Supplementary materials for this article are available online. Please go to http://tandfonline.com/r/JBES

Optimal Forecasts from Markov Switching Models

Tom Boot
University of Groningen, Department of Economics, Econometrics and Finance, 9747 AE Groningen,
The Netherlands (*t.boot@rug.nl*)

Andreas Pick
Econometric Institute, Erasmus University Rotterdam, 3000 DR Rotterdam, The Netherlands; Tinbergen Institute,
1082 MS Amsterdam, The Netherlands; De Nederlandsche Bank, 1017 ZN Amsterdam, The Netherlands
(*andreas.pick@cantab.net*)

We derive forecasts for Markov switching models that are optimal in the mean square forecast error (MSFE) sense by means of weighting observations. We provide analytic expressions of the weights conditional on the Markov states and conditional on state probabilities. This allows us to study the effect of uncertainty around states on forecasts. It emerges that, even in large samples, forecasting performance increases substantially when the construction of optimal weights takes uncertainty around states into account. Performance of the optimal weights is shown through simulations and an application to U.S. GNP, where using optimal weights leads to significant reductions in MSFE. Supplementary materials for this article are available online.

KEY WORDS: Forecasting; GNP forecasting; Markov switching models; Optimal weights.

1. INTRODUCTION

Markov switching models have long been recognized to suffer from a discrepancy between in-sample and out-of-sample performance. In-sample analysis of Markov switching models often leads to appealing results, for example, the identification of business cycles. Out-of-sample performance, in contrast, is frequently inferior to simple benchmark models for standard loss functions. Examples include forecasting exchange rates by Engel (1994), Dacco and Satchell (1999), and Klaassen (2005), forecasting US GNP growth by Clements and Krolzig (1998) and Perez-Quiros and Timmermann (2001), forecasting US unemployment by Deschamps (2008), and forecasting house prices by Crawford and Fratantoni (2003). Additionally, Guidolin (2011) and Rapach and Zhou (2013) provided reviews of the use of Markov switching models in finance.

In this article, we derive minimum mean square forecast error (MSFE) forecasts for Markov switching models by means of optimal weighting schemes for observations. We provide simple, analytic expressions for the weights when the model has an arbitrary number of states and exogenous regressors. We find that forecasts using optimal weights substantially increase forecast precision and, in our application, are more precise than linear alternatives. Additionally, optimal weights lead to insights that help explain why standard Markov switching forecasts are often less precise than linear forecasts.

We start our discussion assuming that the states of the Markov switching model are known and, in a second step, we relax this assumption. When conditioning on the states, the intuition for the optimal weights can easily be seen: a forecast obtained from optimal weights pools all observations and places different weights on observations from different states. This reduces the variance of the forecast but introduces a bias. Optimally weighting all observations ensures that the trade-off is optimal in the MSFE sense. The usual Markov switching forecasts, in contrast, assign nonzero weights only to observations from the state that will govern the forecast period. Conditional on the states of the Markov switching model, the weights mirror those obtained by Pesaran, Pick, and Pranovich (2013), emphasizing a correspondence with the structural break model. The weights depend on the number of observations per regime and the relative differences of the parameter between the regimes.

In the case of three regimes, the weights have interesting properties. For some parameter values, optimal weighting corresponds to equal weighting of observations. For other parameter values, observations from the state prevailing in the forecast period will not be most heavily weighted. However, conditional on the states of the Markov switching model, the optimal weights can be written as $\mathcal{O}(1/T)$ corrections to the usual Markov switching weights, which implies that, conditional on the states, standard Markov switching weights asymptotically achieve the minimum MSFE.

In practice, the states of the Markov switching model are not known with certainty. We therefore relax the assumption that the states are known and derive weights conditional on state probabilities, which is the information used in standard Markov switching forecasts. This results in optimal weights

© 2018 American Statistical Association
Journal of Business & Economic Statistics
October 2018, Vol. 36, No. 4
DOI: 10.1080/07350015.2016.1219264
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jbes.

that no longer correspond to those for the structural break model. Contrasting weights conditional on states with those conditional on state probabilities yield insights into the effect that uncertainty around states has on forecasts. Our findings explain the deterioration of forecast accuracy of the optimal weights in the application of Pesaran, Pick, and Pranovich (2013) because plug-in estimates of the break date substantially shrink optimal weights toward equal weights. Weights conditional on states and the weights implicit in standard Markov switching forecasts downplay the Markov switching nature of the data when estimates of states are plugged in. Weights conditional on state probabilities, in contrast, retain the emphasis on the Markov switching nature of the data. This implies that the forecast accuracy from optimal weights conditional on state probabilities relative to that implied by standard Markov switching forecasts increases in the difference between the states in terms of their parameters and in the variance of the smoothed probabilities. The forecast improvements from using optimal weights do not vanish as the sample size increases as the standard weights and the optimal weights conditional on the state probabilities are not asymptotically equivalent.

We perform Monte Carlo experiments to evaluate the performance of the optimal weights. The results confirm the theoretically expected improvements. The weights that are derived conditional on the states and use the estimated probabilities as plug-in values improve over standard forecasts only for small differences in parameters, which are unlikely to lead to applications of Markov switching models in practice. The weights based on state probabilities, in contrast, produce substantial gains for large differences in parameters between states, uncertainty over the states, and large samples. These settings are likely to be found in many applications, including the one in this article.

We apply our methodology to forecasting quarterly U.S. GNP. Out-of-sample forecasts are constructed for 124 quarters and a range of Markov switching models. At each point, forecasts are made with the Markov switching model that has the best forecasting history using standard weights. With this model, we calculate forecasts based on the standard Markov switching weights and the optimal weights developed in this article. The results suggest that the forecasts using optimal weights significantly outperform the standard Markov switching forecast. We also find that our forecasting schemes lead to improved forecasts compared to a range of linear alternatives. We analyze the sensitivity of the results to the choice of the out-of-sample forecast evaluation period using the tests of Rossi and Inoue (2012), which confirm our findings.

The outline of the article is as follows. Section 2 introduces the model and the standard forecast. In Section 3, we derive the optimal weights for a simple location model and in Section 4 for a model with exogenous regressors. Monte Carlo experiments are presented in Section 5 and an application to U.S. GNP in Section 6. Finally, Section 7 concludes the article. Additional details are presented in a web appendix.

2. MARKOV SWITCHING MODELS AND THEIR FORECASTS

Consider the following m-state Markov switching model

$$y_t = (\mathbf{B}'\mathbf{s}_t)'\mathbf{x}_t + \boldsymbol{\sigma}'\mathbf{s}_t\varepsilon_t, \quad \varepsilon_t \sim \text{iid}(0, 1), \tag{1}$$

where $\mathbf{B} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \dots, \boldsymbol{\beta}_m')'$ is an $m \times k$ matrix, $\boldsymbol{\beta}_i$ is a $k \times 1$ parameter vector for $i = 1, 2, \dots, m$, $\mathbf{x}_t$ is a $k \times 1$ vector of exogenous regressors, $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \dots, \sigma_m)'$ is an $m \times 1$ vectors of error standard deviations, and $\mathbf{s}_t = (s_{1t}, s_{2t}, \dots, s_{mt})'$ is an $m \times 1$ vector of binary state indicators, such that $s_{it} = 1$ and $s_{jt} = 0,\ j \neq i$, if the process is in state i at time t.

This is the standard Markov switching model introduced by Hamilton (1989). The model is completed by a description of the stochastic process governing the states, where $\mathbf{s}_t$ is assumed to be an ergodic Markov chain with transition probabilities

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{m1} \\ p_{12} & p_{22} & \cdots & p_{m2} \\ \vdots & \vdots & & \vdots \\ p_{1m} & p_{2m} & \cdots & p_{mm} \end{bmatrix},$$

where $p_{ij} = \mathrm{P}(s_{jt} = 1 | s_{i,t-1} = 1)$ is the transition probability from state i to state j.

The standard forecast, in this context, would be to estimate $\boldsymbol{\beta}_i$, $i = 1, 2, \dots, m$, as

$$\hat{\boldsymbol{\beta}}_i = \left(\sum_{t=1}^{T} \hat{\xi}_{it}\mathbf{x}_t\mathbf{x}_t'\right)^{-1} \sum_{t=1}^{T} \hat{\xi}_{it}\mathbf{x}_t y_t, \tag{2}$$

where $\hat{\xi}_{it}$ is the estimated probability that observation at time t is from state i using, for example, the smoothing algorithm of Kim (1994). The forecast is then constructed as $\hat{y}_{T+1} = \sum_{i=1}^{m} \hat{\xi}_{i,T+1}\mathbf{x}_{T+1}'\hat{\boldsymbol{\beta}}_i$, where $\hat{\xi}_{i,T+1}$ is the predicted probability of state i in the forecast period, and $\mathbf{x}_{T+1}$ is the vector of regressors in the forecast period, which we assume known at time T. See Hamilton (1994) for an introduction to the Markov switching modeling and forecasting.

In this article, we derive the minimum MSFE forecast for finite samples and different assumptions about the information set that the forecast is based on. We replace the estimated probabilities by general weights w_t for the forecast $\hat{y}_{T+1} = \mathbf{x}_{T+1}'\hat{\boldsymbol{\beta}}(\mathbf{w})$, so that

$$\hat{\boldsymbol{\beta}}(\mathbf{w}) = \left(\sum_{t=1}^{T} w_t\mathbf{x}_t\mathbf{x}_t'\right)^{-1} \sum_{t=1}^{T} w_t\mathbf{x}_t y_t$$

subject to the restriction $\sum_{t=1}^{T} w_t = 1$. The weights are restricted to sum to one as an identifying restriction is required, and we will see in the next section that this is the restriction of the standard Markov switching weights. We do, however, not restrict the weights to be positive. In fact, in Section 3.1.2 we will see that negative weights are a common feature in models with more than two states as they allow the cancellation of biases. The resulting forecasts are then optimal in the sense that the weights will be chosen such that they minimize the expected MSFE.

3. OPTIMAL FORECASTS FOR A SIMPLE MODEL

Initially, consider a simple version of model (1) with $k = 1$ and $x_t = 1$ such that

$$y_t = \boldsymbol{\beta}'\mathbf{s}_t + \boldsymbol{\sigma}'\mathbf{s}_t\varepsilon_t, \quad \varepsilon_t \sim \text{iid}(0, 1), \tag{3}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_m)'$. We use this simple model for ease of exposition but will return to the full model (1) in Section 4.

We can derive the optimal forecast by using a weighted average of the observations with weights that minimize the MSFE. The forecast from weighted observations for (3) is

$$\hat{y}_{T+1} = \sum_{t=1}^{T} w_t y_t \tag{4}$$

subject to $\sum_{t=1}^{T} w_t = 1$.

Note, that the standard forecast can be expressed as (4) with weights

$$w_{\text{MS},t} = \sum_{i=1}^{M} \frac{\hat{\xi}_{i,T+1}\hat{\xi}_{it}}{\sum_{t=1}^{T} \hat{\xi}_{it}}, \tag{5}$$

which only depend on the smoothed and predicted probabilities and have the property that $\sum_{t=1}^{T} w_{\text{MS},t} = 1$. We will call weights (5) the standard Markov switching weights.

To derive the optimal weights, consider the forecast error, which, without loss of generality, is scaled by the error standard deviation of regime m, is

$$\begin{aligned}\sigma_m^{-1} e_{T+1} &= \sigma_m^{-1}(y_{T+1} - \hat{y}_{T+1}) \\ &= \boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1} + \mathbf{q}'\mathbf{s}_{T+1}\varepsilon_{T+1} - \sum_{t=1}^{T} w_t \boldsymbol{\lambda}'\tilde{\mathbf{s}}_t - \sum_{t=1}^{T} w_t \mathbf{q}'\mathbf{s}_t\varepsilon_t,\end{aligned}$$

where

$$\boldsymbol{\lambda} = \begin{pmatrix} (\beta_2 - \beta_1)/\sigma_m \\ (\beta_3 - \beta_1)/\sigma_m \\ \vdots \\ (\beta_m - \beta_1)/\sigma_m \end{pmatrix}, \ \mathbf{q} = \begin{pmatrix} \sigma_1/\sigma_m \\ \sigma_2/\sigma_m \\ \vdots \\ 1 \end{pmatrix}, \text{ and } \tilde{\mathbf{s}}_t = \begin{pmatrix} s_{2t} \\ s_{3t} \\ \vdots \\ s_{mt} \end{pmatrix}.$$

The scaled MSFE is

$$\begin{aligned}\mathrm{E}\left(\sigma_m^{-2} e_{T+1}^2\right) &= \mathrm{E}\left\{\left[\boldsymbol{\lambda}'\left(\tilde{\mathbf{s}}_{T+1} - \sum_{t=1}^{T} w_t \tilde{\mathbf{s}}_t\right)\right]^2\right\} \\ &\quad + \sum_{t=1}^{T} w_t^2 \mathrm{E}[(\mathbf{q}'\mathbf{s}_t)^2] + \mathrm{E}[(\mathbf{q}'\mathbf{s}_{T+1})^2] \\ &= \mathrm{E}(\tilde{\mathbf{s}}_{T+1}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}) - 2\mathbf{w}'\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}) \\ &\quad + \mathbf{w}'\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}})\mathbf{w} + \mathrm{E}[(\mathbf{q}'\mathbf{s}_{T+1})^2] + \mathbf{w}'\mathrm{E}(\mathbf{Q})\mathbf{w} \\ &= \mathbf{w}'[\mathrm{E}(\mathbf{Q}) + \mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}})]\mathbf{w} - 2\mathbf{w}'\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}) \\ &\quad + \mathrm{E}(\tilde{\mathbf{s}}_{T+1}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}) + \mathrm{E}[(\mathbf{q}'\mathbf{s}_{T+1})^2],\end{aligned} \tag{6}$$

where $\tilde{\mathbf{S}} = (\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_T)$, $\mathbf{S} = (s_1, s_2, \ldots, s_T)$, and $\mathbf{Q}$ is a diagonal matrix with typical (t, t)-element $Q_{tt} = \sum_{i=1}^{m} q_i^2 s_{it}$. The first line of (6) contains the squared bias as the first expression on the right-hand side, the variance of the estimated parameters as the second term, and, finally, the variance of the future disturbance term. The weights will trade off the first and second term on the right-hand side to minimize the MSFE. The last term, in contrast, cannot be reduced.

Furthermore, define

$$\mathbf{M} = \mathrm{E}(\mathbf{Q}) + \mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}) \tag{7}$$

and note that $\mathbf{M}$ is invertible as $\mathbf{Q}$ is a diagonal matrix with positive entries and $\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}})$ is positive semidefinite, so that $\mathbf{M}$ is the sum of a positive definite matrix and a positive semidefinite matrix and therefore itself positive definite.

Minimizing (6) subject to $\sum_{t=1}^{T} w_t = 1$ yields the optimal weights

$$\mathbf{w} = \mathbf{M}^{-1}\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}) + \frac{\mathbf{M}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}}[1 - \boldsymbol{\iota}'\mathbf{M}^{-1}\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1})], \tag{8}$$

where $\boldsymbol{\iota} = (1, 1, \ldots, 1)'$ is a vector of ones of length T. We will discuss the properties of the optimal weights in Sections 3.1 and 3.2 under different assumption about the information set. The MSFE given by (6) when applying the optimal weights (8) is

$$\begin{aligned}\mathrm{MSFE}(\mathbf{w}) &= \frac{[1 - \boldsymbol{\iota}'\mathbf{M}^{-1}\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1})]^2}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}} + \mathrm{E}(\tilde{\mathbf{s}}_{T+1}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}) \\ &\quad - \mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1})'\mathbf{M}^{-1}\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}) \\ &\quad + \mathrm{E}[(\mathbf{q}'\mathbf{s}_{T+1})^2].\end{aligned} \tag{9}$$

To proceed, we need to specify the information set that is available to calculate the expectations in (8) and (9). Initially, we will base the weights on the full information set of the DGP, including the state for each observation. Clearly, this information is not available in practice. However, the resulting analysis will prove to be highly informative. The intuition that is gained will prove useful when interpreting the forecast that we will obtain subsequently when allowing for uncertainty around the states. This second step will enable us to analyze the differences between the plug-in estimator for the weights that assume knowledge of the states and optimal weights that are derived under the assumption that the states are uncertain.

Note, that we condition on $\boldsymbol{\lambda}$ throughout our analysis. The reason is that, in a decomposition of the optimal weights for the structural break case, Pesaran, Pick, and Pranovich (2013) showed that the time of the break enters the weights in a term that is of order $\mathcal{O}(1/T)$, whereas the size of the break, $\boldsymbol{\lambda}$, enters the weights in a term that is of order $\mathcal{O}(1/T^2)$. We will show below that the optimal weights for the Markov switching model conditional on the states are equivalent to the weights of Pesaran, Pick, and Pranovich (2013) and their argument therefore carries over to the Markov switching model.

3.1. Weights Conditional on the States

Conditional on the states the expectation operator in (7), (8), and (9) can be omitted such that $\mathbf{M} = \mathbf{Q} + \tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}$ and $\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}) = \tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}$. Given the number of states, weights can now readily be derived.

3.1.1. Two-State Markov Switching Models. In the case of a two-state Markov switching model, $\tilde{\mathbf{s}} = (s_{21}, s_{22}, \ldots, s_{2T})'$ and therefore $\mathbf{M} = \mathbf{Q} + \lambda^2\tilde{\mathbf{s}}\tilde{\mathbf{s}}'$ for which the inverse is given by

$$\begin{aligned}\mathbf{M}^{-1} &= \mathbf{Q}^{-1} - \frac{\lambda^2}{1 + \lambda^2\tilde{\mathbf{s}}'\mathbf{Q}^{-1}\tilde{\mathbf{s}}}\mathbf{Q}^{-1}\tilde{\mathbf{s}}\tilde{\mathbf{s}}'\mathbf{Q}^{-1} \\ &= \mathbf{Q}^{-1} - \frac{\lambda^2}{1 + \lambda^2 T\pi_2}\tilde{\mathbf{s}}\tilde{\mathbf{s}}',\end{aligned}$$

where $\lambda^2 = \frac{(\beta_2-\beta_1)^2}{\sigma_2^2}$ and $\pi_i = \frac{1}{T}\sum_{t=1}^{T} s_{it}$. The elements of the diagonal matrix $\mathbf{Q}$ are $Q_{tt} = q^2 s_{1t} + s_{2t}$ with $q = \frac{\sigma_1}{\sigma_2}$. This yields the following weights:

When $s_{1,T+1} = 1$,

$$w_{(1,1)} = \frac{1}{T}\frac{1+T\lambda^2\pi_2}{\pi_2 q^2 + \pi_1(1+T\pi_2\lambda^2)} \qquad \text{if } s_{1t} = 1 \tag{10}$$

$$w_{(1,2)} = \frac{1}{T}\frac{q^2}{\pi_2 q^2 + \pi_1(1+T\pi_2\lambda^2)} \qquad \text{if } s_{2t} = 1, \tag{11}$$

where $w_{(i,j)}$ is the weight for an observation when $s_{jt} = 1$ while $s_{i,T+1} = 1$.

When $s_{2,T+1} = 1$,

$$w_{(2,1)} = \frac{1}{T}\frac{1}{[\pi_2 q^2 + \pi_1(1+T\pi_2\lambda^2)]} \qquad \text{if } s_{1t} = 1 \tag{12}$$

$$w_{(2,2)} = \frac{1}{T}\frac{q^2+T\lambda^2\pi_1}{[\pi_2 q^2 + \pi_1(1+T\pi_2\lambda^2)]} \qquad \text{if } s_{2t} = 1. \tag{13}$$

Note that, conditional on the state of the future observation, the weights are symmetric under a relabeling of the states. Derivations are provided in a web appendix.

The weights are equivalent to the weights for the break point process developed by Pesaran, Pick, and Pranovich (2013). This implies that, conditional on the states, a Markov switching model is equivalent to a break point model with known break point with the exception that the observations are ordered by the underlying Markov process.

Since the weights $w_{(1,2)}$ and $w_{(2,1)}$ are nonzero, the decrease in the variance of the optimal weights forecast should outweigh the increase in the squared bias that results from using all observations. The expected MSFE under the above weights is

$$\mathrm{E}\left[\sigma_2^{-2}e_{T+1}^2\right]_{\text{opt}} = \begin{cases} q^2(1+w_{(1,1)}) & \text{if } s_{1,T+1} = 1 \\ 1+w_{(2,2)} & \text{if } s_{2,T+1} = 1 \end{cases} \tag{14}$$

while the expected MSFE for standard Markov switching weights is

$$\mathrm{E}\left[\sigma_2^{-2}e_{T+1}^2\right]_{\text{MS}} = \begin{cases} q^2\left(1+\frac{1}{T\pi_1}\right) & \text{if } s_{1,T+1} = 1 \\ 1+\frac{1}{T\pi_2} & \text{if } s_{2,T+1} = 1. \end{cases} \tag{15}$$

It is easy to show that $\mathrm{E}[\sigma_2^{-2}e_{T+1}^2]_{\text{opt}} < \mathrm{E}[\sigma_2^{-2}e_{T+1}^2]_{\text{MS}}$.

Numerical examples of the magnitude of the improvement in MSFE are presented in Table 1, which shows that the improvements scale inversely with the differences in parameters. To gain intuition for these results, consider the case of $\lambda = 0$ and $q = 1$, that is, the case where the two states have identical means, and $s_{1,T+1} = 1$. The standard Markov switching model will use weights $w_{\text{MS},(1,1)} = \frac{1}{T\pi_1}$ and $w_{\text{MS},(1,2)} = 0$, which results in an MSFE of $1 + \frac{1}{T\pi_1}$. The optimal weights, in contrast, are $= w_{\text{opt},(1,1)} = w_{\text{opt},(1,2)} = 1/T$, and the MSFE is $1 + 1/T$. The usual Markov switching forecast disregards the information from the second state even though, in this case, it is highly informative, whereas the optimal weights forecast uses all the observation equally as one would suggest intuitively, given that the states have the same mean.

Table 1. Ratio between the expected MSFEs of optimal and standard MS weights

	$q = 1$			$q = 0.5$		
λ	$\pi_2 = 0.1$	0.2	0.5	0.1	0.2	0.5
0	0.8500	0.9273	0.9808	0.8500	0.9273	0.9808
0.5	0.9294	0.9758	0.9953	0.9268	0.9745	0.9949
1	0.9727	0.9919	0.9986	0.9724	0.9918	0.9985
2	0.9921	0.9978	0.9996	0.9921	0.9978	0.9996

NOTE: Reported are the ratio between (14) and (15) when $s_{2,T+1} = 1$ for different values of λ, the difference in means, and q, the ratio of standard deviations, and π_2, the proportion of observations in state 2. $T = 50$.

As λ increases, the usefulness of the observations in state 2 decreases because the bias introduced by these observations increases. This is reflected in the numbers in Table 1. The same intuition can be gained by increasing or decreasing q away from 1. The difference in MSFE also depends on π_1, that is, the fraction of observations in the state used for forecasting in the standard Markov switching forecast. The fewer observations are available for the standard Markov switching forecast, the more valuable will the observation from the second state be. Finally, as T increases, for a fixed π_1, the parameter estimates will be more precise so that any further gains from using observation in the second state will be less important. In fact, we show below that, asymptotically, the optimal weights and the standard weights are identical. However, as we will show in Section 3.2, the asymptotic equivalence of optimal and standard weights relies on the fact that the states are known with certainty. With uncertainty around the states, the gain from using optimal weights will not disappear with large T.

3.1.2. Three-State Markov Switching Models. If $s_{j,T+1} = 1$, then define $q_i^2 = \sigma_i^2/\sigma_j^2$ and $\lambda_i^2 = (\beta_i - \beta_j)^2/\sigma_j^2$ where $i, j \in \{1, 2, 3\}$. The optimal weights are

$$\begin{aligned} w_{(j,j)} &= \frac{1}{T}\frac{1+T\sum_{i=1}^{3} q_i^{-2}\lambda_i^2\pi_i}{\sum_{i=1}^{3} q_i^{-2}\pi_i + T\sum_{i=1}^{3}\sum_{m=1}^{3} q_i^{-2}q_m^{-2}\pi_i\pi_m\lambda_m(\lambda_m-\lambda_i)} \\ w_{(j,k)} &= \frac{1}{T}\frac{q_k^{-2}+Tq_k^{-2}\sum_{i=1}^{m} q_i^{-2}\lambda_i\pi_i(\lambda_i-\lambda_k)}{\sum_{i=1}^{3} q_i^{-2}\pi_i + T\sum_{i=1}^{3}\sum_{m=1}^{3} q_i^{-2}q_m^{-2}\pi_i\pi_m\lambda_i(\lambda_i-\lambda_m)} \\ w_{(j,l)} &= \frac{1}{T}\frac{q_l^{-2}+Tq_l^{-2}\sum_{i=1}^{m} q_i^{-2}\lambda_i\pi_i(\lambda_i-\lambda_l)}{\sum_{i=1}^{3} q_i^{-2}\pi_i + T\sum_{i=1}^{3}\sum_{m=1}^{3} q_i^{-2}q_m^{-2}\pi_i\pi_m\lambda_m(\lambda_i-\lambda_m)}, \end{aligned} \tag{16}$$

where $j, k, l \in \{1, 2, 3\}$. Derivations are available in the web appendix.

Figure 1 plots weights (16) for $s_{1,T+1} = 1$, that is, the future observation is known to be from the first state. The difference in mean between the first and second state relative to the variance of the first state is set to $\lambda_2 = -2.5$, and the difference in mean between the first and third state, λ_3, varies from -3 to 3. Furthermore, the proportions of observations for the states are $\pi_1 = 0.2, \pi_2 = \pi_3 = 0.4, T = 100$, and the ratio of variances is $q_1 = q_2 = 1$. Each line represents the weight for one representative observation in each state. As 20 observations are in state 1 and 40 in the other two states, it can easily be verified that the weights sum to one. Consider the weights at $\lambda_3 = -2.5$: each observation in state one is weighted with $w_{(1,1)} \approx 0.05$ and the remaining observations with a weight close to zero. As there are 20 observations in state one, the sum of the weights equals

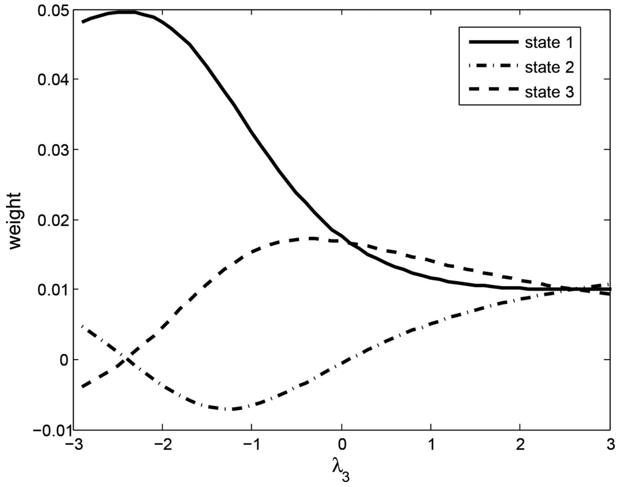

Figure 1. Optimal weights for three-state Markov switching model. The graph depicts the optimal weights (16) for a representative observation in each state when $s_{1,T+1} = 1$, for $\lambda_2 = -2.5$, λ_3 over the range -3 to 3, $T = 100$, $\pi_1 = 0.2$, and $\pi_2 = \pi_3 = 0.4$. The solid line gives the weights for a representative observation where $s_{1t} = 1$, the dash-dotted line a representative observation where $s_{2t} = 1$, and the dashed line a representative observation for $s_{3t} = 1$.

1. Equally, at $\lambda_3 = 2.5$: all observations are equally weighted with a weight of 0.01. As 100 observations are in the sample, the sum of weights equals 1. The standard Markov switching weights are independent of the parameters, $w_{\mathrm{MS},(1,1)} = 0.05$ and $w_{\mathrm{MS},(1,i)} = 0$ for $i \neq 1$, and are not included in Figure 1.

On the left of the graph, where $\lambda_3 = -3$, the observations from state 1 receive nearly all the weight, those from state 2 receive a small positive weight, and those from state 3 a small negative weight. When $\lambda_3 = -2.5$ the weights for $s_{2t} = 1$ and $s_{3t} = 1$ are equal and close to zero. The intuition for the equal weights is that at $\lambda_2 = \lambda_3$ the DGP is essentially a two-state Markov switching model and the observations for the states with equal mean receive the same weight. The large difference between the mean of state 1 and that of the other states induces a potentially large bias when using observations from the other states. As a result, the weights on observations with $s_{2t} = 1$ and $s_{3t} = 1$ are very small.

As λ_3 increases, weights for observations from state 3 increase until, at $\lambda_3 = 0$, they are equal to those for observations with $s_{1t} = 1$. That is, as the third state becomes increasingly similar to the first state, the observations are increasingly useful for forecasting. At $\lambda_3 = 0$, the first and the third state have identical means and the observations therefore receive equal weight. When λ_3 ranges between -2.5 and 0, the weights for the observations from the second state are negative. The intuition is that as the observations from the third state receive an increasingly higher weight, they induce a larger bias, which is in the same direction as the bias due to the observations from the second state. By giving the observations from the second state negative weights, the biases of the observations from the second and third state are of opposite signs and can counteract each other.

As λ_3 increases further and $0 < \lambda_3 < 2.5$, the observations from the third state are weighted heavier than the observations from the first state even though this is the future state. The reason for this at first sight surprising result is that, in this range, the means of observations from state 2 and state 3 have opposite signs. As the bias induced by the observations from the second state is, in absolute terms, larger than that from the third state, the weights on the observations from the third state receive a larger weight to counteract this bias.

At $\lambda_3 = 2.5 = -\lambda_2$, all observations receive the same weight of $\frac{1}{T}$. At this point, the mean of the observations with $s_{1t} = 1$ is between and equally distant to the means of observations with $s_{2t} = 1$ and $s_{3t} = 1$, which implies that with equal weight any biases arising from using observations of the other states cancel. In this case, the optimal weights effectively ignore the Markov switching structure of the model and forecast with equal weights, which is a very different weighting scheme from that suggested by the Markov switching model.

As in the two-state case, when $s_{j,T+1} = 1$ the expected MSFE using the optimal weights is of the form

$$\mathrm{E}\left(\sigma_i^{-2} e_{T+1}^2\right)_{\mathrm{opt}} = \frac{\sigma_j^2}{\sigma_i^2}(1 + w_{(j,j)})$$

with $w_{(j,j)}$ given in (16). For the Markov switching weights, we have

$$\mathrm{E}\left(\sigma_i^{-2} e_{T+1}^2\right)_{\mathrm{MS}} = \frac{\sigma_j^2}{\sigma_i^2}\left(1 + \frac{1}{T\pi_j}\right).$$

Figure 2 displays the ratio of MSFE of the optimal weights relative to that of the standard MSFE forecast for $T = 100, \pi_1 = 0.2, \pi_2 = \pi_3 = 0.4$ for a range of values for λ_2 and λ_3. At $\lambda_2 = \lambda_3 = \pm 3$ the gains from using optimal weights are very small. In this case, the model is essentially a two-state model with a large difference in mean. When λ_2 and λ_3 are of opposite sign, the improvements are the largest. We can therefore expect most gains when the observation to be forecast is in the regime with intermediate location.

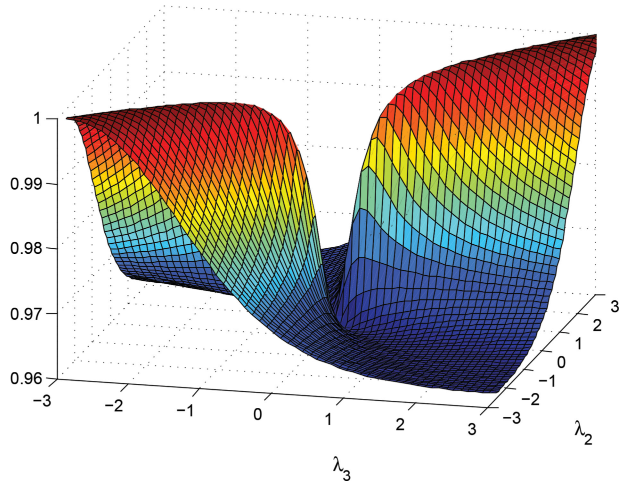

Figure 2. MSFE of optimal weights relative to standard Markov switching weights. The figure displays the ratio of the MSFE of the optimal weights relative to that of the standard MSFE forecast for $T = 100$, $\pi_1 = 0.2$, $\pi_2 = \pi_3 = 0.4$ for a range of values for λ_2 and λ_3.

3.1.3. m-State Markov Switching Models. For $s_{j,T+1} = 1$, we set $\lambda_i = \frac{\beta_i - \beta_j}{\sigma_j}$ and $q_i = \frac{\sigma_i}{\sigma_j}$, which gives for the weights for observations with $s_{l,t} = 1$

$$w_{(j,l)} = \frac{1}{T} \frac{q_l^{-2}(1 + T\sum_{i=1}^{m} q_i^{-2}\lambda_i\pi_i(\lambda_i - \lambda_l))}{\sum_{i=1}^{m} q_i^{-2}\pi_i + T\sum_{i=1}^{m}\sum_{k=1}^{m} q_i^{-2}q_k^{-2}\pi_i\pi_k\lambda_i(\lambda_i - \lambda_k)}. \tag{17}$$

As in the previous cases, the expected MSFE when $s_{j,T+1} = 1$ is

$$\mathrm{E}\left(\sigma_i^{-2}e_{T+1}^2\right)_{\text{opt}} = \frac{\sigma_j^2}{\sigma_i^2}(1 + w_{(j,j)}).$$

The derivation of the weights and the MSFE is in a web appendix. Maximizing the expected MSFE with respect to β_j yields

$$\beta_j = \frac{\sum_{k=1}^{m} q_k^{-2}\pi_k\beta_k}{\sum_{k=1}^{m} q_k^{-2}\pi_k}.$$

Hence, the largest gain occurs when the regime to be forecast is located at the probability and variance weighted average of the other regimes. The minimum MSFE is then

$$\mathrm{E}\left(\sigma_i^{-2}e_{T+1}^2\right) = \frac{1}{\sigma_i^2}\left(\sigma_j^2 + \frac{1}{T}\frac{1}{\sum_{k=1}^{m}\sigma_k^{-2}\pi_k}\right)$$

and when the variances are equal, this reduces to

$$\mathrm{E}\left(\sigma_i^{-2}e_{T+1}^2\right) = 1 + \frac{1}{T}.$$

Thus, the maximum improvement is independent of the number of states when all variances are equal.

3.1.4. Large T Approximation. Interesting results can be obtained when considering the large sample approximation of the two state weights. The optimal weight assigned to an observation is given by

$$\begin{aligned} Tw = {} & s_{1,T+1}\left[\frac{1 + \lambda^2 T\pi_2}{\pi_2 q^2 + \pi_1(1 + \lambda^2 T\pi_2)}s_{1t} \right. \\ & \left. + \frac{q^2}{\pi_2 q^2 + \pi_1(1 + \lambda^2 T\pi_2)}s_{2t}\right] \\ & + s_{2,T+1}\left[\frac{1}{\pi_2 q^2 + \pi_1(1 + \lambda^2 T\pi_2)}s_{1t} \right. \\ & \left. + \frac{q^2 + \lambda^2 T\pi_1}{\pi_2 q^2 + \pi_1(1 + \lambda^2 T\pi_2)}s_{2t}\right]. \end{aligned}$$

We approximate this expression using that $(1 + \frac{\theta}{T})^{-1} = 1 - \frac{\theta}{T} + \mathcal{O}(T^{-2})$, where $\theta = (\pi_2 q^2 + \pi_1)/(\lambda^2\pi_2\pi_1)$. This yields

$$\begin{aligned} Tw = {} & \left(\frac{1}{\pi_1} - \frac{1}{T}\frac{q^2}{\lambda^2\pi_1^2}\right)s_{1t}s_{1,T+1} + \frac{1}{T}\frac{q^2}{\lambda^2\pi_1\pi_2}s_{2t}s_{1,T+1} \\ & + \frac{1}{T}\frac{1}{\lambda^2\pi_1\pi_2}s_{1t}s_{2,T+1} + \left(\frac{1}{\pi_2} - \frac{1}{T}\frac{1}{\lambda^2\pi_2^2}\right)s_{2t}s_{2,T+1} \\ & + \mathcal{O}(T^{-2}). \end{aligned} \tag{18}$$

Hence, the standard Markov switching weights are optimal up to a first-order approximation in T. It is worth noting that this is equivalent to the result obtained by Pesaran, Pick, and Pranovich (2013) for the structural break case where the first-order approximation gives zero weight to prebreak observations and equally weight the post-break observations. This result in (18) also suggests that, in a Markov switching model, accurate estimation of the proportions of the sample in each state is of first-order importance, whereas the differences in means are of second-order importance to obtain a minimal MSFE. This is the motivation for considering the uncertainty around the state estimates, which we turn to now.

3.2. Optimal Weights When States Are Uncertain

We will now contrast the weights conditional on the states with weights that do not assume knowledge of the states. The expectations in (8) can be expressed in terms of the underlying Markov chain. However, it turns out that in this case analytic expressions for the inverse of $\mathbf{M}$ cannot be obtained. In Section 3.3, we will show how numerical values for the inverse can be used to calculate numerical values for the optimal weights.

To analyze the theoretical properties of the optimal weights, analytic expressions for the weights are required, which will allow us to contrast them with the weights that are derived conditional on the states. Such expressions can be obtained by making the simplifying assumption that we can condition on given state probabilities. Estimates of the probabilities are available as output of the estimation of Markov switching models, and this information is also used for the standard forecast from Markov switching models in (2). Note that this is, in fact, more general than the Markov switching model and can accommodate state probabilities from other sources, such as surveys of experts or models other than those considered here.

Denote the probability of state i occurring at time t by ξ_{it}. The expectations in (8) and (9) are then

$$\mathrm{E}(s_{it}s_{j,t+m}) = \begin{cases} \xi_{it} & \text{if } i = j \\ \xi_{it}\xi_{j,t+m} & \text{if } i \neq j, m \geq 0. \end{cases}$$

We will, initially, focus on the two-state case and, subsequently, on m states.

3.2.1. Two-State Markov Switching Models. In a two state model, we have $\tilde{\mathbf{S}} = \mathbf{s}_2 = (s_{21}, s_{22}, \ldots, s_{2T})'$. The matrix $\mathbf{M}$ in (8) is given by

$$\begin{aligned} \mathbf{M} &= \lambda^2\boldsymbol{\xi}\boldsymbol{\xi}' + \lambda^2\mathbf{V} + q^2\mathbf{I} + (1 - q^2)\boldsymbol{\Xi} \\ &= \lambda^2\boldsymbol{\xi}\boldsymbol{\xi}' + \mathbf{D} \end{aligned}$$

with $\boldsymbol{\xi} = (\xi_{21}, \xi_{22}, \ldots, \xi_{2T})$, $\boldsymbol{\Xi} = \mathrm{diag}(\boldsymbol{\xi})$, $\mathbf{V} = \boldsymbol{\Xi}(\mathbf{I} - \boldsymbol{\Xi})$, and $\mathbf{D} = \lambda^2\mathbf{V} + q^2\mathbf{I} + (1 - q^2)\boldsymbol{\Xi}$ and again $q = \sigma_1/\sigma_2$. The inverse of $\mathbf{M}$ is

$$\mathbf{M}^{-1} = \mathbf{D}^{-1} - \frac{\lambda^2}{1 + \lambda^2\boldsymbol{\xi}'\mathbf{D}^{-1}\boldsymbol{\xi}}\mathbf{D}^{-1}\boldsymbol{\xi}\boldsymbol{\xi}'\mathbf{D}^{-1}. \tag{19}$$

Using (8) and (19) yields

$$\mathbf{w} = \lambda^2\xi_{2,T+1}\mathbf{M}^{-1}\boldsymbol{\xi} + \frac{\mathbf{M}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}}\left(1 - \lambda^2\xi_{2,T+1}\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\xi}\right). \tag{20}$$

Denote the typical (t, t)-element of $\mathbf{D}^{-1}$ by d_t, where

$$d_t = \left[\lambda^2\xi_{2,t}(1 - \xi_{2,t}) + q^2 + (1 - q^2)\xi_{2,t}\right]^{-1}.$$

Then, the weight for the observation at time t is given by

$$w_t = \frac{d_t[1 + \lambda^2 \sum_{t'=1}^{T} d_{t'}(\xi_{2t} - \xi_{2t'})(\xi_{2,T+1} - \xi_{2t'})]}{\sum_{t'=1}^{T} d_{t'} + \lambda^2[(\sum_{t'=1}^{T} d_{t'}\xi_{2t'}^2)(\sum_{t'=1}^{T} d_{t'}) - (\sum_{t'=1}^{T} d_{t'}\xi_{2t'})^2]}. \tag{21}$$

The expected MSFE can be calculated from (6) and reduces to

$$\mathrm{E}\left(\sigma_2^{-2}e_{T+1}^2\right) = \left[1 + \lambda^2\xi_{2,T+1}(1 - \xi_{2,T+1})\right](1 + w_{T+1}), \tag{22}$$

where w_{T+1} is given by (21).

When T is large, weights (21) can be written as

$$w_t = \tilde{d}_t \frac{\sum_{t'=1}^{T} \tilde{d}_{t'}(\xi_{2,T+1} - \xi_{2t'})(\xi_{2t} - \xi_{2t'})}{\sum_{t'=1}^{T} \tilde{d}_{t'}(\xi_{2t'} - \sum_{t''=1}^{T} \tilde{d}_{t''}\xi_{2t''})^2} + \mathcal{O}(T^{-2}), \tag{23}$$

where $\tilde{d}_t = d_t/(\sum_{t'=1}^{T} d_t)$. Derivations are provided in a web appendix. While the weights in (21) and (23) provide closed-form solutions, interpretation can be aided by momentarily making the simplifying assumption of constant state variances.

Constant state variance. The interpretation of (21) and (23) is complicated by the fact that ξ_{2t} is a continuous variable in the range [0, 1]—as opposed to the binary variable s_{2t} for the weights conditional on states—so that an infinite number of possible combinations of ξ_{2t} over t is possible. To simplify the interpretation of the weights, we will therefore, for a moment, assume that the variance of the states is constant and denoted as $\sigma_s^2 = \xi_{2t}(1 - \xi_{2t})$.

Summing σ_s^2 over t and solving for σ_s^2 yields

$$\sigma_s^2 = \bar{\xi}_1\bar{\xi}_2 - \frac{1}{T}\sum_t(\xi_{2t} - \bar{\xi}_2)^2, \tag{24}$$

where $\bar{\xi}_1 = \frac{1}{T}\sum_{t=1}^{T}\xi_{1t}$ and $\bar{\xi}_2 = \frac{1}{T}\sum_{t=1}^{T}\xi_{2t}$. Note that the maximum value of σ_s^2 is given by $\bar{\xi}_2\bar{\xi}_1$, which occurs when the probability vector is constant. In the case of a constant σ_s^2, $\tilde{d}_t$ simplifies to $1/T$. Hence, (21) can be written as

$$w_t = \frac{1}{T}\left[1 + \lambda^2 \frac{(\xi_{2,T+1} - \bar{\xi}_2)(\xi_{2t} - \bar{\xi}_2)}{(T\bar{d})^{-1} + \lambda^2(\bar{\xi}_1\bar{\xi}_2 - \sigma_s^2)}\right]$$

and the large T approximation (23) as

$$w_t = \frac{1}{T} + \frac{(\xi_{2,T+1} - \bar{\xi}_2)(\xi_{2t} - \bar{\xi}_2)}{T(\bar{\xi}_1\bar{\xi}_2 - \sigma_s^2)}. \tag{25}$$

The standard Markov switching weights can be expressed as

$$w_{\mathrm{MS},t} = \frac{1}{T} + \frac{(\xi_{2,T+1} - \bar{\xi}_2)(\xi_{2t} - \bar{\xi}_2)}{T\bar{\xi}_1\bar{\xi}_2}, \tag{26}$$

see the web appendix. From a comparison of (25) and (26), it is clear that the two weights differ by the factor σ_s^2 in the denominator and that this difference will not disappear asymptotically. Effectively, the Markov switching weights are more conservative as the optimal weights exploit the regime switching structure more strongly because of the smaller denominator in (25) compared to (26).

The MSFE for the optimal weights and for the standard Markov switching weights under constant state variance are

Table 2. Maximum improvements in a two-state model with $T = 100$

		$\bar{\xi}_2$				
	$\tilde{\sigma}_s^2$	0.1	0.2	0.3	0.4	0.5
$\lambda = 2$	0	1.000	1.000	1.000	1.000	1.000
	0.2	0.993	0.986	0.981	0.979	0.978
	0.4	0.977	0.960	0.950	0.944	0.942
	0.6	0.967	0.946	0.934	0.927	0.926
	0.8	0.974	0.957	0.948	0.944	0.942
$\lambda = 3$	0	1.000	1.000	1.000	1.000	1.000
	0.2	0.982	0.969	0.962	0.958	0.957
	0.4	0.951	0.926	0.913	0.907	0.905
	0.6	0.935	0.908	0.895	0.889	0.887
	0.8	0.949	0.930	0.921	0.917	0.916

NOTE: The table reports the ratio of the MSFE of the optimal weights to that of the Markov switching weights conditional on a constant state variance σ_s^2. $\lambda = (\beta_2 - \beta_1)/\sigma$ denotes the scaled difference between means, $\bar{\xi}_2$ the average probability for state 2, and $\tilde{\sigma}_s^2$ is a negative function of the variance of the state 2 probability.

$$\mathrm{E}\left(\sigma_2^{-2}e_{T+1}^2\right)_{\mathrm{opt}} = [1 + \lambda^2\xi_{2,T+1}(1 - \xi_{2,T+1})] \times\left(1 + \frac{1}{T} + \frac{\lambda^2(\xi_{2,T+1} - \bar{\xi}_2)^2}{1 + \lambda^2\sigma_s^2 + \lambda^2 T(\bar{\xi}_2(1 - \bar{\xi}_2) - \sigma_s^2)}\right) \tag{27}$$

$$\begin{aligned}\mathrm{E}\left(\sigma_2^{-2}e_{T+1}^2\right)_{\mathrm{MS}} &= 1 + \lambda^2\xi_{2,T+1}(1 - \xi_{2,T+1}) + \frac{1}{T}\left(\lambda^2\sigma_s^2 + 1\right) \\ &\quad + \left(\frac{\xi_{2,T+1} - \bar{\xi}_2}{\bar{\xi}_2(1 - \bar{\xi}_2)}\right)^2\left[\frac{1}{T}\left(\bar{\xi}_2(1 - \bar{\xi}_2) - \sigma_s^2\right)\left(\lambda^2\sigma_s^2 + 1\right) + \lambda^2\sigma_s^4\right].\end{aligned} \tag{28}$$

The MSFE for the optimal weights is derived from (22) by substituting in the weights in (21) and using the fact that $\tilde{d}_t = 1/T$ and $d_t = d$, for $t = 1, 2, \ldots, T + 1$, which together with the MSFE for the standard Markov switching weights is derived in a web appendix.

Table 2 displays the improvements in forecast performance expressed as the ratio of (27) over (28) for different values of $\bar{\xi}_2$, $\tilde{\sigma}_s^2 = \sigma_s^2/(\bar{\xi}_2\bar{\xi}_1)$, and λ for $T = 100$. The results indicate that the optimal weights lead to larger gains when λ is large and when $\bar{\xi}_2$ is closer to 0.5. The influence of σ_s^2 is U-shaped with the largest improvement when $\sigma_s^2 = 0.6$. The results in Table 2 show that the improvement can be as large as 11.3% for the range of parameter values considered here.

In this simplified framework, the increase in forecast accuracy does not disappear when the sample size increases. The asymptotic approximation to the MSFE under optimal weights is given by

$$\mathrm{E}\left(\sigma_0^2 e_{T+1}^2\right)_{\mathrm{opt}} = 1 + \lambda^2\xi_{2,T+1}(1 - \xi_{2,T+1}) + \mathcal{O}(T^{-1}) \tag{29}$$

and that under standard Markov switching weights is

$$\mathrm{E}\left(\sigma_0^2 e_{T+1}^2\right)_{\mathrm{MS}} = 1 + \lambda^2\xi_{2,T+1}(1 - \xi_{2,T+1}) + \left(\frac{\xi_{2,T+1} - \bar{\xi}_2}{\bar{\xi}_2\bar{\xi}_1}\right)^2\lambda^2\sigma_s^4 + \mathcal{O}(T^{-1}). \tag{30}$$

The difference between (30) and (29) is positive and does not disappear asymptotically. The relative improvement is expected to be high when λ, σ_s^2, and the difference $\xi_{2,T+1} - \bar{\xi}_2$ are large.

3.2.2. m-State Markov Switching Models. The derivations can be extended to an arbitrary number of states. Note that $\mathbf{M} = \mathrm{E}(\mathbf{Q}) + \mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}})$ and $\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}) = \mathrm{E}(\tilde{\mathbf{S}})'\boldsymbol{\lambda}\boldsymbol{\lambda}'\mathrm{E}(\tilde{\mathbf{S}}) + \mathbf{A}$, where, conditional on the state probabilities, ξ_{jt}, $j = 1, 2, \ldots, m$,

$$\mathbf{A} = \sum_{j=2}^{m} \lambda_j^2 \boldsymbol{\Xi}_j - \left(\sum_{j=2}^{m} \lambda_j \boldsymbol{\Xi}_j\right)^2$$

and $\boldsymbol{\Xi}_j$ is a $T \times T$ diagonal matrix with typical element ξ_{jt}. Define $\tilde{\boldsymbol{\xi}} = \mathrm{E}(\tilde{\mathbf{S}})'\boldsymbol{\lambda}$, which is a $T \times 1$ vector, and $\mathbf{D} = \mathrm{E}(\mathbf{Q}) + \mathbf{A}$. Then,

$$\mathbf{M}^{-1} = \mathbf{D}^{-1} - \frac{1}{1 + \tilde{\boldsymbol{\xi}}\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}}\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}.$$

We can use (8) to derive the weights similar to the case of the two-state weights

$$w_t = \frac{d_t^{(m)}\{1 + [\sum_{t'=1}^{T} d_{t'}^{(m)}(\tilde{\xi}_t - \tilde{\xi}_{t'})(\tilde{\xi}_{T+1} - \tilde{\xi}_{t'})]\}}{\sum_{t'=1}^{T} d_{t'}^{(m)} + (\sum_{t'=1}^{T} d_{t'}^{(m)}\tilde{\xi}_{t'}^2)(\sum_{t'=1}^{T} d_{t'}^{(m)}) - (\sum_{t'=1}^{T} d_{t'}^{(m)}\tilde{\xi}_{t'})^2}, \tag{31}$$

where

$$d_t^{(m)} = \left[\sum_{j=1}^{m} q_j^2 \xi_{jt} + \sum_{j=2}^{m} \lambda_j^2 \xi_{jt} - \left(\sum_{j=2}^{m} \lambda_j \xi_{jt}\right)^2\right]^{-1}$$

$$= \left[\sum_{j=1}^{m} \left(q_j^2 + \lambda_j^2\right)\xi_{jt} - \left(\sum_{j=2}^{m} \lambda_j \xi_{jt}\right)^2\right]^{-1}$$

$$\tilde{\xi}_t = \sum_{j=2}^{m} \xi_{jt}\lambda_j$$

given that $\lambda_1 = 0$.

Examples of weights for a three-state Markov switching model when states are uncertain are plotted in Figure 3. Again, the difference in mean between the first and second state relative to the variance of the first state is set to $\lambda_2 = -2.5$, and the difference in mean between the first and third state, λ_3, varies from -3 to 3. Furthermore, $\pi_1 = 0.2$, $\pi_2 = \pi_3 = 0.4$, $T = 100$, and the ratio of variances is $q_1 = q_2 = 1$. For simplicity of exposition, we assume that the state probabilities are identical for each state in the sense that a prevailing state is given probability $\xi_{it} = 0.8$ and other states $\xi_{jt} = 0.1$. The light gray lines represent the optimal weights (16) that are conditional on the states. The graph on the left plots weights (16) substituting the probabilities ξ_{it} for the states s_{it}, that is, the plug-in estimator of the weights as the black lines. The graph on the right plots the weights (31) as the black lines.

The graph on the left shows how the introduction of the probabilities brings the weights closer to equal weighting compared to the weights for known states. This contrasts with the weights that explicitly take the uncertainty around the states into account. In the plot on the right, these weights are very close to the weights conditional on the states. Hence, using the uncertainty of the states in the derivation of the weights leads to weights that are similar to when the states are known.

An additional difference arises for positive λ_3, where the weights conditional on state probabilities for the future state increase over those conditional on states. The reason is that for λ_2 and λ_3 of opposite sign, the variance of $\boldsymbol{\iota}'\tilde{\boldsymbol{\xi}}$ increases relative to the case of λ's of equal sign, which affects $d_t^{(m)}$ in (31). Hence, the increase of uncertainty about the states leads to an increased reliance on the data that are likely from same state as the future observation.

The relative MSFE of optimal relative to standard weights is displayed in Figure 4. When λ_2 and λ_3 are large and of similar magnitude, optimal weights have a much smaller MSFE as the standard weights are compressed due to the uncertainty around the states. When λ_2 and λ_3 are of opposite signs, the gain is smaller as the compression of the standard weights brings them closer to the optimal weights, which for $\lambda_2 = -\lambda_3$ are equal weights.

3.3. Estimating State Covariances from the Data

Above, we derived weights conditional on the state probabilities, in which case we can write the expectation

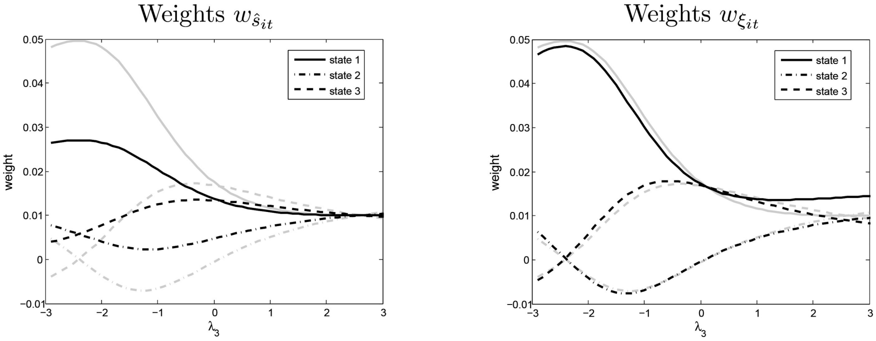

Figure 3. Optimal weights for a three-state Markov switching model. In both plots, the lighter, gray lines depict optimal weights (16), which are conditional on the states, for a representative observation in each state. In the left plot, the darker lines are the optimal weights (16) for a representative observation in each state where the probabilities are used as plug-in values for the states. In the right plot, the darker lines are the weights (31) that are derived conditional on the states under state probabilities $\hat{\boldsymbol{\xi}}_{T+1} = [0.8, 0.1, 0.1]'$ for $\lambda_2 = -2.5$, λ_3 over the range -3 to 3, $T = 100$, $\pi_1 = 0.2$, and $\pi_2 = \pi_3 = 0.4$. The dark, solid line gives the weights when $\hat{\boldsymbol{\xi}}_t = [0.8, 0.1, 0.1]'$, the dark, dash-dotted line when $\hat{\boldsymbol{\xi}}_t = [0.1, 0.8, 0.1]'$, and the dark, dashed line when $\hat{\boldsymbol{\xi}}_t = [0.1, 0.1, 0.8]'$.

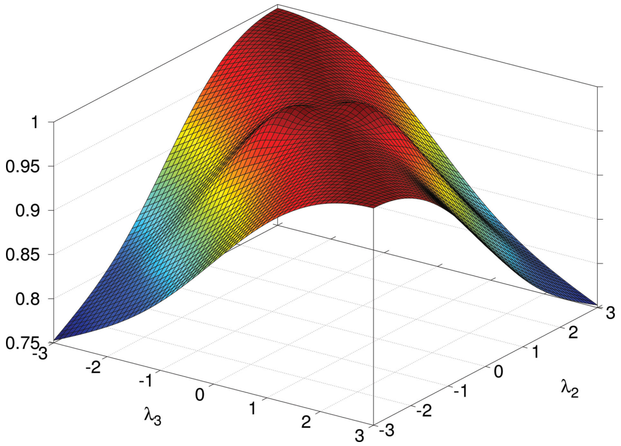

Figure 4. MSFE of optimal weights relative to standard weights when states are uncertain. The figure displays the ratio of the MSFE of the optimal weights relative to that of the standard MSFE forecast. For details of the parameter settings see the footnote of Figure 3.

of the product of two states as $\mathrm{E}(s_{it}s_{j,t+m}) = \xi_{it}\xi_{j,t+m}$. While this assumption allows us to find an explicit inverse of the matrix $\mathbf{M}$ and to obtain analytic expressions for the weights, it does not use the Markov switching nature of the DGP. If one is willing to forgo the convenience of explicit expressions for the weights, it is possible to estimate $\mathbf{M}$ directly from the data.

To estimate $\mathbf{M}$ directly from the data, we now condition on the information setup to time T, denoted Ω_T. Then $\mathrm{E}(s_{it}s_{j,t+m}|\Omega_T) = p(s_{j,t+m} = 1|\Omega_T)p(s_{it} = 1|s_{j,t+m} = 1, \Omega_T)$. The first term is the smoothed probability of being in state j at time $t + m$ as given by an EM-algorithm (Hamilton 1994) or a Markov chain Monte Carlo (MCMC) sampler (Kim and Nelson 1999). The second term can be written as

$$p(s_{it} = 1|s_{j,t+m} = 1, \Omega_T) = \frac{\xi^{i}_{t|t}}{\xi^{j}_{t+m|t+m-1}}\left[\left(\prod_{l=1}^{m-1}\mathbf{P}'\mathbf{A}_{t+l}\right)\mathbf{P}'\right]_{i,j}, \tag{32}$$

where $\mathbf{A}_t$ is a $m \times m$ diagonal matrix with typical i, i-element $\xi_{it|t}/\xi_{it|t-1}$, and $\xi_{it|t}$ and $\xi_{it|t-1}$ denote the filtered and forecast probabilities of state i at time t. The derivation of (32) can be found the web appendix. Using these expressions, we can calculate the expectations in (8). Define

$$\boldsymbol{\Xi}^* = \left[\left(\prod_{l=1}^{k-1}\mathbf{P}'\mathbf{A}_{t+l}\right)\mathbf{P}'\right]_{2:m,2:m}.$$

Then we can write $m - 1 \times m - 1$ matrix of expectations

$$\mathrm{E}(\tilde{\mathbf{s}}_t\tilde{\mathbf{s}}'_{t+k}|\boldsymbol{\Omega}_T) = \boldsymbol{\Xi}_{t|t}\boldsymbol{\Xi}^*(\boldsymbol{\Xi}_{t+k|T} \div \boldsymbol{\Xi}_{t+k|t+k-1}),$$

where $\boldsymbol{\Xi}_{t|t}$ is an $m - 1 \times m - 1$ matrix with typical i, i element $\xi_{it|t}$ is, and $\div$ denotes element-by-element division. Recall $\mathbf{M} = \mathrm{E}(\mathbf{Q}) + \mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}})$. A typical element of the second matrix is given by

$$\begin{aligned}\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}|\boldsymbol{\Omega}_T)_{t,t} &= \boldsymbol{\lambda}'\mathrm{diag}\left[\mathrm{E}\left(\tilde{\mathbf{s}}_t|\,\boldsymbol{\Omega}_T\right)\right]\boldsymbol{\lambda}\\ \mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}|\boldsymbol{\Omega}_T)_{t,t+k} &= \boldsymbol{\lambda}'\mathrm{E}(\tilde{\mathbf{s}}_t\tilde{\mathbf{s}}'_{t+k}|\boldsymbol{\Omega}_T)\boldsymbol{\lambda}.\end{aligned} \tag{33}$$

Using (33) in (8) yields numerical solutions for the weights.

4. MARKOV SWITCHING MODELS WITH EXOGENOUS REGRESSORS

So far, we have considered models that only contain a constant as the regressor. Now, we return to the model with regressors in (1). Rewrite this model as

$$\begin{aligned}\mathbf{y} &= \sum_{i=1}^{m}\mathbf{S}_i(\mathbf{X}\boldsymbol{\beta}_i + \sigma_i\boldsymbol{\varepsilon})\\ &= \mathbf{X}\boldsymbol{\beta}_1 + \sum_{i=1}^{m}\mathbf{S}_i\mathbf{X}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_1) + \sum_{i=1}^{m}\mathbf{S}_i\sigma_i\boldsymbol{\varepsilon},\end{aligned}$$

where $\mathbf{S}_i$ is a $T \times T$ matrix with as its jth diagonal element equal to one if observation j belongs to state i and zero elsewhere, $\mathbf{X}$ a $T \times k$ matrix of exogenous regressors and $\boldsymbol{\beta}_i$ a $k \times 1$ vector of parameters, σ_i the variance of regime i, and we used the fact that $\sum_{i=1}^{m}\mathbf{S}_i = \mathbf{I}$. Also,

$$y_{T+1} = \mathbf{x}'_{T+1}\boldsymbol{\beta}_1 + \sum_{i=2}^{m}s_{i,T+1}\mathbf{x}'_{T+1}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_1) + \sum_{i=1}^{m}s_{i,T+1}\sigma_i\varepsilon_{T+1}.$$

As before, we define the optimally weighted estimator as follows:

$$\boldsymbol{\beta}(\mathbf{w}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y},$$

where $\mathbf{W}$ is a diagonal matrix with diagonal elements $w_1, w_2, \ldots, w_T$. The optimal forecast is then given by $\hat{y}_{T+1} = \mathbf{x}'_{T+1}\boldsymbol{\beta}(\mathbf{w})$.

Define $\boldsymbol{\lambda}_i = (\boldsymbol{\beta}_i - \boldsymbol{\beta}_1)/\sigma_m$, $q_i = \sigma_i/\sigma_m$, and $\boldsymbol{\Lambda}_{ij} = \boldsymbol{\lambda}_i\boldsymbol{\lambda}'_j$. As in the case of structural breaks analyzed by Pesaran, Pick, and Pranovich (2013), large sample approximations to the MSFE are necessary to obtain analytical expressions for the weights. We make the following approximations: $\mathrm{plim}_{T\to\infty}\mathbf{X}'\mathbf{W}\mathbf{X} = \boldsymbol{\Omega}_{XX}$, $\mathrm{plim}_{T\to\infty}\mathbf{X}'\mathbf{S}_i\mathbf{W}\mathbf{X} = \boldsymbol{\Omega}_{XX}\mathbf{w}'\mathbf{s}_i$, $\mathrm{plim}_{T\to\infty}\mathbf{X}'\mathbf{W}^2\mathbf{S}_i\mathbf{X} = \boldsymbol{\Omega}_{XX}\mathbf{w}'\mathbf{S}_i\mathbf{w}$. Then, the MSFE is

$$\begin{aligned}&\mathrm{E}\left(\sigma_m^{-2}e^2_{T+1}\right)\\ &\quad= \sum_{i=1}^{m}\mathrm{E}(s_{i,T+1})\mathbf{x}'_{T+1}\boldsymbol{\Lambda}_{ij}\mathbf{x}_{T+1} + \sum_{i=1}^{m}\mathrm{E}(s_{i,T+1})q_i^2\varepsilon^2_{T+1}\\ &\quad\quad+ \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{w}'\mathrm{E}(\mathbf{s}_i\mathbf{s}'_j)\mathbf{w}\boldsymbol{\Lambda}_{ij}\mathbf{x}_{T+1}\\ &\quad\quad+ \mathbf{x}'_{T+1}\boldsymbol{\Omega}^{-1}_{XX}\sum_{i=1}^{m}q_i^2\mathbf{w}'\mathrm{E}(\mathbf{S}_i)\mathbf{w}\mathbf{x}_{T+1}\\ &\quad\quad- 2\mathbf{x}'_{T+1}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{w}'\mathrm{E}(\mathbf{s}_i s_{j,T+1})\boldsymbol{\Lambda}_{ij}\mathbf{x}_{T+1}.\end{aligned}$$

Maximizing (4) subject to $\boldsymbol{\iota}'\mathbf{w} = 1$ leads to the following optimal weights:

$$\mathbf{w} = \mathbf{M}^{-1}\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\phi}\boldsymbol{\phi}'\tilde{\mathbf{s}}_{T+1}) + \frac{\mathbf{M}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}}[1 - \boldsymbol{\iota}'\mathbf{M}^{-1}\mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\phi}\boldsymbol{\phi}'\tilde{\mathbf{s}}_{T+1})], \tag{34}$$

where $\boldsymbol{\phi}_i = \mathbf{x}'_{T+1}\boldsymbol{\lambda}_i/(\mathbf{x}_{T+1}\boldsymbol{\Omega}^{-1}_{XX}\mathbf{x}_{T+1})^{1/2}$, $\mathbf{M} = \mathrm{E}(\mathbf{Q}) + \mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\phi}\boldsymbol{\phi}'\tilde{\mathbf{S}})$, and $\mathbf{Q}$ a diagonal matrix with typical (t, t)-element $Q_{tt} = \sum_{i=1}^{m}q_i^2 s_{it}$. The results derived for the location model above can, therefore, be straightforwardly extended to allow for exogenous regressors by replacing $\boldsymbol{\lambda}$ with $\boldsymbol{\phi}$.

5. EVIDENCE FROM MONTE CARLO EXPERIMENTS

5.1. Setup of the Experiments

We analyze the forecast performance of the optimal weights in a series of Monte Carlo experiments. Data are generated according to (1) and we consider models with $m = 2$ and $m = 3$ states. We set $\sigma_2^2 = 0.25$ and use a range of values for λ_i and q^2.

The states are generated by a Markov chain with transition probabilities $p_{ij} = \frac{1}{T\pi_i}$, for $i \neq j$, and ergodic probabilities $\pi_i = \pi = 1/m$, $\forall i$, where m is the number of states. The diagonal elements of the transition probability matrix are $p_{ii} = 1 - \sum_{j=1}^{m} p_{ij}$. This creates Markov chains with relatively high persistence. The first state is sampled from the ergodic probability vector, $\mathbf{s}_1 \sim \text{Binomial}(1, \boldsymbol{\pi})$ and subsequent states are drawn as $\mathbf{s}_t \sim \text{Binomial}(1, \mathbf{p}_t)$ where $\mathbf{p}_t = \mathbf{P}\mathbf{s}_{t-1}$. We restrict attention to draws of the data that would be identified as Markov switching models in an application: we require that each regime has at least 10 observations and that regimes are identified empirically in that $\sum_{t=1}^{T} \hat{\xi}_{t|T}^{i} \geq 5$, $\forall i$, which ensures identification of the parameters. The estimation uses the EM algorithm (Dempster et al. 1977) as outlined by Hamilton (1994).

The first set of the Monte Carlo experiments analyzes two state models with a constant only, that is, $k = 1$ and $x_t = 1$ for $T = 200$. A second set of experiments considers three state models for $T = 200$. We also ran experiments for a two-state model with an exogenous regressor. The results do not substantially differ from the results of the mean only model and can be found in the web appendix.

Given the parameter estimates $\hat{\boldsymbol{\beta}}_i$, $\hat{\mathbf{P}}$, $\hat{\sigma}_i$ and the probability vectors with $\hat{\xi}_{t|T}$, $\hat{\xi}_{t|t}$, $\hat{\xi}_{t|t-1}$, we construct the usual Markov switching forecast as

$$\hat{y}_{T+1}^{\text{MS}} = \mathbf{x}_{T+1}' \sum_{i=1}^{m} \hat{\beta}_i \hat{\xi}_{T+1|T}^{i},$$

where $\hat{\beta}_i$ is given in (2).

The optimal weights are calculated as outlined in the sections above. The following notation is used to distinguish the different weights:

- $w_{\hat{s}}$: weights based on known states, operationalized by substituting the smoothed probability vector $\hat{\boldsymbol{\xi}}_{t|T}$ for the states as discussed in Section 3.1.
- $w_{\hat{\xi}}$: weights derived based on state probabilities, with the smoothed probability vector $\hat{\boldsymbol{\xi}}_{t|T}$ as the probabilities as discussed in Section 3.2.
- $w_{\hat{\mathbf{M}}}$: the weights based on state probabilities derived by directly estimating the matrix $\hat{\mathbf{M}}$ as detailed in Section 3.3.

Using these weights, the optimal forecast is constructed as

$$\hat{y}_{T+1}^{\text{opt}} = \mathbf{x}_{T+1}' \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{W}\mathbf{y},$$

where $\mathbf{W}$ is a diagonal matrix with typical diagonal element $w_{\hat{s},t}$, $w_{\hat{\xi},t}$, or $w_{\hat{\mathbf{M}},t}$.

We report ratios of the MSFE of optimally weighted forecasts to that of standard Markov switching forecasts. Additionally, we separated the results by the size of the regime difference, λ_i. Finally, we have seen above that the performance of the weights $w_{\hat{\xi}}$ depends on the variance of the smoothed probability vector. Thus, we separate the results based on the normalized variance of the smoothed probability vector

$$\tilde{\sigma}_{\hat{\xi}}^2 = \frac{\frac{1}{T}\sum_{t=1}^{T} \hat{\xi}_{t|T}^{(i)}(1 - \hat{\xi}_{t|T}^{(i)})}{\frac{1}{T}\sum_{t=1}^{T} \hat{\xi}_{t|T}^{(i)} \frac{1}{T}\sum_{t=1}^{T}(1 - \hat{\xi}_{t|T}^{(i)})}, \tag{35}$$

where i the state that has the minimum normalized variance. Note that in the case of two states for $\frac{1}{T}\sum_{t=1}^{T} \hat{\xi}_{t|T}^{(i)} = \frac{1}{T}\sum_{t=1}^{T}(1 - \hat{\xi}_{t|T}^{(i)}) = 0.5$, the measure $\tilde{\sigma}_{\hat{\xi}}^2$ is analogous to the regime classification measure (RCM) of Ang and Bekaert (2002). The Monte Carlo results are from 10,000 replications.

5.2. Monte Carlo Results

The Monte Carlo results for the two-state model are reported in Table 3, where results for models with switches in mean and homoscedastic errors are in the left panel. The results in Section 3.1 suggest that forecasts from optimal weights conditional on states, $w_{\hat{s}}$, will show the largest gains when the difference between regimes, λ, is small. In contrast, the results in Section 3.2 suggest that the gains for the forecasts from optimal weights conditional on state probabilities, $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$, will be largest for large λ, which is the practically more relevant case.

The results from the simulation confirm the theoretical findings. For small λ, the forecasts from weights, $w_{\hat{s}}$, are more precise than those using standard weights and weights conditional on state probabilities. An additional effect that improves the forecasts using $w_{\hat{s}}$ is that the parameter estimates are biased upward when $\lambda = 1$. In Section 3.1.2, we show that the weights $w_{\hat{s}}$ are shrunk toward equal weights. The upwards bias of $\hat{\lambda}$ will

Table 3. Monte Carlo results: Two states, mean only models

		$q^2 = 1$			$q^2 = 2$		
λ	$\tilde{\sigma}_{\hat{\xi}_{\mid T}}^2$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
1	0.0–0.1	0.997	1.004	1.008	0.998	1.002	1.002
	0.1–0.2	1.000	1.007	1.023	1.000	1.004	1.012
	0.2–0.3	1.000	1.009	1.027	1.000	1.010	1.023
	0.3–0.4	1.001	1.009	1.030	1.001	1.008	1.022
2	0.0–0.1	1.000	1.000	1.024	1.000	1.001	1.018
	0.1–0.2	1.002	0.989	1.024	1.001	0.997	1.034
	0.2–0.3	1.003	0.966	0.998	1.003	0.984	1.011
	0.3–0.4	1.004	0.940	0.967	1.002	0.983	1.004
3	0.0–0.1	1.000	0.999	1.025	1.000	0.999	1.027
	0.1–0.2	1.004	0.959	0.990	1.003	0.975	1.013
	0.2–0.3	1.005	0.903	0.953	1.005	0.950	0.988
	0.3–0.4	1.003	0.845	0.921	1.006	0.889	0.918

NOTE: The table reports the ratio of the MSFE of the optimal weights to that of the Markov switching weights. $y_t = \beta_1 s_{1t} + \beta_2 s_{2t} + (\sigma_1 s_{1t} + \sigma_2 s_{2t})\varepsilon_t$, where $\varepsilon_t \sim \text{N}(0, 1)$, $\sigma_2^2 = 0.25$, $q^2 = \sigma_1^2/\sigma_2^2$. Column labels: $\lambda = (\beta_2 - \beta_1)/\sigma_2$, $\tilde{\sigma}_{\hat{\xi}_{|T}}^2$ is the normalized variance in of the smoothed probability vector (35). $w_{\hat{s}}$: forecasts from weights based on estimated parameters and state probabilities. $w_{\hat{\xi}}$: forecasts from weights conditional on state probabilities. $w_{\hat{\mathbf{M}}}$ are the weights based on numerically inverting $\hat{\mathbf{M}}$. The sample size is $T = 200$ and the results are from $R = 10{,}000$ repetitions.

return the weights closer to the infeasible optimal weights based on the true DGP. The estimated weights conditional on state probabilities are close to the infeasible optimal weights in the absence of a bias, and the bias in $\hat{\lambda}$ will increase them beyond the infeasible optimal weights. The case of $\lambda = 1$ may, however, not be recognized in a given time series as the switches are as large as the disturbance standard deviation. This setting is, therefore, of lesser practical relevance than those with larger λ.

For larger λ the ordering is reversed: the forecasts from optimal weights conditional on states, $w_{\hat{s}}$, are less precise than those of the standard weights. In contrast, the weights conditional on state probabilities, $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$, are substantially more precise. The reason is that in this settings there is a smaller, at times even downward, bias in $\hat{\lambda}$ and the shrinking of the weights $w_{\hat{s}}$ toward equal weights deteriorates the forecasts, whereas the weights conditional on the states benefit from the fact that the weights are close to the infeasible optimal weights.

The theoretical results in Section 3.2 suggest that the relative performance of the weights based on state probabilities, $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$, will increase in the uncertainty around the states. This is because the standard weights and the plug-in weights, $w_{\hat{s}}$, are compressed toward equal weights whereas the optimal weights retain the shape of the weights as if the states where known. Again, the results in Table 3 confirm the finding: the results for weights $w_{\hat{s}}$ are worse when the states are uncertain, the forecasts from the weights conditional on state probabilities improve substantially and lead to large gains. Our application will highlight the practical relevance of large λ and state uncertainty, so that we can expect large gains when using $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$ in practice.

The sample size is the final factor that influences the performance of the forecasts, where weights $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$ improve with the sample size while weights, $w_{\hat{s}}$, deteriorate in the samples size. However, this affect is relevant only for small T such as $T = 50$, which are unlikely to be relevant in practice. Results for $T = 50$ and 100 can be found in the web appendix.

The right panel of Table 3 reports the results for a model with state-dependent mean and variance, where the variance in regime 2 is the same as before but the variance in regime 1 is doubled. This should mute the improvements since the average difference in regimes standardized by the variance decreases. While this decrease is indeed observed, substantial improvements remain in the same parameter regions where the weights under constant variance perform well.

Finally, we investigate forecasts from three state models. The results in Table 4 suggest that the conclusions from two state models carry over to three state models. Sizable improvements are made when using $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$ when $\tilde{\sigma}^2_{\hat{\xi}}$ is large and both differences in parameters, λ_{21} and λ_{31}, are large.

Overall, the findings from the Monte Carlo experiments suggest that optimal weights conditional on states, $w_{\hat{s}}$, work well only for small differences in regimes and when states are estimated with great certainty, which are, arguably, less realistic assumptions in practice. In contrast, optimal weights conditional on state probabilities improve forecasts over standard weights when differences in regimes and uncertainty around states are large, which is the setting most likely found in applications, such as that in Section 6. Using weights that treat states as independent binary variables, $w_{\hat{\xi}}$, avoids the estimation uncertainty around covariances of the state, and in many settings leads to the most precise forecasts. Estimating the full matrix of second moments, $\mathbf{M}$, in the construction of the optimal weights, $w_{\hat{\mathbf{M}}}$, can, however, improve forecasts when the difference between regimes is large while the uncertainty about regimes remains large, too.

Table 4. Monte Carlo results: Three states, intercept only models

$\{\lambda_{31}, \lambda_{21}\}$	$\tilde{\sigma}^2_{\hat{\xi}\mid T}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
{2,1}	0.0–0.1	0.999	1.014	1.027
	0.1–0.2	1.000	1.010	1.024
	0.2–0.3	1.001	1.007	1.031
	0.3–0.4	1.001	0.999	1.019
{3,1}	0.0–0.1	1.000	1.004	1.025
	0.1–0.2	1.001	0.989	1.019
	0.2–0.3	1.002	0.958	0.969
	0.3–0.4	1.002	0.938	0.952
{3.5,2}	0.0–0.1	1.000	1.001	1.024
	0.1–0.2	1.001	0.983	1.021
	0.2–0.3	1.002	0.954	0.960
	0.3–0.4	1.003	0.902	0.918

NOTE: The table reports the ratio of the MSFE of the optimal weights to that of the Markov switching weights for $q^2 = 1$. For details see Table 3.

6. APPLICATION TO U.S. GNP

The U.S. business cycle, which was analyzed by Hamilton (1989), arguably remains one of the most prominent application of Markov switching models. Different variants of such models have been used to model U.S. GNP growth, see, for example, Clements and Krolzig (1998) and Krolzig (1997, 2000). These authors also show that the Markov switching model is frequently outperformed in terms of MSFE by AR models. We use a pseudo-out-of-sample forecast exercise to investigate whether optimal weights improve the forecast accuracy of Markov switching models for U.S. GNP growth, and whether optimal weights improve the forecast accuracy of Markov switching models over that of linear alternatives.

The model by Hamilton (1989) is an example of a Markov Switching in mean model with nonswitching autoregressive regressors. This class of models

$$y_t = \beta_{s_t} + \sum_{i=1}^{p} \phi_i (y_{t-i} - \beta_{s_{t-i}}) + \sigma \varepsilon_t$$

is denoted as MSM(m)-AR(p) by Krolzig (1997), where Hamilton's model takes $m = 2$ and $p = 4$. Here, y_t depends on the current state and on the previous p states. If, in addition, the model contains a state-dependent variance, σ_{s_t}, it is denoted as MSMH(m)-AR(p).

Clements and Krolzig (1998) found that a three-state model with switching intercept instead of switching mean and a state-dependent variance does well in terms of business cycle description and forecast performance. This class of models

$$y_t = \beta_{s_t} + \sum_{i=1}^{p} \phi_i y_{t-i} + \sigma_{s_t} \varepsilon_t$$

is denoted as MSIH(m)-AR(p) by Krolzig (1997) and the model in Clements and Krolzig (1998) takes $m = 3$ and $p = 4$.

Note that, for both models, we can use the optimal weights of the intercept only model because, conditional on the estimated parameters, the state-independent autoregressive component can be moved to the left-hand side. On the right-hand side, only the constant remains and we can use the optimal weights of the intercept only model. We estimate the models using the EM algorithm suggested by Hamilton (1994) with the extensions discussed by Krolzig (1997). We have investigated the performance of optimal weights for such dynamic models in Monte Carlo experiments with details provided in a web appendix. The results indicate that the insights gained from the intercept only model in Section 5 carry over to dynamic models.

In this exercise, we focus on pseudo-out-of-sample forecasts generated by a range of candidate Markov switching models: MSM(m)-AR(p) and MSMH(m)-AR(p) models with $m = 2$ and $p = 0, 1, 2, 3, 4$ and $m = 3$ with $p = 1, 2$, and MSI (m)-AR(p) and MSIH(m)-AR(p) models with $m = 2, 3$ and $p = 0, 1, 2, 3, 4$. We construct expanding window forecasts where for each forecast all models are re-estimated to include all available data at that point in time. We select the Markov switching model that, based on standard weights, delivers the lowest MSFE in a cross-validation sample. Using this model, we then compare the pseudo out-of-sample forecasts using standard weights and optimal weights.

We report the ratio of the MSFE of forecasts from optimal weights relative to those from standard weights together with the Diebold and Mariano (1995) test statistic of equal predictive accuracy. Additionally, we calculate the components of MSFE: the squared biases and variances. We report the differences between the squared bias of the standard weights forecasts and that of the optimal weight forecasts relative to the MSFE of the standard weight forecast, and the differences between the variance of the standard weights forecasts and that of the optimal weight forecasts relative to the MSFE of the standard weight forecast.

The data are (log changes in) U.S. GNP series from 1947Q1 to 2014Q1, which we obtained from the Federal Reserve Economic Data (FRED). The data are seasonally adjusted. In total, the series consists of 269 observations. After accounting for the necessary presample, we start the estimation sample in 1948Q2.

The out-of-sample forecast period is 1983Q2–2014Q1, which amounts to 124 observations and ensures that throughout the forecasting exercise all models are estimated on at least 100 observations. We start evaluating forecasts for model selection purposes with a training period 1973Q2–1983Q1 (40 observations). The model that has the minimum MSFE over this period (using standard weights) is selected as the forecasting model for the observation 1983Q2, and forecasts using the different weights are made with this model. In this way, no information is used that is not available to researchers in real time. Next, we add the next period to our estimation and cross-validation sample, select the minimum MSFE model, and construct the next forecast. Remarkably, in our application, the MSM(3)-AR(1) model is selected throughout.

As mentioned above, the beginning of the out-of-sample forecast period is chosen such that a sufficient amount of observations is available to estimate all Markov switching models. Still, we need to ensure that our results do not critically depend on this choice. In a second step, we therefore check the robustness of our results using the forecast evaluation measures proposed by Rossi and Inoue (2012).

Table 5. GNP forecasts: Forecasting performance

	w_{MS}	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
1983Q2-2014Q1	0.367	1.001	0.970**	0.959***
Subperiods				
1983Q2-1993Q1	0.225	1.002	0.875**	0.898*
1993Q2-2003Q1	0.306	1.000	1.021	0.989
2003Q2-2014Q1	0.553	1.000	0.980*	0.965**
Full sample: 1983Q2-2014Q1				
Square bias	0.008	0.000	0.003	0.005
Variance	0.359	– 0.001	0.028	0.037

NOTE: The second column in the top two panels of the table reports the MSFE based on the best Markov switching model with standard weights. The remaining columns of the table reports the relative MSFE of the optimal weights compared with the Markov switching weights. Asterisks denote significance at the 10% (*), 5% (**), and 1% (***) level using the Diebold–Mariano test statistic. The second column of the last panel reports the square bias and variance of the best Markov switching model with standard weights. The remaining columns give the differences in squared biases and variances between the standard weights and optimal weights forecasts relative to the MSFE of the Markov switching model with standard weights. Positive numbers indicate lower bias/variance.

The forecasting performances of the standard and optimal weights are reported in Table 5. The column with heading w_{MS} reports the MSFE of the best Markov switching model using standard weights. The next three columns report the ratio of MSFE of the optimal weights forecast to the standard weights forecast for the same model. The results in the first line, which are over the full forecast period, show that optimal weights conditional on states, $w_{\hat{s}}$, do not improve forecasts but that, in contrast, weights conditional on state probabilities, $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$, substantially improve the forecast performance over standard weights and that these improvements are significant. The most precise forecasts result from using $w_{\hat{\mathbf{M}}}$. The three state models have an average estimated differences in mean (scaled by the standard deviation) $\hat{\lambda}_{21} = 2.28$ and $\hat{\lambda}_{31} = 4.23$. The average minimum normalized variance of the smoothed probability vector is $\tilde{\sigma}^2_{\hat{\xi}|T} = 0.20$. The size of the improvements over the Markov switching forecast is close to the improvements found in the Monte Carlo simulation for three-state models as presented in Table 4.

It is interesting to also compare forecast performance in subsamples. In the first subsample, 1983Q2–1993Q1, forecasts based on the optimal weights conditional on state probabilities, $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$, improve significantly over the standard weights with gains of more than 10% in forecast accuracy. Forecasts based on the plug-in weights, $w_{\hat{s}}$, in contrast, cannot improve on the standard MS forecasts. In the second subsample, 1993Q2–2003Q1, which largely covers the great moderation, only $w_{\hat{\mathbf{M}}}$ offers a modest improvement. In the last subsample, 2003Q2–2014Q1, again all optimal weights conditional on the state probabilities lead to more precise forecasts than the standard weights and these improvements are again significant.

The optimal weights trade off bias and variance of the forecasts, and it is therefore interesting to consider the magnitude of the bias incurred. The bottom panel of Table 5 reports the squared bias and variance of the forecasts from the standard

Table 6. GNP forecasts: Comparison to linear models

	AR_{dyn}	w_{MS}	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
1983Q2-2014Q1	0.368	0.999	1.000	0.970	0.958
Subperiods					
1983Q2-1993Q1	0.265	0.849**	0.851**	0.743**	0.763**
1993Q2-2003Q1	0.280	1.091	1.091	1.114	1.080
2003Q2-2014Q1	0.540	1.023	1.023	1.003	0.988

NOTE: The second column contains the MSFE of the best linear model. The remaining columns contain the MSFE of the best Markov switching model with different weights relative to that of the linear model. The best Markov switching model is selected based on standard weights. The linear model is the AR(1) model for the first 69 forecasts and AR(2) for the final 55 forecasts. Asterisks denote significance at the 10% (*), 5% (**), and 1% (***) level using the Diebold–Mariano test statistic.

weights forecasts in the second column and, in the subsequent columns, the difference in squared biases and variances of the standard weights and the optimal weights forecasts relative to the MSFE of the standard weights forecasts. It can be seen that the squared bias of the standard weights forecast is very small and only a fraction of the size of the variance. The reduction in MSFE that the optimal weights (based on state probabilities) achieve is therefore for the most part via a reduction in variance. Yet, in this application there appears to be no trade-off in bias as the biases of the optimal weights forecasts are no larger and typically smaller than that of the standard weights forecasts. It appears that the model uncertainty around the Markov switching model induces a bias that the optimal weights mitigate, which leads to improvements of the forecasts in bias and variance.

Having established that the optimal weights improve on the Markov switching model with standard weights, the question remains how the optimal weights forecasts compare to forecasts from linear models, which here are AR(p) models with $p = 1, 2, 3, 4$ and a mean only model. We select the best linear model based on the historic forecast performance in line with the model selection for the Markov switching model. The AR(1) model is selected for the first 69 forecasts and the AR(2) model for the remaining forecasts. The resulting MSFE and relative performance of the different weighting scheme for the selected Markov switching model are reported in Table 6. Over the entire forecast period, the performance of the linear models is very similar to the Markov switching model with standard weights. The same is true for the weights conditional on states. This contrasts with the forecast based on optimal weights conditional on state probabilities that substantially beat the linear models, even if for the full forecast sample the difference is not significant at conventional levels.

The results for the three different subsamples reveal that, in the first subsample, all Markov switching forecasts significantly improve on the linear forecasts. The largest gains are made using the optimal weights conditional on the state probabilities. In the middle subsample, no Markov switching forecast is more precise than the linear model. In the final subsample, optimal weights, $w_{\hat{\mathbf{M}}}$ again yield forecasts with a lower MSFE than the linear model. Comparing these results to those in Table 5 suggests that the optimal weights improve forecasts over the standard weights the most when the data exhibit strong switching behavior. This ties in with the results from our theory in two ways. First, we showed above that the weights conditional on the states are tending toward equal weighting that is in the direction of the linear models, whereas the optimal weights derived conditional on state probabilities emphasize the Markov switching nature of the data. Second, we demonstrated that, in a three-state model, the optimal weights are around $1/T$ when the future regime is the middle regime. This appears to be a distinguishing feature of the subsamples: in the first subsample, the forecast observation is estimated to be, on average, in the middle regime with probability 0.65. In the second and third subsamples, in contrast, the average probabilities are 0.83 and 0.84. Hence, the linear model is more difficult to beat in the second and third subsample as, for many forecast observations, the forecast from the linear model is close to the optimal forecast from the Markov switching model.

To check the robustness of our results to the choice of forecast sample, we additionally use the forecast accuracy tests suggested by Rossi and Inoue (2012). The tests require the calculation of Diebold–Mariano test statistics over a range of possible out-of-sample forecast windows. From these different windows, two tests can be constructed: first, the $\mathcal{A}_T$ test, which is the average of the Diebold–Mariano test statistics, and, second, the $\mathcal{R}_T$ test, which is the supremum of the Diebold–Mariano test statistics. The application of these tests comes with two caveats in our application. First, the relative short first estimation window implied by these tests is problematic as various switches of the Markov chain are required for the estimation of Markov switching models. For the test by Rossi and Inoue (2012), the beginning of the out-of-sample forecast evaluation period is varied over the interval $[\mu T, (1-\mu)T]$ and we set μ to the maximum of 0.35. In contrast, in the baseline application above, the shortest estimation sample is $0.53\,T$. Early forecasts for the Rossi and Inoue test may suffer as a result of a short estimation window. Second, as a further consequence of the shortened estimation sample, we cannot use cross-validation as model selection procedure and therefore consider only the MSM(3)-AR(1) model, which has been selected in our baseline forecast procedure throughout, and for the linear model we use the AR(1) and AR(2) models, which are the models selected in the baseline forecasting exercise.

Table 7 reports the test statistics and associated significance levels. The top panel reports the test statistics of the optimal weights forecasts against the standard weights forecasts. It

Table 7. Rossi and Inoue test of forecast accuracy

	w_{MS}	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
Test against MS weights				
$\mathcal{A}_T$		0.585	– 0.356	– 0.910
$\mathcal{R}_T$		– 0.646	– 1.803	– 2.342**
Test against AR(1)				
$\mathcal{A}_T$	– 0.223	– 0.222	– 0.208	– 0.546
$\mathcal{R}_T$	– 0.954	– 0.951	– 1.071	– 1.575
Test against AR(2)				
$\mathcal{A}_T$	0.372	0.375	0.261	– 0.027
$\mathcal{R}_T$	– 0.469	– 0.477	– 0.621	– 0.928

NOTE: The beginning of the out-of-sample forecast evaluation period is varied between $[\mu T, (1-\mu)T]$ with $\mu = 0.35$ and $T = 264$. $\mathcal{A}_T$ denotes the average and $\mathcal{R}_T$ the supremum of the Diebold–Mariano test statistics over the range of forecast periods. Asterisks denote significance at the 10% (*), 5% (**), and 1% (***) level.

can be seen that the signs of the test statistics are as expected and that the $w_{\hat{\mathbf{M}}}$ weights provide significant improvements on the standard weights according to the $\mathcal{R}_T$ test. The lower two panels of Table 7 report the test statistics when the MSM(3)-AR(1) model is tested against a simple AR(1) and AR(2) model. For the AR(1) model the signs are as expected, although the test statistics do not exceed the critical values reported in Rossi and Inoue (2012). For the AR(2) model, the $\mathcal{A}_T$ test statistic for $w_{\hat{\mathbf{M}}}$ weights remains negative. For these weights the largest negative $\mathcal{R}_T$ test statistic is observed, which it is not significant at conventional levels. This reflects the fact that the linear model is a close approximation to the optimal weights Markov switching model as the forecast sample is dominated by observations that are most likely from the middle regime.

7. CONCLUSION

In this article, we have derived optimal forecasts for Markov switching models and analyzed the effect of uncertainty around states on forecasts based on optimal weights. The importance of uncertainty of the states of the Markov chain is highlighted in the comparison of forecasts from weights conditional on the states and those when the states are not known. The optimal weights for known states share the properties of the weights derived in Pesaran, Pick, and Pranovich (2013) and are asymptotically identical to the Markov switching weights. Improvements in forecasting performance are found when the ratio of the number of observations to the number of estimated parameters is small. This contrasts with the optimal weights for unknown states that are asymptotically different from the Markov switching weights and potential improvements in forecasting accuracy can be considerable for large differences in parameters even in large samples.

The results from theory and the application show that optimal forecasts can differ substantially from standard MS forecasts. Optimal weights emphasize the Markov switching nature of the DGP more than standard weights do. However, in the three-state case, the optimal weights for forecasts in the middle regime lead to weights that effectively ignore the Markov switching nature of the data. This is the case for the GNP forecasts from the great moderation where the vast majority of observations are from the middle regime. This explains the difficulty of Markov switching forecasts to beat linear models, as the optimal forecast from the Markov switching model is essentially the same as the forecast from the linear model.

For practitioners two messages emerge. First, when the observation in the forecast period could likely be from any regime of the Markov switching model, optimal weights conditional on state probabilities will substantially improve forecasts. When the size of the switches is moderate or regime estimates precise, weights that ignore the covariances of the states are more efficient as the additional estimation uncertainty introduced by estimating the covariances of the states dominates the forecasts. When switches are large yet state remain uncertain using the full second moment matrix of the Markov chain leads to more precise forecasts. However, the difference between the two optimal weights is small compared to the overall gains in forecast accuracy. Second, when one expects to forecast predominantly observations from the middle regime in a three-state model, using a linear model will lead to forecasts that are effectively the optimal forecasts from the Markov switching model but with the benefit of substantially reduced estimation uncertainty.

SUPPLEMENTARY MATERIALS

Additional derivations for the optimal weights and additional Monte Carlo experiments are available in the online appendix.

ACKNOWLEDGMENTS

The authors thank Robin Lumbsdaine, Barbara Rossi, Herman van Dijk, and Wendun Wang for insightful discussions, and seminar participants at CORE, DIW, Erasmus School of Economics, ISC-TE Lisbon, University of Copenhagen, University of Manchester, University of St. Andrews, and Tinbergen Institute Amsterdam, and conference participants at Barcelona GSE Summer Forum, ESEM, IAAE annual conference, ISCEF, NESG, RMSE meeting, and V_t Time Series Meeting in Zaragoza for valuable comments. The authors thank SURFsara for access to the Lisa Compute Cluster. The opinions expressed are those of the authors and should not be attributed to DNB.

[Received December 2015. Revised July 2016.]

REFERENCES

Ang, A., and Bekaert, G. (2002), “Regime Switches in Interest Rates,” *Journal of Business & Economic Statistics*, 20, 163–182. [637]

Clements, M. P., and Krolzig, H.-M. (1998), “A Comparison of the Forecast Performance of Markov-Switching and Threshold Autoregressive Models of US GNP,” *Econometrics Journal*, 1, 47–75. [628,638]

Crawford, G. W., and Fratantoni, M. C. (2003), “Assessing the Forecasting Performance of Regime-Switching, ARIMA and GARCH Models of House Prices,” *Real Estate Economics*, 31, 223–243. [628]

Dacco, R., and Satchell, S. (1999), “Why Do Regime-Switching Models Forecast So Badly?” *Journal of Forecasting*, 18, 1–16. [628]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum Likelihood From Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society*, Series B, 39, 1–38. [637]

Deschamps, P. J. (2008), “Comparing Smooth Transition and Markov Switching Autoregressive Models of US Unemployment,” *Journal of Applied Econometrics*, 23, 435–462. [628]

Diebold, F. X., and Mariano, R. S. (1995), “Comparing Predictive Accuracy,” *Journal of Business & Economic Statistics*, 12, 253–263. [639]

Engel, C. (1994), “Can the Markov Switching Model Forecast Exchange Rates?” *Journal of International Economics*, 36, 151–165. [628]

Guidolin, M. (2011), “Markov Switching Models in Empirical Finance,” *Advances in Econometrics*, 27, 1–86. [628]

Hamilton, J. D. (1989), “A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle,” *Econometrica*, 57, 357–384. [629,638]

——— (1994), *Time Series Analysis*, Princeton: Princeton University Press. [629,636,637,639]

Kim, C.-J. (1994), “Dynamic Linear Models With Markov-Switching,” *Journal of Econometrics*, 60, 1–22. [629]

Kim, C.-J., and Nelson, C. R. (1999), “Has the US Economy Become More Stable? A Bayesian Approach Based on a Markov-Switching Model of the Business Cycle,” *Review of Economics & Statistics*, 81, 608–616. [636]

Klaassen, F. (2005), “Long Swings in Exchange Rates: Are They Really in the Data?” *Journal of Business & Economic Statistics*, 23, 87–95. [628]

Krolzig, H.-M. (1997), *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis*, Berlin: Springer Verlag. [638,639]

——— (2000), “Predicting Markov-Switching Vector Autoregressive Processes,” Working Paper, W31, Nuffield College, Oxford. [638]

Perez-Quiros, G., and Timmermann, A. (2001), "Business Cycle Asymmetries in Stock Returns: Evidence From Higher Order Moments and Conditional Densities," *Journal of Econometrics*, 103, 259–306. [628]

Pesaran, M. H., Pick, A., and Pranovich, M. (2013), "Optimal Forecasts in the Presence of Structural Breaks," *Journal of Econometrics*, 177, 134–152. [628,630,631,633,636,641]

Rapach, D., and Zhou, G. (2013), "Forecasting Stock Returns," in *Handbook of Economic Forecasting* (Vol. 2A), eds. G. Elliott, and A. Timmermann, Amsterdam: Elsevier, pp [628]

Rossi, B., and Inoue, A. (2012), "Out-of-Sample Forecast Tests Robust to the Choice of Window Size," *Journal of Business & Economics Statistics*, 30, 432–453. [629,639,640]