Amisano (2012c)) in this work uses the first four moments ($q = 1, 2, 3, 4$) and the first four lagged cross-products ($\ell = 1, 2, 3, 4$) of the normalized PIT z_{jt} ($t = 1, \ldots, 184$) for each constituent $j = 1, \ldots, 7$. For each constituent the asymptotic distribution of the moment and autocorrelation test statistics are each $\chi^2(4)$, and the asymptotic distribution of the joint test statistic is $\chi^2(8)$.

Table 6 reports the results of these tests using FB predictive distributions of the models indicated in the second column for each of the seven time series indicated in the first column. The p-values are based on the simulation sample of size 10^5 described in Section 3.4. For reported values above 0.02 these values are close to those of the asymptotic distributions; for smaller values they are generally larger.

Overall the tests indicate strong evidence against correct model specification. Variation in the results is driven more by variation across constituent series than be variation across models, with the Fed funds rate being by far the most problematic. Among the other series, the right column indicates that the strongest evidence against correct specification arises for consumption growth, the weakest for the income growth and hours worked index.

The evidence against correct specification arises more strongly in the failure to calibrate probabilities correctly on average (the moments test) than in any tendency for realizations to persist on one side of the conditional distribution rather than the other

Table 6: Portmanteu probability integral transform tests

		Moments		Autocorrelation		Joint	
Series	Model	Test	p-value	Test	p-vlaue	Test	p-value
Consumption growth	DFM	104.50	0.0000	3.88	0.4077	108.37	0.0001
	DSGE	32.00	0.0018	17.27	0.0040	49.28	0.0007
	VARD	54.24	0.0003	0.72	0.9441	54.96	0.0005
Investment growth	DFM	39.19	0.0009	12.68	0.0188	51.87	0.0005
	DSGE	10.90	0.0325	11.43	0.0289	22.34	0.0155
	VARD	11.44	0.0285	10.09	0.0459	21.53	0.0178
Income growth	DFM	17.03	0.0104	5.46	0.2358	22.49	0.0151
	DSGE	21.49	0.0057	2.74	0.5837	24.23	0.0113
	VARD	14.81	0.0144	4.96	0.2830	19.77	0.0355
Hours worked index	DFM	5.58	0.1566	9.37	0.0591	14.96	0.0651
	DSGE	14.33	0.0157	14.40	0.0107	28.73	0.0060
	VARD	3.82	0.3111	4.47	0.3340	8.29	0.3225
Inflation	DFM	18.89	0.0080	5.17	0.2629	24.06	0.0116
	DSGE	34.45	0.0014	55.61	0.0000	90.06	0.0001
	VARD	12.95	0.0205	6.53	0.1630	19.48	0.0257
Wage growth	DFM	29.08	0.0026	4.62	0.3177	33.70	0.0033
	DSGE	24.93	0.0039	0.75	0.9397	25.68	0.0092
	VARD	21.92	0.0054	5.50	0.2342	27.41	0.0070
Fed funds rate	DFM	937.17	0.0000	41.51	0.0000	978.69	0.0000
	DSGE	1619.18	0.0000	37.54	0.0000	1656.72	0.0000
	VARD	4130.84	0.0000	47.76	0.0000	4178.60	0.0000

(the autocorrelation test). Variations on these tests that use different combinations of moments (not reported in the table) reveal that the high values of the moments test statistics are driven by higher order moments (especially $q = 4$) than by lower order moments (e.g. $q = 1$). This indicates that the difficulty resides in failure of subjective predictive distributions to be well-calibrated for outlying realizations, which occur more frequently than these distributions imply, rather than in any failure of these distributions to be systematically shifted relative to actual behavior.

In contrast autocorrelation in realizations relative to subjective probability distributions is substantially closer to the PIT paradigm, the Fed funds rate excluded. Indeed, there is little evidence to suggest serial correlation for the DFM and VARD models. The evidence is somewhat stronger for DSGE but this is mild relative to the moments test.

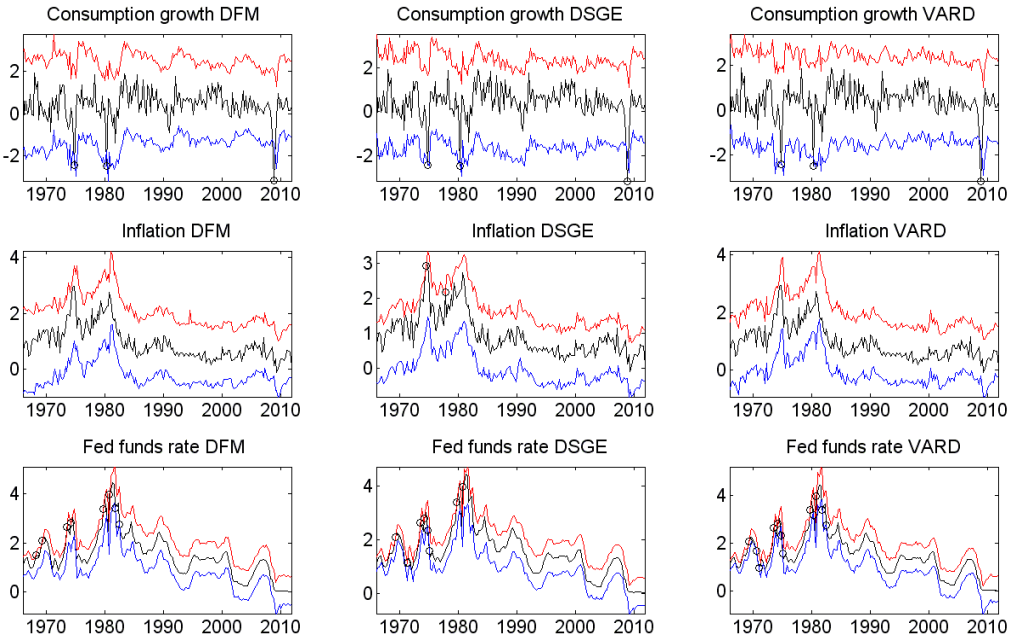

Figure 4: Centered 99% predictive credible intervals (top and bottom lines) and actual series values (middle lines). Actual values outside the intervals have circles.

Figure 4 illustrates some aspects of the behavior of predictive distributions relative to realizations. In each case the upper and lower bands indicate the centered 99% interval for the series in question from the predictive density $p\left(Y_{jt} \mid y_{1:t-1}, A_i\right)$. The line in the center indicates the corresponding realized values y_{jt}. Whenever the realized values are outside the 99% predictive interval the violation is indicated by a circle.

For the consumption growth series (top row of panels) the models all poorly anticipate the sharp drops in the two main recessions of the pre moderation period and in the

global financial crisis in the post moderation period. Credible intervals for the DFM are somewhat shorter than those for the DSGE and VARD models, leading to larger third and fourth sample moments for z_{1t} in the DFM. This produces the relatively high value of the moment test statistic in the first line of Table 6.

For the inflation series (middle row of panels) all three models are better calibrated. In the case of the DFM and VARD models all realizations are within the 99% credible interval. During the pre moderation period actual inflation tends to be persistently to the right side of the predictive distribution. This is especially evident in the DSGE model, which for this series has the largest of all the autocorrelation test statistics in Table 6.

The last row of panels of Figure 4 indicates that the extremely poor calibration of the models for the Fed funds rate conveyed by the last three rows of Table 6 originates in the pre moderation period. The models poorly anticipate the many large upward and downward movements in this period. This is not surprising, given the much more moderate behavior of interest rates prior to 1970 (Table 1), which drives the posterior distribution going into this period. The posterior distributions adjust to this more volatile behavior: note how much wider the predictive intervals are during and after the great moderation than they are in the late 1960's and early 1970's. For the Fed funds rate predictive intervals for the VARD remain somewhat narrower than those for the DFM and DSGE models in the 1970 - 1983 period, leading to larger PIT test statistics in Table 6.

The condition that one of the models includes the data generating process as a special case underlies formal Bayesian inference from multiple models, and is central to many non-Bayesian procedures as well. Any credibility that this condition might have had is refuted by the PIT tests. This suggests that procedures for prediction using several models that do not invoke this condition might produce superior predictions.

5 Model combination

Given several alternative models constructed for the purpose of assigning probabilities to future events, it is natural to investigate whether it is possible to combine models to accomplish this goal more effectively than would be possible with any one model alone. A linear pool of predictive densities (5) is an attractive approach to combination. When future events are functions of several random variables, as is the case here with Y_t, the case for using (5) is compelling: McConway (1981) shows that, under mild regularity conditions, the combination must be of the form (5) if the process of combination is to commute with any possible marginalization of the distributions involved. After summarizing the behavior of some selected static pools (Section 5.1) we turn to three kinds of linear pools: those with equal weights (Section 5.2), pools arising from Bayesian model averaging (Section 5.3), and real time optimal pools (Section 5.4). All of the analysis in this section is based on full Bayesian predictive distributions of the DFM, DSGE and VARD models.

Table 7: Log scores of models (with full Bayesian inference) and the equally weighted pool

	Log score	Relative to EWP			Model values		
Period	EWP	DFM	DSGE	VARD	DFM	DSGE	VARD
Entire	-1036.72	-47.13	-60.31	-85.70	32.10	12.08	12.58
Pre moderation	-526.06	-14.59	-33.68	-82.99	27.68	10.44	-0.67
Great moderation	-411.36	-26.61	-25.19	0.49	4.31	-1.45	12.93
Post moderation	-99.30	-5.93	-1.45	-3.20	0.11	3.09	0.32

5.1 Pools

Let $f(\mathbf{w}_{r:t})$ denote the summation on the right side of (7). This function conveys the performance of any possible linear pool with constant weights over the period from r to t, using the log scoring rule to assess performance. Figure 5 depicts the function for the four periods of interest. The domain is the three-dimensional unit simplex, depicted in two dimensions in the usual way. In all four panels the horizontal axis corresponds to the weight on the DFM model and the vertical axis to the weight on the DSGE model. Thus the value of $f(\mathbf{w}_{r:t})$ at the right vertex corresponds to the log score of the DFM model over the indicated period, at the upper vertex to the DSGE model, and at the origin to the VARD model. These values are indicated in Table 7.

The contours in each panel indicate $[\arg\max_{\mathbf{w}_{rt}} f(\mathbf{w}_{r:t})] - f(\mathbf{w}_{r:t})$, and are chosen to show increments of $0.025\,(t-r+1)$, corresponding to increments of 0.025 in the arithmetic mean of $\log\left[\sum_{i=1}^{n} w_i p(y_s \mid y_{s-1}, A_i\right]$ over the period in question. An increase from one contour to the next corresponds to an increase in the proportion $\exp(0.025) - 1$, or about 2.5%, in the geometric mean of the probability density assigned to observed events. This makes the contours directly comparable across periods of unequal length. Notice that the log score $f(\mathbf{w}_{r:t})$ is much less sensitive to changes in $\mathbf{w}_{t:t}$ near its maximum, indicated by the asterisk in each panel of Figure 5, than it is to changes close to the vertices of the simplex.

5.2 Equally weighted pools

An equally weighted pool, $w_i = 1/3$ $(i = 1, 2, 3)$, is arguably the simplest pool that could be created. These pools are indicated by the $\times$ in each panel of Figure 5. It is evident that such pools improve markedly on the log score of any given model. The log score of the equally weighted pool is also close to the maximum log score indicated by the asterisk. But this maximum log score is unattainable in real time, because the weight vector achieving this maximum is chosen on the basis of all the data for the period in question. This suggests that an equally weighted pool is likely to be a strong competitor for prediction using all three models. Subsequent analysis in this section verifies this.

Table 7 quantifies the gains from pooling with equal weights that is evident in Figure 5. The second column is the log score of the equally weighted pool, indicated by the

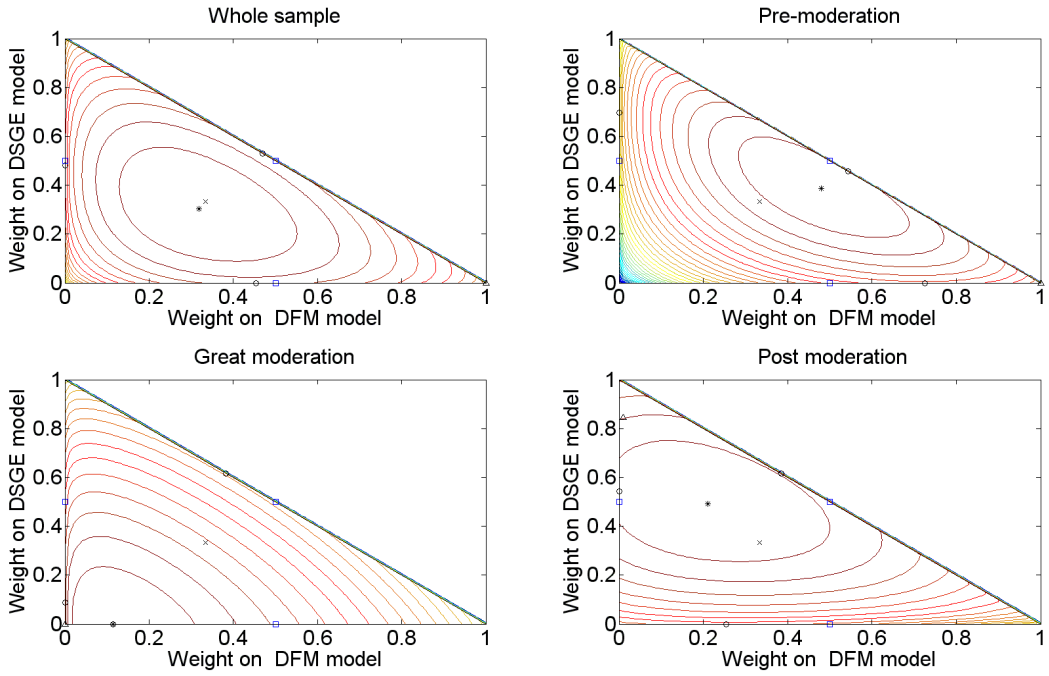

Figure 5: Log scores of pools as a function of the weight vector $\mathbf{w}_{r:t}$ in (7), normalized so that the maximum value is the same. Distance between contours is $0.025\,(t-r+1)$.

$\times$ in each panel of Figure 5. The entries in columns 3 through 5 are the log scores of the FB variants of the indicated models (see Table 3) minus the EWP log scores. Each of the last three columns measures the value of the corresponding model in an equally weighted pool as the difference between the log score of the equally weighted pool with three models and an equally weighted pool composed of the other two models. For example, the entry 34.17 for DFM for the whole period is the difference between the log score of the equally weighted pool (X in the upper left panel of Figure 5), and the log score of the equally weighted pool of the DSGE and VARD models (the square on the vertical axis in this panel). Clearly value measured in this way need not be positive, but it generally is. The DFM has the greatest value in every period except post moderation.

The equally weighted pool also provides a useful benchmark in understanding the gains from pooling and the reason that the DFM is the most valuable contributor to the pool. The left panels of Figure 6 show the model log predictive scores in all 184 quarters. These log scores tend to move together: the correlation coefficient is over 0.9 for all pairs and is driven in large part by extreme events that are assigned low predictive probability by all three models. The right panels of the figure show the difference between model log scores and the log score of the equally weighted pool. This difference cannot exceed $\log(3)$, which is the highest value on the vertical axis in these panels. There is no lower bound.

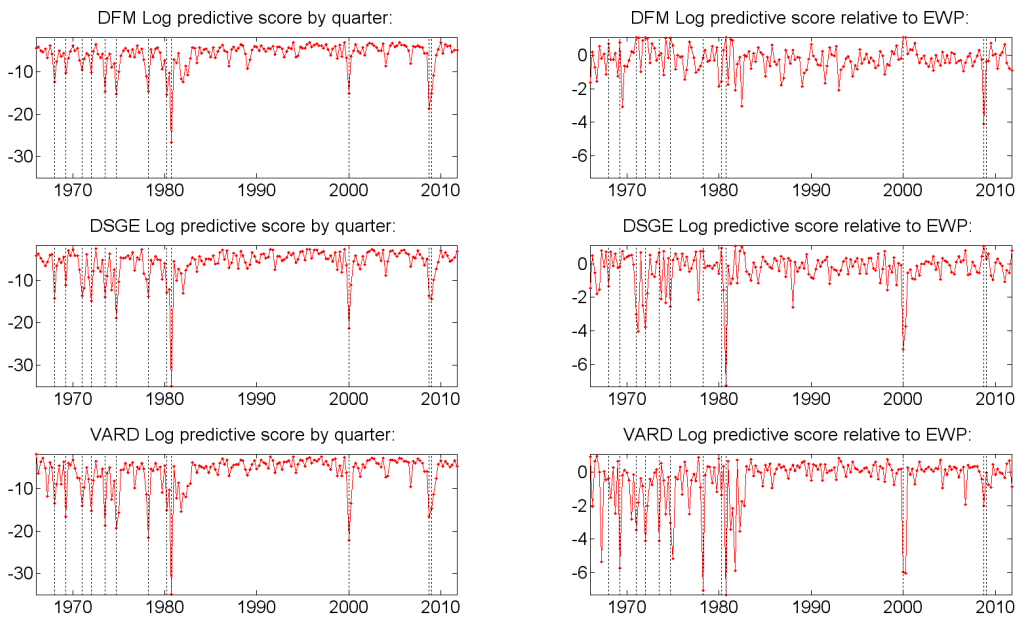

Figure 6: Model predictive log scores by quarter, absolutely (left panels) and relative to the equally weighted pool (right panels). Vertical lines indicate the 12 quarters for which the mean values taken over the model predictive log scores are the smallest.

The vertical lines in Figure 6 denote the twelve quarters in which the mean predictive log score, taken over the three models, was the lowest. In these quarters differences in model log scores also tend to be greatest, because these three models differ substantially in the probabilities they assign to rare events. This can be detected in the left column of panels but is more evident in the right column where the equally weighted pool is used as a common benchmark. In many of these quarters, one of the models substantially outperforms the other two, leading to a log score relative to the equally weighted pool that is close to $\log 3$ for that model, but low values for the other two models. The DFM enjoys this distinction most often. The only notable exception is the final quarter of 2008, where the DSGE outperformed the DFM and VARD.

The DFM contributes the pool in a manner similar to a financial asset that moves against the market. For the differences shown in the right column of panels in Figure 6, the DFM is negatively correlated with both the DSGE (-0.501) and the VARD (-0.313), whereas the DSGE and VARD are positively correlated (0.268). This property of the DFM, together with its higher log score, accounts for the fact that the DFM has the highest value in the equally weighted pool (Table 7).

This analysis tends to obscure the asymmetric behavior of the three models with respect to outlying events that is evident in Figure 6. Figure 7 highlights the different

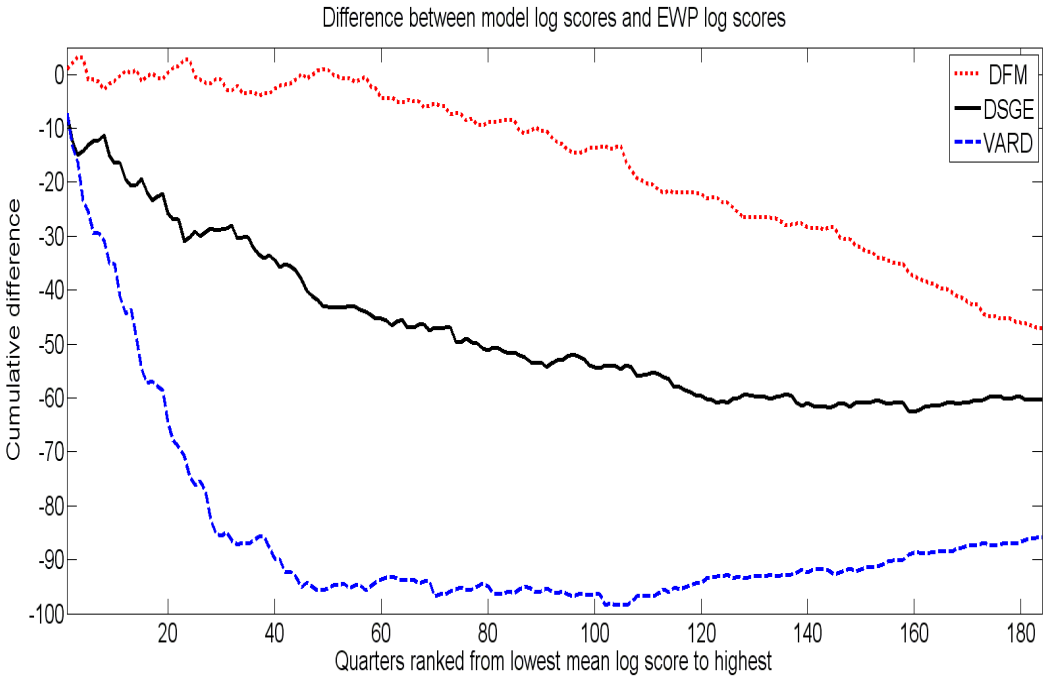

Figure 7: Quarters are ranked by the average log predictive score of the models from lowest to highest (horizontal axis). Then the cumulative differences in model log scores and the log scores of the equally weighted pool (vertical axis) are shown.

properties of the models in this dimension. It cumulates differences between the log predictive scores of the models and the log predictive score of the equally weighted pool, over quarters ordered by the mean of the log predictive score over models, lowest to highest. Thus the twelve quarters highlighted in Figure 6 correspond to the values 1 through 12 on the horizontal axis. For quarter 184 the values are those in the first row of entries in Table 6, columns 3 through 5. Through the 50 quarters with the lowest predictive log scores, the DFM dominates: its log score is very nearly that of the equally weighted pool. These same quarters account all of the deficiency in the predictive log score of the VARD; indeed it more than accounts for this difference, because the value of the VARD performs best, relative to the EWP, in those quarters in which model log scores are highest. This is also consistent with the strong performance of the VARD relative to the other two models during the great moderation (Figure 5).

The improvement in predictive log score achieved in moving from any one model to the equally weighted pool is comparable to the gain in moving from posterior mode to full Bayes predictive densities in each model. For the DFM, DSGE and VARD models the average gain from the former was 64.39(minus the average of relative log scores for the entire period in Table 7) and the average gain from the latter was 75.15 (fifth column of Table 3 for the entire period). Analysis of predictive variance for the pool with equal

Table 8: Decomposition of extrinsic variance in the equally weighted pool

	Fraction of variance extrinsic					
	Adding-up preserved			Unbiasedness preserved		
Series	Within	Between	Total	Within	Between	Total
Consumption growth	0.0507	0.0690	0.1197	0.0491	0.0982	0.1473
Investment growth	0.0523	0.0494	0.1017	0.0508	0.0707	0.1216
Income growth	0.0496	0.0460	0.0956	0.0483	0.0662	0.1145
Hours worked index	0.0489	0.0385	0.0874	0.0477	0.0551	0.1027
Inflation	0.0583	0.0645	0.1228	0.0563	0.0913	0.1476
Wage growth	0.0598	0.0419	0.1017	0.0585	0.0606	0.1191
Fed funds rate	0.0469	0.0695	0.1164	0.0448	0.0950	0.1398

weights (11) leads to the same conclusion, as indicated in the following Table 8.

This table reports two different approximations of the decomposition. The first one (columns 2 through 4) uses the expression $\sum_{i=1}^{n} w_i \left(\mu_{t-1}^i - \mu_{t-1}\right)\left(\mu_{t-1}^i - \mu_{t-1}\right)'$ to approximate $\mathrm{var}_{A_i}\left[\mathrm{E}\left(Y_{t+1} \mid A_i\right)\right]$ just as described in Section 3.3, using the equal weights $w_i = 1/3$ $(i = 1, 2, 3)$. This preserves the identity (11) when the estimates are substituted for the population values. But this also leads to the usual downward bias in the variance estimate, which is severe here because there are only three different models. Columns 5 through 7 use $w_i = 1/2$ $(i = 1, 2, 3)$, which alleviates the bias. The "within" extrinsic variance is that which drove the better performance of FB log scores relative to PM log scores, as argued in Section 4.1; the "between" extrinsic variance is due to differences between models, which drives the improvement in the log predictive scores of the equally weighted pool relative to the individual models. The order of magnitude is similar, supporting the finding that pooling and the use of full Bayes predictive distributions are of comparable importance in improving predictions using several models.

5.3 Bayesian model averaging

As discussed in Section 3.1, differences between models in log scores for full Bayes prediction are closely related to Bayes factors and posterior odds ratios. Precisely,

$$LS_{r:t}\left(A_i\right) = \sum_{s=r}^{t} \log p\left(y_s \mid y_{1:s-1}, A_i\right)$$

is the marginal likelihood in a model for which the prior distribution has density kernel

$$p\left(\theta_i\right) p\left(y_{1:r-1} \mid \theta_i, A_i\right) \tag{13}$$

and the likelihood function is $\prod_{s=r}^{t} p\left(y_s \mid y_{1:s-1}, A_i\right)$. Then $\exp\left[LS_{r:t}\left(A_i\right) - LS_{r:t}\left(A_j\right)\right]$ is the Bayes factor in favor of model A_i over model A_j, and if the prior odds ratio is $1:1$

Table 9: Bayesian model averaging weights and log scores

	End of period				Average over period			
	Model BMA weights			Log	Model BMA weights			Log
Period	DFM	DSGE	VARD	score	DFM	DSGE	VARD	score
Entire	1.0000	0.0000	0.0000	-1083.86	0.9111	0.0656	0.0234	-1084.96
Pre moderation	1.0000	0.0000	0.0000	-540.66	0.7849	0.1585	0.0566	-541.75
Great moderation	0.0000	0.0000	1.0000	-410.87	0.0055	0.0406	0.9540	-411.97
Post moderation	0.0095	0.8446	0.1459	-99.93	0.0910	0.7472	0.1618	-101.68

then this is also the posterior odds ratio. Thus the differences in log scores across models in columns 2 through 4 of Table 2.1 imply large Bayes factors in many cases; e.g. for the entire period the Bayes factor in favor of the DFM model over the DSGE model is 5.24×10^5, and therefore via (2) Bayesian model averaging (BMA) weights are often very close to 0 or 1.

Table 9 provides these weights for the periods studied. In each case, the formal interpretation is that there are three models, each with a prior distribution of the form (13) where $s = 1$ corresponds to 1951:1 and t is the last quarter before the start of the period indicated. At the start of each period the BMA weights are 1/3, and then as predictive likelihoods are accumulated through the period weights move toward 0 or 1 until at the end of the period they have the values shown in the table.

The "End of period" entries show the BMA weights at the end of the period and the log score that results when these weights are applied to the model log scores for the entire period. This is the log score for BMA most commonly reported, but it cannot be realized in real time. The "Average over period" entries are based on BMA weights updated each observation in the period, with average weights shown. The log score is figured by applying the BMA weights updated through period $t-1$ to the predictive densities for period t, which are then evaluated at y_t.

The DFM strongly dominates the entire period and pre moderation; VARD strongly dominates the great moderation; and DSGE weakly dominates the post moderation. In Figure 5 the BMA weights are indicated by the triangle in each panel, and the fifth column of Table 9 shows the corresponding log predictive scores for the Bayesian model averages. They are all lower than the log scores of the equally weighted pool

Suppose a Bayesian econometrician maintained the hypothesis underlying BMA: that the data generating process is exactly $p\left(Y_t \mid Y_{1:t-1}, \theta_i^*, A_i\right)$ for some one of the models A_i and some particular value θ_i^* of that model's parameter vector, though which model and which specific values of the parameter vector are unknown. Were this econometrician to have started to work at the end of 1965, using (13) as the kernel of her prior density, then her BMA weights would have evolved as indicated in the upper panel of Figure 8.[3] The sum of the BMA weights on the DFM and DSGE models drops below 0.1 in

[3]This exercise is consistent with the values in Table 9 for the entire period and the pre moderation period, because these exercises all start in 1966:1. It is not consistent with the values in Table 9 for

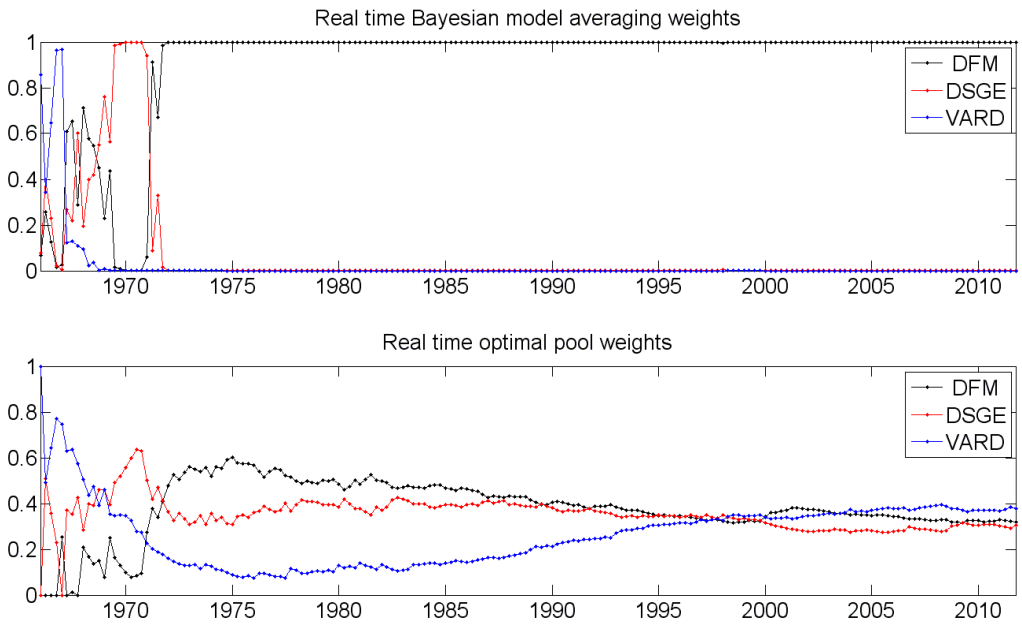

Figure 8: Bayesian model averaging and optimal pool weights updated each quarter

the last quarter of 1971, and below 0.01 the following quarter where it remains for the rest of the entire period. The largest sum of BMA weights on the DFM and DSGE models in the rest of the entire period is 0.00345 in the first quarter of 1998, and for most quarters beyond 1972:1 the sum is less than 10^{-6}. This is all consistent with the typical asymptotic distribution of BMA weights outlined in Section 3.1.

5.4 Optimal pooling

Optimal pools can be constructed for any period as described in Section 3.2, leading to the weight vector $\mathbf{w}^*_{r:t}$ (7), and the corresponding log score for the optimal pool is then given by the summation on the right side of (7) evaluated at the optimal weights. These weights, and related statistics, provide useful summaries of the interaction between models in prediction over particular time periods. However, they could not have been used in real time during the period in question, and any pooling procedure that could be used would lead to a lower log score for the resulting pool. The optimal weights could be used going forward, for example in quarter $t + 1$.

In each panel of Figure 5 the asterisk indicates the weights for the optimal pool formed this way, and columns 2 through 4 of Table 10 provide their values. The value

the great moderation and post moderation periods becuase those calculations begin at the start of the respective periods.

Table 10: Static optimal pools (end of period)

	Optimal pool weights			Log	Model values		
Period	DFM	DSGE	VARD	score	DFM	DSGE	VARD
Entire	0.318	0.304	0.378	-1036.58	32.22	12.02	12.65
Pre moderation	0.480	0.387	0.133	-523.66	28.14	10.02	1.65
Great moderation	0.114	0.000	0.886	-402.62	7.70	0.00	21.21
Post moderation	0.210	0.492	0.298	-99.11	0.29	3.01	0.43

Table 11: Real-time optimal pools (average over period)

	Optimal pool weights			Log	Model values		
Period	DFM	DSGE	VARD	score	DFM	DSGE	VARD
Entire	0.378	0.354	0.268	-1043.41	29.224	8.71	10.15
Pre moderation	0.395	0.385	0.220	-529.97	25.41	6.94	-0.63
Great moderation	0.048	0.009	0.943	-407.58	3.84	-0.24	18.75
Post moderation	0.284	0.521	0.194	-104.73	-3.24	0.24	0.22

of the function at this point is given in column 5 of Table 10.The optimal pool could also be formed eliminating one of the three models. In each panel of Figure 5 the circles indicate the weights in these pools. For example, if the DFM were eliminated from the pool for the entire period then the DSGE model in the resulting optimal pool would have the weight indicated by the circle on the vertical axis of that panel. The log score function at this point is necessarily smaller than the log score function in the three-model optimal pool. It is reasonable to refer to the decrease as the value of the DFM in this pool. The last three columns of Table 10 provide the values of each of the models in each of the periods studied: this is the decrease in the value of the function between the corresponding asterisk in circle in the relevant panel of Figure 5.

For the entire period the DFM is the most valuable model, with a contribution to log score more than 2.5 times as great as the contributions of DSGE and VARD, which are in turn similar. This is due primarily to the fact that the log score of the DFM performs substantially better than do the DSGE and VARD models in quarters where realizations y_t were the least probable under any model, as discussed in Section 5.2. These contributions are concentrated in the pre moderation period, driving the lower values of DFM in the other two periods. Consistent with our findings using other methods of analysis, the VARD is especially valuable during the great moderation. The DSGE dominates the contribution post moderation, although effects in that period are muted by its short duration.

For practical prediction the relevant question is how well optimal pools perform in real time. The natural way for an econometrician to use optimal pools is to compute the optimal weights based on $y_{r:t-1}$ and then attach those weights to the predictive densities

$p(Y_t \mid y_{1:t-1}, A_i)$. Specifically, the econometrician finds

$$\mathbf{w}^*_{r:t-1} = \arg\max_{\mathbf{w}} \sum_{s=r}^{t-1} \log \left[\sum_{i=1}^{n} w_i p(y_s \mid y_{r:s}, A_i) \right]$$

and then uses the predictive density

$$\sum_{i=1}^{n} w^*_{r:t-1,i} p(Y_t \mid y_{r:t-1}, A_i)$$

for quarter t.

Table 11 reports some aspects of the results of this procedure. For period indicated in the table, r is the first quarter in the period. For each period the average weights in columns 2 through 4 are the elements of $(u-r+1)^{-1} \sum_{t=r}^{u} \mathbf{w}^*_{rt}$, where u is the last quarter in the period. The log score in column 5 is the predictive likelihood for this real-time optimal pool,

$$\sum_{t=r}^{u} \log \left[\sum_{i=1}^{n} w^*_{r,t-1,i} p(y_t \mid y_{r:t-1}, A_i) \right], \tag{14}$$

with $w^*_{r,r-1,i} = n^{-1}$. Thus the log scores in this column are directly comparable with those for individual models, equally weighted pools, and Bayesian model averaging: all report results an econometrician could have achieved in real time.

Log scores of real-time optimal pools are, algorithmically, lower than those of static optimal pools for the same period. Comparisons of the corresponding entries of Tables 10 and 11 show that the decrease is between 5 and 7 points in each period. More significantly, the equally weighted pool (Table 9) outperforms the real-time optimal pool in three of the four periods examined. Only in the great moderation, in which the three models display the greatest asymmetry (Figure 5), does the equally weighted pool fall short of the real-time optimal pool.

The last three columns of Table 11 provide real-time model values, parallel to the values for static optimization in Table 10. For each model the real-time optimal pooling exercise is repeated eliminating that model but retaining the others. The value is the difference between the original log predictive likelihood (14) and the corresponding expression with the model eliminated. Such model values are not algorithmically nonnegative, as is the case in the static optimal pool. All model values for all periods are lower in the real-time optimal pool (Table 11) than in the corresponding static optimal pool (Table 10), and the two models with the lowest values in the latter pool have negative values here.

6 Conclusion

The principal conclusion of this work is that predictions are best formed from several macroeconomic models by pooling the Bayesian predictive distributions of the models.

Table 12: Real-time log scores of models and pools, entire period

Model log scores				Model pooling	
Model	PM	FB	FB-PM	Method	Log score
DFM	-1135.10	-1083.86	51.24	BMA	-1084.96
DSGE	-1128.23	-1097.03	31.20	RTOP	-1043.41
VARD	-1265.46	-1122.43	143.03	EWP	-1036.72
Mean	-1176.26	-1101.11	75.15		

This is supported by multiple analyses that contribute to the understanding of this finding and suggest that the conclusions would be reproduced if the work were undertaken with other variants of the three families of models considered. We think it likely that similar findings would also emerge in other data sets, though the postwar US data is unique in extent, continuity and quality; and also with other families of models, either existing or yet to be formulated, of similar intellectual and empirical pedigree.

The procedures used here emulate what could have been done in "real time" by adding the most recent quarter's data and updating posterior distributions accordingly at the end of each quarter for the purposes of predicting the following quarter. The results do not attempt to use the data releases actually at hand each quarter; we doubt that this extension would overturn the main findings in this work. Table 12 summarizes the main quantitative results for the entire postwar US data set.

The first part of the principal conclsion is that gains to using full Bayesian (FB) predictive distributions, as opposed to a "plug in" distribution that replaces the random parameter vector with its value at the posterior mode (PM), are substantial. In Table 12 the posterior mode predictive log score, averaged across the three models, falls short of the average for the full Bayes predictive distributions by 75.15 points. To appreciate this difference, it implies that the full Bayes predictive distributions increase the geometric average of the probabilities $p\left(y_t \mid y_{1:t-1}, A_i\right)$ of the observed Y_t over the period 1966 - 2011 by $100\left[\exp\left(75.15/184\right) - 1\right]\% = 50.4\%$ on average over the three models.

The key deficiency with the posterior mode is its failure to account for parameter uncertainty, and therefore extends to any "plug in" procedure. This is consistent with econometric theory. The analysis in Section 4 linked the results to this theory from a number of analytical perspectives that shed further light on the circumstances that magnify the differences. Parameter uncertainty is increasingly important moving from the DSGE, to the DFM, to the VARD model, gauged either by the crude measure of number of parameters or the more precise measure of the extrinsic fraction of predictive variance, and this accounts for the differences in models in column 4 of Table 12.This interpretation implies that these differences will be most manifest when realizations in the quarter predicted are at the lowest ranges of the predictive densities, and this implication is borne out in the analysis. In the context of a single model, the full Bayesian predictive distribution is not only formally more correct, it is essential to competitive prediction.

The second part of the principal conclusion is that further gains, of a similar order of magnitude, can be accomplished by pooling. This study examined three variants of pooling: Bayesian model averaging (BMA), real-time optimal pooling (RTOP) and equally weighted pools (EWP), all using full Bayesian predictive distributions. Bayesian model averaging puts essentially the entire weight on the DFM model, consistent with the common finding that in large data sets the model with the highest marginal likelihood dominates the other models. The resulting BMA log score is closer to the mean of the individual models using FB than it is to either the RTOP or EWP log scores. Both the RTOP and the EWP improve substantially on the mean log scores of the models using FB, by 57.70 and 64.39 points, respectively. This is comparable with the gains made in moving from PM to FB in the individual models. For the EWP the increase in the geometric average of the probabilities of the observed Y_t over the period 1966-2011 is 41.9%. Combining this with the previous improvement with FB over PM, the EWP increases the geometric average of these probabilities by 113.5% relative to PM predictive distributions in a single model.

BMA conditions on one of the models actually being the data generating process, and the log score of one model often dominates that of the others even in samples of modest size. Taken together, this virtually eliminates any posterior uncertainty about which model is "true." But the condition is not credible. Different models capture different aspects of reality. Their log scores move together, but each bears a distinct relationship to the average log score. In these circumstances there are gains to using a linear combination of predictive densities even when it is quite clear which is highest on average. The situation is analogous (though the math is not identical) to portfolio diversification, in which major gains accrue even when diversification is not optimal.

If there is in fact a true data generating process underlying reality, then given standard side conditions there is a limiting value for the weight vector in the optimal pool (Geweke and Amisano, 2011), and if the econometrician knew that vector then eventually the pool using that vector would outperform the EWP in log score. Of course the is vector is really unknown, and the RTOP uses a weight vector that changes from quarter to quarter and thereby introduces noise, especially when there are few observations. Log scores of pools are not very sensitive to weights until some of the weights approach zero, and consequently it can be the case that fixed reasonable weights outperform an effort to re-optimize the pool each quarter.

Balke N, Gordon R J (1986). Appendix B. Historical Data. In Gordon R J (Ed.), The American business cycle: continuity and change. NBER and University of Chicago Press.

Sargent T, Sims CA (1977). Business cycle modeling without pretending to have too much a-priori economic theory, in: C. Sims et al., eds., New methods in business cycle research (Federal Reserve Bank of Minneapolis, Minneapolis).

References

An S, Schorfheide F (2007). Bayesian analysis of DSGE models. Econometric Reviews 26: 113-172.

Balke N, Gordon R J (1986). Appendix B. Historical Data. In Gordon R J (Ed.), The American business cycle: continuity and change. NBER and University of Chicago Press.

Berkowitz J (2001). Testing density forecasts with applications to risk management. Journal of Business and Economic Statistics 19: 465-474.

Bernanke B, Boivin J (2003). Monetary policy in a data-rich environment. Journal of Monetary Economics 50: 525-546.

Bernardo J, Smith AFM (1993). Bayesian Theory. Oxford: Oxford University Press.

Del Negro M, Schorfheide F (2012). DSGE model-based forecasting. NY FED Staff Report No. 554, March 2012.

Doan T, Litterman R, Sims CA (1984). Forecasting and conditional projection using realistic prior distributions. Econometric Reviews 3: 1–100.

Doucet A, Godsill S, Robert C. (2002). Marginal maximum a posteriori estimation using Markov chain Monte Carlo. Statistics and Computing 12: 77–84.

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: 830-840.

Geweke J (1977). The dynamic factor analysis of economic time series models. in D Aigner and A Goldberger (eds) Latent Variables in Socioeconomic Models 365-383 Amsterdam. North-Holland.

Geweke J, Amisano G (2010). Evaluating the predictive distributions of Bayesian models of asset returns. International Journal of Forecasting 26: 216-230.

Geweke J, Amisano G (2011). Optimal prediction pools. Journal of Econometrics 164:130-141.

Geweke J, Amisano G (2012a). Prediction with misspecified models. American Economic Review Papers and Proceedings, forthcoming.

Geweke J, Amisano G (2012b). Analysis of variance for Bayesian inference. Econometric Reviews, forthcoming.

Geweke J, Amisano G (2012c). Portmanteau probability integral transform tests. Working paper in preparation.

Litterman R (1986). Forecasting With Bayesian vector autoregressions. five years of experience Journal of Business and Economic Statistics 4: 25–38.

McConnell M, Pérez-Quirós G (2000). Output fluctuations in the United States. What has changed since the early 1980s?. American Economic Review. 90: 1464-1476.

McConway KJ (1981). Marginalization and linear opinion pools. Journal of the American Statistical Association 76: 410-414.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. Annals of Mathematical Statistics 23: 470-472.

Sargent T, Sims CA (1977). Business cycle modeling without pretending to have too much a-priori economic theory, in: C. Sims et al., eds., New methods in business cycle research (Federal Reserve Bank of Minneapolis, Minneapolis).

Sims, CA (1980). Macroeconomics and reality. Econometrica 48: 1-48.

Smets F, Wouters R (2003). An estimated dynamic stochastic general equilibrium model of the euro area. Journal of the European Economic Association. 1: 1123–1175.

Smets F, Wouters R (2007). Shocks and frictions in US business cycles: A Bayesian DSGE approach. American Economic Review 97: 586-606.

Smith JQ (1985). Diagnostic checks of non-standard time series models. Journal of Forecasting 4: 283-291.

Stock JH, Watson MW (2002a). Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association. 97: 147-162.

Stock JH, Watson MW (2002b). Has the business cycle changed and why? NBER Macroeconomics Annual 2002, Mark Gertler and Ken Rogoff (eds), MIT Press.

Stock JH, Watson MW (2005). Implications of Dynamic Factor models for VAR analysis. NBER Working Paper 11467.

.