Journal of Business & Economic Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ubes20

Generalized Shrinkage Methods for Forecasting Using Many Predictors

James H. Stock [a] & Mark W. Watson [b]

[a] Department of Economics, Littauer Center, Harvard University, Cambridge, MA

[b] Woodrow Wilson School and Department of Economics, Bendheim Hall, Princeton University, Princeton, NJ

Accepted author version posted online: 03 Aug 2012.Published online: 17 Oct 2012.

To cite this article: James H. Stock & Mark W. Watson (2012) Generalized Shrinkage Methods for Forecasting Using Many Predictors, Journal of Business & Economic Statistics, 30:4, 481-493, DOI: 10.1080/07350015.2012.715956

To link to this article: http://dx.doi.org/10.1080/07350015.2012.715956

Generalized Shrinkage Methods for Forecasting Using Many Predictors

James H. Stock
Department of Economics, Littauer Center, Harvard University, Cambridge, MA (*james_stock@harvard.edu*)

Mark W. Watson
Woodrow Wilson School and Department of Economics, Bendheim Hall, Princeton University, Princeton, NJ (*mwatson@Princeton.edu*)

This article provides a simple shrinkage representation that describes the operational characteristics of various forecasting methods designed for a large number of orthogonal predictors (such as principal components). These methods include pretest methods, Bayesian model averaging, empirical Bayes, and bagging. We compare empirically forecasts from these methods with dynamic factor model (DFM) forecasts using a U.S. macroeconomic dataset with 143 quarterly variables spanning 1960–2008. For most series, including measures of real economic activity, the shrinkage forecasts are inferior to the DFM forecasts. This article has online supplementary material.

KEY WORDS: Dynamic factor models; Empirical Bayes; High-dimensional model.

1. INTRODUCTION

Over the past 10 years, the dynamic factor model (DFM) (Geweke 1977) has been the predominant framework for research on macroeconomic forecasting using many predictors. The conceptual appeal of the DFM is two-fold: methods for estimation of factors in a DFM turn the curse of dimensionality into a blessing (Stock and Watson 1999, 2002a, b; Forni et al. 2000, 2004; Bai and Ng 2002, 2006), and the DFM arises naturally from log-linearized structural macroeconomic models including dynamic stochastic general equilibrium models (Sargent 1989; Boivin and Giannoni 2006). Bai and Ng (2008) and Stock and Watson (2011) surveyed econometric research on DFMs over this period. But the forecasting implications of the DFM—that the many predictors can be replaced by a small number of estimated factors—might not be justified in practice. Indeed, Eickmeier and Ziegler's (2008) meta-study finds mixed performance of DFM forecasts, which suggests considering other ways to handle many predictors. Accordingly, some recent articles have considered whether DFM macro forecasts can be improved upon using other many-predictor methods, including high-dimensional Bayesian vector autoregression (VAR; Andersson and Karlsson 2008; De Mol, Giannone, and Reichlin 2008; Korobilis 2008; Bańbura, Giannone, and Reichlin 2010; Carriero, Kapetanios, and Marcellino 2011), Bayesian model averaging (BMA; Jacobson and Karlsson 2004; Koop and Potter 2004; Wright 2009; Eklund and Karlsson 2007), bagging (BG; Inoue and Kilian 2008), Lasso (Bai and Ng 2007; De Mol, Giannone, and Reichlin 2008), boosting (Bai and Ng 2009), and forecast combination (multiple authors).

One difficulty in comparing these high-dimensional methods theoretically is that their derivations generally rely on specific modeling assumptions (e.g., iid data and strictly exogenous predictors), and it is not clear from those derivations what the algorithms are actually doing when they are applied in settings in which the modeling assumptions do not hold. Moreover, although there have been empirical studies of the performance of many of these methods for macroeconomic forecasting, it is difficult to draw conclusions across methods because of differences in datasets and implementation across studies.

This article therefore has two goals. The first goal is to characterize the properties of some forecasting methods applied to many orthogonal predictors in a time series setting in which the predictors are predetermined but not strictly exogenous. The results cover pretest (PT) and information criterion methods, BMA, empirical Bayes (EB) methods, and BG. It is shown that asymptotically all these methods have the same "shrinkage" representation, in which the weight on a predictor is the ordinary least-squares (OLS) estimator times a shrinkage factor that depends on the t statistic of that coefficient. These representations are a consequence of the algorithms and they hold under weak stationarity and moment assumptions about the actual statistical properties of the predictors; thus these methods can be compared directly using these shrinkage representations.

The second goal is to undertake an empirical comparison of these shrinkage methods using a quarterly U.S. macro dataset that includes 143 quarterly economic time series spanning 49 years. The DFM imposes a strong restriction that there are only a few factors and these factors can supplant the full large dataset for the purpose of forecasting. There are now a number of ways to estimate factors in large datasets, and a commonly used estimator is the first few principal components of the many predictors (ordered by their eigenvalues). The empirical question, then, is whether information in the full dataset, beyond the first few principal components, makes a significant marginal forecasting contribution. There are various ways to approach this question. One could, for example, retain the predictors in their original form, then (by appealing to Frisch–Waugh) consider the marginal predictive power of the part of those predictors

Journal of Business & Economic Statistics
October 2012, Vol. 30, No. 4
DOI: 10.1080/07350015.2012.715956

orthogonal to the factors. Algorithms for averaging or selecting models using the original predictors, which have been used for macro forecasting or closely related problems, include BMA and large VARs. However, we share De Mol, Giannone, and Reichlin's (2008) skepticism about the reliability of any resulting economic interpretation because of the collinearity of the data and the resulting instability of the weights and variable/model selection. Moreover, any economic interpretation that might have been facilitated by using the original series would be obscured by using instead their orthogonal projection on the first few factors. A different approach, the one we adopt, is to retain the perspective of a factor model but to imagine that the number of selected factors is simply smaller than it should be, that is, that the conventional wisdom that a few factors suffice to describe the postwar U.S. data is wrong. Because the principal components are estimates of the factors, this approach leads us to consider forecasts that potentially place nonzero weight on principal components beyond the first few. Because the principal components are orthogonal, shrinkage procedures for orthogonal regressors provide a theoretically well-grounded way to assess the empirical validity of the DFM forecasting restrictions.

We find that, for most macroeconomic time series, among linear estimators the DFM forecasts make efficient use of the information in the many predictors by using only a small number of estimated factors. These series include measures of real economic activity and some other central macroeconomic series, including some interest rates and monetary variables. For these series, the shrinkage methods with estimated parameters fail to provide mean squared error improvements over the DFM. For a small number of series, the shrinkage forecasts improve upon DFM forecasts, at least at some horizons and by some measures, and for these few series, the DFM might not be an adequate approximation. Finally, none of the methods considered here help much for series that are notoriously difficult to forecast, such as exchange rates, stock prices, or price inflation.

The shrinkage representations for forecasts using orthogonal predictors are described in Section 2. Section 3 describes the data and the forecasting experiment. Section 4 presents the empirical results, and Section 5 offers some concluding remarks.

2. SHRINKAGE REPRESENTATIONS OF FORECASTING METHODS

We consider the multiple regression model with orthonormal regressors,

$$Y_t = \delta' P_{t-1} + \varepsilon_t, \quad t = 1, \ldots, T, \quad P_t' P_t / T = I_n, \qquad (1)$$

where P_t is an n-dimensional predictor known at time t with ith element P_{it}, Y_t is the variable to be forecast, and the error ε_t has variance σ^2. It is assumed that Y_t and P_t have sample mean zero. (Extensions to multistep forecasting and including lagged values of Y are discussed below.) For the theoretical development, it does not matter how the regressors are constructed; in our applications and in the recent empirical econometric literature, they are constructed as the first n principal components, dynamic principal components, or a variant of these methods, using an original, potentially larger set of regressors, $\{X_t\}$.

When n is large, there are many regressors and OLS will work poorly. Therefore, we consider forecasting methods that impose and exploit additional structure on the coefficients in Equation (1). We show that all these methods have a shrinkage representation, by which we mean that the forecasts from these methods can all be written as

$$\tilde{Y}_{T+1|T} = \sum_{i=1}^{n} \psi(\kappa t_i) \hat{\delta}_i P_{iT} + o_p(1), \qquad (2)$$

where $\tilde{Y}_{T+1|T}$ is the forecast of Y_{T+1} made using data through time T, $\hat{\delta}_i = T^{-1} \sum_{t=1}^{T} P_{it-1} Y_t$ is the OLS estimator of δ_i (the ith element of δ), $t_j = \sqrt{T}\hat{\delta}_i / s_e$, where $s_e^2 = \sum_{t=1}^{T} (Y_t - \hat{\delta}' P_{t-1})^2 / (T - n)$, and ψ is a function specific to the forecasting method. We consider four classes of forecasting procedures: PT and information criterion methods, Bayesian methods (including BMA), EB, and BG. The factor κ depends on the method. For PT methods and BG, $\kappa = 1$. For the Bayes methods, $\kappa = (s_e / \hat{\sigma})$, where $1/\hat{\sigma}^2$ is the Bayes method's posterior mean of $1/\sigma^2$. This factor arises because the posterior for σ may not concentrate around s_e^2.

Under general conditions, for Bayes, EB, BG, and PT estimators, $0 \leq \psi(x) \leq 1$, so the operational effect of these methods is to produce linear combinations in which the weights are the OLS estimator, shrunk toward zero by the factor ψ. This is the reason for referring to Equation (2) as the shrinkage representation of these forecasting methods.

A key feature of these results is that the proof that the remainder term in Equation (2) is $o_p(1)$ for the different methods relies on much weaker assumptions on the true distribution of (Y, P) than the modeling assumptions used to derive the methods. As a result, these methods can be applied and their performance understood even if they are applied in circumstances in which the original modeling assumptions clearly do not hold, for example, when they are applied to multistep-ahead forecasting.

2.1 Pretest (PT) and Information Criterion Methods

Because the regressors are orthogonal, a hard threshold PT for model selection in (2) corresponds to including those regressors with t statistics exceeding some threshold c. For the PT method, the estimator of the ith coefficient, $\tilde{\delta}_i^{\mathrm{PT}}$, is the OLS estimator, if $|t_i| > c$, and is zero otherwise, that is,

$$\tilde{\delta}_i^{\mathrm{PT}} = 1(|t_i| > c)\hat{\delta}_i. \qquad (3)$$

Expressed in terms of (2), the PT ψ function is

$$\psi^{\mathrm{PT}}(u) = 1(|u| > c). \qquad (4)$$

Under some additional conditions, the PT methods correspond to information criteria methods, at least asymptotically. For example, consider Akaike Information Criterion (AIC) applied sequentially to the sequence of models constructed by sorting the regressors by the decreasing magnitude of their t statistics. If n is fixed, then AIC selection is asymptotically equivalent to the PT selector (4) with $c = \sqrt{2}$.

2.2 Normal Bayes (NB) Methods

For tractability, Bayes methods in the linear model have focused almost exclusively on the case of strictly exogenous regressors and independently distributed homoscedastic (typically

normal) errors. For our purposes, the leading case in which these assumptions are used is the BMA methods discussed in the next section. This modeling assumption is

$$\text{(M1)}\ \{\varepsilon_t\} \perp \{P_t\} \quad \text{and} \quad \varepsilon_t \text{ is iid } N\ (0,\sigma^2).$$

We also adopt the usual modeling assumption of squared error loss. Bayes procedures constructed under assumption (M1) with squared error loss will be called "NB" procedures. Note that we treat (M1) as a modeling tool, where the model is in general misspecified, that is, the true probability law for the data, or data generating process (DGP), is not assumed to satisfy (M1).

Suppose that the prior distribution specifies that the coefficients $\{\delta_i\}$ are iid, that the prior distribution on δ_i given σ^2 can be written in terms of $\tau_i = \sqrt{T}\delta_i/\sigma$, and that $\{\tau_i\}$ and σ^2 have independent prior distributions, respectively, G_τ and G_{σ^2} (where G denotes a generic prior):

$$\text{(M2)}\ \{\tau_i = \sqrt{T}\delta_i/\sigma\} \sim \text{iid } G_\tau, \sigma^2 \sim G_{\sigma^2} \quad \text{and}$$
$$\{\tau_i\} \text{ and } \sigma^2 \text{ are independent.}$$

Under squared error loss, the NB estimator $\tilde{\delta}_i^{\text{NB}}$ is the posterior mean,

$$\tilde{\delta}_i^{\text{NB}} = E_{\delta,\sigma^2}(\delta_i \mid Y, P), \tag{5}$$

where the subscript E_{δ,σ^2} indicates that the expectation is taken with respect to δ (which reduces to δ_i by independence under (M2)) and σ^2. Under (M1), $(\hat{\delta}, s_e^2)$ are sufficient for (δ, σ^2). Moreover, $\hat{\delta}_i$ and $\hat{\delta}_j$ are independently distributed for all $i \neq j$ conditional on (δ, σ^2), and $\hat{\delta}_i|\delta, \sigma^2$ is distributed for $N(\delta_i, \sigma^2/T)$. Thus, (M1) and (M2) imply that, conditional on σ^2, the posterior mean has the so-called simple Bayes form (Maritz and Lwin 1989)

$$\tilde{\delta}_i^{\text{NB}} \mid \sigma^2 = \hat{\delta}_i + \frac{\sigma^2}{T}\ell_\delta(\hat{\delta}_i), \tag{6}$$

where $\ell_\delta(x) = d\ln(m_\delta(x))/dx$, where $m_\delta(x) = \int \phi_{\sigma/\sqrt{T}}(x - \delta)dG_{\delta|\sigma^2}(\delta|\sigma^2)$ is the marginal distribution of an element of $\hat{\delta}$, $G_{\delta|\sigma^2}$ is the conditional prior of an element of δ given σ^2, and ϕ_ω is the pdf of a $N(0,\omega^2)$ random variable.

The shrinkage representation of the NB estimator follows from (6) by performing the change of variables $\tau_i = \sqrt{T}\delta_i/\sigma$. For priors satisfying (M2) and under conditions made precise below, the shrinkage function for the NB estimator is

$$\psi^{\text{NB}}(u) = 1 + \ell(u)/u, \tag{7}$$

where $\ell(u) = d\ln m(u)/du$, $m(u) = \int \phi(u - \tau)dG_\tau(\tau)$, and ϕ is the standard normal density. Integrating over the posterior distribution of σ^2 results in the posterior mean approaching its probability limit, which leads to ψ^{NB} being evaluated at $u = t_i \times \text{plim}(\sigma/\hat{\sigma})$.

It is shown in the online supplementary material that, if the prior density $g_\tau = dG_\tau(u)/du$ is symmetric around zero and is unimodal, then for all u,

$$\psi^{\text{NB}}(u) = \psi^{\text{NB}}(-u) \quad \text{and} \quad 0 \leq \psi^{\text{NB}}(u) \leq 1. \tag{8}$$

2.3 Bayesian Model Averaging (BMA)

Our treatment of BMA with orthogonal regressors follows Clyde, Desimone, and Parmigiani (1996), Clyde (1999a, b), and Koop and Potter (2004). The Clyde, Desimone, and Parmigiani (1996) BMA setup adopts (M1) and a Bernoulli prior model for variable inclusion with a g-prior (Zellner 1986) for δ conditional on inclusion. Specifically, with probability p let $\delta_i|\sigma \sim N(0, \sigma^2/(gT))$ (so $\tau_i \sim N(0, 1/g)$), and with probability $1-p$ let $\delta_i = 0$ (so $\tau_i = 0$). Note that this prior model satisfies (M2). Direct calculations show that, under these priors, the shrinkage representation (7) specializes to

$$\psi^{\text{BMA}}(u) = \frac{pb(g)\phi(b(g)u)}{(1+g)[pb(g)\phi(b(g)u) + (1-p)\phi(u)]}, \tag{9}$$

where $b(g) = \sqrt{g/(1+g)}$ and ϕ is the standard normal density, and where ψ^{BMA} is evaluated at $u = \kappa t_i$, just as in the general case (7). The Bernoulli/normal BMA prior is symmetric and unimodal, so ψ^{BMA} satisfies Equation (8).

2.4 Empirical Bayes (EB)

EB estimation treats the prior G as an unknown distribution to be estimated. Under the stated assumptions, $\{\hat{\delta}_i\}$ constitute n iid draws from the marginal distribution m, which in turn depends on the prior G. Because the conditional distribution of $\hat{\delta}|\delta$ is known under (M1), this permits inference about G. In turn, the estimator of G can be used in Equation (6) to compute the EB estimator. The estimation of the prior can be done either parametrically or nonparametrically. We refer to the resulting EB estimator generically as $\tilde{\delta}_i^{\text{EB}}$. The shrinkage function for the EB estimator is

$$\psi^{\text{EB}}(u) = 1 + \hat{\ell}(u)/u, \tag{10}$$

where $\hat{\ell}(u)$ is the estimate of the score of the marginal distribution of $\{t_i\}$. This score can be estimated directly or can be computed alternatively using an estimated prior $\hat{G}_\tau$, in which case $\hat{\ell}(t) = d\ln\hat{m}(t)/dt$, where $\hat{m}(t) = \int \phi(t - \tau)d\hat{G}_\tau(\tau)$.

2.5 Bagging (BG)

Bootstrap aggregation or "BG" (Breiman 1996) smooths the hard threshold in PT estimators by averaging over a bootstrap sample of PT estimators. Inoue and Kilian (2008) applied BG to a forecasting situation like that considered in this article and reported some promising results; also see Lee and Yang (2006). Bühlmann and Yu (2002) considered BG with a fixed number of strictly exogenous regressors and iid errors, and showed that asymptotically the BG estimator can be represented in the form (2), where (for $u \neq 0$),

$$\psi^{\text{BG}}(u) = 1 - \Phi(u + c) + \Phi(u - c) + t^{-1}[\phi(u - c) - \phi(u + c)], \tag{11}$$

where c is the PT critical value, ϕ is the standard normal density, and Φ is the standard normal cdf. We consider a variant of BG in which the bootstrap step is conducted using a parametric bootstrap under the exogeneity-normality assumption (M1). This algorithm delivers the Bühlmann-Yu (2002) expression (11) under weaker assumptions on the number and properties of the regressors than in Bühlmann and Yu (2002).

2.6 Theoretical Results

We now turn to a formal statement of the validity of the shrinkage representations of the foregoing forecasting methods.

Let P_T denote a vector of predictors used to construct the forecast and let $\{\tilde{\delta}_i\}$ denote the estimator of the coefficients for the method at hand. Then, each method produces forecasts of the form $\tilde{Y}_{T+1|T} = \sum_{i=1}^{p} \tilde{\delta}_i P_{iT}$, with shrinkage approximation $\hat{Y}_{T+1|T} = \sum_{i=1}^{p} \psi(\kappa t_i)\hat{\delta}_i P_{iT}$ for appropriately chosen $\psi(.)$. It follows immediately from the definition of the PT estimator that its shrinkage representation is $\tilde{Y}^{\mathrm{PT}}_{T+1|T} = \sum_{i=1}^{n} \psi^{\mathrm{PT}}(t_i)\hat{\delta}_i P_{iT}$, where $\psi^{\mathrm{PT}}(u) = 1(|u| > c)$ is exact. This section shows that $\tilde{Y}_{T+1|T} - \hat{Y}_{T+1|T} \xrightarrow{m.s.} 0$ for the NB and BG forecasts.

First, consider the NB forecast described in Section 2.2. If σ^2 were known, then Equation (7) implies that the shrinkage representation would hold exactly with $\kappa = s_e/\sigma$. The difference $\tilde{Y}^{\mathrm{NB}}_{T+1|T} - \hat{Y}^{\mathrm{NB}}_{T+1/T}$ is therefore associated with the estimation of σ^2. The properties of the sampling error associated with the estimation of σ^2 depend on the DGP and the modeling assumptions (likelihood and prior) underlying the construction of the Bayes forecast. Assumptions associated with the DGP and Bayes procedures are provided below. Several of these assumptions use the variable $\zeta = \hat{\sigma}^2/\sigma^2$, where $1/\hat{\sigma}^2$ is the posterior mean of $1/\sigma^2$. The assumptions use the expectation operator E, which denotes expectation with respect to the true distribution of Y and P, and E^M, which denotes expectation with respect to the Bayes posterior distribution under the modeling assumptions (M1) and (M2).

The assumptions for the NB forecasts are as follows:

(A1) $\max_i |P_{iT}| \le P_{\max}$, a finite constant.
(A2) $E(T^{-1}\sum_t Y_t^2)^2 \sim O(1)$.
(A3) $n/T \to \upsilon$, where $0 \le \upsilon < 1$.
(A4) $E\{E^M[(\zeta-1)^4 \mid Y,P]\}^4 \sim O(T^{-4-\delta})$ for some $\delta > 0$.
(A5) $E\{E^M[\zeta^{-4} \mid Y,P]\}^4 \sim O(1)$.
(A6) $\sup_u |u^m d^m \psi^{\mathrm{NB}}(u)/du^m| \le M$ for $m = 1, 2$.

Assumptions (A1) and (A2) are restrictions on the DGP, while (A3) is the asymptotic nesting. Assumptions (A4) and (A5) involve both the DGP and the assumed model for the Bayes forecast, and these assumptions concern the rate at which the posterior for σ concentrates around $\hat{\sigma}$. To interpret these assumptions, consider the usual Normal-Gamma conjugate prior (i.e., $\tau_i \sim \mathrm{N}(0,g^{-1})$ and $1/\sigma^2 \sim$ Gamma). A straightforward calculation shows that $E^M[(\zeta-1)^4 \mid Y,P] = 12(\nu+2)/\nu^3$ and $E^M[\zeta^{-4} \mid Y,P] = (\nu/2)^4/[(\nu/2-1)(\nu/2-2)(\nu/2-3)(\nu/2-4)]$, where ν denotes the posterior degrees of freedom. Because $\nu = O(T)$ under (A3), $E\{E^M[(\zeta-1)^4 \mid Y,P]\}^4 \sim O(T^{-8})$, and $E[E^M[\zeta^{-4} \mid Y,P]]^4 \sim O(1)$, so that assumptions (A4) and (A5) are satisfied in this case regardless of the DGP. Assumption (A6) rules out priors that induce mass points in ψ^{NB} or for which $\psi^{\mathrm{NB}}(u)$ approaches 1 very slowly as $u \to \infty$.

With these assumptions, the behavior of $\tilde{Y}^{\mathrm{NB}}_{T+1/T} - \hat{Y}^{\mathrm{NB}}_{T+1/T}$ is characterized in the following theorem:

Theorem 1. Under (A1)–(A6), $\tilde{Y}^{\mathrm{NB}}_{T+1/T} - \hat{Y}^{\mathrm{NB}}_{T+1/T} \xrightarrow{m.s.} 0$.

Proofs are given in the online supplementary material.

An analogous result holds for the BG forecast. To prove this result, we make two additional assumptions:

(A7) $n/B \to 0$.
(A8) $\max_i E(t_i^{12}) < \infty$.

In (A7), B denotes the number of bootstrap replications, and the finite 12th moment assumption in (A8) simplifies the proof of the following theorem:

Theorem 2. Under (A1)–(A3) and (A7)–(A8), $\tilde{Y}^{\mathrm{BG}}_{T+1/T} - \hat{Y}^{\mathrm{BG}}_{T+1/T} \xrightarrow{m.s.} 0$.

Remarks

1. The theorems show that shrinkage factor representations hold under weaker assumptions than those upon which the estimators are derived: the shrinkage factor representations are consequences of the algorithm, not properties of the DGP.
2. Consider the (frequentist) mean square forecast error (MSE) risk of an estimator $\tilde{\delta}$, $R(\tilde{\delta},\delta) = E(\tilde{\delta}-\delta)'(\tilde{\delta}-\delta)$, which is motivated by interest in the prediction problem with orthonormal regressors. Setting $\tilde{\delta}_i = \psi(\kappa t_i)\hat{\delta}_i$, this risk is $E(\tilde{\delta}-\delta)'(\tilde{\delta}-\delta) = \upsilon n^{-1}\sum_{i=1}^{n} E(\psi(\kappa t_i)\sqrt{T}\hat{\delta}_i - \sqrt{T}\delta_i)^2$. Suppose that $\{\sqrt{T}(\hat{\delta}_i - \delta_i)/\sigma\}$ are identically distributed, $i = 1,\ldots,n$, and let $r_\psi(\tau_i) = E(\psi(\kappa t_i)\sqrt{T}\hat{\delta}_i/\sigma - \tau_i)^2$, where $\tau_i = \sqrt{T}\delta_i/\sigma$. Then, $R(\tilde{\delta},\delta) = \upsilon\sigma^2 \int r_\psi(\tau) d\tilde{G}_n(\tau)$, where $\tilde{G}_n$ is the empirical cdf of $\{\tau_t\}$. Thus, the risk depends only on ψ, $\tilde{G}_n$, and the sampling distribution of $(\{\sqrt{T}(\hat{\delta}_i - \delta_i)/\sigma\}, \hat{\sigma}^2/\sigma^2)$. Holding constant this sampling distribution, risk rankings of various estimators depend only on $\tilde{G}_n$. If $\sqrt{T}(\hat{\delta}_i - \delta_i)/\sigma$ is asymptotically normally distributed, then the optimal choice of ψ is ψ^{NB}, with prior distribution equal to (the limit of) G_n (for details see Knox, Stock, and Watson 2004). These considerations provide a justification for thinking that parametric EB estimators will perform well even though the model assumption (M1) used to derive the parametric Bayes estimator does not hold in the time series context of interest here.
3. For EB estimators, the shrinkage function depends on the estimated prior. Under suitable regularity conditions, if the EB estimation step is consistent, then the asymptotic EB shrinkage representation ψ^{EB} is ψ^{NB} with the probability limit of the estimated prior replacing G_τ.
4. These representations permit the extension of these methods to direct multistep forecasting. In a multistep setting, the errors have a moving average structure. However, the forecasting methods can be implemented by substituting heteroscedasticity and autocorrelation consistent (HAC) t statistics into the shrinkage representations.
5. The shrinkage representation of BG allows us to obtain a condition that, if satisfied, implies that BG is asymptotically admissible; this result appears to be unavailable elsewhere. Setting ψ^{BG} equal to ψ^{NB} yields the integral-differential equation,

$$\left.\frac{d\ln\int\phi(z-s)dG_\tau(s)}{dz}\right|_{z=u} = u[\Phi(u-c)-\Phi(u+c)] + \phi(u-c) - \phi(u+c), \quad (12)$$

where both sides are treated as functions of u. If there is a proper prior G_τ that satisfies (12), then this is the prior for which BG is asymptotically Bayes, in which case BG would

be asymptotically admissible. Let G_τ have density g and characteristic function $\tilde{g}(s) = \int e^{ist} g(t)dt$. Then, g satisfies (12), if $\tilde{g}$ satisfies the Fredholm equation of the second kind, $\tilde{g}(s) = \int K(s,t)\tilde{g}(t)dt$, where

$$K(s,t) = 2\frac{e^{-t^2+st}}{s}\left[\frac{\sin(c(s-t))}{(s-t)^2} - c\frac{\cos(c(s-t))}{s-t}\right]. \quad (13)$$

6. Tibshirani (1996, Section 2.2) provides a soft-thresholding or shrinkage representation for the Lasso estimator with orthonormal regressors, derived for strictly exogenous regressors.

3. EMPIRICAL ANALYSIS: DATA AND METHODS

The empirical analysis examines whether the shrinkage methods improve upon DFM forecasts that use only the first few principal components.

3.1 The Data

The dataset consists of quarterly observations on 143 U.S. macroeconomic time series from 1960:II through 2008:IV, for a total of 195 quarterly observations, with earlier observations used for lagged values of regressors as necessary. We have grouped the series into 13 categories that are listed in Table 1. The series are transformed by taking logarithms and/or differencing. In general, first differences of logarithms (growth rates) are used for real quantity variables, first differences are used for nominal interest rates, and second differences of logarithms (changes in rates of inflation) for price series. Let Y^h_{t+h} denote the variable to be forecasted in a h-period ahead forecast. For real activity variables, Y^h_{t+h} is the h-period growth at an annual rate; for interest rates, Y^h_{t+h} is the h-period change; and for nominal price and wage series, Y^h_{t+h} is h-quarter inflation minus current one-quarter inflation (both at annual rates).

Of the 143 series in the dataset, 34 are high-level aggregates that are related by an identity to subaggregates in the dataset. Because including the higher-level aggregates does not add information, only the 109 lower-level disaggregated series were used to compute principal components. All 143 series were used, one at a time, as the dependent variable to be forecasted, using principal components computed from the 109 disaggregates.

The series, their sources, the one- and h-step ahead transformations for each series, and whether the series is one of the 109 series used to estimate the factors are provided in the online supplementary material.

3.2 Methods

This section summarizes the forecasting procedures and the estimation of their parameters and MSE. Estimating the shrinkage parameters by least squares would drive the estimated parameters toward $\psi = 1$, that is, eliminate shrinkage to obtain the least-squares forecast. We therefore instead estimate the shrinkage parameters by minimizing the "leave m out" cross-validation MSE. The performance of the shrinkage methods using the cross-validation estimates of the parameters is then evaluated using two methods: a rolling pseudo out-of-sample forecasting estimate of the MSE and the full-sample cross-validation MSE. The full-sample cross-validation parameters are also used to compare estimated shrinkage functions.

We begin by detailing the forecasting procedures, then describe the computation of the cross-validation MSE and its use in the rolling pseudo out-of-sample forecasting exercise.

Forecasting procedures. We examine six forecasting procedures.

1. *DFM-5*. The DFM-5 forecast uses the first five principal components as predictors, with coefficients estimated by OLS without shrinkage; the remaining principal components are omitted.
2. *PT*. The PT shrinkage function is given by (4) and has one estimated parameter, c.
3. *BG*. The BG shrinkage function is given by (11) and has one estimated parameter, c.
4. *BMA*. The BMA shrinkage function is given by (9) and has two parameters, p and g. Because the parameters are estimated, the BMA method as implemented here is parametric EB.
5. *Logit*. In addition to the methods studied in Section 2, we considered a logit shrinkage function, chosen because it is a conveniently estimated flexible functional form with two

Table 1. Categories of series in the dataset

Group	Brief description	Examples of series	Number of series
1	GDP components	GDP, consumption, investment	16
2	IP	IP, capacity utilization	14
3	Employment	Sectoral and total employment and hours	20
4	Unemployment rate	Unemployment rate, total and by duration	7
5	Housing	Housing starts, total and by region	6
6	Inventories	NAPM inventories, new orders	6
7	Prices	Price indexes, aggregate and disaggregate; commodity prices	37
8	Wages	Average hourly earnings, unit labor cost	6
9	Interest rates	Treasuries, corporate, term spreads, public-private spreads	13
10	Money	M1, M2, business loans, consumer credit	7
11	Exchange rates	Average and selected trading partners	5
12	Stock prices	Various stock price indexes	5
13	Consumer expectations	Michigan consumer expectations	1

parameters, β_0 and β_1:

$$\psi^{\text{logit}}(u) = \frac{\exp(\beta_0 + \beta_1 |u|)}{1 + \exp(\beta_0 + \beta_1 |u|)}. \quad (14)$$

6. *OLS*. For comparison purposes, we also report the OLS forecast based on all principal components (so $\psi^{\text{OLS}} = 1$).

Preliminary investigation showed considerable instability in nonparametric EB estimators, perhaps because the number of observations is too small for nonparametrics, so those methods are not pursued here.

MSE estimation by cross-validation. Consider the h-step ahead series to be predicted, Y^h_{t+h}, let X_t denote the vector of 109 transformed time series, and let $\psi(\tau,\theta)$ denote a candidate shrinkage function with parameter vector θ.

Estimation of the parameters θ and δ and of the MSE for that series/horizon/forecasting method proceeds in three steps. The method is laid out for a sample $t = 1, \ldots, T$, which can be either the full-sample period or a subperiod.

1. Autoregressive dynamics are partialed out by initially regressing Y^h_{t+h} and X_t on 1, Y^1_t, Y^1_{t-1}, Y^1_{t-2}, and Y^1_{t-3}; let $\tilde{Y}^{h,cv}_{t+h}$ and $\tilde{X}^{cv}_t$ denote the residuals from these regressions, standardized to have unit variance in the full sample. The principal components P^{cv}_t of $\tilde{X}^{cv}_t$ are computed using observations $t = 1, \ldots, T$ on the 109 series in the dataset that are not higher-level aggregates. The principal components are ordered according to the magnitude of the eigenvalues with which they are associated, and the first n standardized principal components are retained as P^{cv}_t.
2. Let $\Im^{cv}_t = \{1, \ldots, t - 2h - 3, t + 2h + 3, \ldots, T\}$ be the indices of the dataset dropping the tth observation and $2h + 2$ observations on either side. At each date $t = 1, \ldots, T - h$, the OLS estimators of δ are computed by regressing $\tilde{Y}^{h,cv}_{t+h}$ on P^{cv}_t using observations $t \in \Im^{cv}_t$. Denote these estimators as $\hat{\delta}^{h,cv}_{j,t}$, $j = 1, \ldots, n$. Let $\hat{\tau}^{h,cv}_{j,t}$ denote the conventional OLS t statistic corresponding to $\hat{\delta}^{h,cv}_{j,t}$ (not adjusting for heteroscedasticity or serial correlation).
3. The parameter θ is then estimated by minimizing the sum of squared cross-validation prediction errors:

$$\hat{\theta}^h = \arg\min_{\theta} \text{MSE}^{cv}(\theta), \text{ where } \text{MSE}^{cv}(\theta) = \frac{1}{T-h} \sum_{t=1}^{T-h} \left(\tilde{Y}^{h,cv}_{t+h} - \sum_{i=1}^{100} \psi\left(\hat{\tau}^{h,cv}_{i,t}; \theta\right) \hat{\delta}^{h,cv}_{i,t} P^{cv}_{i,t} \right)^2. \quad (15)$$

Because these are direct forecasts, the estimator $\hat{\theta}^h$ differs by forecast horizon. The estimated shrinkage function for this dependent variable and horizon is $\psi(.;\hat{\theta}^h)$. The cross-validation estimate of the MSE is $\text{MSE}^{cv}(\hat{\theta}^h)$.

All regressions involving P (over all sample periods) impose the moment condition that $P'P/\text{rows}(P) = \text{I}$.

MSE estimation by rolling pseudo out-of-sample forecasts. In the rolling calculation, the forecaster, standing at date t, applies the cross-validation algorithm described above to the most recent 100 observations (with $n = 50$ principal components) to estimate θ for a series/horizon/forecasting method, then uses this estimate of θ to forecast Y^h_{t+h}; this is repeated for the 96-h rolling forecast dates $t = 1985{:}\text{I}, \ldots, 2008{:}\text{IV}–h$. This produces a sequence of rolling pseudo out-of-sample forecasts, $\hat{Y}^{h,\text{rolling}}_{t+h|t}$, computed using the rolling window of length $100 - h$. The rolling estimate of the MSE for a candidate forecast is $\text{MSE}^{\text{rolling}} = (96 - h)^{-1} \sum_{t=1985:\text{I}}^{2008:\text{IV}-h} (Y^h_{t+h} - \hat{Y}^{h,\text{rolling}}_{t+h|t})^2$.

4. EMPIRICAL RESULTS

We begin with results for all series combined, then break the results down by category of series. With the exception of Table 4, all results are presented in terms of root mean square error (RMSE) relative to the DFM-5 benchmark, for example, the relative RMSE for the BMA forecast is $(\text{MSE}^{\text{rolling,BMA}}/\text{MSE}^{\text{rolling,DFM-5}})^{1/2}$.

4.1 Results for all Series: Pseudo Out-of-Sample Forecast Relative RMSEs

Table 2 reports percentiles of the distributions of one-, two-, and four-quarter ahead rolling pseudo out-of-sample RMSEs over the 143 series for the seven forecasting methods, where the RMSEs are relative to the DFM-5 benchmark. For one-quarter

Table 2. Distributions of relative RMSE for 1985–2008 by forecasting method, relative to DFM-5, estimated by rolling forecasts, $h = 1$, 2, and 4

	Percentiles				
Method	0.050	0.250	0.500	0.750	0.950
			(a) $h = 1$		
AR(4)	0.918	0.979	1.007	1.041	1.144
OLS	0.968	1.061	1.110	1.179	1.281
DFM-5	1.000	1.000	1.000	1.000	1.000
Pretest	0.966	1.007	1.048	1.091	1.144
Bagging	0.938	0.996	1.022	1.060	1.104
BMA	0.921	0.993	1.014	1.053	1.103
Logit	0.941	0.999	1.027	1.071	1.120
			(b) $h = 2$		
AR(4)	0.889	0.958	0.990	1.025	1.134
OLS	0.963	1.024	1.087	1.135	1.231
DFM-5	1.000	1.000	1.000	1.000	1.000
Pretest	0.957	1.003	1.030	1.082	1.156
Bagging	0.931	0.982	1.011	1.043	1.106
BMA	0.918	0.976	1.009	1.038	1.106
Logit	0.937	0.988	1.019	1.052	1.116
			(c) $h = 4$		
AR(4)	0.879	0.945	0.980	1.020	1.107
OLS	0.942	1.015	1.066	1.113	1.194
DFM-5	1.000	1.000	1.000	1.000	1.000
Pretest	0.934	1.011	1.048	1.084	1.128
Bagging	0.924	0.984	1.016	1.052	1.094
BMA	0.898	0.979	1.014	1.047	1.086
Logit	0.924	0.982	1.022	1.064	1.120

NOTES: Entries are percentiles of distributions of relative RMSEs over the 143 variables being forecasted, by series, at the two- and four-quarter ahead forecast horizon. RMSEs are relative to the DFM-5 forecast RMSE. All forecasts are direct. RMSEs are calculated using rolling pseudo out-of-sample forecasts over 1985–2008 as described in the text.

Table 3. Median relative RMSE, relative to DFM-5, conditional on the forecasting method improving on AR(4), estimated by rolling forecasts, $h = 1$, 2, and 4

Horizon	OLS	DFM-5	Pretest	Bagging	BMA	Logit
$h = 1$	1.087 (13)	1.000 (85)	1.015 (33)	1.007 (45)	1.009 (51)	1.014 (43)
$h = 2$	1.002 (17)	1.000 (59)	1.008 (32)	0.986 (49)	0.989 (46)	0.998 (39)
$h = 4$	1.000 (18)	1.000 (53)	1.012 (29)	1.007 (40)	1.007 (39)	0.997 (36)

NOTES: Entries are the relative RMSE of the column forecasting method, relative to DFM-5, computed for those series for which the column forecasting method has an RMSE less than the AR(4) forecast. The number of such series appears in parentheses below the relative RMSE. RMSEs are calculated using rolling pseudo out-of-sample forecasts over 1985–2008 as described in the text.

ahead forecasts, the DFM-5 model provides modest forecasting improvements over the AR(4). At horizons $h = 2$ and 4, the DFM-5 improves over the AR(4) for fewer than half the series. These results are in line with results in the literature for U.S. data over this Great Moderation period, during which these series experienced reduced volatility (Kim and Nelson 1999; McConnell and Perez-Quiros 2000; Stock and Watson 2002) and reduced predictability (Stock and Watson 2002c; D'Agostino, Giannone, and Surico 2007).

Our primary interest is in whether the use of principal components beyond the first five improves upon conventional low-dimensional factor model forecasts. As expected, OLS with all 50 principal components in the pseudo out-of-sample experiment results in substantially worse performance at all horizons, relative to the DFM-5. More noteworthy is that the shrinkage methods generally do not improve upon the DFM-5 forecasts: at all horizons, at the median, all hard- and soft-threshold shrinkage methods produce larger RMSEs than the DFM-5, and their upside improvement at the 25th and 5th percentile of RMSEs is nearly always less than their downside at the 75th and 95th percentile of RMSEs, respectively. Of the shrinkage methods, BMA dominates the others in Table 2 in the sense that the distribution of RMSEs is to the left of the RMSE distributions for the other methods, a result that holds at all horizons. The one cell of Table 2 that suggests a role for shrinkage over DFM-5 is that for a small number of series at $h = 4$, BMA improves substantially over DFM-5, with fifth percentile of 0.898.

For many series, principal component methods do not improve upon the AR(4), so it is of interest to focus on those series for which principal component methods appear to be useful. Table 3 therefore reports the median rolling pseudo out-of-sample RMSE, relative to DFM-5, conditional on the candidate method improving upon the AR forecast for that series/horizon combination. For nearly all shrinkage methods and horizons, these medians are quite close to 1, indeed they exceed 1 in 8 of the 12 method/horizon combinations. Even for those series for which the shrinkage method outperforms the AR(4), the forecaster is typically better off just using DFM-5.

Tables 4–6 explore the extent to which the shrinkage forecasts differ from each other and from the DFM-5 forecast. Table 4 presents two measures of similarity of the performance of one-step ahead forecasts: the correlation (over series) among the rolling RMSEs, here relative to the AR(4) forecasts, and the mean absolute difference of these relative RMSEs. The shrinkage forecast relative RMSEs tend to be highly correlated among themselves, with correlations in the range 0.897–0.963; however, the correlations between the shrinkage and DFM-5 relative RMSEs are only approximately 0.67. The mean absolute differences between the RMSEs of DFM-5 and each shrinkage forecast (averaged across series) are also substantial.

Tables 5 and 6 provide some summary statistics about the estimated shrinkage functions for the various methods. To reduce sampling variability, these summary statistics are computed for shrinkage parameters estimated over the full 1960–2008 sample (full-sample cross-validation estimates). Because the estimation period is longer than for the rolling subsamples, these shrinkage functions are evaluated using 100 principal components so that $n/T = 0.51$, approximately the same as the value of 0.50 used in the rolling forecasting exercise. Table 5 reports the distribution across series of the root mean square shrinkage function, $(\sum_{i=1}^{100} \psi(\hat{\tau}_{i,t}^{h,cv}; \hat{\theta}_j^{h,cv})^2/100)^{1/2}$, where $\hat{\theta}_j^{h,cv}$ is the full-sample cross-validation estimated parameter for series j for the row method; because $\psi = 1$ for all principal components for OLS,

Table 4. Two measures of similarity of rolling forecast performance, $h = 1$: correlation (lower left) and mean absolute difference (upper right) of forecast relative MSEs, 1985–2008

	OLS	DFM-5	Pretest	Bagging	BMA	Logit
OLS		0.121	0.077	0.093	0.098	0.090
DFM-5	0.353		0.058	0.044	0.040	0.048
Pretest	0.593	0.670		0.030	0.035	0.028
Bagging	0.617	0.690	0.916		0.013	0.020
BMA	0.620	0.670	0.906	0.963		0.022
Logit	0.551	0.663	0.897	0.930	0.915	

NOTES: Entries below the diagonal are the correlation between the rolling pseudo out-of-sample RMSEs for the row/column forecasting methods, compute over the 143 series being forecasted. Entries above the diagonal are the mean absolute difference between the row/column method RMSEs, averaged across series. For this table, RMSEs are computed relative to the AR(4). Forecasts and RMSEs are calculated using rolling pseudo out-of-sample forecasts over 1985–2008 as described in the text.

Table 5. Distribution of root mean square values of shrinkage function ψ, $h = 1$

	Percentiles				
Method	0.050	0.250	0.500	0.750	0.950
OLS	1.000	1.000	1.000	1.000	1.000
DFM-5	0.224	0.224	0.224	0.224	0.224
Pretest	0.000	0.100	0.141	0.300	0.812
Bagging	0.000	0.100	0.151	0.299	0.697
BMA	0.077	0.118	0.183	0.354	0.639
Logit	0.100	0.141	0.222	0.482	0.769

NOTES: Shrinkage function parameters are estimated by full-sample cross-validation.

Table 6. Distribution of fraction of mean-squared variation of ψ placed on the first five principal components among series with root mean square shrinkage functions >0.05, $h = 1$

		Percentiles					
Method	Number	0.050	0.250	0.500	0.750	0.950	Frac > 0.90
OLS	143	0.050	0.050	0.050	0.050	0.050	0.00
DFM-5	143	1.000	1.000	1.000	1.000	1.000	1.00
Pretest	112	0.000	0.121	0.429	1.000	1.000	0.38
Bagging	119	0.030	0.147	0.359	0.737	1.000	0.13
BMA	136	0.050	0.051	0.215	0.921	1.000	0.26
Logit	138	0.022	0.057	0.233	0.667	1.000	0.21

NOTES: Shrinkage function parameters are estimated by full-sample cross-validation. The final column is the fraction of series for which the row method places at least 90% of the mean square weight on the first five principal components.

for OLS this measure is 1.00 for all series. For DFM-5, $\psi = 1$ for the first five principal components and zero otherwise, so this measure is $\sqrt{5/100} = 0.224$ for all series. Table 6 reports the distribution across series of the average fraction of the mean squared variation in the ψ's attributable to the first five principal components, $\sum_{i=1}^{5} \psi(\hat{\tau}_{i,t}^{h,cv}; \hat{\theta}_j^{h,cv})^2 / \sum_{i=1}^{100} \psi(\hat{\tau}_{i,t}^{h,cv}; \hat{\theta}_j^{h,cv})^2$, among those series for which the root mean square shrinkage function considered in Table 5 is at least 0.05. (A model with shrinkage weight equal to 0.5 for one principal component and equal to 0 for the remaining 99 principal components has a root mean square ψ of 0.05.) The final column of Table 6 reports the fraction of these series for which at least 90% of the mean square weight, for the row model, is placed on the first five principal components.

According to Table 5, the median weight ψ for the shrinkage methods is somewhat less than for DFM-5, but these weights differ substantially across series. Table 6 shows that the fraction of mean square weight placed by the shrinkage methods on the first five principal components also varies considerably across series. For approximately one-quarter of the series (26%), BMA places at least 90% of its mean square weight on the first five principal components, but for one-quarter of the series BMA places only 5.1% of its mean square weight on the first five principal components.

4.2 Results for Cross-Validation RMSEs

The pseudo out-of-sample results in Tables 2–4 pertain to the historically special Great Moderation period. Although these results cannot be extended to the full 1960–2008 period because of the need for a startup window for the rolling forecasts, it is possible to compute the MSEs for the full sample using cross-validation. We therefore computed the counterparts of Tables 2–4 using full-sample cross-validation RMSEs; we summarize the main findings here and provide tables of results in the supplementary material. Three features of the full-sample cross-validation MSEs are noteworthy.

First, the performance of the DFM-5 and shrinkage forecasts, relative to the AR(4), is substantially better when the pre-Great Moderation period is included: based on the full-sample cross-validation RMSEs, the DFM-5 outperforms the AR(4) in more than 75% of series at $h = 1$, as opposed to approximately 50% of series for the RMSEs computed over 1985–2008 by either cross-validation or pseudo out-of-sample forecasts. This result is in keeping with results in the literature documenting reduced predictability in the Great Moderation period. Second, the distributions of RMSEs of the shrinkage methods, relative to DFM-5, are quite similar in the full sample and in the 1985–2008 subsample. There is an increase in dispersion of the RMSE distributions in the 1985–2008 period, compared with the full-sample distributions, but this increase in dispersion is consistent with the shorter period having half as many time series observations. This finding of substantial stability in these distributions of MSEs, relative to DFM-5, is rather surprising given the large documented shifts in the time series properties of these series across the pre-85 and post-85 samples. Third, the overall pattern of correlations in Table 4 (high among shrinkage forecasts, smaller between shrinkage and DFM-5, and smaller still between shrinkage and OLS) is similar when the correlations are computed using the full-sample cross-validation RMSEs, although all the correlations are larger.

Taken together, these results suggest that the shrinkage methods seem to offer little or no improvement over DFM-5, at least on average over all these series. The median cross-validation relative RMSEs are somewhat less than 1 and the median rolling RMSEs are somewhat greater than 1. It is plausible to think that these two estimates bracket the true RMSE: the cross-validation estimates are biased down because they do not include an adjustment for the estimation of θ, while the rolling estimates arguably understate performance because the rolling estimation using 100 observations increases estimation uncertainty for the shrinkage parameters, relative to estimation over the full sample. This bracketing argument suggests that, for this full sample of series, the typical relative RMSE of a shrinkage method to the DFM-5 is quite close to 1 at all horizons considered.

4.3 Results by Category of Series

Table 7 breaks down the results of Table 2 by the 13 categories of series in Table 1. Generally speaking, the categories fall into three groups. The first group consists of series for which the DFM-5 forecasts have the lowest, or nearly the lowest, category-wise median relative RMSE compared with the shrinkage methods, and for which the DFM-5 improves upon the AR(4) benchmark even in the 1985–2008 period. Series in this first group include the major measures of real economic activity (gross domestic product (GDP) components, industrial production (IP), employment, and unemployment rates) and

Table 7. Median RMSE by forecasting method and by category of series, relative to DFM-5, rolling forecast estimates

Category	Brief description	AR(4)	OLS	DFM-5	Pretest	Bagging	BMA	Logit
			(a) $h = 1$					
1	GDP components	1.029	1.107	1.000	1.081	1.046	1.032	1.045
2	IP	1.028	1.182	1.000	1.056	1.031	1.031	1.024
3	Employment	1.022	1.164	1.000	1.068	1.048	1.033	1.062
4	Unemployment rate	1.138	1.142	1.000	1.048	0.989	0.995	1.029
5	Housing	0.973	0.968	1.000	0.973	0.960	0.965	0.995
6	Inventories	1.020	1.126	1.000	0.973	0.938	0.920	0.960
7	Prices	1.000	1.104	1.000	1.033	1.021	1.010	1.026
8	Wages	1.000	1.054	1.000	1.050	1.034	1.008	1.027
9	Interest rates	1.006	1.169	1.000	1.049	1.032	1.020	1.034
10	Money	1.008	1.035	1.000	1.013	0.988	0.997	1.003
11	Exchange rates	0.992	1.105	1.000	1.008	1.002	1.003	1.004
12	Stock prices	0.996	1.049	1.000	1.015	1.010	1.001	1.015
13	Consumer expectations	0.960	1.156	1.000	0.983	0.986	0.985	0.972
	Overall	1.007	1.110	1.000	1.048	1.022	1.014	1.027
			(b) $h = 2$					
1	GDP components	1.009	1.095	1.000	1.050	1.008	1.015	1.024
2	IP	0.990	1.121	1.000	1.061	1.046	1.038	1.053
3	Employment	0.990	1.100	1.000	1.041	1.009	1.011	1.022
4	Unemployment rate	1.235	1.231	1.000	1.018	1.030	1.021	1.009
5	Housing	0.963	0.969	1.000	0.977	0.954	0.965	0.987
6	Inventories	0.938	1.090	1.000	1.012	0.972	0.973	0.961
7	Prices	0.985	1.098	1.000	1.018	0.999	0.992	1.001
8	Wages	0.997	1.062	1.000	1.020	0.994	0.996	1.023
9	Interest rates	0.971	1.057	1.000	1.031	1.019	1.015	1.031
10	Money	0.986	1.010	1.000	1.012	0.997	1.001	0.992
11	Exchange rates	0.987	1.087	1.000	1.050	1.020	1.009	1.023
12	Stock prices	0.996	0.980	1.000	0.978	0.979	0.980	0.981
13	Consumer expectations	0.976	1.111	1.000	1.010	0.994	0.993	1.019
	Overall	0.99	1.087	1.000	1.030	1.011	1.009	1.019
			(c) $h = 4$					
1	GDP components	0.981	1.045	1.000	1.081	1.039	1.033	1.031
2	IP	0.953	1.068	1.000	1.054	1.011	1.008	1.010
3	Employment	0.978	1.048	1.000	1.052	1.014	1.022	1.030
4	Unemployment rate	1.218	1.194	1.000	1.073	1.049	1.056	1.044
5	Housing	0.965	0.972	1.000	1.019	0.990	0.992	1.008
6	Inventories	0.932	1.060	1.000	1.046	1.014	1.021	1.074
7	Prices	0.973	1.100	1.000	1.035	0.998	0.991	0.993
8	Wages	0.975	1.051	1.000	1.037	1.038	1.008	1.029
9	Interest rates	1.010	1.071	1.000	1.052	1.028	1.023	1.035
10	Money	0.985	1.012	1.000	1.011	1.004	0.998	0.998
11	Exchange rates	0.974	1.072	1.000	1.027	1.043	1.018	1.008
12	Stock prices	0.967	0.957	1.000	0.968	0.967	0.958	0.978
13	Consumer expectations	1.001	1.078	1.000	1.114	1.083	1.086	1.043
	Overall	0.980	1.066	1.000	1.048	1.016	1.014	1.022

NOTES: Entries are median RMSEs, relative to DFM-5, for the row category of series. Relative RMSEs are computed as in Table 2.

interest rates. For series in this group, typically the fraction of the mean square weight placed by the shrinkage methods on the first five principal components is large (full-sample cross-validation weights; detailed results are provided in the supplementary material). For these series, the DFM-5 outperforms the AR(4), the shrinkage methods are essentially approximating the DFM-5 model, and the DFM-5 works as well as or better than the shrinkage approximations to it.

Figure 1 presents estimated shrinkage functions for a series in this first group, total employment, at $h = 1$, computed using the full-sample parameter estimates. The upper panel presents the estimated shrinkage functions, and the lower panel plots the weight placed by the various shrinkage functions on each of the 100 ordered principal components. At $h = 1$, the AR(4) rolling RMSE, relative to DFM-5, is 1.098, while the shrinkage estimate rolling RMSEs, relative to DFM-5, range from 1.027 to 1.115; the corresponding full-sample cross-validation relative RMSEs are 1.174 for AR(4) and 1.021–1.044 for the shrinkage methods. All the estimated shrinkage functions are similar, placing substantial weight only on t statistics in excess of approximately 3.2,

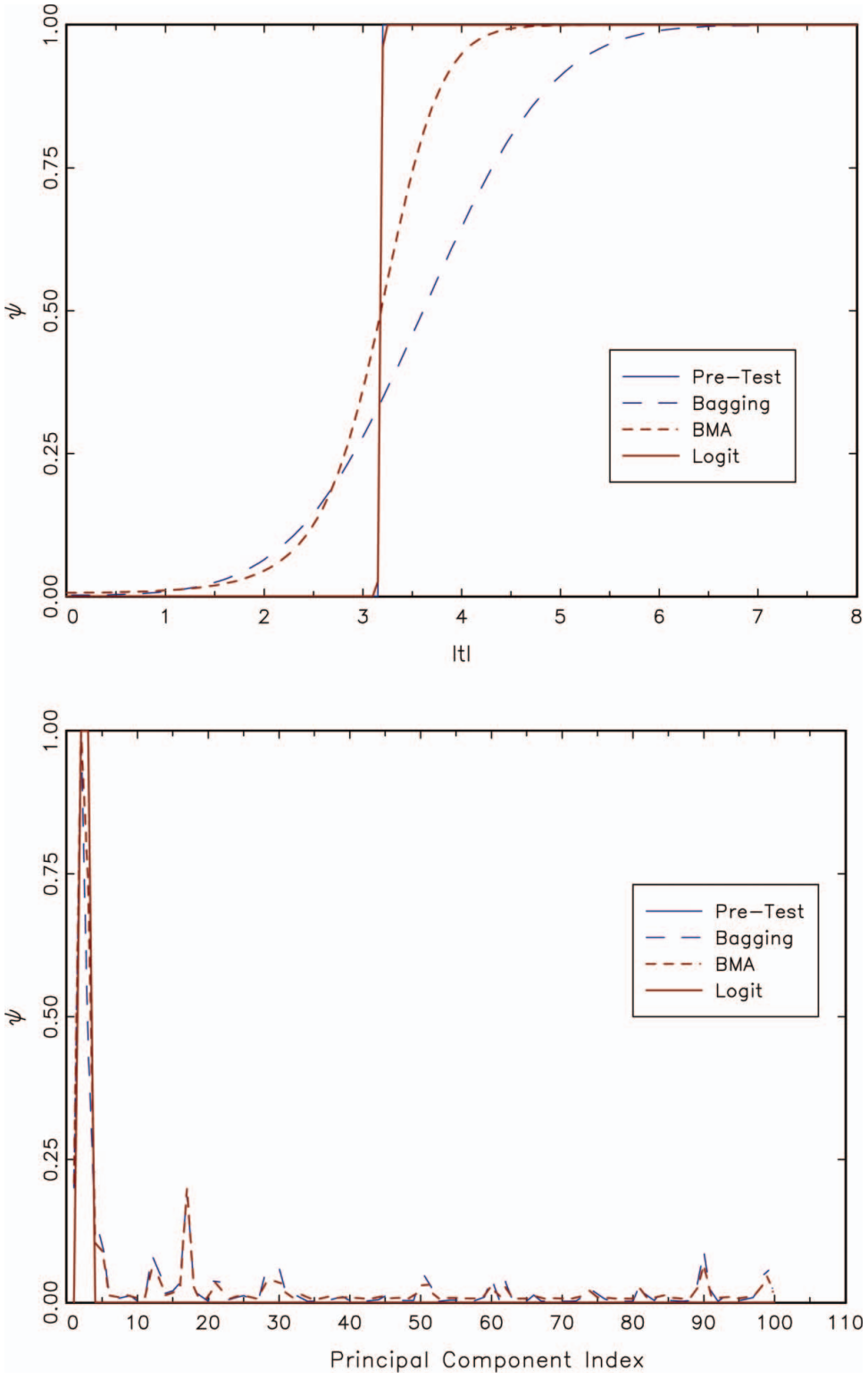

Figure 1. Estimated shrinkage functions (upper panel) and weights $\psi(t_i,\hat{\theta})$ on ordered principal components 1–100: total employment growth, $h = 1$. The online version of this figure is in color.

and the estimated logit and PT shrinkage functions are nearly identical. The shrinkage functions end up placing nearly all the weight on the first few principal components, and only a few higher principal components receive weight exceeding 0.1. For total employment, the shrinkage methods support the DFM-5 restrictions, and relaxing those restrictions increases the RMSE.

There is some evidence of a second, smaller group of series for which one or more shrinkage forecast improves on both the AR and DFM-5 forecasts, but that evidence is delicate and mixed over horizons, among series within categories, and over cross-validation versus rolling RMSEs. Series in this group include real wages and some housing variables. For example, for real

wages in goods producing industries, the median full-sample cross-validation RMSE, relative to DFM-5, is between 0.916 and 0.934 for all four shrinkage methods at the two-quarter horizon, whereas the corresponding relative RMSE for AR(4) is 0.980. These improvements for real wages by shrinkage methods are not found, however, using the rolling RMSEs or in the post-1985 cross-validation subsample.

The final group consists of hard-to-forecast series for which the principal components do not provide meaningful reductions in either rolling or cross-validation RMSEs, relative to AR, using either the DFM-5 or shrinkage forecasts. This group includes price inflation, exchange rates, stock returns, and consumer expectations. The shrinkage parameter objective function (15) is quite flat for many of these series. Figure 2 presents estimated

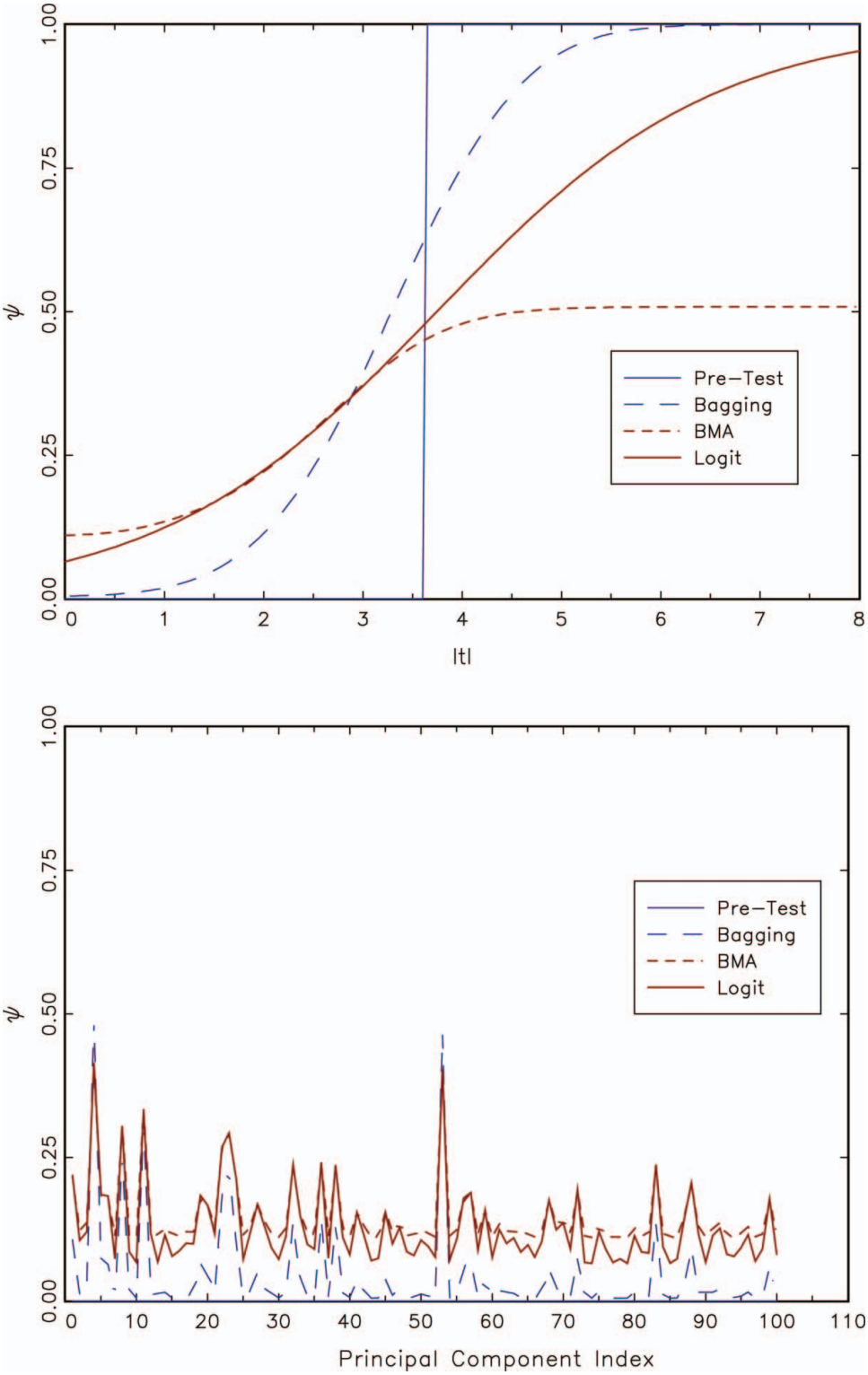

Figure 2. Estimated shrinkage functions (upper panel) and weights $\psi(t_i,\hat{\theta})$ on ordered principal components 1–100: percentage change of S&P 500 Index, $h = 1$. The online version of this figure is in color.

shrinkage functions and weights for a series in this third group, the percentage change in the S&P 500 Index. For all but the PT forecast, most shrinkage methods place a weight of 0.1–0.2 on most of the principal components. For the S&P 500, the rolling RMSE of AR(4), relative to DFM-5, is 1.006 at $h = 1$, and for the shrinkage methods the relative RMSEs range from 1.019 to 1.033; the corresponding full-sample cross-validation RMSEs, relative to DFM-5, are 1.011 for AR(4) and, for shrinkage methods, from 1.005 to 1.011.

4.4 Additional Results and Sensitivity Checks

We also estimated by cross-validation a logit model with a quadratic term to obtain a more flexible parametric specification. The shrinkage function for the quadratic logit model is

$$\psi^{\text{logit-}q}(u) = \frac{\exp(\beta_0 + \beta_1 |u| + \beta_2 u^2)}{1 + \exp(\beta_0 + \beta_1 |u| + \beta_2 u^2)}. \qquad (16)$$

The cross-validation fit of (16) is only marginally better than the linear logit model (14), which we interpret as yielding no meaningful improvement after accounting for the estimation of additional parameter in the quadratic logit.

We also repeated the analysis using Newey-West (1987) standard errors (with a window width of $h + 1$), instead of the homoscedasticity-only OLS standard errors used above, including reestimating (by full-sample cross-validation) the shrinkage parameters using the Newey–West t statistics. There were no substantial changes in the findings discussed above.

5. DISCUSSION

Two points should be borne in mind when interpreting the empirical results. First, we have focused on whether the DFM provides a good framework for macro forecasting. This focus is related to, but different than, asking whether the DFM with a small number of factors explains most of the variation in macro time series; for a discussion of this latter issue, see Giannone, Reichlin, and Sala (2004) and Watson (2004). Second, the DFM forecasting method used here (the first five principal components) was chosen so that it is nested within the shrinkage function framework (2). To the extent that other DFM forecasting methods, such as iterated forecasts based on a high-dimensional state space representation of the DFM (e.g., Doz, Giannone, and Reichlin 2011), improve upon the first five principal components forecasts used here, the results here understate forecasting potential of improved DFM variants.

The facts that some of these shrinkage methods have an interpretation as an EB method and that we have considered a number of flexible functional forms lead us to conclude that it will be difficult to improve systematically upon DFM forecasts using time-invariant linear functions of the principal components of large macro datasets like the one considered here. This conclusion complements Bańbura, Giannone, and Reichlin (2010) and De Mol, Giannone, and Reichlin (2008), who reached a similar conclusion concerning many-predictor models specified in terms of the original variables instead of the factors. This suggests that further forecast improvements over those presented here will need to come from models with nonlinearities and/or time variation, and work in this direction has already begun (e.g., Del Negro and Otrok 2008; Banerjee, Marcellino, and Masten 2009; Stock and Watson 2009; Stevanović 2010a, b).

SUPPLEMENTARY MATERIALS

The supplementary material contains additional analytical results, including proofs of theorems, the definitions of variables used in the empirical analysis, and additional empirical results.

ACKNOWLEDGMENTS

The authors thank Jean Boivin, Domenico Giannone, Lutz Kilian, Serena Ng, Lucrezia Reichlin, Mark Steel, and Jonathan Wright for helpful discussions, Anna Mikusheva for research assistance, and the referees for helpful suggestions. An earlier version of the theoretical results in this article was circulated earlier under the title "An Empirical Comparison of Methods for Forecasting Using Many Predictors." Replication files for the results in this article can be downloaded from *http://www.princeton.edu/∼mwatson*. This research was funded in part by NSF grants SBR-0214131 and SBR-0617811.

[Received October 2009. Revised June 2012.]

REFERENCES

Andersson, M. K., and Karlsson, S. (2008), "Bayesian Forecast Combination for VAR Models," *Advances in Econometrics*, 23, 501–524. [481]

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [481]

——— (2006), "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica*, 74, 1133–1150. [481]

——— (2008), "Large Dimensional Factor Models," *Foundations and Trends in Econometrics*, 3, 89–163. [481]

——— (2009), "Boosting Diffusion Indices," *Journal of Applied Econometrics*, 24, 607–629. [481]

Bańbura, M., Giannone, D., and Reichlin, L. (2010), "Large Bayesian Vector Auto Regressions," *Journal of Applied Econometrics*, 25, 71–92. [481,492]

Banerjee, A., Marcellino, M., and Masten, I. (2009), "Forecasting Macroeconomic Variables Using Diffusion Indexes in Short Samples With Structural Change," in *Forecasting in the Presence of Structural Breaks and Model Uncertainty* (*Frontiers of Economics and Globalization*, Vol. 3), eds. H. Beladi and E. Kwan Choi, Bingley, UK: Emerald Group Publishing Limited, pp. 149–194. [492]

Boivin, J., and Giannoni, M. P. (2006), "DSGE Models in a Data-Rich Environment," NBER Working Paper No. WP12772, National Bureau of Economic Research, Inc. [481]

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 36, 105–139. [483]

Bühlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927–961. [483]

Carriero, A., Kapetanios, G., and Marcellino, M. (2011), "Forecasting Large Datasets With Bayesian Reduced Rank Multivariate Models," *Journal of Applied Econometrics*, 26, 736–761. [481]

Clyde, M. (1999a), "Bayesian Model Averaging and Model Search Strategies" (with discussion), in *Bayesian Statistics* (Vol. 6), eds. J. M. Bernardo, A. P. Dawid, J. O. Berger, and A. F. M. Smith, Oxford: Oxford University Press. [483]

——— (1999b), Comment on "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 401–404. [483]

Clyde, M., Desimone, H., and Parmigiani, G. (1996), "Prediction Via Orthogonalized Model Mixing," *Journal of the American Statistical Association*, 91, 1197–1208. [483]

D'Agostino, A., Giannone, D., and Surico, P. (2007), "(Un)Predictability and Macroeconomic Stability," CEPR Discussion Paper 6594, Centre for Economic Policy Research. [487]

Del Negro, M., and Otrok, C. (2008), "Dynamic Factor Models With Time-Varying Parameters: Measuring Changes in International Business Cycles,"

Federal Reserve Bank of New York Staff Reports No. 326, Federal Reserve Bank of New York. [492]

De Mol, C., Giannone, D., and Reichlin, L. (2008), "Forecasting a Large Number of Predictors: Is Bayesian Regression a Valid Alternative to Principal Components?" *Journal of Econometrics*, 146, 318–328. [481,492]

Doz, C., Giannone, D., and Reichlin, L. (2011), "A Quasi Maximum Likelihood Approach for Large Approximate Dynamic Factor Models," *Review of Economics and Statistics*, forthcoming. [492]

Eickmeier, S., and Ziegler, C. (2008), "How Successful are Dynamic Factor Models at Forecasting Output and Inflation? A Meta-Analytic Approach," *Journal of Forecasting*, 27, 237–265. [481]

Eklund, J., and Karlsson, S. (2007), "Forecast Combination and Model Averaging using Predictive Measures," *Econometric Reviews*, 26, 329–363. [481]

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000), "The Generalized Factor Model: Identification and Estimation," *Review of Economics and Statistics*, 82, 540–554. [481]

——— (2004), "The Generalized Factor Model: Consistency and Rates," *Journal of Econometrics*, 119, 231–255. [481]

Geweke, J. (1977), "The Dynamic Factor Analysis of Economic Time Series," in *Latent Variables in Socio-Economic Models*, eds. D. J. Aigner and A. S. Goldberger, Amsterdam: North-Holland. [481]

Giannone, D., Reichlin, L., and Sala, L. (2004), "Monetary Policy in Real Time," *NBER Macroeconomics Annual*, 2004, 161–200. [492]

Inoue, A., and Kilian, L. . (2008), "How Useful Is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation," *Journal of the American Statistical Association*, 103, 511–522. [481,483]

Jacobson, T., and Karlsson, S. (2004), "Finding Good Predictors for Inflation: A Bayesian Model Averaging Approach," *Journal of Forecasting*, 23, 479–496. [481]

Kim, C.-J., and Nelson, C. R. (1999), "Has the U.S. Economy Become More Stable? A Bayesian Approach Based on a Markov-Switching Model of the Business Cycle," *The Review of Economics and Statistics*, 81, 608–616. [487]

Knox, T., Stock, J. H., and Watson, M. W. (2004), "Empirical Bayes Regression With Many Regressors," unpublished manuscript, Harvard University. [484]

Koop, G., and Potter, S. (2004), "Forecasting in Dynamic Factor Models Using Bayesian Model Averaging," *The Econometrics Journal*, 7, 550–565. [481,483]

Korobilis, D. (2008), "Forecasting in VARs With Many Predictors," *Advances in Econometrics*, 23, 403–431. [481]

Lee, T.-H., and Yang, Y. (2006), "Bagging Binary and Quantile Predictors for Time Series," *Journal of Econometrics*, 135, 465–497. [483]

Maritz, J. S., and Lwin, T. (1989), *Empirical Bayes Methods* (2nd ed.), London: Chapman and Hall. [483]

McConnell, M. M., and Perez-Quiros, G. (2000), "Output Fluctuations in the United States: What has Changed Since the Early 1980's," *American Economic Review*, 90, 1464–1476. [487]

Newey, W. K., and West, K. D. (1987), "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708. [492]

Sargent, T. J. (1989), "Two Models of Measurements and the Investment Accelerator," *Journal of Political Economy*, 97, 251–287. [481]

Stevanović, D. (2010a), "Factor Time Varying Parameter Models," unpublished manuscript, University of Montreal. [492]

——— (2010b), "Common Sources of Parameter Instability in Macroeconomic Models: A Factor-TVP Approach," unpublished manuscript, University of Montreal. [492]

Stock, J. H., and Watson, M. W. (1999), "Forecasting Inflation," *Journal of Monetary Economics*, 44, 293–335. [481]

——— (2002a), "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179. [481,487]

——— (2002b), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20, 147–162. [481]

——— (2002c), "Has the Business Cycle Changed and Why?" *NBER Macroeconomics Annual*, 2002, 159–218. [487]

——— (2009), "Forecasting in Dynamic Factor Models Subject to Structural Instability," in *The Methodology and Practice of Econometrics: Festschrift in Honor of D.F. Hendry* (chap. 7), eds. N. Shephard and J. Castle, Oxford: Oxford University Press. [492]

——— (2011), "Dynamic Factor Models," in *Oxford Handbook of Economic Forecasting*, eds. Michael P. Clements and David F. Hendry,Oxford: Oxford University Press. [481]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [485]

Watson, M. W. (2004), Discussion of "Monetary Policy in Real Time," *NBER Macroeconomics Annual*, 2004, 216–221. [492]

Wright, J. H. (2009), "Forecasting U.S. Inflation by Bayesian Model Averaging," *Journal of Forecasting*, 28, 131–144. [481]

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, eds. P. K. Goel and A. Zellner, Amsterdam: North-Holland, pp. 233–243. [483]