EFWW
7/11/06
9:45 am

Combining Forecasts From Nested Models [*]

Todd E. Clark
Federal Reserve Bank of Kansas City

Michael W. McCracken
Board of Governors of the Federal Reserve System

February 2006

Abstract

Motivated by the common finding that linear autoregressive models forecast better than models that incorporate additional information, this paper presents analytical, Monte Carlo, and empirical evidence on the effectiveness of combining forecasts from nested models. In our analytics, the unrestricted model is true, but as the sample size grows, the DGP converges to the restricted model. This approach captures the practical reality that the predictive content of variables of interest is often low. We derive MSE-minimizing weights for combining the restricted and unrestricted forecasts. In the Monte Carlo and empirical analysis, we compare the effectiveness of our combination approach against related alternatives, such as Bayesian estimation.

JEL Nos.: C53, C52
Keywords: forecast combination, predictability, forecast evaluation

[*] *Clark (corresponding author)*: Economic Research Dept.; Federal Reserve Bank of Kansas City; 925 Grand; Kansas City, MO 64198; `todd.e.clark@kc.frb.org`. *McCracken*: Board of Governors of the Federal Reserve System; 20th and Constitution N.W.; Mail Stop #61; Washington, D.C. 20551; `michael.w.mccracken@frb.gov`. Portions of this paper were written while Michael McCracken was on the economics department faculty of the University of Missouri–Columbia. We gratefully acknowledge helpful comments from Jan Groen, David Hendry, Jim Stock, seminar participants at the Deutsch Bundesbank and Federal Reserve Bank of Kansas City, and participants at the Bank of England Workshop on Econometric Forecasting Models and Methods and the 2005 World Congress of the Econometric Society. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Kansas City, Board of Governors, Federal Reserve System, or any of its staff.

1 Introduction

Forecasters are well aware of the so–called principle of parsimony: "simple, parsimonious models tend to be best for out–of–sample forecasting..." (Diebold (1998)). Although an emphasis on parsimony may be justified on various grounds, parameter estimation error is one key reason. In many practical situations, estimating additional parameters can raise the forecast error variance above what might be obtained with a simple model. Such is clearly true when the additional parameters have population values of zero. But the same can apply even when the population values of the additional parameters are non–zero, if the marginal explanatory power associated with the additional parameters is low enough. In such cases, in finite samples the additional parameter estimation noise may raise the forecast error variance more than including information from additional variables lowers it. For example, simulation evidence in Clark and McCracken (2005b) shows that even though the true model relates inflation to the output gap, in finite samples a simple AR model for inflation will often forecast as well as or better than the true model. Clark and West (2004, 2005) obtain a similar result for some other applications.

As this discussion suggests, parameter estimation noise creates a forecast accuracy tradeoff. Excluding variables that truly belong in the model could adversely affect forecast accuracy. Yet including the variables could raise the forecast error variance if the associated parameters are estimated sufficiently imprecisely. In light of such a tradeoff, combining forecasts from the unrestricted and restricted (or parsimonious) models could improve forecast accuracy. Such combination could be seen as a form of shrinkage, which various studies, such as Stock and Watson (2003), have found to be effective in forecasting.

Accordingly, this paper presents analytical, Monte Carlo, and empirical evidence on the effectiveness of combining forecasts from nested models. Our analytics are based on models we characterize as "weakly" (or, in the terminology of Stock and Watson (2005), "asymptotically") nested: the unrestricted model is the true model, but as the sample size grows large, the DGP converges to the restricted model. This analytic approach captures the practical reality that, in many instances, the predictive content of some variables of interest is quite low. Although we focus the presented analysis on nested linear models, our results could be generalized to nested nonlinear models.

Under the weak nesting specification, we derive weights for combining the forecasts from estimates of the restricted and unrestricted models that are optimal in the sense of minimizing the forecast mean square error (MSE). We then characterize the settings under which the combination forecast will be better than either the restricted or unrestricted forecasts, and the settings in which either the restricted or unrestricted forecast will be most accurate. In the special case in which the coefficients on the extra variables in the unrestricted model are of a magnitude that makes the restricted and unrestricted models equally accurate, the MSE–minimizing forecast is a simple, equally–weighted average of the restricted and unrestricted forecasts.

In the Monte Carlo and empirical analysis, we show that our proposed approach of combining forecasts from nested models works well compared to various alternative methods of forecasting. These alternatives include: using model selection criteria such as the SIC to determine the optimal model (choosing between the restricted and unrestricted, estimated at time t) for forecasting at time $t+1$; Bayesian estimation with priors that push certain coefficients toward zero; and Bayesian model averaging of the restricted and unrestricted models. To ensure the practical relevance of our results, we base our Monte Carlo experiments on DGPs calibrated to actual empirical applications, and, in our empirical work, we consider a wide range of applications. Overall, in both the Monte Carlo and empirical results, two forecast methods seem to work best, in the sense of consistently yielding improvements in MSE: simple averaging of the restricted and unrestricted model forecasts, and Bayesian (Minnesota BVAR) estimation of the unrestricted model.

Our results build on much prior work on forecast combination. Research focused on non–nested models ranges from the early work of Bates and Granger (1969) to recent contributions of Stock and Watson (2003, 2005), Elliott and Timmermann (2004), and Smith and Wallis (2005).[1] Combination of nested model forecasts has been considered only occasionally, in such studies as Filardo (1999), Hendry and Clements (2004), and Goyal and Welch (2003). Forecasts based on Bayesian model averaging as developed in such studies as Wright (2003) could also combine forecasts from nested models. Of course, such Bayesian methods of combination are predicated on model uncertainty. In contrast, our paper provides a theoretical rationale for nested model combination in the absence of model uncertainty. We go on to extend prior work by providing a detailed analysis of the effectiveness of forecast combination in practice.

The paper proceeds as follows. Section 2 provides theoretical results on the possible

[1] A more complete survey of the extensive combination literature is beyond the scope of this paper. For a comprehensive survey, see Timmermann (2004).

gains from combination of forecasts from nested models, including the optimal combination weight. In section 3 we present Monte Carlo evidence on the finite sample effectiveness of our proposed forecast combination methods and various alternatives. Section 4 compares the effectiveness of the forecast methods in a range of empirical applications. Section 5 concludes. Additional details pertaining to theory and data are presented in Appendixes 1 and 2.

2 Theory

We begin by using a simple example to illustrate our essential ideas and results. We then proceed to the more general case. After detailing the necessary notation and assumptions, we provide an analytical characterization of the bias-variance tradeoff, created by weak predictability, involved in choosing among restricted, unrestricted, and combined forecasts. In light of that tradeoff, we then derive the optimal combination weights.

2.1 A simple example

Suppose we are interested in forecasting y_{t+1} from $t = T$ through $T + P - 1$, using a simple model relating y_{t+1} to a constant and a strictly exogenous, scalar variable x_t. Suppose, however, that the predictive content of x_t for y_{t+1} may be weak. To capture this possibility, we model the population relationship between y_{t+1} and x_t using local-to-zero asymptotics, such that, as the sample size grows large, the predictive content of x_t shrinks to zero (assume that, apart from the local element, the model fits in the framework of the usual classical normal regression model, with homoskedastic errors, etc.):

$$y_{t+1} = \beta_0 + \frac{\beta_1}{\sqrt{T}} x_t + u_{t+1}, \;\; E(x_t u_{t+1}) = 0, \;\; E(u_{t+1}^2) = \sigma^2. \tag{1}$$

In light of x's weak predictive content, the forecast from an estimated model relating y_{t+1} to a constant and x_t (henceforth, the *unrestricted* model) could be less accurate than a forecast from a model relating y_{t+1} to just a constant (the *restricted* model). Whether that is so depends on the "signal" and "noise" associated with x_t and its estimated coefficient. Under the local asymptotics incorporated in the DGP (1), the "signal" component is $\beta_1^2 \sigma_x^2$, while the "noise" component is σ^2. The signal-to-noise ratio is then $\beta_1^2 \sigma_x^2 / \sigma^2$. Given σ^2, higher values of the coefficient on x or the variance of x raise the signal relative to the noise; given the other parameters, a higher residual variance σ^2 increases the noise, reducing the signal-to-noise ratio.

In light of the tradeoff considerations described in the introduction, a combination of the unrestricted and restricted model forecasts could be more accurate than either of the individual forecasts. Letting $\hat{y}_{1,t+1}$ denote the forecast from the restricted model and $\hat{y}_{2,t+1}$ represent the unrestricted model's forecast (both based on models estimated by OLS with data through period t), we consider a combined forecast

$$\alpha_t \hat{y}_{1,t+1} + (1 - \alpha_t)\hat{y}_{2,t+1}. \tag{2}$$

[Under our formulation, the optimal combination weight is updated in real time (at each forecast point t), as forecasting moves forward in time.] We then analytically determine the weight α_t^* that yields the forecast with lowest expected squared error in period $t+1$. Our formulation allows for the extreme cases in which the restricted model is best ($\alpha_t^* = 1$) or the unrestricted model is best ($\alpha_t^* = 0$).

As we establish more formally below, the MSE–minimizing (estimated) combination weight α_t^* is a function of the signal–to–noise ratio:

$$\hat{\alpha}_t^* = \left[1 + \left(\frac{\left(\sqrt{t}\ \hat{b}_1\right)^2 \hat{\sigma}_x^2}{\hat{\sigma}^2}\right)\right]^{-1}, \tag{3}$$

where $\hat{b}_1$ denotes the coefficient on x_t ($\sqrt{t}\ \hat{b}_1$ corresponds to an estimate of the local population coefficient β_1), $\hat{\sigma}_x^2$ denotes the variance of x_{t-1}, and $\hat{\sigma}^2$ denotes the error variance of the unrestricted forecast model, all estimated at time t (for forecasting at $t+1$).[2] As this result indicates, if the predictive content of x is such that the signal-to-noise ratio equals 1, then $\hat{\alpha}_t^* = .5$: the MSE–minimizing forecast is a simple average of the restricted and unrestricted model forecasts.

2.2 The general case: environment

In the general case, the possibility of weak predictors is modeled using a sequence of linear DGPs of the form (**Assumption 1**)[3]

$$y_{T,t+\tau} = x'_{T,2,t}\beta_T^* + u_{T,t+\tau} = x'_{T,1,t}\beta_1^* + x'_{T,22,t}(T^{-1/2}\beta_{22}^*) + u_{T,t+\tau}, \tag{4}$$
$$Ex_{T,2,t}u_{T,t+\tau} \equiv Eh_{T,t+\tau} = 0 \text{ for all } t = 1, ..., T, ...T + P - \tau.$$

[2] Clements and Hendry (1998) derive a similar result, for the combination of a forecast based on the unconditional mean and a forecast based on an AR(1) model without intercept, the model assumed to generate the data.

[3] The parameter $\beta_{T,t}^*$ does not vary with the forecast horizon τ since, in our analysis, τ is treated as fixed.

Note that we allow the dependent variable $y_{T,t+\tau}$, the predictors $x_{T,2,t}$ and the error term $u_{T,t+\tau}$ to depend upon T, the initial forecasting origin. This dependence allows the time variation in the parameters to influence their marginal distributions. This is necessary if we want to allow lagged dependent variables to be predictors.

At each origin of forecasting $t = T, ...T+P-\tau$, we observe the sequence $\{y_{T,j}, x'_{T,2,j}\}_{j=1}^{t}$. Forecasts of the scalar $y_{T,t+\tau}$, $\tau \geq 1$, are generated using a $(k \times 1, k = k_1 + k_2)$ vector of covariates $x_{T,2,t} = (x'_{T,1,t}, x'_{T,22,t})'$, linear parametric models $x'_{T,i,t}\beta_i^*$, $i = 1, 2$, and a combination of the two models, $\alpha_t x'_{T,1,t}\beta_1^* + (1-\alpha_t)x'_{T,2,t}\beta_2^*$. The parameters are estimated using OLS (**Assumption 2**) and hence $\hat{\beta}_{i,t} = \arg\min t^{-1}\sum_{s=1}^{t-\tau}(y_{T,s+\tau} - x'_{T,i,s}\beta_i)^2$, $i = 1, 2$, for the restricted and unrestricted, respectively. We denote the loss associated with the τ-step ahead forecast errors as $\hat{u}^2_{i,t+\tau} = (y_{T,t+\tau} - x'_{T,i,t}\hat{\beta}_{i,t})^2$, $i = 1, 2$, and $\hat{u}^2_{W,t+\tau} = (y_{T,t+\tau} - \alpha_t x'_{T,1,t}\hat{\beta}_{1,t} - (1-\alpha_t)x'_{T,2,t}\hat{\beta}_{2,t})^2$ for the restricted, unrestricted, and combined, respectively.

The following additional notation will be used. Let $H_{T,i}(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{T,i,s}u_{T,s+\tau}) = (t^{-1}\sum_{s=1}^{t-\tau} h_{T,i,s+\tau})$, $B_{T,i}(t) = (t^{-1}\sum_{s=1}^{t-\tau} x_{T,i,s}x'_{T,i,s})^{-1}$, and $B_i = \lim_{T\to\infty}(Ex_{T,i,s}x'_{T,i,s})^{-1}$ for $i = 1, 2$. For $U_{T,t} = (h'_{T,2,t+\tau}, vec(x_{T,2,t}x'_{T,2,t})')'$, $V = \sum_{j=-\tau+1}^{\tau-1}\Omega_{11,j}$, where $\Omega_{11,j}$ is the upper block-diagonal element of Ω_j defined below, and $\Rightarrow$ denotes weak convergence. For any $(m \times n)$ matrix A with elements $a_{i,j}$ and column vectors a_j, let: $vec(A)$ denote the $(mn \times 1)$ vector $[a'_1, a'_2, ..., a'_n]'$; $|A|$ denote the max norm; and $tr(A)$ denote the trace. Let $\sup_t = \sup_{T \leq t \leq T+P}$. Finally, we define variable selection matrices and a coefficient vector that appears directly in our key combination results: $J = (I_{k_1\times k_1}, 0_{k_1\times k_2})'$, $J_2 = (0_{k_2\times k_1}, I_{k_2\times k_2})'$ and $\delta = (0_{1\times k_1}, \beta_{22}^{*'})'$.

To derive our general results, we need two more assumptions (in addition to our assumptions (1 and 2) of a DGP with weak predictability and OLS–estimated linear forecasting models).

Assumption 3: (a) $T^{-1}\sum_{t=1}^{[rT]} U_{T,t}U'_{T,t-j} \Rightarrow r\Omega_j$ where $\Omega_j = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} E(U_{T,t}U'_{T,t-j})$ for all $j \geq 0$, (b) $\Omega_{11,j} = 0$ all $j \geq \tau$, (c) $\sup_{T\geq 1, t\leq T+P} E|U_{T,t}|^{2q} < \infty$ some $q > 1$, (d) The zero mean triangular array $U_{T,t} - EU_{T,t} = (h'_{T,2,t+\tau}, vec(x_{T,2,t}x'_{T,2,t} - Ex_{T,2,t}x'_{T,2,t})')'$ satisfies Theorem 3.2 of De Jong and Davidson (2000).

Assumption 4: For $s \in (1, 1+\lambda_P]$, (a) $\alpha_t \Rightarrow \alpha(s) \in [0, 1]$, (b) $\lim_{T\to\infty} P/T = \lambda_P \in (0, \infty)$.

Assumption 3 imposes three types of conditions. First, in (a) and (c) we require that the observables, while not necessarily covariance stationary, are asymptotically mean square stationary with finite second moments. We do so in order to allow the observables to have marginal distributions that vary as the weak predictive ability strengthens along with the sample size but are 'well-behaved' enough that, for example, sample averages converge in probability to the appropriate population means. Second, in (b) we impose the restriction that the τ-step ahead forecast errors are MA($\tau - 1$). We do so in order to emphasize the role that weak predictors have on forecasting without also introducing other forms of model misspecification. Finally, in (d) we impose the high level assumption that, in particular, $h_{T,2,t+\tau}$ satisfies Theorem 3.2 of De Jong and Davidson (2000). By doing so we not only insure (results needed in Appendix 1) that certain weighted partial sums converge weakly to standard Brownian motion, but also allow ourselves to take advantage of various results pertaining to convergence in distribution to stochastic integrals.

Our final assumption is unique: we permit the combining weights to change with time. In this way, we allow the forecasting agent to balance the bias-variance tradeoff differently across time as the increasing sample size provides stronger evidence of predictive ability. Finally, we impose the requirement that $\lim_{T\to\infty} P/T = \lambda_P \in (0, \infty)$ and hence the duration of forecasting is finite but non-trivial.

2.3 Theoretical results on the tradeoff

Our characterization of the bias-variance tradeoff associated with weak predictability is based on $\sum_{t=T}^{T+P-\tau}(\hat{u}_{2,t+\tau}^2 - \hat{u}_{W,t+\tau}^2)$, the difference in the (normalized) MSEs of the unrestricted and combined forecasts. In Appendix 1, we provide a general characterization of the tradeoff, in Theorem 1. But in the absence of a closed form solution for the limiting distribution of the loss differential (the distribution provided in Appendix 1), we proceed in this section to focus on the mean of this loss differential.

From the general case proved in Appendix 1, we first establish the expected value of the loss differential, in the following corollary.

Corollary 1: $E\sum_{t=T}^{T+P}(\hat{u}_{2,t+\tau}^2 - \hat{u}_{W,t+\tau}^2) \to \int_1^{1+\lambda_P} E\xi_W(s) =$
$\int_1^{1+\lambda_P}(1-(1-\alpha(s))^2)s^{-1}tr((-JB_1J' + B_2)V)ds -$
$\int_1^{1+\lambda_P}\alpha^2(s)\delta' B_2^{-1}(-JB_1J' + B_2)B_2^{-1}\delta ds.$

This decomposition implies that the bias-variance tradeoff depends on: (1) the duration of forecasting (λ_P), (2) the dimension of the parameter vectors (through the dimension of δ), (3) the magnitude of the predictive ability (as measured by quadratics of δ), (4) the forecast horizon (via V, the long-run variance of $h_{T,2,t+\tau}$), and (5) the second moments of the predictors ($B_i = \lim_{T\rightarrow\infty}(Ex_{T,i,t}x'_{T,i,t})^{-1}$).

The first term on the right-hand side of the decomposition can be interpreted as the pure "variance" contribution to the mean difference in the unrestricted and combined MSEs. The second term can be interpreted as the pure "bias" contribution. Clearly, when $\delta = 0$ and thus there is no predictive ability associated with the predictors $x_{T,22,t}$, the expected difference in MSE is positive so long as $\alpha(s) \neq 0$. Since the goal is to choose $\alpha(s)$ so that $\int_1^{1+\lambda_P} E\xi_W(s)$ is maximized, we immediately reach the intuitive conclusion that we should always forecast using the restricted model and hence set $\alpha(s) = 1$. When $\delta \neq 0$, and hence there is predictive ability associated with the predictors $x_{T,22,t}$, forecast accuracy is maximized by combining the restricted and unrestricted model forecasts. The following corollary provides the optimal combination weight.[4]

Corollary 2: The pointwise optimal combining weights satisfy

$$\alpha^*(s) = \left[1 + s\left(\frac{\beta'_{22}(Ex_{22,t}x'_{22,t} - Ex_{22,t}x'_{1,t}(Ex_{1,t}x'_{1,t})^{-1}Ex_{1,t}x'_{22,t})\beta_{22}}{tr((-JB_1J' + B_2)V)}\right)\right]^{-1}. \qquad (5)$$

The optimal combination weight is derived by maximizing the arguments of the integrals in Corollary 1 that contribute to the average expected mean square differential over the duration of forecasting — hence our "pointwise optimal" characterization of the weight. In particular, the results of Corollary 2 follow from maximizing

$$(1 - (1 - \alpha(s))^2)s^{-1}tr((-JB_1J' + B_2)V) - \alpha^2(s)\delta' B_2^{-1}(-JB_1J' + B_2)B_2^{-1}\delta \qquad (6)$$

with respect to $\alpha(s)$ for each s.

As is apparent from the formula in Corollary 2, the combining weight is decreasing in the marginal 'signal to noise' ratio

$$\beta'_{22}(Ex_{22,t}x'_{22,t} - Ex_{22,t}x'_{1,t}(Ex_{1,t}x'_{1,t})^{-1}Ex_{1,t}x'_{22,t})\beta_{22}/tr((-JB_1J' + B_2)V).$$

As the marginal 'signal', $\beta'_{22}(Ex_{22,t}x'_{22,t} - Ex_{22,t}x'_{1,t}(Ex_{1,t}x'_{1,t})^{-1}Ex_{1,t}x'_{22,t})\beta_{22}$, increases, we place more weight on the unrestricted model and less on the restricted one. Conversely, as the marginal 'noise', $tr((-JB_1J' + B_2)V)$, increases, we place more weight on the restricted

[4] Note that we have dropped the subscript T from the predictors. In our previous notation, this quantity would be $\lim_{T\rightarrow\infty}(Ex_{T,22,t}x'_{T,22,t} - Ex_{T,22,t}x'_{T,1,t}(Ex_{T,1,t}x'_{T,1,t})^{-1}Ex_{T,1,t}x'_{T,22,t})$. For brevity, we omit this subscript throughout the remainder.

model and less on the unrestricted model. Finally, as the sample size, s, increases, we place increasing weight on the unrestricted model.

In the special case in which the signal–to–noise ratio equals 1, the optimal combination weight is 1/2. In this case, the restricted and unrestricted models are expected to be equally accurate. For example, at time $s = 1$, when

$$\beta_{22}'(Ex_{22,t}x_{22,t}' - Ex_{22,t}x_{1,t}'(Ex_{1,t}x_{1,t}')^{-1}Ex_{1,t}x_{22,t}')\beta_{22} = tr((-JB_1J' + B_2)V), \quad (7)$$

the expected loss differential $E\xi_W(1) = 0$ is 0.

A bit more algebra establishes the determinants of the size of the benefits to combination. If we substitute $\alpha^*(s)$ into (6), we find that $E\xi_W^*(s)$ takes the easily interpretable form

$$\frac{tr((-JB_1J' + B_2)V)^2}{s(s\beta_{22}'(Ex_{22,t}x_{22,t}' - Ex_{22,t}x_{1,t}'(Ex_{1,t}x_{1,t}')^{-1}Ex_{1,t}x_{22,t}')\beta_{22} + tr((-JB_1J' + B_2)V))}. \quad (8)$$

This simplifies even more in the conditionally homoskedastic case, in which $tr((-JB_1J' + B_2)V) = \sigma^2 k_2$. In either case, it is clear that we expect the optimal combination to provide the most benefit when the marginal 'noise', $tr((-JB_1J' + B_2)V)$, is large or when the marginal 'signal', $\beta_{22}'(Ex_{22,t}x_{22,t}' - Ex_{22,t}x_{1,t}'(Ex_{1,t}x_{1,t}')^{-1}Ex_{1,t}x_{22,t}')\beta_{22}$, is small. And again, we obtain the result that, as the sample size increases, any benefits from combination vanish as the parameter estimates become increasingly accurate.

Note, however, that the term $\beta_{22}'(Ex_{22,t}x_{22,t}' - Ex_{22,t}x_{1,t}'(Ex_{1,t}x_{1,t}')^{-1}Ex_{1,t}x_{22,t}')\beta_{22}$ is a function of the local parameters β_{22} and not the global ones we estimate in practice. Moreover, note that these optimal combining weights are not presented relative to an environment in which agents are forecasting in 'real time'. Therefore, for practical use, we suggest a transformed formula. Let $\hat{B}_i$ and $\hat{V}$ denote estimates of B_i and V, respectively, based on data through period t. If we let the estimated global parameter $\hat{\beta}_{22}$ denote an estimate of the local parameter $T^{-1/2}\beta_{22}^*$ and set $s = t/T$, we obtain the following real time estimate of the pointwise optimal combining weight:[5]

$$\hat{\alpha}_t^* = \left[1 + t\left(\frac{\hat{\beta}'_{22}(t^{-1}\sum_{j=1}^t x_{22,j}x_{22,j}' - (t^{-1}\sum_{j=1}^t x_{22,j}x_{1,j}')\hat{B}_1(t^{-1}\sum_{j=1}^t x_{1,j}x_{22,j}'))\hat{\beta}_{22}}{tr((-J\hat{B}_1J' + \hat{B}_2)\hat{V})}\right)\right]^{-1}. \quad (9)$$

[5] We estimate B_i with $\hat{B}_i = (t^{-1}\sum_{j=1}^t x_{i,j}x_{i,j}^{'})^{-1}$, where $x_{i,t}$ is the vector of regressors in the forecasting model (supposing the MSE stationarity assumed in the theoretical analysis). In the Monte Carlo experiments, we impose conditional homoskedasticity in computing the noise term as $tr((-J\hat{B}_1J' + \hat{B}_2)\hat{V}) = k_2\hat{\sigma}^2$, where k_2 is the number of additional regressors in the unrestricted model and $\hat{\sigma}^2$ is the estimated residual variance of the unrestricted forecasting model estimated with data from 1 to t. In the empirical applications, we allow for conditional heteroskedasticity and compute the noise term using $\hat{V} = t^{-1}\sum_{j=1}^t \hat{u}_{2,j}^2 x_{2,j}x_{2,j}^{'}$.

In doing so, though, we acknowledge that the estimates of the global parameters are not consistent estimates of the local parameters on which our theoretical derivations (Corollary 2 and (9)) are based. The local asymptotics allow us to derive closed–form solutions for the optimal combination weights, but local parameters cannot be estimated consistently. We therefore simply use global magnitudes to estimate (inconsistently) the assumed local magnitudes and optimal combining weights. Below we use Monte Carlo experiments and empirical examples to determine whether the estimated quantities perform well enough to be a valuable tool for forecasting.

Conceptually, our proposed combination (9) might be expected to have some relationship to Bayesian methods. In the very simple case of the example of section 2.1, the proposed combination forecast corresponds to a forecast from an unrestricted model with Bayesian posterior mean coefficients estimated with a prior mean of 0 and variance proportional to the signal–noise ratio.[6] More generally, our proposed combination could correspond to the Bayesian model averaging considered in such studies as Wright (2003) and Stock and Watson (2005). Indeed, in the scalar environment of Stock and Watson (2005), setting their weighting function to t-stat^2/(1 + t-stat^2) yields our combination forecast. In the more general case, we have been unable to derive a simple shrinkage prior that would yield a Bayesian model averaging forecast equal to our combination forecast. However, there is likely to be some prior (that is, some specification of the shrinkage parameter ϕ of Wright (2003)) that makes a Bayesian average of the restricted and unrestricted forecasts very similar or identical to the combination forecast based on (9). Note, however, that the underlying rationale for Bayesian averaging is quite different from the combination rationale developed in this paper. Bayesian averaging is generally founded on model uncertainty. In contrast, our combination rationale is based on the bias–variance tradeoff associated with parameter estimation error, in an environment without model uncertainty.

Instead of using our approximation (9) to the optimal combination, one might instead consider using a Bates and Granger (1969) combination approach, based on regression estimates. That is, consider that at time T we estimate the optimal combining weight using a sequence of N existing pseudo-out-of-sample forecast errors $\hat{u}_{i,t+\tau} = (y_{T,t+\tau} - x'_{T,i,t}\hat{\beta}_{i,t})$, $t = R....R+N = T-\tau$, and the OLS estimated regression $\hat{u}_{2,t+\tau} = \alpha(\hat{u}_{2,t+\tau} - \hat{u}_{1,t+\tau}) + \eta_{t+\tau}$.[7]

[6] Specifically, using a prior variance of the signal–noise ratio times the OLS variance yields a posterior mean forecast equivalent to the combination forecast.

[7] This combination regression is obtained from the general regression $y_{T+\tau} = \alpha_{BG}\hat{y}_{1,T+\tau} + (1 - \alpha_{BG})\hat{y}_{2,T+\tau} + \eta_{t+\tau}$ by: (1) subtracting $\hat{y}_{2,T+\tau}$ from both sides and combining the remaining terms on

Under Assumptions 1-4, we can show that the resulting estimator $\hat{\alpha}_{BG}$ is inappropriate when the forecasts are from nested rather than non-nested models. In particular, if we define $\lim_{T\to\infty} N/R = \pi \in (0,\infty)$, let W_0 and W_1 denote independent $(k \times 1)$ standard normal vectors, and (for analytical tractability) restrict attention to fixed scheme pseudo-out-of-sample forecasts (so that $\hat{\beta}_{i,t} = \hat{\beta}_{i,R}$ $t = R....R+N = T-\tau$), we obtain the following result on the limiting behavior of the estimated combining coefficient from a Bates–Granger regression.

Proposition 1: $\hat{\alpha}_{BG} \to_d 1-\pi^{-1}\left(\frac{(W_0+\frac{\pi}{1+\pi}V^{-1/2}B_2^{-1}\delta)'[V^{1/2}(-JB_1J'+B_2)V^{1/2}](W_1+\frac{1}{1+\pi}V^{-1/2}B_2^{-1}\delta)}{(W_1+\frac{1}{1+\pi}V^{-1/2}B_2^{-1}\delta)'[V^{1/2}(-JB_1J'+B_2)V^{1/2}](W_1+\frac{1}{1+\pi}V^{-1/2}B_2^{-1}\delta)}\right)$.

Proposition 1 establishes that a Bates–Granger regression yields a combination estimate that is not only inconsistent for our optimal combination weight but also converges in distribution rather than in probability. In unreported simulations of DGP 1 described in Section 3, we find that while the support of the asymptotic distribution of $\hat{\alpha}_{BG}$ contains the value of our optimal combining weight, it has a large variance, often yielding values of $\hat{\alpha}_{BG}$ that are much larger or much smaller than the optimal combining weight derived in Corollary 2. The apparent suboptimality of this approach reflects the fact that the original motivation for the regression was based upon combination for non-nested rather than nested models. As shown in Clark and McCracken (2001) and McCracken (2004), out-of-sample methods designed for the comparison of non-nested models need not be applicable for the comparison of nested models.

3 Monte Carlo Evidence

We use Monte Carlo simulations of bivariate data-generating processes to evaluate the finite–sample performance of the combination methods described above. In these experiments, the DGPs relate the predictand y to lagged y and lagged x, with the coefficients on lagged x set at various values. Forecasts of y are generated with the combination approaches considered above, along with some related methods that are used or might be used in practice, such as Bayesian estimation. Performance is evaluated using simple summary statistics of the distribution of each forecast's MSE: the average MSE across Monte Carlo draws (medians yield similar results), and the probability of equaling or beating the restricted model's forecast MSE.

the right–hand side; (2) substituting $\hat{u}_{2,t+\tau}$ for $y_{T+\tau} - \hat{y}_{2,T+\tau}$; and (3) substituting $\hat{u}_{2,t+\tau} - \hat{u}_{1,t+\tau}$ for $\hat{y}_{1,T+\tau} - \hat{y}_{2,T+\tau}$.

3.1 Experiment design

In light of the considerable practical interest in the out–of–sample predictability of inflation (see, for example, Stock and Watson (1999, 2003), Atkeson and Ohanian (2001), Fisher, et al. (2002), Orphanides and van Norden (2005), and Clark and McCracken (2005b)), we present results for DGPs broadly based on estimates of quarterly inflation models. In particular, we consider models based on the relationship of the change in core PCE inflation to lags of the change in inflation, the output gap, and, in some cases, the growth rate of unit labor costs and import price inflation.[8] With prior results in the inflation forecasting literature sufficiently mixed as to suggest the predictive content of the output gap and other variables may be weak, we consider various values of the coefficients (corresponding to our theoretical β_{22}) on these variables, ranging from zero to quite large values. We compare forecasts from an unrestricted model that corresponds to the DGP to forecasts from a restricted model that takes an AR form (that is, a model that drops from the unrestricted model all but the constant and lags of the dependent variable). Although not presented in the interest of brevity, we obtained qualitatively similar results with a DGP based on estimates of a model relating the (quarterly) excess return on the S&P 500 to the dividend–price ratio and a short–term (relative) interest rate (in those applications, the null forecasting model related y to just a constant).

In each experiment, we conduct 10,000 simulations of data sets of 160 observations (not counting the initial observations necessitated by the lag structure of the DGP). In our reported results, with quarterly data in mind, we use an "in–sample" size of $T = 80$, and evaluate forecast accuracy over forecast periods of various lengths: $P = 1$, 20, 40, and 80, corresponding to $\lambda_P = .0125$, .2, .5, and 1. We obtained very similar results with $T = 120$ and have omitted those results in the interest of brevity.

The first DGP, based on the empirical relationship between the change in core inflation (y_t) and the output gap ($x_{1,t}$), takes the form

$$\begin{aligned} y_t &= -.40y_{t-1} - .16y_{t-2} + b_{11}x_{1,t-1} + u_t \\ x_{1,t} &= 1.18x_{1,t-1} - .06x_{1,t-2} - .20x_{1,t-3} + v_{1,t} \\ &\text{var}\begin{pmatrix} u_t \\ v_{1,t} \end{pmatrix} = \begin{pmatrix} .73 & \\ .02 & .59 \end{pmatrix}. \end{aligned} \tag{10}$$

We consider various experiments with different settings of b_{11}, the coefficient on the "output

[8] See Appendix 2's description of applications 6 and 7 for data details.

gap." As becomes clear when we describe below the competing forecasting models, b_{11} corresponds to our theoretical construct $\beta_{22}/\sqrt{T}$. The baseline value of b_{11} is the one that, in population, makes the null and alternative models equally accurate (in expectation) in forecast period $T+1$ — the value that satisfies (7). Given the population moments implied by the DGP parameterization, this value is $b_{11} = .327/\sqrt{T} = .037$. The second setting we consider is the empirical value: $b_{11} = .10$. To illustrate how each method fares if the predictive content of $x_{1,t}$ is truly non–existent, we also report results from an experiment with $b_{11} = 0$.

The second DGP, based on estimated relationships among inflation (y_t), the output gap ($x_{1,t}$), growth in unit labor costs ($x_{2,t}$), and import price inflation ($x_{3,t}$), takes the form:

$$\begin{aligned}
y_t &= -.40y_{t-1} - .16y_{t-2} + b_{11}x_{1,t-1} + b_{21}x_{2,t-1} + b_{22}x_{2,t-2} + b_{31}x_{3,t-1} + b_{32}x_{3,t-2} + u_t \\
x_{1,t} &= 1.18x_{1,t-1} - .06x_{1,t-2} - .20x_{1,t-3} + v_{1,t} \\
x_{2,t} &= 1.54x_{1,t-1} - 1.13x_{1,t-2} + .31x_{2,t-1} + .37x_{2,t-2} + v_{2,t} \qquad (11) \\
x_{3,t} &= .39x_{2,t-1} - .06x_{2,t-2} + .55x_{3,t-1} + .05x_{3,t-2} + v_{3,t} \\
&\text{var}\begin{pmatrix} u_t \\ v_{1,t} \\ v_{2,t} \\ v_{3,t} \end{pmatrix} = \begin{pmatrix} .73 & & & \\ .02 & .59 & & \\ .36 & -1.72 & 11.90 & \\ 1.37 & .43 & 1.10 & 27.14 \end{pmatrix}.
\end{aligned}$$

As with DGP 1, we consider experiments with three different settings of the set of b_{ij} coefficients, which correspond to the elements of $\beta_{22}/\sqrt{T}$. One setting is based on empirical estimates: $b_{11} = .10$, $b_{21} = .03$, $b_{22} = -.02$, $b_{31} = .05$, $b_{32} = -.03$. We take as the baseline experiment one in which all of these empirical values of the b_{ij} coefficients are multiplied by a constant less than one, such that, in population, the null and alternative models are expected to be equally accurate in forecast period $T+1$. With $T = 80$, this multiplying constant is .527. Finally, we also report results for a DGP with all of the b_{ij} coefficients set to zero.

3.2 Forecast approaches

In the case of DGP 1, forecasts of y_{t+1}, $t = T, \ldots, T+P$, are formed from various combinations of estimates of the following forecasting models:

$$y_t = \delta_0 + \delta_1 y_{t-1} + \delta_2 y_{t-2} + u_{1,t} \qquad (12)$$

$$y_t = \gamma_0 + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \gamma_3 x_{1,t-1} + u_{2,t}. \qquad (13)$$

In the case of DGP 2, the unrestricted forecasting model is augmented to include $x_{2,t-1}$, $x_{2,t-2}$, $x_{3,t-1}$, and $x_{3,t-2}$:

$$y_t = \gamma_0 + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \gamma_3 x_{1,t-1} + \gamma_4 x_{2,t-1} + \gamma_5 x_{2,t-2} + \gamma_6 x_{3,t-1} + \gamma_7 x_{3,t-2} + u_{2,t}. \quad (14)$$

Note that, with these specifications, $k_2 = 1$ for DGP 1 and $k_2 = 5$ for DGP 2.

The forecasts or methods we consider, detailed in Table 1, include those described above, as well as some natural alternatives. In particular, we examine the accuracy of forecasts from: (1) OLS estimates of the restricted model (12); (2) OLS estimates of the unrestricted model ((13) in DGP 1 simulations and (14) in DGP 2 simulations); (3) the "known" optimal linear combination of the restricted and unrestricted forecasts, using the weight implied by equation (8) and population moments implied by the DGP; (4) the estimated optimal linear combination of the restricted and unrestricted forecasts, using the weight given in (9) and estimated moments of the data; and (5) a simple average of the restricted and unrestricted forecasts (as noted above, weights of 1/2 are optimal if the signal associated with the x variables equals the noise, making the models equally accurate at $T + 1$).

We also consider forecasts based on common model–selection procedures applied as forecasting moves forward in time. One such approach, suggested in Bossaerts and Hillion (1999) and Inoue and Kilian (2004b), is to use the model with a lower SIC score as of time t to forecast y_{t+1}. That is, at each forecast origin t, estimate both the restricted and unrestricted models, and then use the model with the lower SIC score to construct the $t+1$ forecast. We consider this real–time SIC approach, as well as a corresponding real–time AIC method. Many studies, such as Marcellino, Stock, and Watson (2004) and Orphanides and van Norden (2005), have similarly used the AIC or SIC to determine the lag orders of forecasting models.

Finally, we also consider select Bayesian forecasting methods that may be seen as natural alternatives to the combination methods proposed in this paper. Doan, Litterman, and Sims (1984) suggest that conventional Bayesian estimation (specifically, the prior) provides a flexible method for balancing the tradeoff between signal and parameter estimation noise. Accordingly, we construct one forecast based on Bayesian estimation of the unrestricted forecasting model ((13) in DGP 1 simulations and (14) in DGP 2 simulations), using Minnesota–style priors as described in Litterman (1986). For our particular applications, we use a prior mean of zero for all coefficients, with prior variances that are tighter for longer lags than shorter lags and tighter for lags of $x_{i,t}$ than y_t. In the notation of

Litterman, we use the following parameter settings in determining the prior variances: $\lambda = .2$ and $\theta = .5$.[9]

We construct another forecast by applying Bayesian model averaging (BMA) to the restricted and unrestricted models, using the BMA approach of Wright (2003). In particular, we use Bayesian methods simply to weight OLS estimates of the two models. The prior probability on each model, Prob(M_i), i = 1,2, is just 1/2. In calculating the posterior probabilities of each model, Prob(M_i|data), we set the prior on the coefficients to zero. At each forecast origin t, we then calculate the posterior probability of each model using

$$\begin{aligned}
\text{Prob}(M_i|\text{data}) &= \frac{\text{Prob}(\text{data}|M_i) \times \text{Prob}(M_i)}{\sum_i \text{Prob}(\text{data}|M_i) \times \text{Prob}(M_i)} \qquad (15)\\
\text{Prob}(\text{data}|M_i) &\propto (1+\phi)^{-p_i/2} S_i^{-(t+1)} \\
\phi &= \text{parameter determining rate of shrinkage toward the restricted model} \\
p_i &= \text{the number of explanatory variables in model } i \\
S_i^2 &= (Y - X_i\hat{\Gamma}_i)'(Y - X_i\hat{\Gamma}_i) + \frac{1}{1+\phi}\hat{\Gamma}_i' X_i' X_i \hat{\Gamma}_i \\
X_i &= \text{matrix of regressors in model } i \\
\hat{\Gamma}_i &= \text{vector of OLS estimates of the coefficients of model } i.
\end{aligned}$$

We report results for two different settings of the shrinkage parameter ϕ, one relatively high ($\phi = 2$) and one low ($\phi = .2$). Lower values of ϕ are associated with greater shrinkage toward the restricted model.

3.3 Simulation results

In our Monte Carlo comparison of methods, we primarily base our evaluation on average MSEs over a range of forecast samples. For simplicity, in presenting average MSEs, we only report actual average MSEs for the restricted model (12). For all other forecasts, we report the ratio of a forecast's average MSE to the restricted model's average MSE. To capture potential differences in MSE distributions, we also present some evidence on the probabilities of equaling or beating the restricted model.

3.3.1 Simple combination forecasts

We begin with the case in which the coefficients b_{ij} (elements of β_{22}) on the lags of x_{it} (elements of x_{22}) in the DGPs (10) and (11) are set such that the restricted and unrestricted

[9] For the intercept of each model, we follow the example of Robertson and Tallman (1999) and use a prior mean of 0 and standard deviation of .3 times the standard error of an estimated AR model for y.

model forecasts for period $T+1$ are expected to be equally accurate — because the signal and noise associated with the x_{it} variables are equalized. In this setting, the optimally combined forecast should, on average, be more accurate than either the restricted or unrestricted forecasts.

The average MSE results for DGPs 1 and 2 reported in the left panels of Table 2 confirm the theoretical implications. With DGP 1, the ratio of the unrestricted model's average MSE to the restricted model's average MSE is very close to 1.000 for all forecast samples. The same is true with DGP 2, except that, with a forecast sample of just $P = 1$, the unrestricted model's average squared forecast error is slightly larger than the restricted model's (MSE ratio of 1.013).

A combination of the restricted and unrestricted forecasts has a lower average MSE, although only trivially so in the DGP 1 experiment, in which the restricted model omits only one variable (in the DGP 2 experiment, though, the restricted model omits five variables). Using the known optimal combination weight α_t^* yields an MSE ratio of about .995 in the case of DGP 1 and .975 in the case of DGP 2. These gains are in line with those indicated by the theoretical results in section 2. For these particular experiments (in which the forecast errors are conditionally homoskedastic and the restricted and unrestricted models are expected to be equally accurate as of T), the expected gain (8) as a percentage of the residual variance (σ^2) simplifies to $\frac{k_2}{2s}$. The resulting theoretic gains are 0.5 percent for DGP 1 and 2.5 percent for DGP 2, in line with the gains in the experiments.

Not surprisingly, having to estimate the optimal combination weight tends to slightly reduce the gains to combination. For example, in the case of DGP 2 and $P = 40$, the MSE ratio for the estimated optimal combination forecast is .980, compared to .973 for the known optimal combination forecast. The simple average of the restricted and unrestricted forecasts performs about as well as the known optimal combination, because, with signal = noise at least as of period T, the optimal combination weight is 1/2. As forecasting moves forward in time, though, the known optimal combination weight declines, because as more and more data become available for estimation, the signal-to-noise ratio rises (e.g., in the case of DGP 2, the known optimal weight for the forecast of the 80th observation in the prediction sample is about .33). But the declines aren't great enough to cause the performance of the simple average to deteriorate materially relative to the known optimal combination, for the forecast samples considered.

Combination continues to perform well in DGPs with larger b_{ij} (β_{22}) coefficients — that is, coefficient values set to those obtained from empirical estimates of inflation models. With these larger coefficients, the signal associated with the x_{it} (x_{22}) variables exceeds the noise, such that the unrestricted model is expected to be more accurate than the restricted model. In this setting, too, our asymptotic results imply the optimal combination forecast should be more accurate than the unrestricted model forecast, on average. However, the gains to combination should be smaller than in DGPs with smaller b_{ij} coefficients. The results for DGPs 1 and 2 reported in the right panels of Table 2 broadly confirm these theoretical implications, although, in some cases, the estimated optimal combination's average accuracy is no greater than the unrestricted model's average accuracy. Compared to the restricted model's MSE, the unrestricted model's average MSE is about 7 percent lower in the case of DGP 1 (MSE ratio of about .93) and 11-12 percent lower in the case of DGP 2 (MSE ratio of .88-.89).

Combination using the known optimal combination weight α_t^* improves accuracy further, more noticeably in the DGP 2 experiments, which involve a more richly parameterized unrestricted forecasting model. For example, with DGP 2 and $P = 40$, the MSE ratio is .874 for the known α_t^*–combined forecast, compared to the unrestricted forecast's MSE ratio of .884. In these experiments, the combination forecast based on the estimated α_t^* performs about as well as that based on the known α_t^*: in the same example, the MSE ratio for the opt. combination: $\hat{\alpha}_t^*$ forecast is .878. Finally, combination in the form of a simple average of the restricted and unrestricted forecasts yields a forecast that is about as accurate, although not quite, as the unrestricted model's forecast or the optimally combined forecast. For example, with DGP 2 and $P = 40$, the MSE ratio of the simple average forecast is .889, compared to .878 for the estimated optimal combination and .884 for the unrestricted model.

Unreported results for DGP 1 with $b_{11} = .20$ — with an "output gap" coefficient twice its estimated magnitude — confirm that the same basic patterns hold as the predictive content of the variables of interest becomes quite high. But, not surprisingly, as signal becomes high relative to noise, the performance of a simple average forecast deteriorates (the average forecast has an MSE ratio of about .84, while the unrestricted and optimal combination forecasts have MSE ratios of about .8). Of course, when the signal–to–noise ratio is high, the optimal combination weight is close to 1, so a simple average does not perform as well.

With predictive content often found to be weak in many practical settings, the coefficients of interest could actually be zero (zero signal), rather than just close to zero (small, non-zero signal). Accordingly, in Table 3, we report results for DGPs in which all b_{ij} coefficients (β_{22}) equal 0. In this setting, of course, the restricted model will be more accurate than the unrestricted model, in terms of average MSE, with the accuracy difference increasing in the number of variables in x_{22}. Indeed, as shown in the table, the average MSE of the unrestricted model is 1-2 percent higher than that of the restricted model in the case of DGP 1 and 5-8 percent higher in the case of DGP 2. The estimated optimal combination forecast is considerably better than the unrestricted forecast, although not quite as good as the restricted forecast. For example, with $P = 40$, the MSE ratio of the estimated optimal combination forecast is 1.006 for DGP 1 and 1.019 for DGP 2. The simple average forecast is slightly better than the optimal combination, with MSE ratios of 1.003 (DGP 1) and 1.015 (DGP 2) for $P = 40$. Thus, even if the variables of interest have no true predictive content, combination can greatly limit the losses relative to the optimal restricted model forecast.

In addition to helping to lower the average forecast MSE, combination of restricted and unrestricted forecasts helps to tighten the distribution of relative accuracy — for example, the MSE relative to the MSE of the restricted model. In particular, the results in Tables 4 and 5 indicate that combination — especially simple averaging — often increases the probability of beating the MSE of the restricted model, often by more than it lowers average MSE. As shown in Table 4, for instance, with DGP 1 parameterized such that signal = noise as of time T (with $b_{11} = .037$), the frequency with which the unrestricted model's MSE is less than the restricted model's MSE is 42.2 percent for $P = 40$. The frequency with which the known optimal combination forecast's MSE is below the restricted model's MSE is 49.3 percent. Although the estimated combination does not fare as well (probability of 40.1 percent in the sample example), a simple average fares even better, beating the MSE of the restricted model in 50.2 percent of the simulations. By this probability metric, the simple average also fares well in the experiment with DGP 1 and $b_{11} = .10$ (signal > noise). Again using the $P = 40$ example, the probabilities of beating the restricted model's MSE are 77.2, 78.7, and 87.3 percent, respectively, for the unrestricted, estimated optimal combination, and simple combination forecasts. Results in Tables 4 and 5 for other experiments (DGP

1 with $b_{11} = 0$, DGP 2 with all $b_{ij} = 0$ and coefficients scaled to make signal=noise, and DGP 2 with empirical coefficients) confirm the same basic patterns: (i) compared to the unrestricted forecast, simple averaging improves the chances of beating the accuracy of the restricted model's forecast; (ii) although the known optimal combination can also offer a material gain (not always as large as simple combination), estimating the combination weight reduces the gain, sometimes materially.

3.3.2 Comparison to other methods

As noted above, our proposed combination procedure has a number of natural alternatives, related to procedures used in practice: forecasting y_{t+1} with the period t–estimated model (restricted or unrestricted) that the SIC or AIC indicates to be superior; Bayesian shrinkage of estimates of the unrestricted model, using BVAR techniques; or Bayesian model averaging of the restricted and unrestricted models.

Of these alternative methods, the Bayesian approaches seem to work best in our experiments — and about as well as our simple combination approaches. BVAR estimation delivers an average MSE ratio that is quite similar to those obtained with our feasible combination approaches. In the case of DGP 1 with $b_{11} = .10$ (so signal > noise) and P=40, the MSE ratio of the BVAR forecast is .932, compared to the estimated optimal combination and simple average ratios of .936 and .945 (Table 2). In the case of DGP 2 with estimated b_{ij} coefficients (signal > noise) and P=40, the BVAR's MSE ratio is .889, while the estimated combination forecast's ratio is .878 and the simple average's is .889 (Table 2). With DGP 2's b_{ij} coefficients set to 0, the BVAR forecast's average MSE is 1.016, about the same as those for the estimated optimal and simple average forecasts (Table 3). In terms of probability of beating the restricted model in MSE, the BVAR generally falls somewhere between the estimated optimal combination and the simple average. But when the b_{ij} coefficients are truly zero, the BVAR typically has the highest probability of beating the restricted model (but still less than 50 percent).

The BMA approaches also perform comparably to our proposed simple combination approaches, although more so in terms of average MSE than probability of beating the restricted model's MSE. For example, using DGP 2 with b_{ij} coefficients set to make signal equal to noise, and P=40, the BMA: $\phi = .2$ ($\phi = 2$) forecast's MSE ratio is .977 (.988), compared to the ratios of .980 for the estimated combination and .973 for the simple average (Table 2). With P=40, the probability of beating the restricted model's MSE is 62.2 percent for the ϕ=.2 BMA forecast and 53.8 percent for the ϕ=2 forecast, compared to the BVAR and simple average probabilities of 62.4 and 67.8 percent (Table 5). Clearly, using more shrinkage in the Bayesian model averaging (lower ϕ) tightens the relative MSE distribution. In the case of DGP 2 with estimated b_{ij} coefficients (signal > noise) and P=40, the BMA: $\phi = .2$ ($\phi = 2$) forecast's MSE ratio is .879 (.886), compared to the ratios of .889, .878, and .889 for the BVAR, estimated combination, and simple average forecasts, respectively (Table 2). The likelihoods of beating the accuracy of the restricted model follow the same ordering given in the prior example: simple average (94.0), BVAR (90.3), BMA: $\phi = .2$ (88.1), and BMA: $\phi = 2$ (81.5) (Table 5).

Although the SIC and AIC model selection methods work well in some instances, overall, these methods that base the forecast at $t + 1$ on a single model selected at each t don't perform as well as the simple combination and Bayesian methods. In some settings, to be sure, these selection methods can perform as well as the combination methods, but the selection methods are never better, and they can be worse.[10] Consider, for example, the DGP 2 simulations, with P=40. In the (Table 2) experiment with the b_{ij} coefficients set to make signal equal to noise, the AIC approach yields an average MSE ratio of 1.006. The SIC approach, which selects the unrestricted model with a lower frequency, yields a slightly lower MSE ratio, of 1.002. But both methods fall short of the simple average forecast, which has an average MSE ratio of .973. In the (Table 2) experiment with estimated b_{ij} coefficients (signal > noise), the AIC often results in the selection of the unrestricted model, so it yields an average MSE ratio (.890) that is essentially the same as that of the unrestricted model (.884) and simple average forecast (.889). Because the more parsimonious SIC less frequently selects the unrestricted model, the SIC yields a higher average MSE ratio, of .947.

Overall, the Monte Carlo evidence shows simple forecast combination and Bayesian shrinkage to be useful tools for improving forecast accuracy. Simple combination — either in the form of an optimal combination estimated with the approach developed in section 2 or an average — improves average forecast accuracy. Combination, especially simple averaging, can also significantly increase the odds of improving on the accuracy of the benchmark restricted model. Bayesian shrinkage, especially of the type associated with with Minnesota–style BVAR model estimation, offers comparable benefits.

[10] In line with our findings, Cecchetti (1995) reports that, across a range of bivariate inflation models, in–sample SIC values have little correlation with forecast RMSEs.

4 Empirical Applications

To evaluate the empirical performance of the various forecast methods, we follow the spirit of Stock and Watson (1996, 2003, 2005) in considering a wide range of applications and forecast performance over two periods, 1978-91 and 1992-2004. For a number of the applications, other studies have found some evidence of weak predictability of the information included in an unrestricted model of interest but not a benchmark restricted model. In line with common empirical practice (see, for example, the aforementioned work of Stock and Watson), our presented results are simple MSEs for one-step ahead forecasts, presented in the form of ratios relative to the restricted model's MSE. We consider the same forecast methods included in the Monte Carlo analysis, except that, by necessity, we drop the combination forecast based on the known optimal weight.

4.1 Applications

The predictands in the 12 applications listed below are widely–studied, broad economic indicators for the United States. We describe below the specification of each unrestricted forecasting model (although not mentioned explicitly below, all models include a constant). Unless otherwise noted, the restricted model takes the form of an AR model, using the lag order of the unrestricted model. As the list below indicates, the applications include both monthly and quarterly examples. Appendix 2 provides additional detail on the data and estimation samples.

1. Forecasting monthly growth in industrial production with six lags of production and six lags of the Federal Reserve Bank of Chicago's factor index of the national business cycle (examples: Stock and Watson (2002, 2005) and Shintani (2005)).[11]

2. Predicting monthly growth in real disposable personal income using models of the form described in (1) (same examples).

3. Predicting monthly growth in real manufacturing and trade sales using models of the form described in (1) (same examples).

4. Forecasting monthly growth in employment using models of the form described in (1) (same examples).

5. Forecasting the change in the monthly unemployment rate with six lags, six lags of new claims for unemployment insurance, and six lags of continuing claims for insurance

[11] For this and the other three factor index applications, including one lag of three factors rather than six lags of just the first factor yields qualitatively similar results.

(examples: Montgomery, et al. (1998) and Gavin and Kliesen (2002)).

6. Predicting the change in quarterly core PCE inflation with two lags and one lag of the output gap, defined as the log of the ratio of actual GDP to the CBO's estimate of potential GDP (examples: Stock and Watson (1999, 2003), Brave and Fisher (2004), and Clark and McCracken (2005b)).

7. Predicting the change in quarterly core PCE inflation with two lags, one lag of the output gap (defined as above), two lags of growth in unit labor costs, and two lags of import price inflation (examples: Mehra (1990) and Brayton, et al. (1999)).

8. Forecasting quarterly GDP growth with a VAR(2) in GDP growth, the log consumption/GDP ratio, growth in aggregate hours worked, the change in core PCE inflation, and the change in the three-month Treasury bill rate (examples: Litterman (1986), Rotemberg (1996), Rotemberg and Woodford (1996), and Bernanke and Boivin (2003)).

9. Predicting the change in the three-month Treasury bill rate with a VAR(2) in the same variables.

10. Forecasting the quarterly excess return on the S&P 500 with one lag of the dividend–price ratio and one lag of a quarterly risk–free interest rate relative to its average over the prior year (examples: Pesaran and Timmermann (2002) and Campbell and Thompson (2005)). In this case, the restricted model includes just a constant.

11. Forecasting the monthly change in the U.S. dollar-Swiss franc exchange rate with interest rate differentials at 1, 3, 6, and 12 months (examples: Clarida and Taylor (1997) and Clarida, et al. (2003)). The restricted model includes just a constant.

12. Forecasting the monthly change in the U.S. dollar-U.K. sterling exchange rate with interest rate differentials at 1, 3, 6, and 12 months (same examples).

4.2 Results

In one broad respect, the application results in Table 6 seem reasonably reflective of patterns common in published forecasting research: across the applications, the variables included in the unrestricted model but not the restricted only sometimes improve forecast accuracy. Across the 24 cases represented in Table 6 (12 applications × two sample periods), the unrestricted model's MSE is lower in 13 of them (in the one case in which the reported MSE ratio is 1.000, the unrestricted model's MSE is actually a bit lower than the restricted's). In some cases, such as the 1978-91 employment growth application, the unrestricted model is much better than the restricted, with an MSE ratio of .827. In other cases, such as the

1992-04 stock return application, the unrestricted model is much worse, with an MSE ratio of 1.175.

Consistent with the Monte Carlo evidence, the application results suggest that simple forecast combination and Bayesian shrinkage generally improve forecast accuracy, while the forecast model selection approaches (real-time SIC and AIC) have more mixed consequences. In all 24 cases, the estimated optimal combination forecast is effectively as good as or better than the unrestricted model forecast, in terms of MSE. For the most part, the instances in which the estimated optimal combination is materially better than the unrestricted forecast are those in which the unrestricted model is inferior to the restricted — examples include disposable income over 1978-91 (unrestricted and optimal combination MSE ratios of 1.140 and 1.087) and GDP growth over 1992-2004 (MSE ratios of 1.066 and 1.034). But there are instances in which the combination improves upon an unrestricted forecast that is better than the restricted model's forecast, such as with GDP growth over 1978-91 (unrestricted and optimal combination MSE ratios of .849 and .809).

By various criteria, a simple average seems to be even better than the estimated optimal combination. Of the 24 cases covered in Table 6, the simple average forecast has the lowest MSE in seven of them (as we detail below, by this simple count, only the BVAR forecast fares as well) — note that, for each case, the best MSE performance is distinguished by a bold font. In general, when the unrestricted or optimal combination forecast is more accurate than the restricted model's forecast, the simple average generally does just about as well. For example, for 1978-91 forecasts of employment growth, the MSE ratio of the simple combination is .847, compared to the MSE ratios of .832 for the unrestricted forecast and .826 for the estimated optimal combination. For 1978-91 forecasts of inflation based on the output gap, unit labor costs, and import price inflation, the simple average approach yields an MSE ratio of .953, compared to the unrestricted model's MSE ratio of .973 and the optimal combination's ratio of .975. The greater advantage of the simple average comes in those cases in which the restricted forecast beats the unrestricted model's projection. To again use the disposable income example, over 1978-91 the simple average forecast's MSE ratio is 1.021, well below that of the optimal combination's 1.066 and unrestricted model's 1.140. The same applies in the case of GDP growth over 1992-2004 (simple average MSE ratio of .972, compared to 1.032 (optimal combination) and 1.082 (unrestricted)).

Some Bayesian methods perform about as well as simple forecast combination. In our particular results, BVAR estimation fares better than Bayesian model averaging, although BMA with the higher degree of shrinkage ($\phi = .2$) may be seen as comparable. Of course, it is possible that using other shrinkage settings could improve the BMA outcomes (although the same applies to our BVAR implementation, in which we have considered a single, pre–determined setting of the priors). Of Table 6's 24 cases, the BVAR forecast (like the simple average forecast) yields the lowest MSE in seven of them. In several of these cases, the BVAR forecast materially improves on the unrestricted and simple average forecasts, as well as the restricted forecast (disposable income, 1978-91; unemployment, 1978-91; and T-bill rate, 1978-91 and 1992-04). Like the simple average, in those cases in which the restricted forecast is more accurate than the unrestricted, the BVAR forecast's MSE is reasonably comparable to the restricted forecast's. In the case of disposable income forecasts for 1978-91, the BVAR forecast's MSE ratio is .946, compared to the simple average's ratio of 1.021 and the unrestricted's 1.140. For 1978-91 forecasts of the dollar–Swiss franc exchange rate, the BVAR yields an MSE ratio of 1.036, compared to those of 1.034 for the simple average and 1.105 for the unrestricted forecast.

Of the two BMA forecasts, the projection using higher shrinkage ($\phi = .2$) is always (in effect) at least as good as the projection using lower shrinkage ($\phi = 2$). This better BMA forecast is sometimes about as good as or better than the BVAR (examples: industrial production, 1992-04; and employment, 1978-91), but sometimes not as good, and sometimes quite inferior (disposable income, 1978-91; and unemployment, 1992-04). Similarly, the better BMA forecast sometimes matches or beats the simple average (for instance, manufacturing and trade sales, 1992-04) and sometimes falls short (stock returns, 1992-04).

As to the model selection methods (SIC, etc.), our reading of the application evidence is that, on balance, these methods that base the forecast at $t+1$ on a single model selected at each t don't perform as well as the simple combination and Bayesian methods. Take, for example, the 1978-91 forecasts of growth in industrial production. The forecasts derived from AIC model selection are nearly as accurate as the unrestricted model's forecast: the AIC yields an MSE ratio of .918, compared to the unrestricted model's ratio of .889 and the simple average forecast's MSE ratio of .897. In these cases, the AIC usually, but not always, selects the unrestricted model. The forecasts derived from SIC selection are less accurate, with a MSE ratio of 1.024, because the more parsimonious SIC select the unrestricted model with a lower frequency. However, in the exchange rate applications — cases in which predictive content appears to be especially weak — the model selection methods, especially the SIC, fare better, because they virtually always select the restricted model.

To more formally summarize the performance of the various methods, we compare across the 24 cases in Table 6 each method's MSE against the forecast MSE that is best in each case. That is, for each case, we determine the lowest MSE, and then, for each method, form the ratio of its MSE to the lowest MSE. For the resulting 24 MSE ratios, we then tabulate the mean, median, standard deviation, and spread between the 90th and 10th percentiles. Our idea is that, for a forecast method to be recommended for practical use, it should be relatively close, on average, to the best forecast, and rarely deviate too much from the best forecast.

The comparison statistics in Table 7 confirm that, across our applications, the simple average and BVAR forecasts perform best. On average, the simple average forecast's MSE is just 2.5 percent larger than the best forecast's MSE; the BVAR's forecast is, on average, 3.0 percent larger than the best MSE. The dispersion of these two forecast approaches around the best forecast is also relatively low, with a 90-10%ile spread of 7.5 percent (simple average) and 7.3 percent (BVAR). All of the other forecasts have higher average MSE ratios and suffer more dispersion relative to the best forecast. For example, the estimated optimal combination forecast's MSE exceeds the best MSE by an average of 4.3 percent, with a dispersion of 12.0 percent. The restricted and unrestricted forecasts yield MSEs that exceed the best MSE by an average of 8.4 and 6.9 percent, respectively. Finally, the forecasts based on SIC and AIC model selection yield MSEs that exceed the best MSE by an average of 7.1 and 7.6 percent, respectively, with dispersion of roughly 17-19 percent.

5 Conclusion

As reflected in the principle of parsimony, when some variables are truly but weakly related to the variable being forecast, having the additional variables in the model may detract from forecast accuracy, because of parameter estimation error. Focusing on such cases of weak predictability, we show that combining the forecasts of the parsimonious and larger models can improve forecast accuracy. We first derive, theoretically, the optimal combination weight and combination benefit. In the special case in which the coefficients on the variables of interest are of a magnitude that makes the restricted and unrestricted models equally accurate, the MSE–minimizing forecast is a simple, equally–weighted average of the restricted and unrestricted forecasts. With a range of Monte Carlo and empirical examples, we show that our proposed approach of combining forecasts from nested models works well compared to various alternative methods of forecasting. Overall, in both the Monte Carlo and empirical results, two forecast methods seem to work best, in the sense of consistently yielding improvements in MSE: simple averaging of the restricted and unrestricted model forecasts, and Bayesian (Minnesota BVAR) estimation of the unrestricted model.

6 Appendix 1: Theory Details

Note that, in the notation below, $W(\cdot)$ denotes a standard $(k \times 1)$ Brownian motion.

Theorem 1: $\sum_{t=T}^{T+P}(\hat{u}^2_{2,t+\tau} - \hat{u}^2_{W,t+\tau}) \to_d \int_1^{1+\lambda_P} \xi_W(s) =$
$\{-2\int_1^{1+\lambda_P} \alpha(s)s^{-1}W'(s)V^{1/2}(-JB_1J' + B_2)V^{1/2}dW(s)$
$+ \int_1^{1+\lambda_P} (1-(1-\alpha(s))^2)s^{-2}W'(s)V^{1/2}(-JB_1J' + B_2)V^{1/2}W(s)ds\}$
$+ 2\{-\int_1^{1+\lambda_P} \alpha(s)\delta' B_2^{-1}(-JB_1J' + B_2)V^{1/2}dW(s)$
$+ \int_1^{1+\lambda_P} \alpha^2(s)s^{-1}\delta' B_2^{-1}(-JB_1J' + B_2)B_2^{-1}JB_1J'V^{1/2}W(s)ds$
$+ \int_1^{1+\lambda_P} \alpha(s)(1-\alpha(s))s^{-1}\delta' B_2^{-1}(-JB_1J' + B_2)V^{1/2}W(s)ds\}$
$+ \{-\int_1^{1+\lambda_P} \alpha(s)^2\delta' B_2^{-1}(-JB_1J' + B_2)B_2^{-1}\delta ds\}.$

Proof of Theorem 1: The proof is provided in two stages. In the first stage we provide an asymptotic expansion. In the second we apply a functional central limit theorem and a weak convergence to stochastic integrals result, both from De Jong and Davidson (2000).

In the first stage we show that

$$\begin{aligned}
&\sum_{t=T}^{T+P}(\hat{u}^2_{2,t+\tau} - \hat{u}^2_{W,t+\tau}) && (16)\\
= \quad &\{-2\sum_{t=T}^{T+P}\alpha_t(T/t)(T^{-1/2}h'_{T,2,t+\tau})(-JB_1J' + B_2)(T^{1/2}H_{T,2}(t))\\
&+T^{-1}\sum_{t=T}^{T+P}(1-(1-\alpha_t)^2)(T/t)^2(T^{1/2}H'_{T,2}(t))(-JB_1J' + B_2)(T^{1/2}H_{T,2}(t))\}\\
&+2\{-\sum_{t=T}^{T+P}\alpha_t\delta' B_2^{-1}(-JB_1J' + B_2)(T^{-1/2}h_{T,2,t+\tau})\\
&+T^{-1}\sum_{t=T}^{T+P}\alpha_t^2(T/t)\delta' B_2^{-1}(-JB_1J' + B_2)B_2^{-1}JB_1J'(T^{1/2}H_{T,2}(t))\\
&+T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)\delta' B_2^{-1}(-JB_1J' + B_2)(T^{1/2}H_{T,2}(t))\}\\
&+\{-T^{-1}\sum_{t=T}^{T+P}\alpha_t^2(T/t)\delta' B_2^{-1}(-JB_1J' + B_2)B_2^{-1}\delta\} + o_p(1).
\end{aligned}$$

To do so first note that straightforward algebra reveals that

$$\begin{aligned}
&\sum_{t=T}^{T+P}(\hat{u}^2_{2,t+\tau} - \hat{u}^2_{W,t+\tau}) && (17)\\
= \quad &\{-2\sum_{t=T}^{T+P}\alpha_t(T/t)(T^{-1/2}h'_{T,2,t+\tau})(-JB_1(t)J' + B_2(t))(T^{1/2}H_{T,2}(t))\\
&+T^{-1}\sum_{t=T}^{T+P}(1-(1-\alpha_t)^2)(T/t)^2(T^{1/2}H'_{T,2}(t))B_2(t)x_{T,2,t}x'_{T,2,t}B_2(t)(T^{1/2}H_{T,2}(t))\\
&-T^{-1}\sum_{t=T}^{T+P}\alpha_t^2(T/t)^2(T^{1/2}H'_{T,2}(t))JB_1(t)J'x_{T,2,t}x'_{T,2,t}JB_1(t)J'(T^{1/2}H_{T,2}(t))\\
&-2T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)^2(T^{1/2}H'_{T,2}(t))B_2(t)x_{T,2,t}x'_{T,2,t}JB_1(t)J'(T^{1/2}H_{T,2}(t))\}\\
&+2\{-\sum_{t=T}^{T+P}\alpha_t\delta' B_2^{-1}(t)(-JB_1(t)J' + B_2(t))(T^{-1/2}h_{T,2,t+\tau})\\
&+T^{-1}\sum_{t=T}^{T+P}\alpha_t^2(T/t)\delta' B_2^{-1}(t)(-JB_1(t)J' + B_2(t))x_{T,2,t}x'_{T,2,t}JB_1(t)J'(T^{1/2}H_{T,2}(t))\\
&+T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)\delta' B_2^{-1}(t)(-JB_1(t)J' + B_2(t))x_{T,2,t}x'_{T,2,t}B_{T,2}(t)(T^{1/2}H_{T,2}(t))\}\\
&+\{-T^{-1}\sum_{t=T}^{T+P}\alpha_t^2(T/t)\delta' B_2^{-1}(t)(-JB_1(t)J' + B_2(t))x_{T,2,t}x'_{T,2,t}(-JB_1(t)J' + B_2(t))B_2^{-1}(t)\delta\}\\
&+o_p(1).
\end{aligned}$$

We must then show that each bracketed term from (16) corresponds to that in (17). For brevity we will show this in detail only for the first bracketed term. The second and third follow from similar arguments.

Consider the first bracketed term in (17). If we add and subtract $-JB_1J' + B_2$ in the first component, and rearrange terms we obtain

$$\begin{aligned}
& -2\sum\nolimits_{t=T}^{T+P} \alpha_t(T/t)(T^{-1/2}h'_{T,2,t+\tau})(-JB_1(t)J' + B_2(t))(T^{1/2}H_{T,2}(t)) \\
= & -2\sum\nolimits_{t=T}^{T+P} \alpha_t(T/t)(T^{-1/2}h'_{T,2,t+\tau})(-JB_1J' + B_2)(T^{1/2}H_{T,2}(t)) \\
& -2T^{-1/2}\sum\nolimits_{t=T}^{T+P} \alpha_t(T/t)[(T^{1/2}H'_{T,2}(t)) \otimes (T^{-1/2}h'_{T,2,t+\tau})]vec(T^{1/2}[(-JB_1(t)J' + B_2(t)) - (-JB_1J' + B_2)]).
\end{aligned}$$

The first right-hand side term is the desired result. For the second right-hand side term first note that Assumptions 3 and 4 suffice for each of $\alpha(t)$, T/t, $T^{1/2}H'_{T,2}(t)$ and $vec(T^{1/2}[(-JB_1(t)J' + B_2(t)) - (-JB_1J' + B_2)])$ to converge weakly. Applying Theorem 3.2 of de Jong and Davidson (2000) then implies that the second right-hand side term is $O_p(T^{1/2})$ and the proof is complete.

For the second, third and fourth components of the first bracketed term note that adding and subtracting B_2, B_2^{-1}, B_1 and B_1^{-1} provides

$$\begin{aligned}
& T^{-1}\sum\nolimits_{t=T}^{T+P} (1-(1-\alpha_t)^2)(T/t)^2(T^{1/2}H'_{T,2}(t))B_2(t)x_{T,2,t}x'_{T,2,t}B_2(t)(T^{1/2}H_{T,2}(t)) \\
= & T^{-1}\sum\nolimits_{t=T}^{T+P} (1-(1-\alpha_t)^2)(T/t)^2(T^{1/2}H'_{T,2}(t))B_2(T^{1/2}H_{T,2}(t)) \\
& +2T^{-1}\sum\nolimits_{t=T}^{T+P} (1-(1-\alpha_t)^2)(T/t)^2(T^{1/2}H'_{T,2}(t))(B_2(t)-B_2)(T^{1/2}H_{T,2}(t)) \\
& +T^{-1}\sum\nolimits_{t=T}^{T+P} (1-(1-\alpha_t)^2)(T/t)^2(T^{1/2}H'_{T,2}(t))B_2(x_{T,2,t}x'_{T,2,t} - B_2^{-1})B_2(T^{1/2}H_{T,2}(t)) \\
& +2T^{-1}\sum\nolimits_{t=T}^{T+P} (1-(1-\alpha_t)^2)(T/t)^2(T^{1/2}H'_{T,2}(t))B_2(x_{T,2,t}x'_{T,2,t} - B_2^{-1})(B_2(t)-B_2)(T^{1/2}H_{T,2}(t)) \\
& +T^{-1}\sum\nolimits_{t=T}^{T+P} (1-(1-\alpha_t)^2)(T/t)^2(T^{1/2}H'_{T,2}(t))(B_2(t)-B_2)(x_{T,2,t}x'_{T,2,t} - B_2^{-1})(B_2(t)-B_2)(T^{1/2}H_{T,2}(t)) \\
& +2T^{-1}\sum\nolimits_{t=T}^{T+P} (1-(1-\alpha_t)^2)(T/t)^2(T^{1/2}H'_{T,2}(t))(B_2(t)-B_2)B_2^{-1}(B_2(t)-B_2)(T^{1/2}H_{T,2}(t)),
\end{aligned} \tag{18}$$

$$\begin{aligned}
& T^{-1}\sum\nolimits_{t=T}^{T+P} \alpha_t^2(T/t)^2(T^{1/2}H'_{T,2}(t))JB_1(t)J'x_{T,2,t}x'_{T,2,t}JB_1(t)J'(T^{1/2}H_{T,2}(t)) \\
= & T^{-1}\sum\nolimits_{t=T}^{T+P} \alpha_t^2(T/t)^2(T^{1/2}H'_{T,2}(t))JB_1J'(T^{1/2}H_{T,2}(t)) \\
& +2T^{-1}\sum\nolimits_{t=T}^{T+P} \alpha_t^2(T/t)^2(T^{1/2}H'_{T,2}(t))JB_1J'B_2^{-1}J(B_1(t)-B_1)J'(T^{1/2}H_{T,2}(t)) \\
& +T^{-1}\sum\nolimits_{t=T}^{T+P} \alpha_t^2(T/t)^2(T^{1/2}H'_{T,2}(t))JB_1J'(x_{T,2,t}x'_{T,2,t} - B_2^{-1})JB_1J'(T^{1/2}H_{T,2}(t)) \\
& +2T^{-1}\sum\nolimits_{t=T}^{T+P} \alpha_t^2(T/t)^2(T^{1/2}H'_{T,2}(t))JB_1J'(x_{T,2,t}x'_{T,2,t} - B_2^{-1})J(B_1(t)-B_1)J'(T^{1/2}H_{T,2}(t)) \\
& +T^{-1}\sum\nolimits_{t=T}^{T+P} \alpha_t^2(T/t)^2(T^{1/2}H'_{T,2}(t))J(B_1(t)-B_1)J'(x_{T,2,t}x'_{T,2,t} - B_2^{-1})J(B_1(t)-B_1)J'(T^{1/2}H_{T,2}(t)) \\
& +T^{-1}\sum\nolimits_{t=T}^{T+P} \alpha_t^2(T/t)^2(T^{1/2}H'_{T,2}(t))J(B_1(t)-B_1)J'B_2^{-1}J(B_1(t)-B_1)J'(T^{1/2}H_{T,2}(t)),
\end{aligned} \tag{19}$$

$$\begin{aligned}
& T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)^2(T^{1/2}H'_{T,2}(t))B_2(t)x_{T,2,t}x'_{T,2,t}JB_1(t)J'(T^{1/2}H_{T,2}(t)) \qquad (20)\\
= \; & T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)^2(T^{1/2}H'_{T,2}(t))JB_1J'(T^{1/2}H_{T,2}(t))\\
& +T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)^2(T^{1/2}H'_{T,2}(t))B_2(x_{T,2,t}x'_{T,2,t}-B_2^{-1})JB_1J'(T^{1/2}H_{T,2}(t))\\
& +T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)^2(T^{1/2}H'_{T,2}(t))B_2(x_{T,2,t}x'_{T,2,t}-B_2^{-1})J(B_1(t)-B_1)J'(T^{1/2}H_{T,2}(t))\\
& +T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)^2(T^{1/2}H'_{T,2}(t))(B_2(t)-B_2)B_2^{-1}JB_1J'(T^{1/2}H_{T,2}(t))\\
& +T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)^2(T^{1/2}H'_{T,2}(t))(B_2(t)-B_2)B_2^{-1}J(B_1(t)-B_1)J'(T^{1/2}H_{T,2}(t))\\
& +T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)^2(T^{1/2}H'_{T,2}(t))J(B_1(t)-B_1)J'(T^{1/2}H_{T,2}(t))\\
& +T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)^2(T^{1/2}H'_{T,2}(t))(B_2(t)-B_2)(x_{T,2,t}x'_{T,2,t}-B_2^{-1})J(B_1(t)-B_1)J'(T^{1/2}H_{T,2}(t))\\
& +T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)^2(T^{1/2}H'_{T,2}(t))(B_2(t)-B_2)(x_{T,2,t}x'_{T,2,t}-B_2^{-1})JB_1J'(T^{1/2}H_{T,2}(t)).
\end{aligned}$$

Note that the weighted sum of the first right-hand side term of each of (18) – (20) gives us

$$\begin{aligned}
& T^{-1}\sum_{t=T}^{T+P}(1-(1-\alpha_t)^2)(T/t)^2(T^{1/2}H'_{T,2}(t))B_2(T^{1/2}H_{T,2}(t))\\
& -T^{-1}\sum_{t=T}^{T+P}\alpha_t^2(T/t)^2(T^{1/2}H'_{T,2}(t))JB_1J'(T^{1/2}H_{T,2}(t))\\
& -2T^{-1}\sum_{t=T}^{T+P}\alpha_t(1-\alpha_t)(T/t)^2(T^{1/2}H'_{T,2}(t))JB_1J'(T^{1/2}H_{T,2}(t))\\
= \; & T^{-1}\sum_{t=T}^{T+P}(1-(1-\alpha_t)^2)(T/t)^2(T^{1/2}H'_{T,2}(t))(-JB_1J'+B_2)(T^{1/2}H_{T,2}(t))
\end{aligned}$$

the second right-hand side term in (17). We must therefore show that all of the remaining right-hand side terms in (18)-(20) are $o_p(1)$. The proof of each is very similar. For example, taking the absolute value of the fifth right-hand side term in (18) provides

$$\begin{aligned}
& |T^{-1}\sum_{t=T}^{T+P}(1-(1-\alpha_t)^2)(T/t)^2(T^{1/2}H'_{T,2}(t))(B_2(t)-B_2)(x_{T,2,t}x'_{T,2,t}-B_2^{-1})(B_2(t)-B_2)(T^{1/2}H_{T,2}(t))|\\
\leq \; & k^4(\sup_t|T^{1/2}H_{T,2}(t)|)^2(\sup_t|B_2(t)-B_2|)^2(T^{-1}\sum_{t=T}^{T+P}|x_{T,2,t}x'_{T,2,t}-B_2^{-1}|).
\end{aligned}$$

Since assumptions 3 and 4 suffice for $T^{-1}\sum_{t=T}^{T+P}|x_{T,2,t}x'_{T,2,t}-B_2^{-1}|=O_p(1)$, $\sup_t|T^{1/2}H_{T,2}(t)|=O_p(1)$ and $\sup_t|T^{1/2}H_{T,2}(t)|=o_p(1)$ we obtain the desired result.

For the second stage of the proof we show that the expansion in (17) converges in distribution to the term provided in the Theorem. To do so recall that Assumption 4 implies $\alpha_t \Rightarrow \alpha(s)$ and $(t/T) \Rightarrow s$. Also, Assumptions 3 (a) - (d) imply $T^{1/2}H_{T,2}(t) \Rightarrow s^{-1}V^{1/2}W(s)$. Continuity then provides the desired results for the second contribution to the first bracketed term, for the second and third contributions to the second bracketed term and the third bracketed term.

The remaining two contributions (the first in each of the first two bracketed terms), are each weighted sums of increments $h_{T,t+\tau}$. Consider the first contribution to the second bracketed term.

Since this increment satisfies Assumption 3 (d) and has an associated long-run variance V, we can apply Theorem 4.1 of De Jong and Davidson (2000) directly to obtain the desired convergence in distribution

$$-\sum\nolimits_{t=T}^{T+P} \alpha_t \delta' B_2^{-1}(-JB_1J' + B_2)(T^{-1/2}h_{T,2,t+\tau}) \to_d \\ -\int_1^{1+\lambda_P} \alpha(s)\delta' B_2^{-1}(-JB_1J' + B_2)V^{1/2}dW(s).$$

For the first contribution to the first bracketed term additional care is needed. Again, since the increments satisfy Assumption 3 (d) with long-run variance V we can apply Theorem 4.1 of De Jong and Davidson (2000) to obtain

$$-2\sum\nolimits_{t=T}^{T+P} \alpha_t (T/t)(T^{-1/2}h'_{T,2,t+\tau})(-JB_1J' + B_2)(T^{1/2}H_{T,2}(t)) \\ \to_d -2\int_1^{1+\lambda_P} \alpha(s)s^{-1}W'(s)V^{1/2}(-JB_1J' + B_2)V^{1/2}dW(s) + \Lambda.$$

Note the addition of the drift term Λ. To obtain the desired result we must show that this term is zero. A detailed proof is provided in Lemma A6 of Clark and McCracken (2005a) – albeit under the technical conditions provided in Hansen (1992) rather than those provided here. Rather than repeat the proof we provide an intuitive argument. Note that $H_{T,2}(t) = t^{-1}\sum_{s=1}^{t-\tau} h_{T,2,s+\tau}$. In particular note the range of summation. Since Assumption 3 (b) maintains that the increments of the stochastic integral $h_{T,2,t+\tau}$ form an MA($\tau - 1$) we find that $h_{T,2,t+\tau}$ is uncorrelated with every element of $H_{T,2}(t)$. Since Λ captures the contribution to the mean of the limiting distribution due to covariances between the increments $h_{T,2,t+\tau}$ and the elements of $H_{T,2}(t)$ we know that $\Lambda = 0$ and the proof is complete.

Proof of Corollary 1: Note that both the second bracketed term and the first component of the first bracketed term are zero mean and moreover, the third bracketed term is nonstochastic. Taking expectations we then obtain

$$\begin{aligned}
& E\{\int_1^{1+\lambda_P} \xi_W(s)\} \\
= \ & \{0 + \int_1^{1+\lambda_P} (1-(1-\alpha(s))^2)s^{-2}E[W'(s)V^{1/2}(-JB_1J' + B_2)V^{1/2}W(s)]ds\} \\
& +\{0\} - \int_1^{1+\lambda_P} \alpha^2(s)\delta' B_2^{-1}(-JB_1J' + B_2)B_2^{-1}\delta ds \\
= \ & \int_1^{1+\lambda_P} (1-(1-\alpha(s))^2)s^{-2}tr(E[W(s)W'(s)](-JB_1J' + B_2)V)ds \\
& - \int_1^{1+\lambda_P} \alpha^2(s)\delta' B_2^{-1}(-JB_1J' + B_2)B_2^{-1}\delta ds \\
= \ & \int_1^{1+\lambda_P} (1-(1-\alpha(s))^2)s^{-1}tr((-JB_1J' + B_2)V)ds \\
& - \int_1^{1+\lambda_P} \alpha^2(s)\delta' B_2^{-1}(-JB_1J' + B_2)B_2^{-1}\delta ds.
\end{aligned}$$

Proof of Corollary 2: We obtain our pointwise optimal combining weight by maximizing, for each fixed s, the argument of the integral in Corollary 1. That is we choose $\alpha(s)$ to maximize

$$(1-(1-\alpha(s))^2)s^{-1}tr((-JB_1J'+B_2)V)-\alpha^2(s)\delta'B_2^{-1}(-JB_1J'+B_2)B_2^{-1}\delta \qquad (21)$$

Differentiating (21) with respect to α we obtain

$$\begin{aligned} FOC\ \alpha &: \quad 2(1-\alpha(s))s^{-1}tr((-JB_1J'+B_2)V)-2\alpha(s)\delta'B_2^{-1}(-JB_1J'+B_2)B_2^{-1}\delta \\ SOC\ \alpha &: \quad -2\alpha(s)s^{-1}tr((-JB_1J'+B_2)V)-2\delta'B_2^{-1}(-JB_1J'+B_2)B_2^{-1}\delta. \end{aligned}$$

Setting the FOC to zero and solving for $\alpha(s)$ provides the formula from the Corollary. Algebra reveals that the SOC is negative at this solution and we obtain the desired result.

Proof of Proposition 1: Straightforward algebra reveals that

$$\hat{\alpha}_{BG}=\frac{(N^{-1}\sum_{s=R}^{T-\tau}\hat{u}_{2,t+\tau}x'_{2,t})(J\hat{\beta}_{1,R}-\hat{\beta}_{2,R})}{(J\hat{\beta}_{1,R}-\hat{\beta}_{2,R})'(N^{-1}\sum_{s=R}^{T-\tau}x_{2,t}x'_{2,t})(J\hat{\beta}_{1,R}-\hat{\beta}_{2,R})}$$

where $J\hat{\beta}_{1,R}-\hat{\beta}_{2,R}=-(-JB_1(R)J'+B_2(R))H_2(R)-T^{-1/2}(-JB_1(R)J'+B_2(R))B_2^{-1}(R)\delta$ and $N^{-1}\sum_{s=R}^{T-\tau}\hat{u}_{2,t+\tau}x'_{2,t}=N^{-1}\sum_{s=R}^{T-\tau}u_{2,t+\tau}x'_{2,t}-(N^{-1}\sum_{s=R}^{T-\tau}x_{2,t}x'_{2,t})B_2(R)H_2(R)$. Note that Assumption 3 implies $B_i(T)\rightarrow_p B_i$, $i=1,2$ and $N^{-1}\sum_{s=R}^{T-\tau}x_{2,t}x'_{2,t}\rightarrow_p B_2^{-1}$ while Assumption 4 suffices for $R^{1/2}H_2(R)\rightarrow_d V^{1/2}W_1$ and $N^{-1/2}\sum_{s=R}^{T-\tau}u_{t+\tau}x_{2,t}\rightarrow_d V^{1/2}W_0$ where Assumption 3 (b) (and the range of summation) implies that W_0 and W_1 are uncorrelated and hence independent. Let $\tilde{B}\equiv(-JB_1J'+B_2)$. Taking limits and rearranging terms we then obtain

$$\begin{aligned} \hat{\alpha}_{BG} &= \frac{-(N^{-1}\sum_{s=R}^{T-\tau}u_{t+\tau}x'_{2,t})\tilde{B}H_2(R)+H'_2(R)\tilde{B}H_2(R)-T^{-.5}(N^{-1}\sum_{s=R}^{T-\tau}u_{t+\tau}x'_{2,t})\tilde{B}B_2^{-1}\delta+T^{-.5}H'_2(R)\tilde{B}B_2^{-1}\delta+o_p(T^{-.5})}{H'_2(R)\tilde{B}H_2(R)+2T^{-.5}H'_2(R)\tilde{B}B_2^{-1}\delta+T^{-1}\delta'B_2^{-1}\tilde{B}B_2^{-1}\delta+o_p(T^{-.5})} \\ &= 1-\frac{(N^{-1}\sum_{s=R}^{T-\tau}u_{t+\tau}x_{2,t}+T^{-.5}B_2^{-1}\delta)'\tilde{B}(H_2(R)+T^{-.5}B_2^{-1}\delta)+o_p(T^{-.5})}{(H_2(R)+T^{-.5}B_2^{-1}\delta)'\tilde{B}(H_2(R)+T^{-.5}B_2^{-1}\delta)+o_p(T^{-.5})} \\ &= 1-(\frac{N}{R})^{-1}\frac{(N^{-.5}V^{-.5}\sum_{s=R}^{T-\tau}u_{t+\tau}x_{2,t}+(\frac{N}{T})^{1/2}V^{-.5}B_2^{-1}\delta)'[V^{1/2}\tilde{B}V^{1/2}](R^{1/2}V^{-.5}H_2(R)+(\frac{R}{T})^{1/2}V^{-.5}B_2^{-1}\delta)}{(R^{1/2}V^{-.5}H_2(R)+(\frac{R}{T})^{1/2}V^{-.5}B_2^{-1}\delta)'[V^{1/2}\tilde{B}V^{1/2}](R^{1/2}V^{-.5}H_2(R)+(\frac{R}{T})^{1/2}V^{-.5}B_2^{-1}\delta)}+o_p(1) \\ &\rightarrow {}^d 1-\pi^{-1}\left(\frac{(W_0+\frac{\pi}{1+\pi}V^{-.5}B_2^{-1}\delta)'[V^{1/2}\tilde{B}V^{1/2}](W_1+\frac{1}{1+\pi}V^{-.5}B_2^{-1}\delta)}{(W_1+\frac{1}{1+\pi}V^{-.5}B_2^{-1}\delta)'[V^{1/2}\tilde{B}V^{1/2}](W_1+\frac{1}{1+\pi}V^{-.5}B_2^{-1}\delta)}\right) \end{aligned}$$

and the proof is complete.

7 Appendix 2: Application Details

Unless otherwise noted, all data are taken from the FAME database of the Federal Reserve Board of Governors. All data end in 2004:Q4 or December 2004. *Start point* refers to the beginning of the regression sample, determined by the availability of the raw data, any differencing, and lag orders. The *predictand* column provides the data frequency and transformation of the predictand. Inflation rates are calculated as log changes. In all cases the forecasting models include a constant in the set of predictors. Note that *CFNAI* refers to the Federal Reserve Bank of Chicago's national activity index, a factor index of the business cycle obtained from the bank's website.

application	**predictand**	**predictors**	**data notes**
Ind. prod. & factor index	production (monthly, 1200 $\Delta\ln$)	6 lags of production; 6 lags of CFNAI	start point: 1967:9
Disp. income & factor index	income (monthly, 1200 $\Delta\ln$)	6 lags of income; 6 lags of CFNAI	start point: 1967:9
M&T sales & factor index	sales (monthly, 1200 $\Delta\ln$)	6 lags of sales; 6 lags of CFNAI	start point: 1967:9
Employ. & factor index	employment (monthly, 1200 $\Delta\ln$)	6 lags of employ.; 6 lags of CFNAI	start point: 1967:9
Unemp. & ins. claims	unemployment rate (monthly, Δ)	6 lags of unemp.; 6 lags of new claims for unemp. insurance; 6 lags of continuing claims for unemp. insurance	start point: 1967:7
Inflation & output gap	annualized core PCE inflation (quarterly, Δ)	2 lags of inflation; 1 lag of output gap (log of actual GDP/CBO potential GDP)	start point: 1960:1 CBO series from St. Louis Fed's FRED database
Infl. & gap, ULC, imports	annualized core PCE inflation (quarterly, Δ)	2 lags of infl.; 1 lag of output gap; 2 lags of unit labor costs ($\Delta\ln$); 2 lags of import price inflation	start point: 1967:4
GDP & VAR(2)	GDP (quarterly, 400 $\Delta\ln$)	2 lags of: GDP; ln(consumption /GDP); aggregate hours ($\Delta\ln$); core PCE inflation (Δ); and the 3-month T-bill rate (Δ)	start point: 1964:4
3mo. T-bill & VAR(2)	3-month T-bill rate (quarterly, Δ)	same as in GDP application	start point: 1964:4
Stock return & d/p, int. rate	excess return on S&P 500 (quarterly, percentage points: nominal return less safe interest rate)	1 lag of the dividend-price ratio; 1 lag of interest rate relative to prior year average	start point: 1953:2 d/p ratio based on average of dividends over prior year; dividend data from Global Insight; interest rate (compounded over three months of quarter) from Kenneth French's website
U.S.-Switz. ex. rate & int. diff.	exchange rate (monthly, 1200 $\Delta\ln$)	interest differential at 1, 3, 6, and 12 months, 1 lag each	start point: 1973:7 interest rates from Global Insight
U.S.-U.K. ex. rate & int. diff.	exchange rate (monthly, 1200 $\Delta\ln$)	interest differentials at 1, 3, 6, and 12 months, 1 lag each	start point: 1973:10

References

Atkeson, Andrew, and Lee E. Ohanian (2001). "Are Phillips Curves Useful for Forecasting Inflation?" Federal Reserve Bank of Minneapolis *Quarterly Review* 25, 2-11.

Bates, J.M., and Clive W.J. Granger (1969), "The Combination of Forecasts," *Operations Research Quarterly* 20, 451-468.

Bernanke, Ben S., and Jean Boivin (2003), "Monetary Policy in a Data-Rich Environment," *Journal of Monetary Economics* 50, 525-46.

Bossaerts, Peter, and Pierre Hillion (1999), "Implementing Statistical Criteria to Select Return Forecasting Models: What Do We Learn?" *Review of Financial Studies* 12, 405-28.

Brave, Scott, and Jonas D.M. Fisher (2004), "In Search of a Robust Inflation Forecast," Federal Reserve Bank of Chicago *Economic Perspectives*, Fourth Quarter, 12-13.

Brayton, Flint, John M. Roberts, and John C. Williams (1999), "What's Happened to the Phillips Curve?" Board of Governors of the Federal Reserve System Finance and Economics Discussion Paper No. 1999-49.

Campbell, John Y., and Samuel B. Thompson (2005), "Predicting the Equity Premium Out of Sample: Can Anything Beat the Historical Average?" manuscript, Harvard University.

Cecchetti, Stephen G. (1995), "Inflation Indicators and Inflation Policy," *NBER Macroeconomics Annual*, 189-219.

Clarida, Richard H., and Mark P. Taylor (1997), "The Term Structure of Forward Exchange Premiums and the Forecastability of Spot Exchange Rates: Correcting the Errors," *Review of Economics and Statistics* 79, 353-61.

Clarida, Richard H., Lucia Sarno, Mark P. Taylor, and Giorgio Valente (2003), "The Out-of-Sample Success of Term Structure Models as Exchange Rate Predictors: a Step Beyond," *Journal of International Economics* 60, 61-83.

Clark, Todd E., and Michael W. McCracken (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105, 85-110.

Clark, Todd E., and Michael W. McCracken (2005a), "Evaluating Direct Multistep Forecasts," *Econometric Reviews* 24, 369-404.

Clark, Todd E., and Michael W. McCracken (2005b), "The Predictive Content of the Output Gap for Inflation: Resolving In–Sample and Out–of–Sample Evidence," *Journal of Money, Credit, and Banking*, forthcoming.

Clark, Todd E., and Kenneth D. West (2004), "Using Out–of–Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis," *Journal of Econometrics*, forthcoming.

Clark, Todd E., and Kenneth D. West (2005), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," manuscript, University of Wisconsin.

Clements, Michael P., and David F. Hendry (1998), *Forecasting Economic Time Series*, Cambridge University Press, Cambridge.

de Jong, Robert M., and James Davidson (2000), "The Functional Central Limit Theorem and Weak Convergence to Stochastic Integrals I: Weakly Dependent Processes," *Econometric Theory* 16, 621-42.

Diebold, Francis X. (1998), *Elements of Forecasting*, Cincinnati: South-Western College Publishing.

Doan, Thomas, Robert Litterman, and Christopher Sims (1984), "Forecasting and Conditional Prediction Using Realistic Prior Distributions," *Econometric Reviews* 3, 1-100.

Elliott, Graham, and Allan Timmermann (2004), "Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions," *Journal of Econometrics* 122, 47-79.

Filardo, Andrew J., 1999, "To Combine or Not To Combine Inflation Forecasts," manuscript, Federal Reserve Bank of Kansas City.

Fisher, Jonas D.M., C.T. Liu, and R. Zhou (2002). "When Can We Forecast Inflation?" Federal Reserve Bank of Chicago *Economic Perspectives* 26, First Quarter, 30-42.

Gavin, William T., and Kevin L. Kliesen (2002), "Unemployment Insurance Claims and Economic Activity," Federal Reserve Bank of St. Louis *Economic Review*, May/June, 15-27.

Goyal, Amit, and Ivo Welch (2003), "Predicting the Equity Premium with Dividend Ratios," *Management Science* 49, 639-54.

Hansen, Bruce E. (1992), "Convergence to Stochastic Integrals for Dependent Heterogeneous Processes, *Econometric Theory* 8, 489-500.

Hendry, David F., and Michael P. Clements (2004), "Pooling of Forecasts," *Econometrics Journal* 7, 1-31.

Inoue, Atsushi, and Lutz Kilian, 2004(b), "On the Selection of Forecasting Models," *Journal of Econometrics*, forthcoming.

Litterman, Robert B. (1986), "Forecasting with Bayesian Vector Autoregressions — Five Years of Experience," *Journal of Business and Economic Statistics* 4, 25-38.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson (2004), "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time

Series," *Journal of Econometrics*, forthcoming.

McCracken, Michael W. (2004), "Asymptotics for Out–of–Sample Tests of Causality," manuscript, University of Missouri.

Mehra, Yash P. (1990), "Real Output and Unit Labor Costs as Predictors of Inflation," Federal Reserve Bank of Richmond *Economic Review*, July/August, 31-39.

Montgomery, Alan L., Victor Zarnowitz, Ruey S. Tsay, and George C. Tiao (1998), "Forecasting the U.S. Unemployment Rate," *Journal of the American Statistical Association* 93, 478-93.

Orphanides, Athanasios, and Simon van Norden (2005), "The Reliability of Inflation Forecasts Based on Output Gap Estimates in Real Time," *Journal of Money, Credit, and Banking* 37, 583-601.

Pesaran, M. Hashem, and Allan Timmermann, (2002), "Market Timing and Return Prediction Under Model Instability," *Journal of Empirical Finance* 9, 495-510.

Robertson, John, and Ellis Tallman (1999), "Vector Autoregressions: Forecasting and Reality," Federal Reserve Bank of Atlanta *Economic Review*, First Quarter, 4-18.

Rotemberg, Julio J. (1996), "Prices, Output, and Hours: An Empirical Analysis Based on a Sticky Price Model," *Journal of Monetary Economics* 37, 505-33.

Rotemberg, Julio J., and Michael Woodford (1996), "Real-Business-Cycle Models and the Forecastable Movements in Output, Hours, and Consumption," *American Economic Review* 86, 71-111.

Shintani, Mototsugu (2005), "Nonlinear Forecasting Analysis Using Diffusion Indexes: An Application to Japan," *Journal of Money, Credit, and Banking* 37, 517-38.

Smith, Jeremy, and Kenneth F. Wallis (2005), "Combining Point Forecasts: The Simple Average Rules, OK?" manuscript, University of Warwick.

Stock, James H., and Mark W. Watson (1996), "Evidence on Structural Stability in Macroeconomic Time Series Relations," *Journal of Business and Economic Statistics* 14, 11-30.

Stock, James H., and Mark W. Watson (1999), "Forecasting Inflation," *Journal of Monetary Economics* 44, 293-335.

Stock, James H., and Mark W. Watson (2002), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics* 20, 147-62.

Stock, James H., and Mark W. Watson (2003), "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature* 41, 788-829.

Stock, James H., and Mark W. Watson (2005), "A Unified Treatment of Methods for Forecasting Using Many Predictors," manuscript, Harvard University.

Timmermann, Allan (2004), "Forecast Combinations," *Handbook of Forecasting*, forthcoming.

Wright, Jonathan H. (2003), "Forecasting U.S. Inflation by Bayesian Model Averaging," manuscript, Board of Governors of the Federal Reserve System.

Table 1: **Summary of Forecast Approaches**

approach to forecasting y_{t+1}	**details**
1. restricted	OLS estimates of model omitting x terms
2. unrestricted	OLS estimates of full model
3. opt. combination: known α_t^*	α_t^* × restricted + $(1-\alpha_t^*)$ × unrestricted, with α_t^* computed according to (5), using the known features of the DGP
4. opt. combination: $\hat{\alpha}_t^*$	$\hat{\alpha}_t^*$ × restricted + $(1-\hat{\alpha}_t^*)$ × unrestricted, with $\hat{\alpha}_t^*$ computed according to (8), using moments estimated from the data
5. simple average	.5 × restricted + .5 × unrestricted
6. SIC	forecast with model (restricted or unrestricted) with lower SIC at t
7. AIC	forecast with model (restricted or unrestricted) with lower AIC at t
8. BVAR	forecast with posterior mean estimate of unrestricted model, using Minnesota–type prior (all prior means equal 0; prior variances pinned down by hyperparameters $\lambda = .2$ and $\theta = .5$)
9. BMA: $\phi = .2$	Bayesian model averaging of restricted and unrestricted forecasts, using prior mean of equal weights and shrinkage parameter $\phi = .2$
10. BMA: $\phi = 2$	Bayesian model averaging of restricted and unrestricted forecasts, using prior mean of equal weights and shrinkage parameter $\phi = 2$

Table 2: Monte Carlo Results, Average MSEs

(for restricted model, average MSE; for other forecasts,
ratio of average MSE to restricted model's average MSE)

	DGP 1: signal=noise				**DGP 1: signal>noise**			
method/model	P=1	P=20	P=40	P=80	P=1	P=20	P=40	P=80
restricted	.762	.768	.765	.761	.821	.823	.820	.816
unrestricted	1.004	1.002	1.000	.998	.933	.935	.934	.931
opt. combination: known α_t^*	.995	.995	.994	.993	.931	.933	.933	.930
opt. combination: $\hat{\alpha}_t^*$	1.001	.999	.998	.996	.936	.937	.936	.933
simple average	.995	.995	.994	.993	.943	.945	.945	.944
SIC	1.007	1.004	1.003	1.002	.956	.953	.950	.943
AIC	1.009	1.004	1.003	1.001	.940	.940	.938	.934
BVAR	1.001	.997	.997	.995	.931	.932	.932	.930
BMA: $\phi = .2$	.996	.995	.995	.994	.939	.940	.940	.937
BMA: $\phi = 2$	1.000	.998	.997	.996	.940	.940	.939	.935
	DGP 2: signal=noise				**DGP 2: signal>noise**			
	P=1	P=20	P=40	P=80	P=1	P=20	P=40	P=80
restricted	.794	.802	.798	.793	.901	.907	.903	.897
unrestricted	1.013	1.006	1.000	.991	.893	.889	.884	.876
opt. combination: known α_t^*	.975	.974	.973	.970	.878	.876	.874	.868
opt. combination: $\hat{\alpha}_t^*$	.985	.982	.980	.976	.883	.881	.878	.872
simple average	.975	.974	.973	.970	.890	.890	.889	.887
SIC	1.003	1.002	1.002	1.001	.960	.955	.947	.929
AIC	1.015	1.010	1.006	.999	.904	.897	.890	.880
BVAR	.975	.974	.973	.971	.892	.891	.889	.884
BMA: $\phi = .2$	.980	.978	.977	.973	.883	.882	.879	.873
BMA: $\phi = 2$	.993	.990	.988	.984	.894	.891	.886	.878

Notes:

1. DGPs 1 and 2 are defined in equations (11) and (12). In the DGP 1 simulations, the coefficient b_{11} is set to .037 in the signal = noise experiment and .10 in the signal > noise experiment. In the DGP 2 simulations, the coefficients b_{ij} are set at empirically–based values ($b_{11} = .10$, $b_{21} = .03$, $b_{22} = -.02$, $b_{31} = .05$, $b_{32} = -.03$) in the signal > noise experiment and .527 times the empirical values in the signal = noise experiment.
2. The forecast approaches are defined in Table 1.
3. The total number of observations generated for each experiment is 160 (not counting the initial observations equal to the number of lags in the DGP). Forecasting begins with observation 81. Results are reported for forecasts evaluated over the following samples: 81 (P=1); 81-100 (P=20); 81-120 (P=40); and 81-160 (P=80).
4. The table entries are based on averages of forecast MSEs across 10,000 Monte Carlo simulations.

Table 3: Monte Carlo Results, Average MSEs in DGPs with $\beta_{22} = 0$

(for restricted model, average MSE; for other forecasts,
ratio of average MSE to restricted model's average MSE)

	DGP 1: signal=0				**DGP 2: signal=0**			
method/model	P=1	P=20	P=40	P=80	P=1	P=20	P=40	P=80
restricted	.751	.758	.756	.752	.748	.757	.754	.750
unrestricted	1.019	1.014	1.012	1.010	1.075	1.065	1.059	1.049
opt. combination: known α_t^*	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
opt. combination: $\hat{\alpha}_t^*$	1.010	1.007	1.006	1.005	1.025	1.021	1.019	1.016
simple average	1.006	1.004	1.003	1.003	1.018	1.016	1.015	1.012
SIC	1.005	1.003	1.003	1.002	1.001	1.000	1.000	1.000
AIC	1.008	1.008	1.007	1.006	1.025	1.015	1.013	1.011
BVAR	1.015	1.009	1.008	1.007	1.018	1.017	1.016	1.015
BMA: $\phi = .2$	1.006	1.004	1.004	1.003	1.021	1.017	1.016	1.013
BMA: $\phi = 2$	1.008	1.005	1.005	1.004	1.019	1.014	1.013	1.011

Notes:

1. See the notes to Table 2.
2. In these experiments, the DGP coefficients b_{ij} are set to 0.

Table 4: Monte Carlo Probabilities of Equaling or Beating Restricted Model's MSE, DGP 1

	DGP 1: signal=noise							
	P=1		P=20		P=40		P=80	
method/model	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$
unrestricted	.000	.499	.000	.436	.000	.422	.000	.446
opt. combination: known α_t^*	.000	.509	.000	.491	.000	.493	.000	.539
opt. combination: $\hat{\alpha}_t^*$	.005	.502	.000	.407	.000	.401	.000	.437
simple average	.000	.509	.000	.494	.000	.502	.000	.558
SIC	.861	.064	.780	.081	.733	.091	.655	.113
AIC	.660	.161	.512	.205	.432	.225	.319	.277
BVAR	.000	.491	.000	.481	.000	.497	.000	.516
BMA: $\phi = .2$	.000	.508	.000	.483	.000	.485	.000	.537
BMA: $\phi = 2$	.000	.507	.000	.457	.000	.448	.000	.481
	DGP 1: signal>noise							
	P=1		P=20		P=40		P=80	
	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$
unrestricted	.000	.542	.000	.701	.000	.772	.000	.875
opt. combination: known α_t^*	.000	.547	.000	.723	.000	.798	.000	.894
opt. combination: $\hat{\alpha}_t^*$	.000	.548	.000	.709	.000	.787	.000	.889
simple average	.000	.565	.000	.792	.000	.873	.000	.953
SIC	.359	.342	.217	.518	.146	.606	.067	.761
AIC	.165	.452	.074	.635	.040	.724	.011	.848
BVAR	.000	.538	.000	.706	.000	.793	.000	.892
BMA: $\phi = .2$	.000	.559	.000	.763	.000	.840	.000	.933
BMA: $\phi = 2$	.000	.549	.000	.714	.000	.786	.000	.887
	DGP 1: signal=0							
	P=1		P=20		P=40		P=80	
	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$
unrestricted	.000	.475	.000	.319	.000	.262	.000	.205
opt. combination: known α_t^*	1.000	.000	1.000	.000	1.000	.000	1.000	.000
opt. combination: $\hat{\alpha}_t^*$	.008	.478	.000	.275	.000	.223	.000	.183
simple average	.000	.484	.000	.357	.000	.315	.000	.276
SIC	.957	.015	.923	.018	.906	.014	.885	.014
AIC	.830	.078	.724	.076	.666	.073	.587	.068
BVAR	.000	.479	.000	.427	.000	.402	.000	.369
BMA: $\phi = .2$	.000	.483	.000	.351	.000	.306	.000	.265
BMA: $\phi = 2$	.000	.483	.000	.336	.000	.283	.000	.231

Notes:
1. See the notes to Tables 2 and 3.
2. The table entries are frequencies (percentages of 10,000 Monte Carlo simulations) with which each forecast approach yields a forecast MSE less than or equal to the restricted model's MSE.

Table 5: Monte Carlo Probabilities of Equaling or Beating Restricted Model's MSE, DGP 2

	DGP 2: signal=noise							
	P=1		P=20		P=40		P=80	
method/model	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$
unrestricted	.000	.492	.000	.470	.000	.478	.000	.527
opt. combination: known α_t^*	.000	.520	.000	.609	.000	.658	.000	.745
opt. combination: $\hat{\alpha}_t^*$	.000	.512	.000	.548	.000	.582	.000	.671
simple average	.000	.520	.000	.616	.000	.678	.000	.776
SIC	.968	.015	.935	.022	.910	.029	.863	.044
AIC	.560	.215	.388	.261	.291	.298	.169	.385
BVAR	.000	.523	.000	.570	.000	.624	.000	.709
BMA: $\phi = .2$	.000	.515	.000	.577	.000	.622	.000	.717
BMA: $\phi = 2$	.000	.513	.000	.518	.000	.538	.000	.600
	DGP 2: signal>noise							
	P=1		P=20		P=40		P=80	
	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$
unrestricted	.000	.541	.000	.720	.000	.804	.000	.917
opt. combination: known α_t^*	.000	.558	.000	.786	.000	.871	.000	.957
opt. combination: $\hat{\alpha}_t^*$	.000	.557	.000	.779	.000	.864	.000	.955
simple average	.000	.579	.000	.859	.000	.940	.000	.990
SIC	.632	.199	.457	.338	.335	.451	.168	.660
AIC	.106	.483	.036	.680	.015	.777	.002	.907
BVAR	.000	.562	.000	.808	.000	.903	.000	.973
BMA: $\phi = .2$	.000	.564	.000	.797	.000	.881	.000	.962
BMA: $\phi = 2$	.000	.549	.000	.732	.000	.815	.000	.923
	DGP 2: signal=0							
	P=1		P=20		P=40		P=80	
	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$	$= MSE_1$	$< MSE_1$
unrestricted	.000	.461	.000	.264	.000	.183	.000	.110
opt. combination: known α_t^*	1.000	.000	1.000	.000	1.000	.000	1.000	.000
opt. combination: $\hat{\alpha}_t^*$	.000	.479	.000	.320	.000	.254	.000	.192
simple average	.000	.479	.000	.361	.000	.309	.000	.255
SIC	.999	.001	.998	.000	.997	.000	.996	.000
AIC	.893	.045	.812	.036	.771	.032	.719	.023
BVAR	.000	.484	.000	.397	.000	.364	.000	.300
BMA: $\phi = .2$	.000	.479	.000	.348	.000	.291	.000	.231
BMA: $\phi = 2$	.000	.487	.000	.322	.000	.251	.000	.178

Notes:

1. See the notes to Tables 2-4.

Table 6: Application Results

RMSE for restricted forecast, and MSE ratios for other forecasts

	Ind. prod. & factor index		**Disp. income & factor index**		**M&T sales & factor index**		**Employ. & factor index**	
method/model	**78-91**	**92-04**	**78-91**	**92-04**	**78-91**	**92-04**	**78-91**	**92-04**
restricted	8.561	6.030	7.997	9.242	13.631	9.756	2.119	1.096
unrestricted	.889	.925	1.140	.995	.940	1.000	.832	1.057
opt. combination: $\hat{\alpha}_t^*$	**.888**	.923	1.066	.991	**.936**	.986	**.826**	1.022
simple average	.897	.936	1.021	.991	.943	**.965**	.847	**.925**
SIC	1.024	.925	1.038	1.000	.992	1.000	.853	1.057
AIC	.918	.925	1.140	.995	.940	1.000	.832	1.057
BVAR	.906	.934	**.946**	**.990**	.940	.982	.879	.995
BMA: $\phi = .2$	.892	**.922**	1.078	.992	.936	.991	.838	1.007
BMA: $\phi = 2$	.894	.925	1.137	.995	.940	1.000	.828	1.057
	Unemp. & ins. claims		**Inflation & output gap**		**Infl. & gap, ULC, imports**		**GDP & VAR(2)**	
	78-91	**92-04**	**78-91**	**92-04**	**78-91**	**92-04**	**78-91**	**92-04**
restricted	.189	.136	1.155	.568	1.164	.572	3.805	1.824
unrestricted	.903	.849	**.937**	.962	.973	1.048	.849	1.082
opt. combination: $\hat{\alpha}_t^*$	.877	.838	.941	.958	.975	.982	.814	1.034
simple average	.869	.854	.954	.957	**.953**	**.959**	**.809**	**.972**
SIC	.966	.849	**.937**	.962	1.000	1.000	1.000	1.001
AIC	.903	.849	**.937**	.962	1.140	1.048	.911	1.082
BVAR	**.832**	**.817**	.960	1.001	.975	1.021	.907	1.036
BMA: $\phi = .2$	.886	.848	.951	**.955**	.967	.977	.811	1.004
BMA: $\phi = 2$	.903	.849	.946	.962	1.013	1.021	.854	1.079
	3mo. T-bill & VAR(2)		**Stock return & d/p, int. rate**		**U.S.-Switz. ex. rate & int. diff.**		**U.S.-U.K. ex. rate & int. diff.**	
	78-91	**92-04**	**78-91**	**92-04**	**78-91**	**92-04**	**78-91**	**92-04**
restricted	1.221	.419	8.118	**7.826**	**4.069**	**3.126**	3.509	**2.602**
unrestricted	1.042	1.015	.993	1.175	1.105	1.015	1.010	1.044
opt. combination: $\hat{\alpha}_t^*$	1.020	.928	.977	1.151	1.048	1.004	.998	1.026
simple average	1.000	.903	**.933**	1.068	1.032	1.002	.997	1.018
SIC	1.000	1.000	.993	1.175	**1.000**	**1.000**	1.000	**1.000**
AIC	1.022	1.015	.993	1.175	1.086	**1.000**	1.007	**1.000**
BVAR	**.927**	**.790**	.937	1.100	1.036	1.010	**.982**	1.040
BMA: $\phi = .2$	1.010	.933	.960	1.123	1.042	1.003	.997	1.021
BMA: $\phi = 2$	1.038	.999	.993	1.172	1.054	1.003	.998	1.019

Notes:

1. The table presents MSE results for 1–step ahead forecasts over the indicated samples. In each case, the forecast with the lowest MSE is indicated with a bold font.
2. Details of the applications (data, forecast model specification, etc.) are provided in Section 4.1 and Appendix 2.
3. The forecast approaches are defined in Table 1.

Table 7: Comparison of Each Forecast's MSE to Best MSE,
Summary Statistics Across 12 Applications and Two Sample Periods

	mean	*median*	*st. dev.*	*90%ile – 10%ile*
restricted	1.084	1.062	.083	.220
unrestricted	1.069	1.041	.075	.162
opt. combination: $\hat{\alpha}_t^*$	1.043	1.024	.052	.120
simple average	1.025	1.013	.036	.075
SIC	1.071	1.041	.078	.171
AIC	1.076	1.052	.079	.189
BVAR	1.030	1.019	.035	.073
BMA: $\phi = .2$	1.040	1.020	.049	.112
BMA: $\phi = 2$	1.065	1.046	.071	.160

Notes:
1. See the notes to Table 6.
2. The presented results are summary statistics, across 24 (12 applications × two sample periods) sets of application results, for the ratio of each forecast's MSE to the best MSE in the given case (application and sample period).