Optimal forecasts from Markov switching models

Tom Boot* Andreas Pick†

October 31, 2014

Abstract

We derive optimal weights for Markov switching models by weighting observations such that forecasts are optimal in the MSFE sense. 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 US GNP, where using optimal weights leads to significant reductions in MSFE.
JEL codes: C25, C53, E37
Keywords: Markov switching models, forecasting, optimal weights, GNP forecasting

*Erasmus University Rotterdam, boot@ese.eur.nl

†Erasmus University Rotterdam and De Nederlandsche Bank, andreas.pick@cantab.net
We would like to 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, and the Tinbergen Institute Amsterdam, and conference participants at the IAAE annual conference, the Barcelona GSE Summer Forum, NESG, and ESEM for valuable comments. We thank SURFsara (www.surfsara.nl) for the support in using the Lisa Compute Cluster. The opinions expressed are those of the authors and should not be attributed to DNB.

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. 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) provides a recent review of the use of Markov switching models in finance.

In this paper, 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.

Initially, we assume that the states of the Markov switching model are known and, in a second step, relax this assumption. Conditional on the states of the Markov switching model, the weights mirror those obtained by Pesaran et al. (2013), which emphasizes the correspondence of structural break and Markov switching models for forecasting purposes. The weights depend on the number of observations per regime and the relative differences of the parameter between the regimes. While, conditional on the states, the usual Markov switching forecasts assign non-zero weights only to observations from the same state as that of the forecast period, the optimal weights assign non-zero weights to observations from all states. Using all observations reduces the variance of the forecast but introduces a bias, and optimally weighting all observations ensures that the trade-off is optimal in the MSFE sense. In the case of three regimes, the weights have interesting properties: for some parameter values, the optimal weights correspond to equal weighting of observations; for another range of parameter values, observations in regimes other than that of the future observation will 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 suggests that, conditional on the states, standard Markov switching weights achieve the minimum MSFE asymptotically.

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. Contrasting weights conditional on states with those conditional on state probabilities leads to

interesting insights into the role uncertainty around states plays for forecasting. While 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 retain the emphasis on the Markov switching nature of the data. This results in relative forecast performances where optimal weights conditional on state probabilities perform the better the larger the difference between the regimes in terms of their parameters and the larger the variance of the estimated smoothed probabilities. The reason is that in these scenarios the quality of the Markov switching forecast deteriorates, whereas the MSFE from the optimal weights conditional on state probabilities remains largely unaffected. 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.

These findings provide insights into the performance of the optimal weights proposed by Pesaran et al. (2013) for structural breaks. These authors show that, in the case of a structural break, the time of the break is of first order importance for the optimal weights. This corresponds to the membership of the observations to the states in the Markov switching model. The results from our analysis suggest that, in the structural break case, the plug-in estimator of the weights in Pesaran et al. (2013), which is derived conditional on break dates, will result in weights that are too close to equal weights. This explains the relatively poor performance of these weights in the empirical application reported by Pesaran et al. (2013).

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 improve for small values of the break size and small samples. The weights based on state probabilities produce substantial gains for large break sizes and a large variance in the smoothed probability vector, and these improvements increase with the sample size.

We apply the methodology to forecasting quarterly US GNP. Out-of-sample forecasts are constructed over 124 quarters for 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 paper. The results suggest that the forecasts using optimal weights significantly outperform the standard Markov switching forecast. We compare our forecasting schemes to a range of linear alternatives and find that they lead to improved forecasts. 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 paper 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 US GNP in Section 6. Finally, Section 7 concludes the paper. Details of the derivations are presented in the Appendix.

2 Markov switching models and their forecasts

Consider the following m-state Markov switching model

$$y_t = \boldsymbol{\beta}'_{s_t}\mathbf{x}_t + \sigma_{s_t}\varepsilon_t, \quad \varepsilon_t \sim iid(0,1) \tag{1}$$

where $\boldsymbol{\beta}_{s_t} = \mathbf{B}'\mathbf{s}_t$, $\mathbf{B} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_m)'$ is an $m \times k$ matrix, $\boldsymbol{\beta}_i$ is a $k \times 1$ parameter vector, $\mathbf{x}_t$ is a $k \times 1$ vector of exogenous regressors, $\sigma_{s_t} = \boldsymbol{\sigma}'\mathbf{s}_t$, $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_m)'$ are $m \times 1$ vectors of error standard deviations, and $\mathbf{s}_t = (s_{1t}, s_{2t}, \ldots, s_{mt})'$ 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 $\boldsymbol{\beta}_i$, $i = 1, 2, \ldots, m$, as

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

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 $\hat{y}_{T+1} = \sum_{i=1}^{m} \hat{\xi}_{i,T+1}\mathbf{x}'_{T+1}\hat{\boldsymbol{\beta}}_i$, see Hamilton (1994).

In this paper, 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 forecasts are 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

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

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

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

$$y_{T+1} = \sum_{t=1}^{T} w_t y_t \tag{4}$$

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

The forecast error, which, without loss of generality, is scaled by the error standard deviation of regime m, is

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

where

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

and the scaled MSFE is

$$\begin{aligned}
\mathrm{E}\left(\sigma_m^{-2} e_{T+1}^2\right) &= \mathrm{E}\left[\left(\boldsymbol{\lambda}'\left(\tilde{\mathbf{s}}_{T+1} - \sum_{t=1}^{T} w_t \tilde{\mathbf{s}}_t\right)\right)^2\right] + \mathrm{E}\left[(\mathbf{q}'\mathbf{s}_{T+1})^2\right] - \sum_{t=1}^{T} w_t^2 \mathrm{E}\left[(\mathbf{q}'\boldsymbol{s}_t)^2\right] \\
&= \mathrm{E}\left[\tilde{\mathbf{s}}_{T+1}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right] - 2\mathbf{w}'\mathrm{E}\left[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right] \\
&\quad + \mathbf{w}'\mathrm{E}\left[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}\right]\mathbf{w} + \mathrm{E}\left[(\mathbf{q}'\mathbf{s}_{T+1})^2\right] - \mathbf{w}'\mathrm{E}[\mathbf{Q}]\mathbf{w} \\
&= \mathbf{w}'\left[\mathrm{E}[\mathbf{Q}] + \mathrm{E}\left[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}\right]\right]\mathbf{w} - 2\mathbf{w}'\mathrm{E}\left[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right] \\
&\quad + \mathrm{E}\left[\tilde{\mathbf{s}}_{T+1}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right] + \mathrm{E}\left[(\mathbf{q}'\mathbf{s}_{T+1})^2\right]
\end{aligned} \tag{5}$$

where $\tilde{\mathbf{S}} = (\tilde{\boldsymbol{s}}_1, \tilde{\boldsymbol{s}}_2, \ldots, \tilde{\boldsymbol{s}}_T)$, $\mathbf{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_T)$ and $\mathbf{Q}$ is a diagonal matrix with typical (t,t)-element $Q_{tt} = \sum_{i=1}^{m} q_i^2 s_{it}$.

Furthermore, define

$$\mathbf{M} = \mathrm{E}[\mathbf{Q}] + \mathrm{E}\left[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}\right] \tag{6}$$

and note that $\mathbf{M}$ is invertible as $\mathbf{Q}$ is a diagonal matrix with positive entries and $\mathrm{E}\left[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}\right] = \mathrm{Cov}(\tilde{\mathbf{S}}'\boldsymbol{\lambda}) + \mathrm{E}(\tilde{\mathbf{S}}'\boldsymbol{\lambda})\mathrm{E}[\boldsymbol{\lambda}'\tilde{\mathbf{S}}]$, so that $\mathbf{M}$ is the sum of a positive definite matrix and a positive semi-definite matrix and therefore itself positive definite.

Minimizing (5) subject to $\sum_{t=1}^{T} w_t = 1$ yields the optimal weights

$$\mathbf{w} = \mathbf{M}^{-1}\mathrm{E}\left[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right] + \frac{\mathbf{M}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}}\left(1 - \boldsymbol{\iota}'\mathbf{M}^{-1}\mathrm{E}\left[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right]\right) \tag{7}$$

The MSFE given by (5) when applying the optimal weights (7) is

$$\begin{aligned}\mathrm{MSFE}(\mathbf{w}) = {} & \frac{\left(1 - \boldsymbol{\iota}'\mathbf{M}^{-1}\mathrm{E}\left[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right]\right)^2}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}} + \mathrm{E}\left[\tilde{\mathbf{s}}_{T+1}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right] \\ & -\mathrm{E}\left[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right]'\mathbf{M}^{-1}\mathrm{E}\left[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right] + \mathrm{E}\left[(\mathbf{q}'\mathbf{s}_{T+1})^2\right]\end{aligned} \tag{8}$$

In order to proceed, we need to specify the information set that is available to calculate the expectations in (7) and (8). 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 practical applications. However, the resulting analysis will prove to be highly informative. In a second step we will allow for uncertainty around the states. This 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, as Pesaran et al. (2013) show, the importance of the time of the break (or in our case, states) is of order $\mathcal{O}(1/T)$ for the optimal weights whereas that of λ is of order $\mathcal{O}(1/T^2)$.

3.1 Weights conditional on the states

Conditional on the states the expectation operator in (6), (7) and (8) can be omitted such that $\mathbf{M} = \mathbf{Q} + \tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}$ and $\mathrm{E}\left(\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}\right) = \tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{s}}_{T+1}$. 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 $s_{1,T+1} = 1$,

$$w_{11} = \frac{1}{T} \frac{1 + T\lambda^2 \pi_2}{\pi_2 q^2 + \pi_1 (1 + T\pi_2 \lambda^2)} \quad \text{if } s_{1t} = 1 \tag{9}$$

$$w_{12} = \frac{1}{T} \frac{q^2}{\pi_2 q^2 + \pi_1 (1 + T\pi_2 \lambda^2)} \quad \text{if } s_{2t} = 1 \tag{10}$$

where $w_{ij} = w(s_{i,T+1} = 1, s_{jt} = 1)$.

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

$$w_{21} = \frac{1}{T} \frac{1}{[\pi_2 q^2 + \pi_1 (1 + T\pi_2 \lambda^2)]} \quad \text{if } s_{1t} = 1 \tag{11}$$

$$w_{22} = \frac{1}{T} \frac{q^2 + T\lambda^2 \pi_1}{[\pi_2 q^2 + \pi_1 (1 + T\pi_2 \lambda^2)]} \quad \text{if } s_{2t} = 1 \tag{12}$$

Note that, conditional on the state of the future observation, the weights are symmetric under a relabeling of the states. Derivations are provided in Appendix A.1.1.

The weights are equivalent to the weights for the break point process developed by Pesaran et al. (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_{12} and w_{21} 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

$$\mathrm{E}[\sigma_2^{-2} e_{T+1}^2]_{\text{opt}} = \begin{cases} q^2(1 + w_{11}) & \text{if } s_{1,T+1} = 1 \\ 1 + w_{22} & \text{if } s_{2,T+1} = 1 \end{cases} \tag{13}$$

Table 1: Ratio between the expected MSFE for optimal weights and for standard MS weights, $T = 50$

λ	$q = 1$			$q = 0.5$		
	$\pi_2 = 0.1$	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 (13) and (14) 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.

We can compare this to the expected MSFE for standard Markov switching weights, which is given by

$$\mathrm{E}[\sigma_2^{-2} e_{T+1}^2]_{\mathrm{MS}} = \begin{cases} q^2(1 + \frac{1}{T\pi_1}) & \text{if } s_{1,T+1} = 1 \\ 1 + \frac{1}{T\pi_2} & \text{if } s_{2,T+1} = 1 \end{cases} \tag{14}$$

It is easy to show that $\mathrm{E}[\sigma_2^{-2} e_{T+1}^2]_{\mathrm{opt}} < \mathrm{E}[\sigma_2^{-2} e_{T+1}^2]_{\mathrm{MS}}$. Numerical examples of the magnitude of the improvement in MSFE is presented in Table 1, which shows that the improvements scale inversely with the break size. The intuition for this result is that the observations of the respective other state are increasingly useful for forecasting the smaller the difference between states. In fact, it is easy to show that the difference between (13) and (14) is maximized when $\lambda = 0$.

3.1.2 Three-state Markov switching models

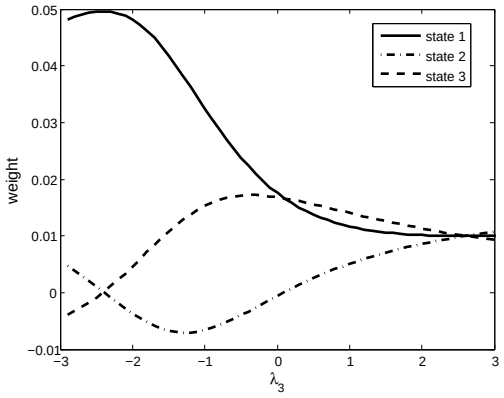
If $s_{j,T+1} = 1$, then define $q_i^2 = \sigma_i^2/\sigma_j^2$ and $\lambda_i^2 = (\beta_i - \beta_j)^2/\sigma_j^2$ where $i, j \in 1, 2, 3$. The weights are

$$\begin{aligned} w_{jj} &= \frac{1}{T} \frac{1 + T\sum_{i=1}^3 q_i^{-2}\lambda_i^2\pi_i}{\sum_{i=1}^3 q_i^{-2}\pi_i + T\sum_{i=1}^3\sum_{m=1}^3 q_i^{-2}q_m^{-2}\pi_i\pi_m\lambda_m(\lambda_m - \lambda_i)} \\ w_{jk} &= \frac{1}{T} \frac{q_k^{-2} + Tq_k^{-2}\sum_{i=1}^m q_i^{-2}\lambda_i\pi_i(\lambda_i - \lambda_k)}{\sum_{i=1}^3 q_i^{-2}\pi_i + T\sum_{i=1}^3\sum_{m=1}^3 q_i^{-2}q_m^{-2}\pi_i\pi_m\lambda_i(\lambda_i - \lambda_m)} \\ w_{jl} &= \frac{1}{T} \frac{q_l^{-2} + Tq_l^{-2}\sum_{i=1}^m q_i^{-2}\lambda_i\pi_i(\lambda_i - \lambda_l)}{\sum_{i=1}^3 q_i^{-2}\pi_i + T\sum_{i=1}^3\sum_{m=1}^3 q_i^{-2}q_m^{-2}\pi_i\pi_m\lambda_m(\lambda_i - \lambda_m)} \end{aligned} \tag{15}$$

Derivations are in Appendix A.1.2.

Figure 1 plots weights (15) for λ_3 over the range -3 to 3, and $\lambda_2 = -2.5$, $\pi_1 = 0.2$, $\pi_2 = \pi_3 = 0.4$, $T = 100$ and $q_1 = q_2 = 1$ for $s_{1,t+1} = 1$, that is,

Figure 1: Optimal weights for three state Markov switching model



Note: The graph depicts the optimal weights (15) when $s_{1,T+1} = 1$, for λ_3 over the range -3 to 3, $\lambda_2 = -2.5$, $T = 100$, $\pi_1 = 0.2$, and $\pi_2 = \pi_3 = 0.4$. The solid line gives the weights for the observations where $s_{1t} = 1$, the dash-dotted line those where $s_{2t} = 1$, and the dashed line those for $s_{3t} = 1$.

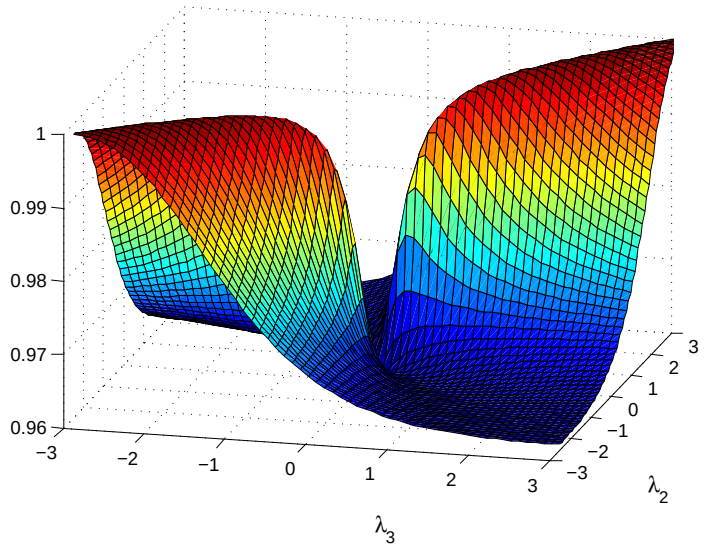
the future observation is known to be from the first state. The standard Markov switching weights are independent of the break size with $w_{11} = 0.05$ and $w_{1i} = 0$ for $i \neq 1$ and therefore not included in Figure 1.

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 relatively large difference between the mean of state 1 and that of the other states is due to the fact that the observations from the other states induce a large bias, and therefore 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 and the observations increasingly useful for forecasting. At $\lambda_3 = 0$, the first and the third state have identical means and the observations therefore receive equal weight.

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 state 1 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

Figure 2: MSFE of optimal weights relative to standard Markov switching weights



Note: 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.

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

$$\mathrm{E}[\sigma_i^{-2} e_{T+1}^2]_{\mathrm{opt}} = \frac{\sigma_j^2}{\sigma_i^2}(1 + w_{jj}) \tag{16}$$

with w_{jj} given in (15). For the Markov switching weights we have

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

Table 2: Maximum improvements in a three state model with $s_{j,T+1} = 1$

π_j	$T = 50$	100	200
0.1	0.8500	0.9182	0.9571
0.2	0.9273	0.9619	0.9805
0.5	0.9808	0.9902	0.9950

Note: The table reports the maximum improvement in relative MSFE (17).

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 between the states. 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.

The conditions under which the optimal weights result in the largest gains can be established formally. For given β_l and β_k, the value of β_j that implies the largest improvement in forecasts can be found by maximizing (16) with respect to β_j, which yields

$$\beta_j = \frac{q_k^2 \pi_l \beta_l + q_l^2 \pi_k \beta_k}{q_k^2 \pi_l + q_l^2 \pi_k}$$

Hence, the largest gain occurs when the regime to be forecast is located at the probability and variance weighted average of the other two regimes. The reason is that, in this case, the means of the other regimes are located such that they are optimally used to reduce variance and bias at the same time.

As an example consider $q_l = q_k = 1$, the weight w_{jj} is then equal to $1/T$ and the maximal expected improvement in MSFE of the optimal weights compared to the usual Markov switching forecast is given by

$$\frac{\mathrm{E}[\sigma_j^{-2} e_{T+1}^2]_{\mathrm{opt}}}{\mathrm{E}[\sigma_j^{-2} e_{T+1}^2]_{\mathrm{MS}}} = \frac{1 + \frac{1}{T}}{1 + \frac{1}{T\pi_j}} \tag{17}$$

where π_j is the percentage of observations in the regime to be forecast. Numerical values of (17), given in Table 2, show that optimal weights lead to larger improvements for smaller T and π_j. It is interesting to note that the maximum improvement is the same as in the two state case.

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 $s_{l,t} = 1$

$$w_{jl} = \frac{1}{T} \frac{q_l^{-2} \left(1 + T \sum_{i=1}^m q_i^{-2} \lambda_i \pi_i (\lambda_i - \lambda_l)\right)}{\sum_{i=1}^m q_i^{-2} \pi_i + T \sum_{i=1}^m \sum_{k=1}^m q_i^{-2} q_k^{-2} \pi_i \pi_k \lambda_i (\lambda_i - \lambda_k)} \tag{18}$$

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

$$\mathrm{E}[\sigma_i^{-2} e_{T+1}^2]_{\text{opt}} = \frac{\sigma_j^2}{\sigma_i^2}(1 + w_{jj})$$

The derivation of the weights and the MSFE is in Appendix A.1.2. The maximum gain is realized when the mean β_j satisfies

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

The minimum MSFE is then

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

and when the variances are equal this reduces to

$$\mathrm{E}[\sigma_i^{-2} e_{T+1}^2] = 1 + \frac{1}{T}$$

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

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

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

$$\begin{aligned} Tw =& \left(\frac{1}{\pi_1} - \frac{1}{T} \frac{q^2}{\lambda^2 \pi_1^2} \right) s_{1t} s_{1,T+1} + \frac{1}{T} \frac{q^2}{\lambda^2 \pi_1 \pi_2} s_{2t} s_{1,T+1} + \\ &+ \frac{1}{T} \frac{1}{\lambda^2 \pi_1 \pi_2} s_{1t} s_{2,T+1} + \left(\frac{1}{\pi_2} - \frac{1}{T} \frac{1}{\lambda^2 \pi_2^2} \right) s_{2t} s_{2,T+1} + \mathcal{O}(T^{-2}) \end{aligned} \tag{19}$$

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 et al. (2013) for the structural break case where the first order approximation gives zero weight to pre-break observations and equally weight the post-break observations. This result in (19) 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 the weights that do not assume knowledge of the states. The expectations in (7) 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.

In order to analyze the theoretical properties of the optimal weights, we need analytic expressions for the weights, 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, however, 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 outside the one under consideration.

Denote the probability of state i occurring at time t by ξ_{it}. The expectations in (7) and (8) are then

$$\mathrm{E}[s_{it}s_{j,t+m}] = \begin{cases} \xi_{it} & \text{if } i = j \\ \xi_{it}\xi_{j,t+m} & \text{if } i \neq j, m \geqslant 0 \end{cases}$$

We will initially focus on the two state case, but we will extend the analysis to m states below.

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 (7) 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 $\boldsymbol{\xi} = (\xi_{21}, \xi_{22}, \ldots, \xi_{2T})$, $\boldsymbol{\Xi} = \text{diag}(\boldsymbol{\xi})$, $\mathbf{V} = \boldsymbol{\Xi}(\mathbf{I} - \boldsymbol{\Xi})$, 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

$$\mathbf{M}^{-1} = \mathbf{D}^{-1} - \frac{\lambda^2}{1 + \lambda^2\boldsymbol{\xi}'\mathbf{D}^{-1}\boldsymbol{\xi}}\mathbf{D}^{-1}\boldsymbol{\xi}\boldsymbol{\xi}'\mathbf{D}^{-1} \tag{20}$$

Using (7) and (20) yields

$$\mathbf{w} = \lambda^2\xi_{2,T+1}\mathbf{M}^{-1}\boldsymbol{\xi} + \frac{\mathbf{M}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}}\left[1 - \lambda^2\xi_{2,T+1}\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\xi}\right] \tag{21}$$

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

$$w_t = \frac{d_t\left[1 + \lambda^2\sum_{t'=1}^{T} d_{t'}(\xi_{2t} - \xi_{2t'})(\xi_{2T+1} - \xi_{2t'})\right]}{\sum_{t'=1}^{T} d_{t'} + \lambda^2\left[\left(\sum_{t'=1}^{T} d_{t'}\xi_{2t'}^2\right)\left(\sum_{t'=1}^{T} d_{t'}\right) - \left(\sum_{t'=1}^{T} d_{t'}\xi_{2t'}\right)^2\right]} \tag{22}$$

The expected MSFE can be calculated from (5) and reduces to

$$\text{E}[\sigma_2^{-2}e_{T+1}^2] = \left(1 + \lambda^2\xi_{2,T+1}(1-\xi_{2,T+1})\right)(1 + w_{T+1}) \tag{23}$$

where w_{T+1} is given by (22).

When T is large, weights (22) can be written as

$$w_t = \tilde{d}_t\frac{\sum_{t'=1}^{T}\tilde{d}_{t'}\left(\xi_{2,T+1} - \xi_{2t'}\right)\left(\xi_t - \xi_{2t'}\right)}{\sum_{t'=1}^{T}\tilde{d}_{t'}\left(\xi_{t'} - \sum_{t''=1}^{T}\tilde{d}_{t''}\xi_{2t''}\right)^2} + \mathcal{O}(T^{-2}) \tag{24}$$

where $\tilde{d}_t = d_t/(\sum_{t'=1}^{T} d_t)$. Derivations are provided in Appendix A.2.1. While the weights in (22) and (24) provide closed form solutions, interpretation can be aided by momentarily making the simplifying assumption of constant state variances.

Constant state variance The interpretation of (22) and (24) 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. In order 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

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

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, (22) can be written as

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

and the large T approximation (24) as

$$w_t = \frac{1}{T} + \frac{(\xi_{2,T+1} - \bar{\xi}_2)(\xi_{2t} - \bar{\xi}_2)}{T(\bar{\xi}_1\bar{\xi}_2 - \sigma_s^2)} \tag{26}$$

The standard Markov switching weights can be expressed as

$$w_t^{\text{MS}} = \frac{1}{T} + \frac{(\xi_{2,T+1} - \bar{\xi}_2)(\xi_{2,t} - \bar{\xi}_2)}{T\bar{\xi}_1\bar{\xi}_2} \tag{27}$$

see Appendix A.2.2. From a comparison of (26) and (27) 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 (26) compared to (27).

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

$$\begin{aligned}
\mathrm{E}[\sigma_2^{-2}e_{T+1}^2]_{\text{opt}} &= \left[1 + \lambda^2\xi_{2,T+1}(1 - \xi_{2,T+1})\right] \\
&\quad \times \left(1 + \frac{1}{T} + \frac{\lambda^2(\xi_{2,T+1} - \bar{\xi}_2)^2}{1 + \lambda^2\sigma_s^2 + \lambda^2 T(\bar{\xi}_2(1 - \bar{\xi}_2) - \sigma_s^2)}\right) \qquad (28) \\
\mathrm{E}[\sigma_2^{-2}e_{T+1}^2]_{\text{MS}} &= 1 + \lambda^2\xi_{2,T+1}(1 - \xi_{2,T+1}) + \frac{1}{T}(\lambda^2\sigma_s^2 + 1) \qquad (29) \\
&\quad + \left(\frac{\xi_{2,T+1} - \bar{\xi}_2}{\bar{\xi}_2(1 - \bar{\xi}_2)}\right)^2 \left[\frac{1}{T}(\bar{\xi}_2(1 - \bar{\xi}_2) - \sigma_s^2)\left(\lambda^2\sigma_s^2 + 1\right) + \lambda^2\sigma_s^4\right]
\end{aligned}$$

The MSFE for the optimal weights is derived from (23) by substituting in the weights in (22) and using the fact that $\tilde{d}_t = 1/T$ and $d_t = d$, for $t = 1, \ldots, T+1$. The MSFE for the Markov switching weights is derived in Appendix A.2.2.

Table 3 displays the improvements in forecast performance expressed as the ratio of (28) over (29) 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 3 show that the improvement can be as large as 11.3% for the range of parameter values considered here.

Table 3: 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.

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

$$\mathrm{E}[\sigma_0^2 e_{T+1}^2]_{\text{opt}} = 1 + \lambda^2 \xi_{2,T+1}(1 - \xi_{2,T+1}) + \mathcal{O}(T^{-1}) \tag{30}$$

and that under standard Markov switching weights is

$$\mathrm{E}[\sigma_0^2 e_{T+1}^2]_{\text{MS}} = 1 + \lambda^2 \xi_{2,T+1}(1 - \xi_{2,T+1}) + \left(\frac{\xi_{2,T+1} - \bar{\xi}_2}{\bar{\xi}_2 \bar{\xi}_1}\right)^2 \lambda^2 \sigma_s^4 + \mathcal{O}(T^{-1}) \tag{31}$$

The difference between (31) and (30) 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 that we can write

$$\mathrm{E}[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}] = \mathrm{E}[\tilde{\mathbf{S}}]'\boldsymbol{\lambda}\boldsymbol{\lambda}'\mathrm{E}[\tilde{\mathbf{S}}] + \mathbf{A}$$

where, conditional on the state probabilities, ξ_{jt}, $j = 1, 2, \ldots, m$,

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

and $\mathbf{\Xi}_j$ 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 the inverse of $\mathbf{M}$ is

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

We can use (7) to derive the weights similar to the case of the two-state weights

$$w_t = \frac{d_t^{(m)} \left[1 + \left(\sum_{t'=1}^{T} d_{t'}^{(m)} (\tilde{\xi}_t - \tilde{\xi}_{t'})(\tilde{\xi}_{T+1} - \tilde{\xi}_{t'}) \right) \right]}{\sum_{t'=1}^{T} d_{t'}^{(m)} + \left(\sum_{t'=1}^{T} d_{t'}^{(m)} \tilde{\xi}_{t'}^2 \right) \left(\sum_{t'=1}^{T} d_{t'}^{(m)} \right) - \left(\sum_{t'=1}^{T} d_{t'}^{(m)} \tilde{\xi}_{t'} \right)^2} \tag{32}$$

where now we have

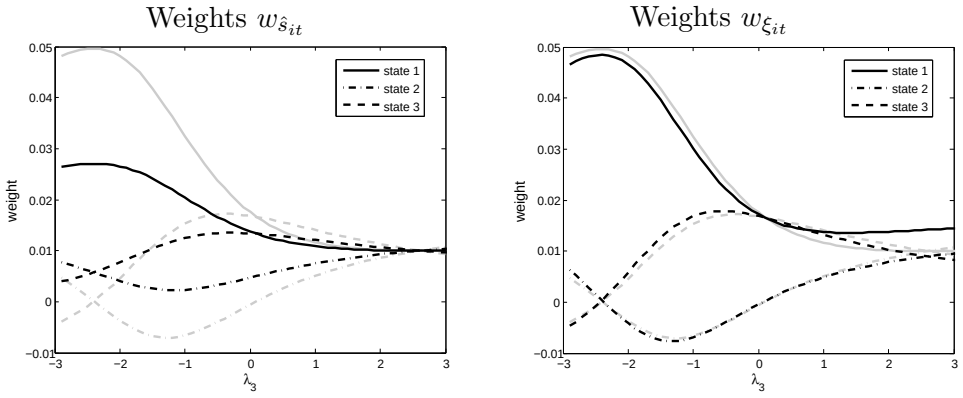
$$\begin{aligned} d_t^{(m)} &= \left[\sum_{j=1}^{m} q_j^2 \xi_{jt} + \sum_{j=2}^{m} \lambda_j^2 \xi_{jt} - \left(\sum_{j=2}^{m} \lambda_j \xi_{jt} \right)^2 \right]^{-1} \\ &= \left[\sum_{j=1}^{m} (q_j^2 + \lambda_j^2) \xi_{jt} - \left(\sum_{j=2}^{m} \lambda_j \xi_{jt} \right)^2 \right]^{-1} \\ \tilde{\xi}_t &= \sum_{j=2}^{m} \xi_{jt} \lambda_j \end{aligned}$$

and where we have used the fact that $\lambda_1 = 0$.

Examples of weights for a three state Markov switching model over a range of λ_3 for $T = 100$, $\pi_1 = 0.2, \pi_2 = \pi_3 = 0.4$ and $\lambda_2 = -2.5$ are plotted in Figure 3. For simplicity of exposition, we assume that the state probabilities are identical for each state in the sense that a prevailing state has $\xi_{it} = 0.8$ and other states $\xi_{jt} = 0.1$. The light gray lines represents the optimal weights (15) that are conditional on the states. The graph on the left plots weights (15) 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 (32) 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

Figure 3: Optimal weights for three state Markov switching model



Note: The graphs depicts the optimal weights (15) using s_{it} in both plots as the lighter, gray lines. In the left plot the darker lines are the optimal weights (15) using ξ_{it} in place of s_{it}. In the right plot the darker lines are the weights (32), when $\hat{\boldsymbol{\xi}}_{T+1} = [0.8, 0.1, 0.1]'$ for λ_3 over the range -3 to 3, $\lambda_2 = -2.5$, $T = 100$, $\pi_1 = 0.2$, and $\pi_2 = \pi_3 = 0.4$. The solid line gives the weights for the observations where $\hat{\boldsymbol{\xi}}_t = [0.8, 0.1, 0.1]'$, the dashed line those where $\hat{\boldsymbol{\xi}}_t = [0.1, 0.8, 0.1]'$, and the dash-dotted line those for $\hat{\boldsymbol{\xi}}_t = [0.1, 0.1, 0.8]'$.

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 (32). 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.

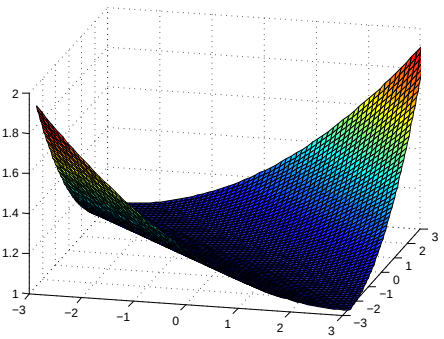
The MSFE for both the Markov switching and the optimal weights is displayed in Figure 4. As might be expected based on the weights shown in Figure 3, the optimal weights achieve an MSFE, displayed in Figure 4(b), that closely corresponds to the MSFE from the conditional weights in Figure 2. This contrasts sharply with the MSFE for standard Markov switching weights in Figure 4(a). When λ_2 and λ_3 are large and nearly equal, the MSFE shows a sharp increase towards values that are almost twice that of the MSFE of the optimal weights. Hence, for these value the relative MSFE, displayed in Figure 4(c), shows substantial improvements.

3.3 Estimating state covariances from the data

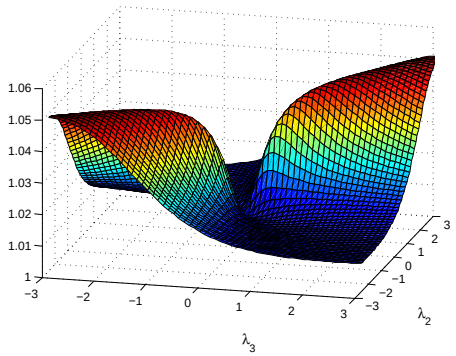
Above, we derived weights conditional on the state probabilities, in which case we can write the expectation of the product of two states as $\mathrm{E}[s_{it}s_{j,t+m}] = \xi_{it}\xi_{j,t+m}$. While this assumption allows us to find an explicit inverse of the

Figure 4: (Relative) MSFE under Markov switching and optimal weights

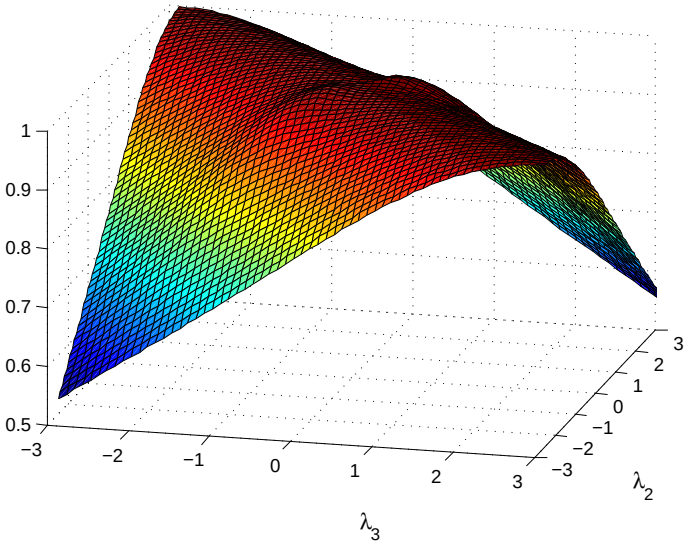
(a) MSFE under standard weights



(b) MSFE under optimal weights



(c) Relative MSFE



Note: Figure (a) displays the MSFE of the standard Markov switching weights and Figure (b) that of the optimal weights conditional on the probabilities for $T = 100$, $\pi_1 = 0.2$, $\pi_2 = \pi_3 = 0.4$ for a range of values for λ_2 and λ_3. Figure (c) displays the ratio of the MSFE of the optimal weights relative to that of the standard MSFE forecast.

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 $\hat{\mathbf{M}}$ directly from the data.

To estimate $\hat{\mathbf{M}}$ directly from the data, we now condition on the information set up 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 MCMC sampler Kim and Nelson (1999). The second term can be written as

$$p(s_{it} = 1|s_{j,t+m} = 1, \Omega_T) = \frac{\xi^i_{t|t}}{\xi^j_{t+m|t+m-1}} \left[\left(\prod_{l=1}^{m-1} (\mathbf{P}'\mathbf{A}_{t+l}) \right) \mathbf{P}' \right]_{i,j} \tag{33}$$

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 (33) can be found in Appendix A.2.4. Using these expressions we can calculate the expectations in (7). 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

$$\mathrm{E}[\tilde{\mathbf{s}}_t\tilde{\mathbf{s}}'_{t+k}] = \boldsymbol{\Xi}_{t|t}\boldsymbol{\Xi}^* \left(\boldsymbol{\Xi}_{t+k|T} \div \boldsymbol{\Xi}_{t+k|t+k-1} \right)$$

where $\boldsymbol{\Xi}_{t|t}$ is an $m-1 \times m-1$ matrix with typical i,i element $\hat{\xi}_{it|t}$ 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

$$\begin{aligned} \mathrm{E}[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}]_{t,t} &= \boldsymbol{\lambda}'\mathrm{diag}(\mathrm{E}[\tilde{\mathbf{s}}_t])\boldsymbol{\lambda} \\ \mathrm{E}[\tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}}]_{t,t+k} &= \boldsymbol{\lambda}'\mathrm{E}[\tilde{\mathbf{s}}_t\tilde{\mathbf{s}}'_{t+k}]\boldsymbol{\lambda} \end{aligned} \tag{34}$$

Using (34) in (7) 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 j-th 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

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

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 $\boldsymbol{\Lambda}_{ij} = \boldsymbol{\lambda}_i\boldsymbol{\lambda}'_j$. The expected MSFE is given by

$$\begin{aligned}
\mathrm{E}[\sigma_m^{-2}e_{T+1}^2] &= \sum_{i=1}^{m} \mathrm{E}[s_{i,T+1}]\mathbf{x}'_{T+1}\boldsymbol{\Lambda}_{ij}\mathbf{x}_{T+1} + \sum_{i=1}^{m} \mathrm{E}[s_{i,T+1}]q_i^2\varepsilon_{T+1}^2 \qquad (35)\\
&+\mathbf{x}'_{T+1}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\sum_{i=1}^{m}\sum_{j=1}^{m} \mathrm{E}[(\mathbf{X}'\mathbf{W}\mathbf{S}_i\mathbf{X})\boldsymbol{\Lambda}_{ij}(\mathbf{X}'\mathbf{S}_j\mathbf{W}\mathbf{X})](\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_{T+1}\\
&+\mathbf{x}'_{T+1}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\sum_{i=1}^{m} q_i^2\mathbf{X}'\mathbf{W}\mathrm{E}[\mathbf{S}_i]\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_{T+1}\\
&-2\mathbf{x}'_{T+1}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\sum_{i=1}^{m}\sum_{j=1}^{m} \mathrm{E}[\mathbf{X}'\mathbf{W}\mathbf{S}_i\mathbf{X}\boldsymbol{\Lambda}_{ij}s_{j,T+1}]\mathbf{x}_{T+1}
\end{aligned}$$

As in the case of structural breaks analyzed by Pesaran et al. (2013), large sample approximations to (35) 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, (35) reduces to

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

Maximizing (4) subject to $\boldsymbol{\iota}'\mathbf{w} = 1$ leads to the following first order condi-

tions for $\mathbf{w}$,

$$\frac{\partial \mathrm{E}\left[\sigma_m^{-2} e_{T+1}^2\right]}{\partial \mathbf{w}} = 2\sum_{i=1}^{m}\sum_{j=1}^{m} \mathbf{x}_{T+1}' \mathbf{\Lambda}_{ij} \mathbf{x}_{T+1} \mathrm{E}[\mathbf{s}_i \mathbf{s}_j'] \mathbf{w} + 2\sum_{i=1}^{m} \mathbf{x}_{T+1}' \mathbf{\Omega}_{XX}^{-1} \mathbf{x}_{T+1} q_i^2 \mathrm{E}[\mathbf{S}_i] \mathbf{w} - 2\sum_{i=1}^{m}\sum_{j=1}^{m} \mathbf{x}_{T+1}' \mathbf{\Lambda}_{ij} \mathbf{x}_{T+1} \mathrm{E}[\mathbf{s}_i s_{j,T+1}] + \theta \boldsymbol{\iota} = 0$$

Define $\boldsymbol{\phi}_i = \mathbf{x}_{T+1}' \boldsymbol{\lambda}_i / (\mathbf{x}_{T+1} \mathbf{\Omega}_{XX}^{-1} \mathbf{x}_{T+1})^{1/2}$, solving for the weights yields

$$\mathbf{w} = \left(\left[\mathrm{E}[\tilde{\mathbf{S}}' \boldsymbol{\phi} \boldsymbol{\phi}' \tilde{\mathbf{S}}] + \mathrm{E}[\mathbf{Q}] \right] \right)^{-1} \mathrm{E}[\tilde{\mathbf{S}} \boldsymbol{\phi} \boldsymbol{\phi}' \tilde{\mathbf{s}}_{T+1}] - \theta \boldsymbol{\iota}$$

which is identical to (7) with the exception that $\boldsymbol{\lambda}$ is replaced by $\boldsymbol{\phi}$. Hence,

$$\mathbf{w} = \mathbf{M}^{-1} \mathrm{E}\left[\tilde{\mathbf{S}}' \boldsymbol{\phi} \boldsymbol{\phi}' \tilde{\mathbf{s}}_{T+1}\right] + \frac{\mathbf{M}^{-1} \boldsymbol{\iota}}{\boldsymbol{\iota}' \mathbf{M}^{-1} \boldsymbol{\iota}} \left(1 - \boldsymbol{\iota}' \mathbf{M}^{-1} \mathrm{E}\left[\tilde{\mathbf{S}}' \boldsymbol{\phi} \boldsymbol{\phi}' \tilde{\mathbf{s}}_{T+1}\right]\right) \tag{36}$$

with $\mathbf{M} = \mathrm{E}[\mathbf{Q}] + \mathrm{E}\left(\tilde{\mathbf{S}}' \boldsymbol{\phi} \boldsymbol{\phi}' \tilde{\mathbf{S}}\right)$ and $\mathbf{Q}$ a diagonal matrix with typical (t,t)-element $Q_{tt} = \sum_{i=1}^{m} q_i^2 s_{it}$ and

$$\mathrm{E}[\sigma_m^{-2} e_{T+1}^2]_{\mathrm{opt}} = \frac{\left(1 - \boldsymbol{\iota}' \mathbf{M}^{-1} \mathrm{E}\left[\tilde{\mathbf{S}}' \boldsymbol{\phi} \boldsymbol{\phi}' \tilde{\mathbf{s}}_{T+1}\right]\right)^2}{\boldsymbol{\iota}' \mathbf{M}^{-1} \boldsymbol{\iota}} + \mathrm{E}\left[\tilde{\mathbf{s}}_{T+1}' \boldsymbol{\phi} \boldsymbol{\phi}' \tilde{\mathbf{s}}_{T+1}\right] - \mathrm{E}\left[\tilde{\mathbf{S}}' \boldsymbol{\phi} \boldsymbol{\phi}' \tilde{\mathbf{s}}_{T+1}\right]' \mathbf{M}^{-1} \mathrm{E}\left[\tilde{\mathbf{S}}' \boldsymbol{\phi} \boldsymbol{\phi}' \tilde{\mathbf{s}}_{T+1}\right] + \mathrm{E}\left[(\mathbf{q}' \mathbf{s}_{T+1})^2\right]$$

All formulae derived for the location model above can be straightforwardly extended to allow for exogenous regressors by replacing $\boldsymbol{\lambda}$ with $\boldsymbol{\phi}$.

5 Evidence from Monte Carlo experiments

5.1 Set up 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 with $m = 2$ and $m = 3$ states. We set $\sigma_2^2 = 0.25$ and use a range of values for λ_i. We will distinguish experiments based on the size of the switches, λ_i.

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 $\pi_i = \pi = 1/m$ equal for all states, 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, $\mathbf{s}_1 \sim \text{binomial}(1, \boldsymbol{\pi})$. Subsequent states are drawn as $\mathbf{s}_t \sim \text{binomial}(1, \mathbf{p}_t)$ where $\mathbf{p}_t = \mathbf{P}\mathbf{s}_{t-1}$. In order to identify all

parameters in the models, we require that at least 10% of the observations occupies each regime.

The first set of the Monte Carlo experiments analyzes two state models with only a constant, so that $k = 1$ and $x_t = 1$. To investigate the influence of the sample size T on the results we present results for $T = 50$ and $T = 100$. We then add an exogenous regressor to a two state model, such that $\mathbf{x}_t = [1, z_t]'$ where $z_t \sim \mathrm{N}(0, 0.25)$. The variance of z_t is chosen such that the centered R^2 is roughly equal to a model with no exogenous regressors. We then continue with a three state model, where we restrict the analysis to the simple mean only model for computational efficiency.

The estimation is performed using the EM algorithm (Dempster et al. 1977) as outlined in Hamilton (1994). The algorithm stops when the increase in log-likelihood falls below 10^{-8}. In order to avoid situations where the EM algorithm assigns all probability to one state vector, in which case at least one of the parameters β_i is not identified, we impose $\frac{1}{T}\sum_{t=1}^{T} \hat{\xi}^i_{t|T} > 0.05$ for all i. If the restriction is not satisfied we simulate a new state vector and generate a new data set.

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

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

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

Using these weights the optimal forecast is constructed as

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

where $\mathbf{W}$ is a diagonal matrix with typical diagonal element $w_{\cdot,t}$ where $w_{\cdot,t} \in \left\{w_{\hat{s},t}, w_{\hat{\xi},t}, w_{\hat{\mathbf{M}},t}\right\}$. The results are presented as the ratio of the MSFE of the optimally weighted forecast to that of the standard Markov switching forecast.

The results will be separated for different values of λ_i to show the effect of the break size. Furthermore, the performance of the weights $w_{\hat{\xi}}$ has

been shown to depend on the variance of the smoothed probability vector. Thus, we also separate the results based on the normalized variance of the smoothed probability vector given by

$$\tilde{\sigma}^2_{\hat{\xi}} = \frac{\frac{1}{T}\sum_{t=1}^{T}\hat{\xi}^{(i)}_{t|T}(1-\hat{\xi}^{(i)}_{t|T})}{\frac{1}{T}\sum_{t=1}^{T}\hat{\xi}^{(i)}_{t|T}\frac{1}{T}\sum_{t=1}^{T}(1-\hat{\xi}^{(i)}_{t|T})} \tag{37}$$

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

5.2 Monte Carlo results

5.2.1 Monte Carlo results for two state models

The Monte Carlo results for the simple model with two states are reported in Table 4. The top panel concentrates on models with a break in mean only. Weights $w_{\hat{s}}$ should improve the most when the break size is small. This is supported by the simulation. We see that this improvement is largest when the uncertainty around the states is small. The induced estimation uncertainty outweighs the benefits of the optimal weights when λ takes larger values. This contrasts with the results for the weights $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$. When $\lambda = 1$, the estimation uncertainty in the parameters outweighs the potential improvement in MSFE but as λ increases the improvements are quite substantial, especially when the variance in the smoothed probability vector is high. While the differences between the forecast performance of $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$ are small, in these settings the forecasts from $w_{\hat{\xi}}$ generally beat those from $w_{\hat{\mathbf{M}}}$.

Theoretically, the weights $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$ are expected to perform better when the sample size is larger. These findings are supported by the results for $T = 100$ that are reported in Table 4. The results show that the improvements increase with large λ and high $\tilde{\sigma}^2_{\hat{\xi}}$ and that these improvements are more pronounced in larger data sets. The weights $w_{\hat{s}}$ lead to forecasts that improve less on the standard weights for the larger data set, which confirms the theoretical results above that, asymptotically, these weights are identical.

The lower panel of Table 4 reports the results for a model that also contains a break in the variance. This break is such that the variance in regime 1 is the same as before, but the variance in regime 2 is increased. This should decrease the improvements, since the average break size standardized with the variance decreases. This decrease is indeed observed, but substantial improvements remain in the same parameter regions where the weights under constant variance perform well.

Table 4: Monte Carlo results: two states, intercept only models

		$T = 50$			$T = 100$		
λ	$\tilde{\sigma}^2_{\hat{\xi}_{\|T}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
				$q^2 = 1$			
1	0.0-0.1	0.983	1.004	1.005	0.993	1.005	1.005
	0.1-0.2	0.990	1.021	1.024	0.997	1.013	1.022
	0.2-0.3	0.996	1.027	1.033	0.999	1.019	1.032
	0.3-0.4	0.998	1.028	1.035	1.000	1.024	1.037
2	0.0-0.1	0.995	1.008	1.015	0.999	1.005	1.023
	0.1-0.2	1.001	1.005	1.018	1.002	0.994	1.034
	0.2-0.3	1.003	0.987	0.999	1.003	0.977	1.004
	0.3-0.4	1.004	0.982	0.991	1.004	0.961	0.973
3	0.0-0.1	1.000	0.998	1.013	1.000	0.997	1.022
	0.1-0.2	1.005	0.974	0.988	1.005	0.961	0.993
	0.2-0.3	1.006	0.957	0.965	1.007	0.920	0.944
	0.3-0.4	1.006	0.957	0.956	1.007	0.892	0.912
				$q^2 = 2$			
1	0.0-0.1	0.985	1.000	1.001	0.992	1.001	1.001
	0.1-0.2	0.990	1.009	1.011	0.996	1.009	1.013
	0.2-0.3	0.996	1.022	1.027	0.999	1.014	1.021
	0.3-0.4	0.998	1.017	1.021	1.001	1.018	1.026
2	0.0-0.1	0.993	1.006	1.008	0.998	1.005	1.019
	0.1-0.2	0.999	1.012	1.023	1.002	0.999	1.030
	0.2-0.3	1.003	1.001	1.015	1.003	0.992	1.021
	0.3-0.4	1.004	0.993	0.999	1.003	0.987	1.003
3	0.0-0.1	0.998	1.003	1.009	1.000	0.999	1.027
	0.1-0.2	1.003	0.986	1.010	1.003	0.980	1.025
	0.2-0.3	1.007	0.958	0.977	1.007	0.946	0.962
	0.3-0.4	1.010	0.942	0.943	1.007	0.920	0.939

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 $\varepsilon_t \sim \mathrm{N}(0,1)$, $\sigma_2^2 = 0.25$, $q^2 = \sigma_1^2/\sigma_2^2$. Column labels: $\lambda = (\beta_2 - \beta_1)/\sigma_2$, $\tilde{\sigma}^2_{\hat{\xi}_{|T}}$ is the normalized variance in of the smoothed probability vector (37). $w_{\hat{s}}$: 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 $\hat{\mathbf{M}}$.

Table 5: Monte Carlo results: two states, models with exogenous regressors

		$T = 50$			$T = 100$		
λ	$\tilde{\sigma}^2_{\hat{\xi}_{\vert T}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
1	0.0-0.1	0.962	0.988	0.986	0.986	1.002	1.002
	0.1-0.2	0.973	1.021	1.001	0.993	1.014	1.018
	0.2-0.3	0.991	1.025	1.021	0.999	1.023	1.028
	0.3-0.4	0.995	1.030	1.028	1.000	1.026	1.032
2	0.0-0.1	0.990	1.000	1.002	0.999	1.003	1.013
	0.1-0.2	1.004	1.008	1.016	1.006	0.997	1.031
	0.2-0.3	1.011	0.999	1.013	1.011	0.978	1.009
	0.3-0.4	1.012	0.986	0.999	1.019	0.956	0.991
3	0.0-0.1	1.005	1.004	1.013	1.005	1.001	1.027
	0.1-0.2	1.018	0.998	1.026	1.020	0.979	1.033
	0.2-0.3	1.031	0.983	1.010	1.043	0.935	1.008
	0.3-0.4	1.020	0.969	0.991	1.051	0.919	0.958

Note: The table reports the ratio of the MSFE of the optimal asymptotic weights to that of the Markov switching weights. DGP: $y_t = x_t'\beta_1 + \sigma\left(x_t'\lambda s_{2t} + \varepsilon_t\right)$ where $\varepsilon_t \sim \text{NID}(0,1)$. Also $\sigma^2 = 0.25$, $\beta_1 = 1$ and $x_t = [1, z_t]$ where $z_t \sim \text{N}(0, 0.25)$. For the column labels see the footnote of Table 4.

In Table 5 we display the results for models that include an exogenous regressor. The optimal forecast are obtained by using an asymptotic approximation to the covariance matrix as in (36). As the ratio of parameters to estimate versus the number of observations increases, the performance of the weights $w_{\hat{s}}$ increases even if improvements are small. The improvements for weights $w_{\hat{\mathbf{M}}}$ are less marked.

From the results in Table 6 we see that the conclusions from the two state models carry over to the three state models. Again, sizeable improvements are made for $w_{\hat{\xi}}$ and $w_{\hat{\mathbf{M}}}$ when $\tilde{\sigma}^2_{\hat{\xi}}$ is large and both break sizes λ_{21} and λ_{31} are large. These improvements increase when the sample size increases from $T = 50$ to $T = 100$.

6 Application to US GNP

Modeling the US business cycle was the original application of the Markov switching model by Hamilton (1989), and business cycle analysis has remained one of the most important applications. Different variants of Markov

Table 6: Monte Carlo results: three states, intercept only models

		$T = 50$			$T = 100$		
$\{\lambda_{31}, \lambda_{21}\}$	$\tilde{\sigma}^2_{\hat{\xi}_{\mid T}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
$\{2, 1\}$	0.0-0.1	0.996	1.029	1.026	0.998	1.025	1.027
	0.1-0.2	0.997	1.031	1.035	0.999	1.027	1.046
	0.2-0.3	0.999	1.028	1.035	1.000	1.012	1.027
	0.3-0.4	1.001	1.013	1.017	1.001	1.007	1.018
$\{3, 1\}$	0.0-0.1	0.998	1.016	1.014	0.999	1.011	1.026
	0.1-0.2	1.000	1.008	1.013	1.001	0.998	1.013
	0.2-0.3	1.002	0.990	0.993	1.002	0.971	0.986
	0.3-0.4	1.004	0.969	0.966	1.003	0.939	0.953
$\{4, 2\}$	0.0-0.1	0.999	1.011	1.011	1.000	1.005	1.019
	0.1-0.2	1.001	0.997	0.999	1.001	0.987	1.007
	0.2-0.3	1.003	0.973	0.971	1.003	0.943	0.963
	0.3-0.4	1.003	0.954	0.951	1.004	0.882	0.876

Note: The table reports the ratio of the MSFE of the optimal weights to that of the Markov switching weights. For details see Table 4.

switching models have been used to analyze business cycles, and we will use the classification scheme of Krolzig (1997). Applications of different models to US GNP can be found in Clements and Krolzig (1998), Krolzig (1997) and Krolzig (2000), which show that the Markov switching model is frequently outperformed in terms of MSFE by a simple linear AR model. We use a pseudo-out-of-sample forecast exercise to analyze whether optimal weights improve the forecast accuracy of Markov switching models for US GNP, and whether using optimal weights improves the forecasts of Markov switching models over those from linear alternatives.

The model by Hamilton (1989) is an example of a Markov Switching in mean model with non-switching autoregressive regressors. This class of models is denoted as MSM(m)-AR(p) by Krolzig (1997), where Hamilton's model takes $m = 2$ and $p = 4$:

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

Here, y_t depends on the current state but also on the previous p states. If the model has a state dependent variance σ_{s_t} it is denoted by MSMH(m)-AR(p).

Clements and Krolzig (1998) find that a three state model which has a switching intercept instead of a switching mean, and a state dependent variance also performs well in terms of business cycle description and forecast

performance. This class of models is denoted by MSIH(m)-AR(p) and the model in Clements and Krolzig (1998) takes $m = 3$ and $p = 4$:

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

Note that both these models fit in the framework of the intercept only model by simply moving the autoregressive regressors to the left hand side after we estimated the coefficients. On the right hand side remains the constant and we can use the finite sample expressions derived for the intercept only model. Estimation is performed using the EM algorithm, which uses the implementation of Hamilton (1994) with the extensions to estimate the MSM models suggested by Krolzig (1997). We have investigated the performance of the optimal weights for these models in Monte Carlo experiments and the results from the intercept only model in Section 5 carry over to these models. The results are reported in Appendix B.

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 at each step all models are re-estimated to include all the available data at that point in time. We select the Markov switching model that delivers the best forecast, where the selection is based on the historic forecast performance under standard weights as measured by the MSFE. Using this model, we then compare the pseudo out-of-sample forecasts from the standard weights to those from the optimal weights. We report the ratio of the MSFE of the optimal weights relative to the standard weights together with the Diebold and Mariano (1995) test statistic of equal predictive accuracy.

The data we use are (log changes in) the US GNP series between 1947Q1 and 2014Q1 obtained from the Federal Reserve Economic Data (FRED). The data is seasonal adjusted. In total, the series consists of 269 observations. Because we analyze log changes, we lose one observation. To keep the sample size the same for models of all lag lengths, we start the estimation in 1948Q2 so that the models that use lagged dependent variables can all be initialized based on available data.

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 based on 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 all weights are made with this model. In this way no information is used that is not available to

Table 7: 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**

Note: The second column 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.

a researcher at that point in time. Then we add the next period to our estimation and cross-validation sample, select the minimum MSFE model, and construct the next forecast. Based on the model selection procedure the MSM(3)-AR(1) model is selected for the entire forecast period.

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).

The forecasting performances of the standard and optimal weights are reported in Table 7. 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 break size $\hat{\lambda}_{21} = 2.28$ and $\hat{\lambda}_{31} = 4.23$. The average minimum normalized variance of the smoothed probability vector $\tilde{\sigma}^2_{\hat{\xi}_{|T}} = 0.20$. 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 6.

It is interesting to also compare forecast performance in subperiods. In

Table 8: 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.

the first subperiod, 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 subperiod, 1993Q2–2003Q1, which largely covers the great moderation, only $w_{\hat{\mathbf{M}}}$ offers a modest improvement. In the last subperiod, 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.

Further insight can be gained by comparing the accuracy of the Markov switching forecasts with that from linear models, which here are the random walk and AR(p) models with $p = 1, 2, 3, 4$. We select the best AR(p) 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 8. 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 the states. This contrasts with the forecast based on optimal weights conditional on state probabilities that beat the linear models, even if the difference is not significant at conventional levels.

The forecasts over the subperiods reveal that in the first subperiod, 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 subperiod no Markov switching forecast is more precise than the linear model. In the final subperiod optimal weights,

Table 9: 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.

$w_{\hat{\mathbf{M}}}$ yield forecasts with a lower MSFE than the linear model. Comparing these results to those in Table 7, 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 towards 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 distinguish feature of the subperiods: in the first subperiod the middle regime has an average probability of 0.65 whereas in the second and third subperiods it given a probability of 0.83 and 0.84. Hence, the linear model is more difficult to beat in the last two subperiods as for many forecasts it is close to the optimal forecasting model.

In order to check the robustness of our results to the choice of forecast sample, we additionally use the forecast 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 could be an issue 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 model selected in the baseline forecasting exercise.

Table 9 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 can be seen that the signs of the test statistics are as we would have expected them 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 9 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 paper, we have derived optimal forecasts for Markov switching models and analyzed the effect of uncertainty around states on forecasts based on optimal weights. Applying the methodology to Markov switching models helps tightening the well documented gap between in-sample and out-of-sample performance of these models. The importance of uncertainty around the timing of the switches between states is shown by comparing optimal forecasts when the states of the Markov chain are assumed to be known with optimal forecasts when they are not known. The optimal weights for known states share the properties of the weights derived in Pesaran et al.

(2013). They are asymptotically identical to the Markov switching weights and improvements in forecasting performance are found when the ratio of the number of observations to the number of estimated parameters is small. In contrast, the optimal weights for unknown states are asymptotically different from the Markov switching weights and potential improvements in forecasting accuracy can be considerable for large break sizes 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 forecasts from the great moderation and explains the difficulty of Markov switching forecasts to beat linear models.

A Mathematical details

A.1 Derivations conditional on states

A.1.1 Weights for two-state Markov switching model

In order to derive weights (9)–(12), define $\lambda = \frac{\beta_2 - \beta_1}{\sigma_2}$ and $q = \frac{\sigma_1}{\sigma_2}$, $\pi_1 = \frac{1}{T}\sum_{t=1}^{T} s_{1t}$, and $\pi_2 = \frac{1}{T}\sum_{t=1}^{T} s_{2t}$. Then we have

$$\begin{aligned}\mathbf{M} &= \mathbf{Q} + \tilde{\mathbf{S}}'\boldsymbol{\lambda}\boldsymbol{\lambda}'\tilde{\mathbf{S}} \\ &= q^2\mathbf{S}_1 + \mathbf{S}_2 + \lambda^2\mathbf{s}_2\mathbf{s}_2'\end{aligned}$$

where $\mathbf{S}_i$ is a $T \times T$ diagonal matrix with typical t,t-element $s_{i,t}$. The inverse of $\mathbf{M}$ is

$$\begin{aligned}\mathbf{M}^{-1} &= (q^2\mathbf{S}_1 + \mathbf{S}_2)^{-1} - \frac{\lambda^2(q^2\mathbf{S}_1 + \mathbf{S}_2)^{-1}\mathbf{s}_2\mathbf{s}_2'(q^2\mathbf{S}_1 + \mathbf{S}_2)^{-1}}{1 + \lambda^2\mathbf{s}_2'(q^2\mathbf{S}_1 + \mathbf{S}_2)^{-1}\mathbf{s}_2} \\ &= \frac{1}{q^2}\mathbf{S}_1 + \mathbf{S}_2 - \frac{\lambda^2(\frac{1}{q^2}\mathbf{S}_1 + \mathbf{S}_2)\mathbf{s}_2\mathbf{s}_2'(\frac{1}{q^2}\mathbf{S}_1 + \mathbf{S}_2)}{1 + \lambda^2\mathbf{s}_2'(\frac{1}{q^2}\mathbf{S}_1 + \mathbf{S}_2)\mathbf{s}_2} \\ &= \frac{1}{q^2}\mathbf{S}_1 + \mathbf{S}_2 - \frac{\lambda^2\mathbf{s}_2\mathbf{s}_2'}{1 + \lambda^2 T\pi_2}\end{aligned}$$

The weights are given by

$$\mathbf{w} = \lambda^2\mathbf{M}^{-1}\mathbf{s}_2 s_{2,T+1} + \frac{\mathbf{M}^{-1}\boldsymbol{\iota}}{\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota}}\left(1 - \lambda^2\boldsymbol{\iota}'\mathbf{M}^{-1}\mathbf{s}_2 s_{2,T+1}\right)$$

The various components needed to calculate the weights are given by

$$\begin{aligned}\mathbf{M}^{-1}\mathbf{s}_2 &= \mathbf{s}_2 - \frac{\lambda^2 T\pi_2}{1+\lambda^2 T\pi_2}\mathbf{s}_2 \\ &= \frac{1}{1+\lambda^2 T\pi_2}\mathbf{s}_2 \\ \mathbf{M}^{-1}\boldsymbol{\iota} &= \frac{1}{q^2}\mathbf{s}_1 + \mathbf{s}_2 - \frac{\lambda^2 T\pi_2}{1+\lambda^2 T\pi_2}\mathbf{s}_2 \\ &= \frac{\mathbf{s}_1(1+\lambda^2 T\pi_2)+q^2\mathbf{s}_2}{q^2(1+\lambda^2 T\pi_2)}\end{aligned}$$

and

$$\boldsymbol{\iota}'\mathbf{M}^{-1}\mathbf{s}_2 = \frac{T\pi_2}{1+\lambda^2 T\pi_2}, \quad \boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota} = T\frac{\pi_1+\lambda^2 T\pi_1\pi_2+q^2\pi_2}{q^2(1+\lambda^2 T\pi_2)}$$

This yields the weights

$$\begin{aligned}\mathbf{w} &= \lambda^2\frac{1}{1+\lambda^2 T\pi_2}\mathbf{s}_2 s_{2,T+1} + \frac{1}{T}\frac{\mathbf{s}_1(1+\lambda^2 T\pi_2)+q^2\mathbf{s}_2}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\left(1-\lambda^2\frac{T\pi_2 s_{2,T+1}}{1+\lambda^2 T\pi_2}\right) \\ &= \frac{1}{1+\lambda^2 T\pi_2}\left(\mathbf{s}_2 s_{2,T+1} + \frac{1}{T}\frac{\mathbf{s}_1(1+\lambda^2 T\pi_2)+q^2\mathbf{s}_2}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\left(1+\lambda^2 T\pi_2(1-s_{2,T+1})\right)\right)\end{aligned}$$

Suppose $s_{2,T+1} = s_{2,t} = 1$, then

$$\begin{aligned}w_{22} &= \frac{1}{1+\lambda^2 T\pi_2}\left(\lambda^2 + \frac{1}{T}\frac{q^2}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\right) \\ &= \frac{1}{1+\lambda^2 T\pi_2}\frac{1}{T}\frac{1}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\left(q^2(1+\lambda^2 T\pi_2)+\lambda^2 T\pi_1(1+\lambda^2 T\pi_2)\right) \\ &= \frac{1}{T}\frac{q^2+\lambda^2 T\pi_1}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\end{aligned}$$

when $s_{2,T+1} = 1$, $s_{2,t} = 0$, then

$$\begin{aligned}w_{21} &= \frac{1}{1+\lambda^2 T\pi_2}\left(\frac{1}{T}\frac{1+\lambda^2 T\pi_2}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\right) \\ &= \frac{1}{T}\frac{1}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\end{aligned}$$

when $s_{2,T+1} = 0, s_{2,t} = 1$, then

$$\begin{aligned}w_{12} &= \frac{1}{1+\lambda^2 T\pi_2}\left(\frac{1}{T}\frac{q^2(1+\lambda^2 T\pi_2)}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\right) \\ &= \frac{1}{T}\frac{q^2}{\pi_2 q^2+\pi_1(1+T\pi_2\lambda^2)}\end{aligned}$$

finally, when $s_{2,T+1} = 0, s_{2,t} = 0$, then

$$\begin{aligned} w_{11} &= \frac{1}{1+\lambda^2 T\pi_2}\left(\frac{1}{T}\frac{(1+\lambda^2 T\pi_2)^2}{\pi_2 q^2 + \pi_1(1+T\pi_2\lambda^2)}\right) \\ &= \frac{1}{T}\frac{1+\lambda^2 T\pi_2}{\pi_2 q^2 + \pi_1(1+T\pi_2\lambda^2)} \end{aligned}$$

In order to show the symmetry of the weights, consider the definition of λ and q conditional on the regime $s_{i,T+1}$. If $s_{2,T+1} = 1$, define $\lambda = \frac{\beta_2-\beta_1}{\sigma_2}$ and $q = \frac{\sigma_1}{\sigma_2}$, but if $s_{1,T+1} = 1$, define $\lambda_* = \frac{\beta_1-\beta_2}{\sigma_1}$ and $q_* = \frac{\sigma_2}{\sigma_1}$. Then, $\lambda^2 = \lambda_*^2/q_*^2$ and we have for w_{12} and w_{11}

$$\begin{aligned} w_{12} &= \frac{1}{T}\frac{q^2}{\pi_2 q^2 + \pi_1(1+T\pi_2\lambda^2)} \\ &= \frac{1}{T}\frac{1/q_*^2}{\pi_2/q_*^2 + \pi_1(1+1/q_*^2 T\pi_2\lambda_*^2)} \\ &= \frac{1}{T}\frac{1}{\pi_1 q_*^2 + \pi_2(1+T\pi_1\lambda_*^2)} \end{aligned}$$

$$\begin{aligned} w_{11} &= \frac{1}{T}\frac{1+\lambda^2 T\pi_2}{\pi_2 q^2 + \pi_1(1+T\pi_2\lambda^2)} \\ &= \frac{1}{T}\frac{1+1/q_*^2\lambda_*^2 T\pi_2}{\pi_2/q_*^2 + \pi_1(1+1/q_*^2 T\pi_2\lambda_*^2)} \\ &= \frac{1}{T}\frac{q_*^2+\lambda_*^2 T\pi_2}{\pi_1 q_*^2 + \pi_2(1+T\pi_1\lambda_*^2)} \end{aligned}$$

The symmetry of the weights is a natural consequence of the fact that the Markov Switching model is invariant under a relabeling of the states.

A.1.2 Weights and MSFE for m-state Markov switching model

To derive weights for an m-state Markov switching model, we will concentrate on $s_{k,T+1} = 1$ as we have shown above that the weights are symmetric. In this case, define $\lambda_i = (\beta_i - \beta_k)/\sigma_k$ and $q_i = \sigma_i/\sigma_k$. The model is given by

$$\begin{aligned} y_t &= \sum_{i=1}^m \beta_i s_{it} + \sum_{i=1}^m \sigma_i s_{it}\varepsilon_t \\ &= \beta_k + \sum_{i=1}^m (\beta_i - \beta_k)s_{it} + \sum_{i=1}^m \sigma_i s_{it}\varepsilon_t \\ &= \sigma_k\left(\frac{\beta_k}{\sigma_k} + \sum_{i=1}^m \lambda_i s_{it} + \sum_{i=1}^m q_i s_{it}\varepsilon_t\right) \end{aligned}$$

For the observation at $T+1$ we have

$$\frac{1}{\sigma_k} y_{T+1} = \frac{\beta_k}{\sigma_k} + \varepsilon_{T+1}$$

The forecast error is

$$\frac{1}{\sigma_k}(y_{T+1} - \mathbf{w}'\mathbf{y}) = \varepsilon_{T+1} - \sum_{i=1}^{m} \lambda_i \mathbf{w}'\mathbf{s}_i - \sum_{i=1}^{m} q_i \mathbf{w}'\mathbf{S}_i \boldsymbol{\varepsilon}$$

Squaring and taking expectations gives

$$\mathrm{E}\left[\sigma_k^{-2}(y_{T+1} - \mathbf{w}'\mathbf{y})^2\right] = 1 + \sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j \mathbf{w}'\mathbf{s}_i\mathbf{s}_j'\mathbf{w} + \sum_{i=1}^{m} q_i^2 \mathbf{w}\mathbf{S}_i\mathbf{w}$$

Implementing the constraint $\sum_{t=1}^{T} w_t = 1$ by a Lagrange multiplier and taking the derivative gives

$$\begin{aligned} \mathbf{w} &= \left(\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j \mathbf{w}'\mathbf{s}_i\mathbf{s}_j' + \sum_{i=1}^{m} q_i^2 \mathbf{w}\mathbf{S}_i \right)^{-1} (-\theta\boldsymbol{\iota}) \\ &= -\theta \mathbf{M}^{-1}\boldsymbol{\iota} \end{aligned} \tag{38}$$

The inverse can be expressed analytically through the Sherman Morrison formula as

$$\begin{aligned} \mathbf{M}^{-1} &= \sum_{i=1}^{m} \frac{1}{q_i^2}\mathbf{S}_i - \frac{\left(\sum_{i=1}^{m} \frac{1}{q_i^2}\mathbf{s}_i\right)\left(\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i\lambda_j \mathbf{s}_i\mathbf{s}_j'\right)\left(\sum_{i=1}^{m} \frac{1}{q_i^2}\mathbf{s}_i\right)}{1 + \left(\sum_{j=1}^{m} \lambda_j \mathbf{s}_j'\right)\left(\sum_{i=1}^{m} \frac{1}{q_i^2}\mathbf{s}_i\right)\left(\sum_{i=1}^{m} \lambda_j \mathbf{s}_i\right)} \\ &= \sum_{i=1}^{m} \frac{1}{q_i^2}\mathbf{S}_i - \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\lambda_i}{q_i^2}\frac{\lambda_j}{q_j^2}\mathbf{s}_i\mathbf{s}_j'}{1 + T\sum_{i=1}^{m} \frac{\lambda_i^2}{q_i^2}\pi_i} \end{aligned}$$

Multiplying with $\boldsymbol{\iota}$ as in equation (38) gives

$$\mathbf{M}^{-1}\boldsymbol{\iota} = \sum_{i=1}^{m} \frac{1}{q_i^2}\mathbf{s}_i - \frac{T\sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\lambda_i}{q_i^2}\frac{\lambda_j}{q_j^2}\pi_j\mathbf{s}_i}{1 + T\sum_{i=1}^{m} \frac{\lambda_i^2}{q_i^2}\pi_i}$$

Since the weights should sum up to one, we have

$$\begin{aligned} \boldsymbol{\iota}'\mathbf{w} &= \left[T\sum_{i=1}^{m} \frac{1}{q_i^2}\pi_i - \frac{T^2\sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\lambda_i}{q_i^2}\frac{\lambda_j}{q_j^2}\pi_j\pi_i}{1 + T\sum_{i=1}^{m} \frac{\lambda_i^2}{q_i^2}\pi_i} \right](-\theta) \\ &= 1 \end{aligned}$$

which gives

$$\theta = \frac{1 + T\sum_{j=1}^{m}\frac{\lambda_i^2}{q_i^2}\pi_i}{T}\left[\sum_{i=1}^{m}\frac{1}{q_i^2}\pi_i + T\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\frac{1}{q_i^2}\frac{\lambda_j}{q_j^2}\pi_i\pi_j - \frac{\lambda_i}{q_i^2}\frac{\lambda_j}{q_j^2}\pi_j\pi_i\right)\right]^{-1}$$
$$= \frac{1 + T\sum_{j=1}^{m}\frac{\lambda_i^2}{q_i^2}\pi_i}{T}\left[\sum_{i=1}^{m}\frac{1}{q_i^2}\pi_i + T\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{q_i^2}\frac{1}{q_j^2}\pi_i\pi_j\lambda_j(\lambda_j - \lambda_i)\right]^{-1}$$

The weights are then given by

$$\mathbf{w} = \frac{1}{T}\frac{\sum_{i=1}^{m}\frac{1}{q_i^2}\mathbf{s}_i + T\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{q_i^2}\frac{1}{q_j^2}\pi_j\lambda_j(\lambda_j - \lambda_i)\mathbf{s}_i}{\sum_{i=1}^{m}\frac{1}{q_i^2}\pi_i + T\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{q_i^2}\frac{1}{q_j^2}\pi_i\pi_j\lambda_j(\lambda_j - \lambda_i)}$$

So that if $s_{lt} = 1$ the weight at time t is

$$w_t = \frac{1}{T}\frac{\frac{1}{q_l^2} + T\sum_{j=1}^{m}\frac{1}{q_l^2}\frac{1}{q_j^2}\pi_j\lambda_j(\lambda_j - \lambda_l)}{\sum_{i=1}^{m}\frac{1}{q_i^2}\pi_i + T\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{1}{q_i^2}\frac{1}{q_j^2}\pi_i\pi_j\lambda_j(\lambda_j - \lambda_i)}$$

The MSFE is easy to derive by noting that we can substitute the first order condition for the weights

$$\begin{aligned}\mathrm{E}\left[\sigma_k^{-2}(y_{T+1} - \mathbf{w}'\mathbf{y})^2\right] &= 1 + \sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\mathbf{w}'\mathbf{s}_i\mathbf{s}_j'\mathbf{w} + \sum_{i=1}^{m}q_i^2\mathbf{w}\mathbf{S}_i\mathbf{w}\\ &= 1 - \theta\\ &= 1 + w_{kk}\end{aligned}$$

where w_{kk} is the weight when $s_{k,T+1} = s_{kt} = 1$.

A.2 Derivations conditional on state probabilities

A.2.1 Large T approximation for optimal weights

Rewrite (22) as

$$w_t = \frac{1}{T}\frac{d_t[\frac{1}{T} + \lambda^2\frac{1}{T}\sum_{t'=1}^{T}d_{t'}(\xi_{2,T+1} - \xi_{2t'})(\xi_{2t} - \xi_{2t'})]}{\frac{1}{T}\left(\frac{1}{T}\sum_{t'=1}^{T}d_{t'}\right) + \lambda^2\left[\frac{1}{T}\sum_{t'=1}^{T}d_{t'}\xi_{2t'}^2\frac{1}{T}\sum_{t'=1}^{T}d_{t'} - \left(\frac{1}{T}\sum_{t'=1}^{T}d_{t'}\xi_{2t'}\right)^2\right]} \tag{39}$$

where

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

To perform the large sample approximation we need to establish that $\frac{1}{T}\sum_{t=1}^{T}d_t < \infty$, $\frac{1}{T}\sum_{t=1}^{T}\xi_{2t}d_t < \infty$ and $\frac{1}{T}\sum_{t=1}^{T}\xi_{2t}^2d_t < \infty$. Proving the first of these relations implies the other two, since $0 \leqslant \xi_{2t} \leqslant 1$. Define $a_t = \frac{1}{d_t}$. We then

need to prove that $a_t > 0$. The only scenario where $a_t = 0$ is when $\xi_{2t} = 0$ and $q^2 = 0$, so the only restriction that we must impose to obtain $a_t > 0$ is that $q^2 > 0$. Then

$$\frac{1}{T}\sum_{t=1}^{T} d_t = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{a_t} \leqslant \frac{1}{T}T\frac{1}{a_{\min}} = \frac{1}{a_{\min}} < \infty$$

where $a_{\min}$ is the minimum value of a_t over $t = 1, 2, \ldots, T$.

Denote $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$, $\overline{d\xi} = \frac{1}{T}\sum_{t=1}^{T} d_t\xi_{2t}$, and $\overline{d\xi^2} = \frac{1}{T}\sum_{t=1}^{T} d_t\xi_{2t}^2$, then (39) can be written as

$$\begin{aligned}
w_t &= \frac{1}{T}d_t\left[\frac{\frac{1}{T}}{\frac{1}{T}\bar{d} + \lambda^2\left[\overline{d\xi^2}\bar{d} - \overline{d\xi}^2\right]} + \frac{\lambda^2(\xi_{2t}\xi_{2,T+1}\bar{d} - \xi_{2t}\overline{d\xi} - \xi_{2,T+1}\overline{d\xi} + \overline{d\xi^2})}{\frac{1}{T}\bar{d} + \lambda^2\left[\overline{d\xi^2}\bar{d} - \overline{d\xi}^2\right]}\right] \\
&= \frac{1}{T}d_t\left[\frac{1}{T}\frac{1}{\lambda^2\left[\overline{d\xi^2}\bar{d} - \overline{d\xi}^2\right]}\frac{1}{1+\frac{\theta}{T}} + \frac{\lambda^2(\xi_{2t}\xi_{2,T+1}\bar{d} - \xi_{2t}\overline{d\xi} - \xi_{2,T+1}\overline{d\xi} + \overline{d\xi^2})}{\lambda^2\left[\overline{d\xi^2}\bar{d} - \overline{d\xi}^2\right]}\frac{1}{1+\frac{\theta}{T}}\right] \\
&= \frac{1}{T}d_t\frac{\lambda^2(\xi_{2t}\xi_{2,T+1}\bar{d} - \xi_{2t}\overline{d\xi} - \xi_{2,T+1}\overline{d\xi} + \overline{d\xi^2})}{\lambda^2\left[\overline{d\xi^2}\bar{d} - \overline{d\xi}^2\right]} + \mathcal{O}\left(T^{-2}\right)
\end{aligned}$$

where $\theta = \frac{\bar{d}}{\lambda^2\left[\overline{d\xi^2}\bar{d} - \overline{d\xi}^2\right]} = \frac{1}{\lambda^2\sum_{t=1}^{T}\tilde{d}_t(\xi_{2t} - \frac{1}{T}\sum_{t'=1}^{T}\tilde{d}_{t'}\xi_{2t'})^2}$ where $\tilde{d}_t = d_t/\sum_{t'} d_{t'}$. The numerator is nonzero unless for the trivial case when ξ_{2t} is constant for all t. Using this and the result that $\bar{d}$, $\overline{d\xi}$ and $\overline{d\xi^2}$ are finite for any T proves that we can apply the expansion in terms of θ/T. Dividing w_t by $\sum_{t=1}^{T} d_t$ yields (24).

A.2.2 Weights and MSFE for standard Markov switching model

The Markov switching weights can be written as

$$\begin{aligned}
\mathbf{w}_{\text{MS}} &= \frac{\xi_{1,T+1}\boldsymbol{\xi}_1}{\sum_{t=1}^{T}\xi_{1t}} + \frac{\xi_{2,T+1}\boldsymbol{\xi}_2}{\sum_{t=1}^{T}\xi_{2t}} \\
&= \frac{1}{T}\frac{\xi_{2,T+1}\boldsymbol{\xi}_2}{\bar{\xi}_2} + \frac{1}{T}\frac{(1-\xi_{2,T+1})(\boldsymbol{\iota} - \boldsymbol{\xi}_2)}{(1-\bar{\xi}_2)} \\
&= \frac{1}{T}\frac{1}{\bar{\xi}_2(1-\bar{\xi}_2)}(\xi_{2,T+1}\boldsymbol{\xi}_2(1-\bar{\xi}_2+\bar{\xi}_2) + \bar{\xi}_2\boldsymbol{\iota} - \bar{\xi}_2\xi_{2,T+1}\boldsymbol{\iota} - \bar{\xi}_2\boldsymbol{\xi}_2) \\
&= \frac{1}{T}\frac{1}{\bar{\xi}_2(1-\bar{\xi}_2)}(\xi_{2,T+1} - \bar{\xi}_2)(\boldsymbol{\xi}_2 - \bar{\xi}_2\boldsymbol{\iota}) + \bar{\xi}_2(1-\bar{\xi}_2) \\
&= \frac{1}{T} + \frac{1}{T}\frac{(\xi_{2,T+1} - \bar{\xi}_2)(\boldsymbol{\xi}_2 - \bar{\xi}_2\boldsymbol{\iota})}{\bar{\xi}_2(1-\bar{\xi}_2)} \qquad (40)
\end{aligned}$$

For a general vector of weights $\mathbf{w}$, subject to $\sum_{t=1}^{T} w_t = 1$, and assuming

a constant error variance, we have the following MSFE

$$\begin{aligned}\mathrm{E}[\sigma^{-2}e_{T+1}^2] &= 1+\lambda^2\xi_{2,T+1}+\mathbf{w}'\mathbf{M}\mathbf{w}-2\lambda^2\mathbf{w}'\boldsymbol{\xi}\xi_{2,T+1}\\ &= 1+\lambda^2\xi_{2,T+1}+\lambda^2(\mathbf{w}'\boldsymbol{\xi})^2+\mathbf{w}'\mathbf{D}\mathbf{w}-2\lambda^2\mathbf{w}'\boldsymbol{\xi}\xi_{2,T+1}\end{aligned} \tag{41}$$

where $\mathbf{D}=(1+\lambda^2\sigma_\xi^2)\mathbf{I}$.

Using (40) we have that

$$\begin{aligned}\mathbf{w}_{\mathrm{MS}}'\boldsymbol{\xi} &= \bar{\xi}_2+\frac{\xi_{2,T+1}-\bar{\xi}_2}{(1-\bar{\xi}_2)\bar{\xi}_2}\left(\frac{1}{T}\sum_{t=1}^{T}\xi_t^2-T\bar{\xi}_2^2\right)\\ &= \bar{\xi}_2+\frac{\xi_{2,T+1}-\bar{\xi}_2}{(1-\bar{\xi}_2)\bar{\xi}_2}(\bar{\xi}_2(1-\bar{\xi}_2)-\sigma_\xi^2)\\ &= \xi_{2,T+1}-\frac{\xi_{2,T+1}-\bar{\xi}_2}{\bar{\xi}_2(1-\bar{\xi}_2)}\sigma_\xi^2\end{aligned}$$

where we have used (25), and

$$\mathbf{w}_{\mathrm{MS}}'\mathbf{D}\mathbf{w}_{\mathrm{MS}}=(1+\lambda^2\sigma_\xi^2)\left(\frac{1}{T}+\frac{(\xi_{2,T+1}-\bar{\xi}_2)^2}{T\bar{\xi}_2^2(1-\bar{\xi}_2)^2}(\bar{\xi}_2(1-\bar{\xi}_2)-\sigma_\xi^2)\right)$$

So that the MSFE is

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

A.2.3 MSFE for Markov switching model using optimal weights

Equation (23) for an arbitrary number of states is derived as follows

$$\begin{aligned}
\mathrm{E}[\sigma^{-2}e_{T+1}^2] &= (\boldsymbol{\iota}'\mathbf{M}^{-1}\boldsymbol{\iota})^{-1}(1-\boldsymbol{\iota}'\mathbf{M}^{-1}\tilde{\boldsymbol{\xi}}\tilde{\xi}_{T+1})^2 + \\
&\quad + \sum_{j=2}^{m}\lambda_j^2\xi_{j,T+1} - \tilde{\xi}_{T+1}^2\tilde{\boldsymbol{\xi}}'\mathbf{M}^{-1}\tilde{\boldsymbol{\xi}} + \sum_{j=1}^{m}q_j^2\xi_{j,T+1} \\
&= \frac{1+\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}}{\boldsymbol{\iota}'\mathbf{D}^{-1}\boldsymbol{\iota}(1+\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}) - (\boldsymbol{\iota}\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}})^2}\left[1+\frac{\tilde{\xi}_{T+1}^2(\boldsymbol{\iota}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}})^2}{(1+\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}})^2} + \right. \\
&\quad \left. -2\frac{\tilde{\xi}_{T+1}\boldsymbol{\iota}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}}{1+\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}}\right] + \tilde{\xi}_{T+1}^2 - \frac{\tilde{\xi}_{T+1}^2\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}}{1+\tilde{\boldsymbol{\xi}}'\mathbf{D}^{-1}\tilde{\boldsymbol{\xi}}} + \frac{1}{d_{T+1}} \\
&= \frac{1+\sum_{t=1}^{T}\tilde{\xi}_t^2 - 2\tilde{\xi}_{T+1}\sum_{t=1}^{T}d_t\tilde{\xi}_t + \tilde{\xi}_{T+1}^2\sum_{t=1}^{T}d_t}{\sum_{t=1}^{T}d_t(1+\sum_{t'=1}^{T}d_{t'}\tilde{\xi}_{t'}^2) - (\sum_{t=1}^{T}d_t\tilde{\xi}_t)^2} + \frac{1}{d_{T+1}} \\
&= \frac{w_{T+1}}{d_{T+1}} + \frac{1}{d_{T+1}} \\
&= \frac{1}{d_{T+1}}(1+w_{T+1})
\end{aligned}$$

A.2.4 Derivation of (33)

To save on notation, in the following we use $p(s_{jt}|s_{i,t+m},\Omega_T)$ to write $p(s_{jt}=1|s_{i,t+m}=1,\Omega_T)$. To derive (33), take for example a three state model and calculate

$$\begin{aligned}
p(s_{jt}|s_{i,t+3},\Omega_T) &= \sum_{k=0}^{2}p(s_{jt}|s_{k,t+1},s_{i,t+3},\Omega_T)p(s_{k,t+1}|s_{i,t+3},\Omega_T) \\
&= \sum_{k=0}^{2}p(s_{jt}|s_{k,t+1},\Omega_t)\sum_{l=0}^{2}p(s_{k,t+1}|s_{l,t+2},\Omega_{t+1})p(s_{l,t+2}|s_{i,t+3},\Omega_{t+2}) \\
&= \sum_{k=0}^{2}\frac{p_{jk}p(s_{jt}|\Omega_t)}{p(s_{k,t+1}|\Omega_t)}\sum_{l=0}^{2}\frac{p_{kl}p(s_{k,t+1}|\Omega_{t+1})}{p(s_{l,t+2}|\Omega_{t+1})}\frac{p_{li}p(s_{l,t+2}|\Omega_{t+2})}{p(s_{i,t+3}|\Omega_{t+2})} \\
&= \frac{p(s_{jt}|\Omega_t)}{p(s_{i,t+3}|\Omega_{t+2})}\sum_{k=0}^{2}\sum_{l=0}^{2}p_{jk}a_{t+1}^k p_{kl}a_{t+2}^l p_{li} \\
&= \frac{p(s_{jt}|\Omega_t)}{p(s_{i,t+3}|\Omega_{t+2})}\left(\mathbf{P}'\mathbf{A}_{t+1}\mathbf{P}'\mathbf{A}_{t+2}\mathbf{P}'\right)_{j,i}
\end{aligned}$$

where $a_{t+1}^k = \frac{p(s_{k,t+1}=1|\Omega_{t+1})}{p(s_{k,t+1}=1|\Omega_t)}$. On the second line we use that the regime s_t depends on future observations only through s_{t+1}.

Table 10: Monte Carlo results: MSI and MSM models

		$T = 50$			$T = 100$		
λ	$\tilde{\sigma}^2_{\hat{\xi}_{\mid T}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$	$w_{\hat{s}}$	$w_{\hat{\xi}}$	$w_{\hat{\mathbf{M}}}$
MSI							
1	0.0-0.1	0.988	1.008	1.002	0.994	1.006	1.006
	0.1-0.2	0.994	1.019	1.016	0.997	1.016	1.020
	0.2-0.3	0.997	1.018	1.018	0.999	1.017	1.026
2	0.0-0.1	0.997	1.005	1.006	0.999	1.003	1.020
	0.1-0.2	1.000	1.005	1.017	1.002	0.994	1.030
	0.2-0.3	1.003	0.993	1.007	1.003	0.985	1.018
3	0.0-0.1	1.000	0.999	1.004	1.000	0.999	1.012
	0.1-0.2	1.004	0.983	1.026	1.004	0.972	1.020
	0.2-0.3	1.005	0.970	0.986	1.005	0.944	0.981
MSM							
1	0.0-0.1	0.991	1.010	1.008	0.994	1.019	1.020
	0.1-0.2	0.994	1.023	1.017	0.996	1.033	1.042
	0.2-0.3	0.995	1.029	1.037	0.998	1.033	1.043
2	0.0-0.1	0.996	1.011	1.009	0.999	1.012	1.028
	0.1-0.2	0.998	1.015	1.019	1.000	1.010	1.034
	0.2-0.3	0.999	1.015	1.022	1.001	1.007	1.024
3	0.0-0.1	0.999	1.004	1.004	1.000	1.002	1.015
	0.1-0.2	1.000	1.002	1.013	1.002	0.991	1.012
	0.2-0.3	1.000	1.006	1.007	1.003	0.974	0.983

Note: The table reports the ratio of the MSFE of the optimal weights to that of the Markov switching weights. DGP MSI: $y_t = \beta_1 s_{1t} + \beta_2 s_{2t} + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \sigma\varepsilon_t$ where $\varepsilon_t \sim \mathrm{N}(0,1)$. DGP MSM: $y_t = \beta_1 s_{1t} + \beta_2 s_{2t} + \phi_1(y_{t-1} - \beta_{s_{t-1}}) + \phi_2(y_{t-2} - \beta_{s_{t-2}}) + \sigma\varepsilon_t$, $\sigma^2 = 0.25$, $\phi_1 = 0.4$, $\phi_2 = -0.3$. Column labels as in Table 4.

B Monte Carlo results for MSI and MSM models

Table 10 presents Monte Carlo results for the models that are frequently used in empirical applications. These models are the m-state Markov switching in intercept (MSI) and Markov switching in mean (MSM) models which include p lags of the dependent variable. We analyze the performance of the optimal weights for an MSI(2)-AR(2) and MSM(2)-AR(2) model. For both

models, Table 10 shows that the improvements by using optimal weights are consistent with the results for the Markov switching model with no lagged dependent variables. However, the additional parameter estimates imply noise that leads to slightly less pronounced differences in MSFE compared to the intercept only model.

References

Ang, A. and Bekaert, G. (2002). Regime switches in interest rates. *Journal of Business & Economic Statistics*, 20(2):163–182.

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(1):47–75.

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(2):223–243.

Dacco, R. and Satchell, S. (1999). Why do regime-switching models forecast so badly? *Journal of Forecasting*, 18(1):1–16.

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, B*, 39(1):1–38.

Deschamps, P. J. (2008). Comparing smooth transition and Markov switching autoregressive models of US unemployment. *Journal of Applied Econometrics*, 23(4):435–462.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 12:253–263.

Engel, C. (1994). Can the Markov switching model forecast exchange rates? *Journal of International Economics*, 36(1):151–165.

Guidolin, M. (2011). Markov switching models in empirical finance. *Advances in Econometrics*, 27:1–86.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, Princeton.

Kim, C.-J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(1):1–22.

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(4):608–616.

Klaassen, F. (2005). Long swings in exchange rates: Are they really in the data? *Journal of Business & Economic Statistics*, 23(1):87–95.

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

Krolzig, H.-M. (2000). Predicting Markov-switching vector autoregressive processes. Nuffield College *Working Paper*, W31.

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(1):259–306.

Pesaran, M. H., Pick, A., and Pranovich, M. (2013). Optimal forecasts in the presence of structural breaks. *Journal of Econometrics*, 177(2):134–152.

Rossi, B. and Inoue, A. (2012). Out-of-sample forecast tests robust to the choice of window size. *Journal of Business & Economics Statistics*, 30(3):432–453.