Optimal forecasts from Markov switching models and the effect of uncertain break dates

Tom Boot*   Andreas Pick†
April 24, 2014

Abstract

We study the effect of uncertain break dates on forecasts based on optimal weighting of observations. We focus on Markov switching models where break dates correspond to switches between states. It emerges that the forecasting performance increases drastically when the construction of the optimal weights takes uncertainty around the time of breaks into account. Analytic expressions for the weights are provided both under exact knowledge of the break points and when the break points are uncertain. In the latter case, forecasting improvements are substantial even in large samples. The performance of the optimal weights is shown through simulations and an application to forecasting U.S. GNP over 30 years. In the application we find that using the optimal weights leads to a significant reduction of the 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
1 Introduction

Models that allow for breaks in the parameters are in widespread use in macroeconomics and finance for in-sample analysis and for out-of-sample forecasting. Recently, Pesaran et al. (2013) developed optimal forecasts for models with parameter instability by weighting observations such that the mean square forecast error (MSFE) is minimized. However, the optimal weights perform less well in applications as estimates of the model’s parameters are required. It emerges from their analysis that the timing of breaks is of first order importance and the uncertainty surrounding the break dates are therefore likely to be a major cause of the deterioration of the performance of the optimal weights in applications. For this reason, Pesaran et al. (2013) suggest robust optimal weights where the parameters are integrated out.

In this paper, we analyze the effect of break date uncertainty on the optimal weights. As the finite sample distribution of break dates is difficult to obtain, we concentrate on Markov switching models where break dates correspond to switches between states, and state probabilities are estimated along with other parameters in the EM estimation. We show that for the purpose of forecasting with optimal weights Markov switching models are equivalent to models with discrete breaks with the exception that in the Markov switching model the observations are ordered by the underlying Markov process.

Beyond the aim of improving the understanding of optimal weights, this paper also provides an insight into the stylized fact that Markov switching models provide poor out-of-sample forecasts despite their attractive in-sample properties. Examples include Markov switching models to forecast exchange rates by Engel (1994), Dacco and Satchell (1999) and Klaassen (2005), US GNP growth by Clements and Krolzig (1998) and Perez-Quiros and Timmermann (2001), US unemployment by Deschamps (2008), and house prices by Crawford and Fratantoni (2003). A recent review on the use of Markov switching models in finance is provided by Guidolin (2011). To improve the forecast performance, the model can be extended by including the possibility of out-of-sample breaks as, for example, in the work of Pesaran et al. (2006). However, the optimal forecasts derived in this paper can improve the forecast performance without increasing the model complexity.

We start by deriving the optimal weighting scheme conditional on the states, which will serve as a benchmark to analyze the influence of uncertainty around states on the forecast performance. We find that conditional on the states of the Markov switching model, the weights mirror those obtained by Pesaran et al. (2013), which emphasizes the correspondence of the structural break and Markov switching models. In the case of three regimes, we show that 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. We show that, under specific assumptions, the maximum improvement is independent of the number of states. Finally, the optimal weights can be written as a $O(1/T)$ correction to the usual Markov switching weights, which suggests that, conditional on the states, standard Markov switching weights asymptotically achieve the minimum MSFE.

Introducing uncertainty around the states in the derivation results in an optimal weighting scheme that is close to the weighting scheme conditional on the states even when these are not known. We show that this leads to a weighting scheme that more strongly emphasizes the regime switches compared to standard Markov switching weights. The key determinants for the relative forecast performance of the optimal weights compared with the Markov switching weights are the break size and the variance in the estimated smoothed probabilities, with the largest improvements being obtained when both are large. The reason for this is that the quality of the Markov switching forecast deteriorates in these scenarios, whereas the MSFE for the optimal weights forecasts remains largely unaffected. In contrast to the weights conditional on the states, the optimal weights conditional on state probabilities are not asymptotically equivalent to the standard weights. The forecast improvements therefore do not vanish as the sample size increases.

We perform Monte Carlo experiments to evaluate the performance of the optimal weights. The results confirm the theoretically expected improvements. The conditional on the states weights improve for small values of the break size. The improvements increase when the number of estimated parameters to the number of observations increases. The weights that use only state probabilities produce significant 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 in an empirical example on quarterly U.S. GNP. Out-of-sample forecasts are constructed over 120 quarters for a range of Markov switching business cycle 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 weights developed in this paper. The results suggest that the forecasts using optimal weights outperform the standard Markov switching forecast. We also compare our forecasting schemes to a range of linear alternatives and find that they lead to improved forecasts.

The outline of the paper is as follows. Section 2 introduces the model and the standard forecast. Section 3 and 4 derives optimal weights for different settings. Monte Carlo experiments are presented in Section 5 and an application 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 Markov switching model

\[ y_t = \beta_s' \mathbf{x}_t + \sigma_s \varepsilon_t, \quad \varepsilon_t \sim iid(0, 1) \]  

where \( \beta_s = \mathbf{B}' \mathbf{s}_t \), \( \mathbf{B} = (\beta'_1, \beta'_2, \ldots, \beta'_m)' \) is an \( m \times k \) matrix, \( \beta_i \) is a \( k \times 1 \) parameter vector, \( \mathbf{x}_t \) is a \( k \times 1 \) vector of exogenous regressors, \( \sigma_s = \mathbf{\sigma}' \mathbf{s}_t \), \( \mathbf{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_m)' \) 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 as 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

\[ P = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{m1} \\ p_{12} & p_{22} & \cdots & p_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1m} & p_{2m} & \cdots & p_{mm} \end{bmatrix} \]

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

The standard forecast in this context would be to estimate \( \beta_i, i = 1, 2, \ldots, m \), as

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

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 (1993). The forecast is then constructed as \( \hat{y}_{T+1} = \sum_{i=1}^{m} \hat{\xi}_{i,T+1} \mathbf{x}_{T+1}' \hat{\beta}_i \); see, for example, 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 \( \hat{y}_{T+1} = \mathbf{x}_{T+1}' \hat{\beta}(\mathbf{w}) \), so that

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

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
\[
y_t = \beta^\prime s_t + \sigma^\prime s_t \varepsilon_t, \quad \varepsilon_t \sim iid(0, 1)
\]
(3)
where \( \beta = (\beta_1, \beta_2, \ldots, \beta_m)^\prime \). We initially 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
\[
y_{T+1} = \sum_{t=1}^{T} w_t y_t
\]
subject to \( \sum_{t=1}^{T} w_t = 1 \).

The forecast error, scaled by the error standard deviation is
\[
\sigma^{-1}_m e_{T+1} = \sigma^{-1}_m (y_{T+1} - \tilde{y}_{T+1})
\]
\[
= \lambda^\prime s_{T+1} + q^\prime s_{T+1} \varepsilon_{T+1} - \sum_{t=1}^{T} w_t \lambda^\prime \tilde{s}_t - \sum_{t=1}^{T} w_t q^\prime s_t \varepsilon_t
\]
where
\[
\lambda = \begin{pmatrix}
(\beta_2 - \beta_1)/\sigma_m \\
(\beta_3 - \beta_1)/\sigma_m \\
\vdots \\
(\beta_m - \beta_1)/\sigma_m
\end{pmatrix}, \quad q = \begin{pmatrix}
\sigma_1/\sigma_m \\
\sigma_2/\sigma_m \\
\vdots \\
1
\end{pmatrix}
\]
and \( \tilde{s}_t = \begin{pmatrix}
s_{2t} \\
s_{3t} \\
\vdots \\
s_{mt}
\end{pmatrix} \)

and the scaled MSFE is
\[
E(\sigma^{-2}_m e_{T+1}^2) = E \left[ \lambda^\prime \left( s_{T+1} - \sum_{t=1}^{T} w_t \tilde{s}_t \right) \right]^2 + (q^\prime s_{T+1})^2 - \sum_{t=1}^{T} w_t^2 (q^\prime s_t)^2
\]
\[
= E (\tilde{s}_{T+1}^\prime \lambda^\prime \lambda \tilde{s}_{T+1}) - 2w^\prime E (\tilde{s}^\prime \lambda \lambda^\prime \tilde{s}_{T+1})
\]
\[
+ w^\prime E (\tilde{s}^\prime \lambda \lambda^\prime \tilde{s}) w + (q^\prime s_{T+1})^2 - w^\prime Q w
\]
\[
= w^\prime \left[ Q + E (\tilde{s}^\prime \lambda \lambda^\prime \tilde{s}) \right] w - 2w^\prime E (\tilde{s}^\prime \lambda \lambda^\prime \tilde{s}_{T+1})
\]
\[
+ E (\tilde{s}_{T+1}^\prime \lambda^\prime \lambda \tilde{s}_{T+1}) + (q^\prime s_{T+1})^2
\]
(5)
where \( \tilde{S} = (\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_T) \), \( S = (s_1, s_2, \ldots, s_T) \). The matrix \( Q \) is a diagonal matrix with typical \( t \)-element \( Q_{tt} = \sum_{i=1}^{m} q^2_i s_{it} \).

Furthermore, define
\[
M = \left[ Q + E (\tilde{s}^\prime \lambda \lambda^\prime \tilde{s}) \right]
\]
(6)
and note that $M$ is invertible as $Q$ is a diagonal matrix with positive entries and $E \left( \tilde{S}' \lambda \lambda' \tilde{S} \right) = \text{Cov}(\tilde{S}' \lambda) + E(\tilde{S}' \lambda)E(\lambda' \tilde{S})$, so that $M$ 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

$$w = M^{-1}E \left( \tilde{S}' \lambda \lambda' \tilde{S}_{t+1} \right) + \frac{M^{-1}_t}{\sum_{t=1}^{T} M^{-1}_t} \left[ 1 - \frac{M^{-1}_t}{M^{-1}} E \left( \tilde{S}' \lambda \lambda' \tilde{S}_{T+1} \right) \right] \quad (7)$$

MSFE (5) when applying weights (7) is

$$\text{MSFE}(w) = \frac{1}{\sum_{t=1}^{T} M^{-1}_t} \left( 1 - \frac{M^{-1}_t}{M^{-1}} E \left( \tilde{S}' \lambda \lambda' \tilde{S}_{T+1} \right) \right)^2 + E \left( \tilde{s}'_{T+1} \lambda \lambda' \tilde{s}_{T+1} \right) -E \left( \tilde{s}'_{T+1} \lambda \lambda' \tilde{s}_{T+1} \right) \left( q' \tilde{s}_{T+1} \right)^2 \quad (8)$$

In order to proceed, we need to specify the information set available. Initially, we will base the weights on the full information of the DGP, including the state for each observation. Clearly, this information is not available in practical applications and usually estimates are used for the coefficients and for the state probabilities to obtain plug-in estimates of the optimal weights. Yet, the resulting analysis will prove to be informative. In a second step we will allow for uncertainty around the states.

### 3.1 Weights conditional on the states

Conditional on the states we have $M = Q + \tilde{S}' \lambda \lambda' \tilde{S}$ and $E \left( \tilde{S}' \lambda \lambda' \tilde{S}_{T+1} \right) = \tilde{S}' \lambda \lambda' \tilde{S}_{T+1}$. 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, $\tilde{s} = (s_{21}, s_{22}, \ldots, s_{2T})'$ and therefore $M = Q + \lambda^2 \tilde{s} \tilde{s}'$ for which the inverse is given by

$$M^{-1} = \frac{\lambda^2}{1 + \lambda^2 \tilde{s} \tilde{s}' Q Q^{-1} \tilde{s} \tilde{s}' Q^{-1}} - Q^{-1} \frac{\lambda^2}{1 + \lambda^2 T \pi_2 \tilde{s} \tilde{s}'}$$

where $\lambda^2 = \frac{(\beta_2 - \beta_1)^2}{\sigma_2^2}$ and $q_1 = \frac{\sigma_1}{\sigma_2}$, This yields the following weights: When $s_{1,T+1} = 1$,

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

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

6
where \( \pi_2 = \frac{1}{T} \sum_{t=1}^{T} s_{2t} \) and \( w_{ij} = w(s_{i,T+1} = 1, s_{jt} = 1) \).

When \( s_{1,T+1} = 0 \),
\[
\begin{align*}
  w_{21} &= \frac{1}{T} \frac{1}{\left[ \pi_2 q^2 + (1 - \pi_2)(1 + T \pi_2 \lambda^2) \right]} & \text{if } s_{1t} = 1 \quad (11) \\
  w_{22} &= \frac{1}{T} \frac{q^2 + T \lambda^2(1 - \pi_2)}{\left[ \pi_2 q^2 + (1 - \pi_2)(1 + T \pi_2 \lambda^2) \right]} & \text{if } s_{2t} = 1 \quad (12)
\end{align*}
\]

Note that, conditional on the state of the future observation, the weights are symmetric. 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.

The expected MSFE under the previously derived weights is
\[
E[\sigma^2_{T+1} \mid \text{opt}] = \begin{cases} 
  q^2(1 + w_1) & \text{if } s_{1,T+1} = 1 \\
  (1 + w_2) & \text{if } s_{2,T+1} = 1 
\end{cases} \quad (13)
\]

We can compare this to the MS expected mean squared forecast error given by
\[
E[\sigma^2_{T+1} \mid \text{MS}] = \begin{cases} 
  q^2(1 + \frac{1}{T \pi_i}) & \text{if } s_{1,T+1} = 1 \\
  (1 + \frac{1}{T \pi_i}) & \text{if } s_{2,T+1} = 1 
\end{cases} \quad (14)
\]

It is then easy to show that \( \text{MSFE}_{\text{opt}} < \text{MSFE}_{\text{MS}} \). 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

Set \( s_{j,T+1} = 1 \), then define \( q_i^2 = \sigma_i^2/\sigma_j^2 \) and \( \lambda_i^2 = (\beta_i - \beta_j)^2/\sigma_j^2 \). The weights are
\[
\begin{align*}
  w_{jj} &= \frac{1}{T} \frac{1 + T \sum_{i=1}^{3} q_i^{-2} \pi_i}{\sum_{i=1}^{3} q_i^{-2} \pi_i} & \text{if } s_{1t} = 1 \\
  w_{jk} &= \frac{1}{T} \frac{q_k^{-2} + T q_k^{-2} \sum_{i=1}^{m} q_i^{-2} \pi_i \lambda_m (\lambda_m - \lambda_i)}{\sum_{i=1}^{3} q_i^{-2} \pi_i} & \text{if } s_{2t} = 1 \\
  w_{jl} &= \frac{1}{T} \frac{q_l^{-2} + T q_l^{-2} \sum_{i=1}^{m} q_i^{-2} \pi_i \lambda_m (\lambda_i - \lambda_m)}{\sum_{i=1}^{3} q_i^{-2} \pi_i} & \text{if } s_{3t} = 1
\end{align*}
\]

\( (15) \)
Table 1: Ratio between the expected MSFE for the optimal weights and the MS weights for $T = 50$

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$q = 1$</th>
<th>$q = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi = 0.1$</td>
<td>0.8500</td>
<td>0.9273</td>
</tr>
<tr>
<td>0.5</td>
<td>0.9294</td>
<td>0.9758</td>
</tr>
<tr>
<td>1</td>
<td>0.9727</td>
<td>0.9919</td>
</tr>
<tr>
<td>2</td>
<td>0.9921</td>
<td>0.9978</td>
</tr>
</tbody>
</table>

Note: Reported are the ratio between (13) and (14) for different values of $\lambda$, the difference in mean, and $q$, the ratio of standard deviations, and $\pi_2$, the proportion of observations in state 2.

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

$$E[\sigma^{-2}_i e^2_{T+1}]_{\text{opt}} = \frac{\sigma^2_j}{\sigma^2_i} (1 + w_{jj}) \quad (16)$$

For the Markov switching weights we have

$$E[\sigma^{-2}_i e^2_{T+1}]_{\text{MS}} = \frac{\sigma^2_j}{\sigma^2_i} (1 + \frac{1}{T\pi_j})$$

Figure 1 plots the weights (15) when $s_{1,T+1} = 1$, for $\lambda_3$ over the range $-3$ to 3, $\lambda_2 = -2.5$, $\pi_1 = 0.2$, $\pi_2 = \pi_3 = 0.4$, $T = 100$ and $q_1 = q_2 = 1$. On the left, where $\lambda_3$ is $-3$ the observations where $s_{1t} = 1$ receive nearly all the weight, those where $s_{2t} = 1$ receive a small positive weight and those where $s_{3t} = 1$ 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 the other state implies that the observations from the other state induce a large bias. For this reason, the weight on the observations with $s_{2t} = 1$ and $s_{3t} = 1$ is very small.

As $\lambda_3$ increases, the weights for observations with $s_{3t} = 3$ 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 second state have identical mean and the observations therefore equal weight. When $0 < \lambda_3 < 2.5$, the observations with $s_{3t} = 1$ are weighted heavier than the observations for $s_{1t} = 1$ even though this is the state of the
future observation to be forecast. The reason is that in this range the means of observations with \( s_{2t} = 1 \) and \( s_{3t} = 1 \) have opposite sign and, because the bias induced by the observations from the second state are larger than those from the third state in absolute terms, the weights on the observations from the third state need to receive a large weight to counteract the 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.

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 \( \lambda_2 \) and \( \lambda_3 \). 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 \( \lambda_2 \) and \( \lambda_3 \) are of opposite sign, the increases 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 results in the largest gains can be established formally. For given \( \beta_l \) and \( \beta_k \), the \( \beta_j \) that implies the largest improvement in forecasts can be found by maximizing (16) with
Figure 2: Excess MSFE of optimal weight forecast

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 $\lambda_2$ and $\lambda_3$.

respect to $\beta_j$, 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 for a bias correction.

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

$$\frac{E[\sigma_j^{-2} e_{T+1}^2]}{E[\sigma_j^{-2} e_{T+1}^2]}_{\text{opt}} = \frac{T + 1}{T + \frac{1}{\pi_j}} \quad (17)$$

where $\pi_j$ 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 $\pi_j$. It is interesting to note that the maximum improvement is the same as in the two state case.
Table 2: Maximum improvements in a three state model with $s_{j,T+1} = 1$

<table>
<thead>
<tr>
<th>$\pi_j$</th>
<th>$T = 50$</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.7286</td>
<td>0.8417</td>
<td>0.9136</td>
</tr>
<tr>
<td>0.1</td>
<td>0.8500</td>
<td>0.9182</td>
<td>0.9571</td>
</tr>
<tr>
<td>0.2</td>
<td>0.9273</td>
<td>0.9619</td>
<td>0.9805</td>
</tr>
<tr>
<td>0.5</td>
<td>0.9808</td>
<td>0.9902</td>
<td>0.9950</td>
</tr>
</tbody>
</table>

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

3.1.3 $m$-state Markov switching models

For $s_{j,T+1} = 1$ and $s_{1,t} = 1$ we set $\lambda = \frac{\beta_i - \beta_j}{\sigma_j}$ and $q_i = \frac{\sigma_i}{\sigma_j}$, which gives for the weights

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

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

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

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

$$\beta_j = \frac{\sum_k q_k^{-2} \pi_k \beta_k}{\sum_k q_k^{-2} \pi_k}$$

The minimum MSFE is then

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

and when the variances are equal this reduces to

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

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
The observation is given by

\[
T_w = s_{1,T+1} \left[ \frac{q^2 + \phi^2 T \pi_1}{\pi_2 + \pi_1 (q^2 + \phi^2 T \pi_2)} s_{1t} + \frac{1}{\pi_2 + \pi_1 (q^2 + \phi^2 T \pi_2)} s_{2t} \right] + s_{2,T+1} \left[ \frac{\phi^2}{\pi_2 + \pi_1 (q^2 + \phi^2 T \pi_2)} s_{1t} + \frac{1 + \phi^2 T \pi_1}{\pi_2 + \pi_1 (q^2 + \phi^2 T \pi_2)} s_{2t} \right]
\]

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

\[
T'_w = \left( \frac{1}{\pi_1} - \frac{1}{T \phi^2 \pi_1^2 \pi_2^2} \right) s_{1t} s_{1,T+1} + \frac{1}{T \phi^2 \pi_1 \pi_2} s_{2t} s_{1,T+1} + \left( \frac{1}{\pi_2} - \frac{1}{T \phi^2 \pi_2^2} \right) s_{2t} s_{2,T+1} + O(T^{-2})
\]

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) 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, to which we turn to now.

### 3.2 Optimal weights when the states are uncertain

We will now discuss optimal weights conditional on the state probabilities. Estimates of the probabilities are available to the researcher, and this information is used for the standard forecast from Markov switching models in (2).

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

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

We will initially focus on the two state case, 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 \(S = s_2 = (s_2, s_2, \ldots, s_{2T})'\). The matrix \(M\) in (7) is given by

\[
M = \lambda^2 \xi \xi' + \lambda^2 V + q_1^2 I + (1 - q_1^2) \Xi
\]

\[
= \lambda^2 \xi \xi' + D
\]

12
with $\xi = (\xi_{21}, \xi_{22}, \ldots, \xi_{2T})$, $\Xi = \text{diag}(\xi)$, $V = \Xi(I - \Xi)$, and $D = \lambda^2V + q_1^2I + (1 - q_1^2)\Xi$. The inverse of $M$ is then

$$M^{-1} = D^{-1} - \frac{\lambda^2}{1 + \lambda^2\xi^T D^{-1}\xi} D^{-1} \xi \xi^T D^{-1}$$

The weights follow from (7) as

$$w = \lambda^2 \xi_{2T+1} M^{-1} \xi + \frac{M^{-1}_{t,t}}{\xi^T M^{-1}_t} [1 - \lambda^2 \xi_{2T+1} \xi^T M^{-1}_t]$$  \hspace{1cm} (20)

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

$$d_t = [\lambda^2 \xi_{2,t}(1 - \xi_{2,t}) + q_1^2 + (1 - q_1^2)\xi_{2,t}]^{-1}$$

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

$$w_t = \frac{d_t \left[ 1 + \lambda^2 \sum_{t'=1}^{T} d_{t'} (\xi_{2t} - \xi_{2t'}) (\xi_{2t+1} - \xi_{2t'}) \right]}{\sum_{t'=1}^{T} d_{t'} + \lambda^2 \left( \sum_{t'=1}^{T} d_{t'} \xi_{2t'}^2 \right)}$$  \hspace{1cm} (21)

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

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

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

**Large $T$ approximation** When $T$ is large, weights (21) can be written as

$$w_t = \frac{\tilde{d}_t \sum_{t'=1}^{T} \tilde{d}_{t'} (\xi_{2,T+1} - \xi_{2t'}) (\xi_t - \xi_{2t'})}{\sum_{t'=1}^{T} \tilde{d}_{t'} \left( \xi_{t'} - \sum_{t'=1}^{T} \tilde{d}_{t'} \xi_{2t'} \right)^2} + O(T^{-2})$$  \hspace{1cm} (23)

where $\tilde{d}_t = d_t / (\sum_{t'=1}^{T} d_{t'})$, and derivations are provided in Appendix A.2.1.

**Constant state variance** The interpretation of (21) and (23) is complicated by the fact that $\xi_{2t}$ 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 possible combinations of $\xi_{2t}$ 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

$$\sigma_s^2 = \xi_{2,t}(1 - \xi_{2,t})$$

Summing $\sigma_s^2$ over $t$ and solving for $\sigma_s^2$ yields

$$\sigma_s^2 = \bar{\xi}_1 \bar{\xi}_2 - \frac{1}{T} \sum_t (\xi_{2t} - \bar{\xi}_2)^2$$  \hspace{1cm} (24)
where $\bar{\xi}_1 = \frac{1}{T} \sum_{t=1}^{T} \xi_{1t}$ and $\bar{\xi}_2 = \frac{1}{T} \sum_{t=1}^{T} \xi_{2t}$. Note that the maximum value of $\sigma_s^2$ is given by $\xi_2 \xi_1$, which occurs when the probability vector is constant. In the case of a constant $\sigma_s^2$, $\tilde{d}_t$ simplifies to $1/T$. After some straightforward algebra, (21) can be written as

$$w_t = \frac{1}{T} \left( 1 + \frac{\lambda^2 (\xi_{2,T+1} - \bar{\xi}_2)(\xi_{2,t} - \bar{\xi}_2)}{(Td)^{-1} + \lambda^2 (\xi_1 \xi_2 - \sigma_s^2)} \right)$$

and the large $T$ approximation (23) as

$$w_t = \frac{1}{T} + \frac{(\xi_{2,T+1} - \bar{\xi}_2)(\xi_{2,t} - \bar{\xi}_2)}{T(\xi_1 \xi_2 - \sigma_s^2)} \quad (25)$$

The standard Markov switching weights can be expressed as

$$w_t^{MS} = \frac{1}{T} + \frac{(\xi_{2,T+1} - \bar{\xi}_2)(\xi_{2,t} - \bar{\xi}_2)}{T \xi_1 \xi_2} \quad (26)$$

see Appendix A.2.2. From a comparison of (25) and (26) it is clear that the two weights differ by the factor $\sigma_s^2$ in the denominator and that this difference will not disappear asymptotically. Effectively, the Markov switching weights are more conservative as the optimal weights exploit the regime switching structure more strongly because of the smaller denominator in (25) compared to (26).

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

$$E[\sigma_{2,T+1}^2]^{opt} = \left[ 1 + \frac{\lambda^2 \xi_{2,T+1}(1 - \xi_{2,T+1})}{1 + \frac{\lambda^2 \xi_{2,T+1} - \bar{\xi}_2}{\sigma_s^2} \left( 1 + \frac{\lambda^2 \xi_{2,T+1} - \bar{\xi}_2}{\sigma_s^2} \right)^{-1}} \right]$$

$$E[\sigma_{2,T+1}^2]^{MS} = 1 + \frac{\lambda^2 \xi_{2,T+1}(1 - \xi_{2,T+1})}{1 + \frac{\lambda^2 \xi_{2,T+1} - \bar{\xi}_2}{\sigma_s^2} \left( 1 + \frac{\lambda^2 \xi_{2,T+1} - \bar{\xi}_2}{\sigma_s^2} \right)^{-1}} + \frac{\xi_{2,T+1} - \bar{\xi}_2}{\xi_{2}(1 - \xi_2)} \left( \frac{\lambda^2 \xi_{2}^2}{\sigma_s^2} + 1 \right) \quad (27)$$

The MSFE for the optimal weights is derived from (22) by substituting in the weights in (21) and using the fact that $\tilde{d}_t = 1/T$ and $d_i = d$, $\forall t$. 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 (27) over (28) for different values of $\bar{\xi}_2$ against different values of $\sigma_s^2 = \sigma_s^2/\xi_2 \xi_1$ for $T = 100$. The results indicate that the optimal weights lead to larger gains when $\lambda$ is large and when $\bar{\xi}_2$ is closer to 0.5. The influence of $\sigma_s^2$ 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 $T = 100$

<table>
<thead>
<tr>
<th>$\tilde{\sigma}_s^2$</th>
<th>$\xi_2$</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.993</td>
<td>0.986</td>
<td>0.981</td>
<td>0.979</td>
<td>0.978</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.977</td>
<td>0.960</td>
<td>0.950</td>
<td>0.944</td>
<td>0.942</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.967</td>
<td>0.946</td>
<td>0.934</td>
<td>0.927</td>
<td>0.926</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.974</td>
<td>0.957</td>
<td>0.948</td>
<td>0.944</td>
<td>0.942</td>
<td></td>
</tr>
<tr>
<td>$\lambda = 3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.982</td>
<td>0.969</td>
<td>0.962</td>
<td>0.958</td>
<td>0.957</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.951</td>
<td>0.926</td>
<td>0.913</td>
<td>0.907</td>
<td>0.905</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>0.935</td>
<td>0.908</td>
<td>0.895</td>
<td>0.889</td>
<td>0.887</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.949</td>
<td>0.930</td>
<td>0.921</td>
<td>0.917</td>
<td>0.916</td>
<td></td>
</tr>
</tbody>
</table>

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 $\sigma_s^2$. $\lambda = (\beta_2 - \beta_1)\sigma$ denotes the difference between regimes, $\bar{\xi}_2$ 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

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

and that under standard Markov switching weights is

$$E[\sigma_0^2 e_{T+1}^2]_{\text{MS}} = 1 + \lambda^2 \xi_{2,T+1}(1 - \xi_{2,T+1}) + \left(\frac{\xi_{2,T+1} - \xi_2}{\xi_2 \xi_1}\right)^2 \lambda^2 \sigma_s^4 + O(T^{-1})$$

The difference between (30) and (29) is positive and does not disappear asymptotically. The relative improvement is expected to be high when $\lambda$, $\sigma_s^2$, and the difference $\xi_{2,T+1} - \xi_2$ are large.

### 3.2.2 $m$-state Markov switching models

The derivation of the two-state weights can be extended to an arbitrary number of states. Note that $M = Q + E[\tilde{S}' \lambda \lambda' \tilde{S}]$ and that we can write

$$E[\tilde{S}' \lambda \lambda' \tilde{S}] = E[\tilde{S}' \lambda \lambda' E[\tilde{S}]] + A$$
where, conditional on the state probabilities, \( \xi_{jt}, j = 1, 2, \ldots, m, \)

\[
A = \sum_{j=2}^{m} \lambda_j^2 \Xi_j - \left( \sum_{j=2}^{m} \lambda_j \Xi_j \right)^2
\]

and \( \Xi_j \) is a \( T \times T \) diagonal matrix with typical element \( \xi_{jt} \). Define \( \tilde{\xi} = E[S] \lambda \), which is a \( T \times 1 \) vector, and \( D = Q + A \). Then the inverse of \( M \) is

\[
M^{-1} = D^{-1} - \frac{1}{1 + \xi D^{-1} \xi} D^{-1} \tilde{\xi} \tilde{\xi}^\top D^{-1}
\]

Then we can use (7) to derive the weights similar to the case of the two-state weights and resulting in weights

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

where now we have

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

where we have used the fact that \( \lambda_1 = 0 \).

Examples of weights for a three state Markov switching model over a range of \( \lambda_2 \) for \( T = 100, \pi_1 = 0.2, \pi_2 = \pi_3 = 0.4 \) and \( \lambda_1 = -2.5 \) 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 \( \xi_{it} \) for the states \( s_{it} \), that is, the plug-in estimator of the weights as the black lines. The graph on the right plots the weights (31) as the black lines.

The graph on the left shows how the introduction of the probabilities brings the weights closer to equal weighting compared to the weights for known states. This contrasts with the weights conditional on the probabilities in the plot on the right that are very close to the weights conditional on the states. Hence, using the uncertainty of the states in the derivation of the weights leads to essentially same weights as if the states were known.
Figure 3: Optimal weights for three state Markov switching model

Note: The graphs depict 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 \( \xi_{it} \) in place of \( s_{it} \). In the right plot the darker lines are the weights (31), when \( \hat{\xi}_T = [0.8, 0.1, 0.1]' \) for \( \lambda_3 \) over the range \(-3 \) to \( 3 \), \( \lambda_2 = -2.5 \), \( T = 100 \), \( \pi_0 = 0.2 \), and \( \pi_1 = \pi_2 = 0.4 \). The solid line gives the weights for the observations where \( \hat{\xi}_t = [0.8, 0.1, 0.1]' \), the dashed line those where \( \hat{\xi}_t = [0.1, 0.8, 0.1]' \), and the dash-dotted line those for \( \hat{\xi}_t = [0.1, 0.1, 0.8]' \).

The MSFE for both, the Markov switching and the optimal weights, are 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 \( \lambda_2 \) and \( \lambda_3 \) are large and nearly equal the MSFE shows a sharp increase, towards values that are almost two times as high as the MSFE for the optimal weights. Hence, for these values 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 \( E[s_{it}s_{j,t+m}] = \xi_{it}\xi_{j,t+m} \). While this assumption allows us to find an explicit inverse of the matrix \( M \) 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 \( \hat{M} \) directly from the data. One can then use \( \hat{M} \) as a plug-in estimator for \( M \).

We now condition on the information set up to time \( T \), denoted \( \Omega_T \). Then \( E[s_{it}s_{j,t+m}|\Omega_T] = p(s_{j,t+m} = 1|\Omega_T)p(s_{it} = 1|s_{j,t+m} = 1, \Omega_T) \). 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
Figure 4: (Relative) MSFE under Markov switching and optimal weights

(a) MSFE under MS 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 $\lambda_2$ and $\lambda_3$. Figure (c) displays the ratio of the MSFE of the optimal weights relative to that of the standard MSFE forecast.
Nelson 1998). The second term can be written as

\[ p(s_{it} = 1|s_{j,t+m} = 1, \Omega_T) = \frac{\xi_{it}^i}{\xi_{j,t+m|t+m-1}} \left[ \prod_{l=1}^{m-1} (P'A_{t+l}) \right]_{i,j} \]  

(32)

where \( A \) is a \( m \times m \) diagonal matrix with typical \( i,i \)-element \( \xi_{it} \), and \( \xi_{it} \) and \( \xi_{it-1} \) denote the filtered and forecast probabilities of state \( i \) at time \( t \). The derivation of (32) can be found in Appendix A.2.4. Using these expressions we can calculate the expectations in (7). Define

\[ \Xi^s = \left[ \left( \prod_{l=1}^{k-1} (P'A_{t+l}) \right) \right]_{2:m,2:m} \]

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

\[ E[\tilde{s}_t \tilde{s}_{t+k}] = \Xi_{t|t} \Xi^s (\Xi_{t+k|t} \div \Xi_{t+k|t+k-1}) \]

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

\[ E[\tilde{S}'\lambda'\tilde{S}]_{t,t} = \lambda' \text{diag}(E[\tilde{s}_t]) \lambda \]

\[ E[\tilde{S}'\lambda'\tilde{S}]_{t,t+k} = \lambda' E[\tilde{s}_t \tilde{s}_{t+k}] \lambda \]  

(33)

Using (33) 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

\[ y = \sum_{i=1}^{m} S_i (X\beta_i + \sigma_i \varepsilon) \]

\[ = X\beta_1 + \sum_{i=1}^{m} S_i X(\beta_i - \beta_1) + \sum_{i=1}^{m} S_i \sigma_i \varepsilon \]

where \( S_i \) 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, \( X \) a \( T \times k \) matrix of exogenous regressors and \( \beta_i \) a \( k \times 1 \) vector of parameters, \( \sigma_i \) the variance of regime \( i \), and we used the fact that \( \sum_{i=1}^{m} S_i = I \). Also,

\[ y_{T+1} = x_{T+1}' \beta_1 + \sum_{i=2}^{m} s_{i,T+1} x_{T+1}' (\beta_i - \beta_1) + \sum_{i=1}^{m} s_{i,T+1} \sigma_i \varepsilon_{T+1} \]
As before, we define the optimally weighted estimator as follows

$$\beta(w) = (X'WX)^{-1}X'Wy$$

The optimal forecast is then given by $\hat{y}_{T+1} = x'_{T+1}\beta(w)$.

Define $\lambda_i = (\beta_i - \beta_j) / \sigma_m$, $q_i = \sigma_i / \sigma_m$ and $\Lambda_{ij} = \lambda_i \lambda'_j$. The expected MSFE is then given by

$$E(\sigma_m^{-2} \epsilon^2_{T+1}) = \sum_{i=1}^m E[s_{i,T+1}]x'_{T+1} \Lambda_{ij} x_{T+1} + \sum_{i=1}^m E[s_{i,T+1}]q_i^2 \epsilon^2_{T+1}$$

$$+ x'_{T+1}(X'WX)^{-1} \sum_{i=1}^m \sum_{j=1}^m E[(X'WS_jX)\Lambda_{ij}(X'S_jWX)](X'WX)^{-1}x_{T+1}$$

$$+ x'_{T+1}(X'WX)^{-1} \sum_{i=1}^m q_i^2 X'WE[S_j]WX(X'WX)^{-1}x_{T+1}$$

$$- 2x'_{T+1}(X'WX)^{-1} \sum_{i=1}^m \sum_{j=1}^m E[X'WS_jX\Lambda_{ij}s_j,T+1]x_{T+1}$$

As in the case of structural breaks analyzed by Pesaran et al. (2013), we use large sample approximations to (34) to obtain analytical expressions for the weights. We make the following approximations: $\text{plim}_{T \to \infty} X'WX = \Omega_{XX}$, $\text{plim}_{T \to \infty} X'S_jWX = \Omega_{XX} w's_i$, $\text{plim}_{T \to \infty} X'W^2S_jX = \Omega_{XX} w'Xw$. Then, (34) reduces to

$$E(\sigma_m^{-2} \epsilon^2_{T+1}) = \sum_{i=1}^m E[s_{i,T+1}]x'_{T+1} \Lambda_{ij} x_{T+1} + \sum_{i=1}^m E[s_{i,T+1}]q_i^2 \epsilon^2_{T+1}$$

$$+ \sum_{i=1}^m \sum_{j=1}^m w'E[s_{i}s'_j]w\Lambda_{ij} x_{T+1} + x'_{T+1}\Omega_{XX}^{-1} \sum_{i=1}^m q_i^2 w'E[S_j]wx_{T+1}$$

$$- 2x'_{T+1} \sum_{i=1}^m \sum_{j=1}^m w'E[s_{i}s_j,T+1]\Lambda_{ij} x_{T+1}$$

Maximizing (35) subject to $\ell'w = 1$ leads to the following first order conditions for $w$,

$$\frac{\partial E(\sigma_m^{-2} \epsilon^2_{T+1})}{\partial w} = 2 \sum_{i=1}^m \sum_{j=1}^m x'_{T+1} \Lambda_{ij} x_{T+1} E[s_{i}s'_j]w + 2 \sum_{i=1}^m x'_{T+1} \Omega_{XX}^{-1} x_{T+1} q_i^2 E[S_i]w$$

$$- 2 \sum_{i=1}^m \sum_{j=1}^m x'_{T+1} \Lambda_{ij} x_{T+1} E[s_{i}s_j,T+1] + \theta \ell = 0$$

Define $\phi_i = x'_{T+1} \lambda_i / (x'_{T+1} \Omega_{XX}^{-1} x_{T+1})^{1/2}$, solving for the weights yields

$$w = \left(\left[E[S'\phi\phi'S] + Q\right]^{-1} E[S\phi\phi'S_{T+1}] - \theta \ell$$
which is identical to (7) with the exception that $\lambda$ is replaced by $\phi$. Hence,

$$w = M^{-1}E\left(\tilde{S}'\phi\tilde{S}_{T+1}\right) + \frac{M^{-1}t}{\nu'M^{-1}t} \left[1 - \nu'M^{-1}E\left(\tilde{S}'\phi\tilde{S}_{T+1}\right)\right]$$

(36)

with $M = \left[ Q + E\left(\tilde{S}'\phi\tilde{S}\right) \right]$ and $Q = E[\text{diag}(S'qq'S)]$, and

$$E(\sigma_m^2e_T^2)_{opt} = \frac{\left(1 - \nu'M^{-1}E\left(\tilde{S}'\phi\tilde{S}_{T+1}\right)\right)^2}{\nu'M^{-1}t} + E\left(\tilde{S}_{T+1}^t\phi\tilde{S}_{T+1}\right) - E\left(\tilde{S}'\phi\tilde{S}_{T+1}\right)'M^{-1}E\left(\tilde{S}'\phi\tilde{S}_{T+1}\right) + E\left(\left(q's_{T+1}\right)^2\right)$$

All the formula derived above under the assumption of known states or when there is uncertainty around the states can be straightforwardly extended to allow for exogenous regressors by replacing $\lambda$ with $\phi$.

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 $\sigma^2 = 0.25$ and use a range of values for $\lambda_i$. We will distinguish experiments based on the size of the switches, $\lambda_i$.

The states are generated by a Markov chain with transition probabilities $p_{ij} = \frac{1}{\tau_i}$ for $i \neq j$, which creates Markov chains with relatively high persistency. Furthermore, the ergodic probabilities are assumed to be equal for all states such that $\pi_i = \pi = 1/m$, with $m$ again the number of states. The diagonal elements of the transition probability matrix are $p_{ii} = 1 - \sum_{j=1}^{m} p_{ij}$. The first state is sampled from the ergodic probability vector, $s_1 \sim \text{binomial}(1, \pi)$. Subsequent states are drawn as $s_t \sim \text{binomial}(1, p_t)$ where $p_t = Ps_{t-1}$. 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 subsequently for $T = 100$. We then add an exogenous regressor to a two state model, such that $x_t = [1, z_t]'$ where $z_t \sim N(0, 0.25)$. 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 for computational efficiency we restrict the analysis to the simple mean only model.
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 $\beta_i$ is not identified, we impose $\frac{1}{T} \sum_{t=1}^{T} \xi^i_t > 0.05$ 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 $\hat{\beta}_i$, $\hat{P}$, $\hat{\sigma}$ and the probability vectors $\hat{\xi}_t|T$, $\hat{\xi}_t|y$, $\hat{\xi}_t|t-1$ we construct the usual Markov switching forecast as

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

The optimal weights are calculated as outlined in the sections above. The following notation is used to distinguish the weights derived under different assumptions on the content of the information set.

- $w^{\hat{\xi}}$: plug-in weights that use estimates of the parameters and substitutes the smoothed probability vector $\hat{\xi}_t|T$ for the states.
- $w^{\hat{\xi}_t}$: weights conditional of the state probabilities. This probability is taken to be the smoothed probability vector $\hat{\xi}_t|T$.
- $w^{\hat{\xi}_{asy}}$: the asymptotic limit of the weights conditional on state probabilities, $\lim_{T \to \infty} w^{\hat{\xi}}$.
- $w^{\hat{M}}$: the weights derived by directly estimating the matrix $\hat{M}$ as in Section 3.3.

Using these weights the optimal forecast is constructed as

$$\hat{y}_{T+1}^{\text{opt}} = x_{T+1}' (X'WX)^{-1} X'Wy$$

where $W$ is a diagonal matrix with typical diagonal element $w_{i,t}$ where $w_{i,t} \in \{w^{\hat{\beta}}, w^{\hat{\xi}}, w^{\hat{\xi}_{asy}}, w^{\hat{M}} \}$. The results are presented as the ratio of the MSFE of the optimally weighted forecast to the usual Markov switching forecast.

### 5.2 Monte Carlo results

The results will be separated for different values of $\lambda_i$ 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. We
therefore also separate the results based on the normalized variance of the
smoothed probability vector given by

\[
\hat{\sigma}^2_{\xi} = \frac{\frac{1}{T} \sum_{t=1}^{T} \hat{\xi}_{i|T}(1 - \hat{\xi}_{i|T})}{\frac{1}{T} \sum_{t=1}^{T} \hat{\xi}_{i|T} \frac{1}{T} \sum_{t=1}^{T} (1 - \hat{\xi}_{i|T})}
\]

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

All results are obtained by generating 10,000 data sets that have the corresponding \(\lambda\) and \(\hat{\sigma}^2_{\xi}\).

### 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. When using weights \(w_{\hat{\xi}}\), which are based on the parameter estimates and smoothed probabilities in place of the known states, we expect to 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 \(\lambda\) takes larger values. This contrasts with the results for the weights \(w_{\hat{\xi}}, w_{\hat{\xi}|\text{asy}}\) and \(w_{\hat{\xi}|\text{M}}\). When \(\lambda = 1\) the estimation uncertainty in the parameters outweighs the potential improvement in MSFE but when \(\lambda\) increases the improvements are quite substantial, especially when the variance in the smoothed probability vector is high. The differences between the forecast performance of \(w_{\hat{\xi}}, w_{\hat{\xi}|\text{asy}}\) and \(w_{\hat{\xi}|\text{M}}\) are small. The forecasts from \(w_{\hat{\xi}|\text{asy}}\) lead to the best but also to the worst forecast, whereas forecasts from \(w_{\hat{\xi}}\) are more conservative and generally beat those from \(w_{\hat{\xi}|\text{M}}\).

Theoretically, the weights \(w_{\hat{\xi}}, w_{\hat{\xi}|\text{asy}}\) and \(w_{\hat{\xi}|\text{M}}\) 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 \(\lambda\) and high \(\hat{\sigma}^2_{\xi}\) and that these improvements are more pronounced in larger data sets. The weights \(w_{\hat{\xi}}\) lead for 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
Table 4: Monte Carlo results: two states, intercept only models

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\tilde{\sigma}^2_{\xi T}$</th>
<th>$T = 50$</th>
<th>$T = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$w_\beta$</td>
<td>$w_\xi$</td>
<td>$w_\xi</td>
</tr>
<tr>
<td>$q^2 = 1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0-0.1</td>
<td>0.983</td>
<td>1.004</td>
</tr>
<tr>
<td></td>
<td>0.1-0.2</td>
<td>0.990</td>
<td>1.021</td>
</tr>
<tr>
<td></td>
<td>0.2-0.3</td>
<td>0.996</td>
<td>1.027</td>
</tr>
<tr>
<td></td>
<td>0.3-0.4</td>
<td>0.998</td>
<td>1.028</td>
</tr>
<tr>
<td>2</td>
<td>0.0-0.1</td>
<td>0.995</td>
<td>1.008</td>
</tr>
<tr>
<td></td>
<td>0.1-0.2</td>
<td>1.001</td>
<td>1.005</td>
</tr>
<tr>
<td></td>
<td>0.2-0.3</td>
<td>1.003</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td>0.3-0.4</td>
<td>1.004</td>
<td>0.982</td>
</tr>
<tr>
<td>3</td>
<td>0.0-0.1</td>
<td>1.000</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>0.1-0.2</td>
<td>1.005</td>
<td>0.974</td>
</tr>
<tr>
<td></td>
<td>0.2-0.3</td>
<td>1.006</td>
<td>0.957</td>
</tr>
<tr>
<td></td>
<td>0.3-0.4</td>
<td>1.006</td>
<td>0.957</td>
</tr>
<tr>
<td>$q^2 = 2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0-0.1</td>
<td>0.985</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>0.1-0.2</td>
<td>0.990</td>
<td>1.009</td>
</tr>
<tr>
<td></td>
<td>0.2-0.3</td>
<td>0.996</td>
<td>1.022</td>
</tr>
<tr>
<td></td>
<td>0.3-0.4</td>
<td>0.998</td>
<td>1.017</td>
</tr>
<tr>
<td>2</td>
<td>0.0-0.1</td>
<td>0.993</td>
<td>1.006</td>
</tr>
<tr>
<td></td>
<td>0.1-0.2</td>
<td>0.999</td>
<td>1.012</td>
</tr>
<tr>
<td></td>
<td>0.2-0.3</td>
<td>1.003</td>
<td>1.001</td>
</tr>
<tr>
<td></td>
<td>0.3-0.4</td>
<td>1.004</td>
<td>0.993</td>
</tr>
<tr>
<td>3</td>
<td>0.0-0.1</td>
<td>0.999</td>
<td>1.002</td>
</tr>
<tr>
<td></td>
<td>0.1-0.2</td>
<td>1.004</td>
<td>0.986</td>
</tr>
<tr>
<td></td>
<td>0.2-0.3</td>
<td>1.007</td>
<td>0.965</td>
</tr>
<tr>
<td></td>
<td>0.3-0.4</td>
<td>1.010</td>
<td>0.942</td>
</tr>
</tbody>
</table>

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 $\varepsilon_t \sim N(0, 1)$, $\sigma_1^2 = 0.25$, $q^2 = \sigma_1^2 / \sigma_2^2$. Column labels: $\lambda = (\beta_2 - \beta_1) / \sigma_1$, $\tilde{\sigma}^2_{\xi T}$ is the normalized variance in of the smoothed probability vector. $w_\beta$: forecasts from weights based on estimated parameters and state probabilities. $w_\xi$: forecasts from weights conditional on state probabilities. $w_\xi|_{asy}$: forecasts from weights based on numerically inverting $\tilde{\sigma}^2_{\xi T}$. $w_M$: the weights based on numerically inverting $\tilde{M}$. $w_M$.
Table 5: Monte Carlo results: two states, models with exogenous regressors

<table>
<thead>
<tr>
<th>λ</th>
<th>$\tilde{\sigma}^2_{\xi,T}$</th>
<th>$T = 50$</th>
<th>$T = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$w_\delta$</td>
<td>$w_\xi$</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.962</td>
<td>0.986</td>
</tr>
<tr>
<td>1.0-0.1</td>
<td>0.973</td>
<td>1.021</td>
<td>1.001</td>
</tr>
<tr>
<td>0.2-0.3</td>
<td>0.991</td>
<td>1.025</td>
<td>1.021</td>
</tr>
<tr>
<td>0.3-0.4</td>
<td>0.995</td>
<td>1.030</td>
<td>1.028</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.990</td>
<td>1.000</td>
</tr>
<tr>
<td>0.1-0.2</td>
<td>1.004</td>
<td>1.008</td>
<td>1.016</td>
</tr>
<tr>
<td>0.2-0.3</td>
<td>1.011</td>
<td>0.999</td>
<td>1.013</td>
</tr>
<tr>
<td>0.3-0.4</td>
<td>1.012</td>
<td>0.986</td>
<td>0.999</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1.005</td>
<td>1.004</td>
</tr>
<tr>
<td>0.1-0.2</td>
<td>1.018</td>
<td>0.998</td>
<td>1.026</td>
</tr>
<tr>
<td>0.2-0.3</td>
<td>1.031</td>
<td>0.983</td>
<td>1.010</td>
</tr>
<tr>
<td>0.3-0.4</td>
<td>1.020</td>
<td>0.969</td>
<td>0.991</td>
</tr>
</tbody>
</table>

Note: The table reports the ratio of the MSFE of the optimal asymptotic weights to that of the Markov switching weights. DGP: $y_t = x_t^T \beta_1 + \sigma (x_t^T \lambda s_t + \varepsilon_t)$ 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.

Substantial improvements remain in the same parameter regions where the weights under constant variance perform well.

In Table 5 we display the results when we add an exogenous regressor to the model. The optimal forecast can be obtained by using an asymptotic approximation as in (36). As the ratio of parameters to estimate versus the number of observations increases we expect an increase in the performance of the weights $w_\delta$. This is indeed the case, although improvements are small. The improvements for the weights $w_\xi$ and $w_{\xi|\text{asy}}$ are almost equal. The weights $w_\tilde{\xi}$, in contrast, perform worse than the Markov switching weights in almost all parameter regions.

5.2.2 Monte Carlo results for three state models

From the results in Table 6 we see that for three states the improvement in MSFE for $w_\delta$, $w_{\xi|\text{asy}}$ and $w_\tilde{\xi}$ are somewhat less pronounced than in two states. This confirms the theoretical results obtained above which suggests that the improvements are largest for two states. In the Monte Carlo experiment we find that only for $0.2 < \tilde{\sigma}^2_\xi < 0.3$ sizable improvements are made.
Table 6: Monte Carlo results: three states, intercept only models

| \{\lambda_{31}, \lambda_{21}\} | \hat{\sigma}^2_{\xi|T} | w_\hat{\delta} | w_\hat{\xi} | w_{\hat{\xi}|\text{asy}} | w_M | w_\hat{\delta} | w_\hat{\xi} | w_{\hat{\xi}|\text{asy}} | w_M |
|-----------------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| \{2,1\} | 0.0-0.1 | 0.996 | 1.029 | 1.036 | 1.026 | 0.998 | 1.022 | 1.026 | 1.027 |
| | 0.1-0.2 | 0.997 | 1.031 | 1.039 | 1.035 | 0.999 | 1.022 | 1.025 | 1.042 |
| | 0.2-0.3 | 0.999 | 1.028 | 1.036 | 1.035 | 1.000 | 1.018 | 1.022 | 1.032 |
| | 0.3-0.4 | 1.001 | 1.013 | 1.018 | 1.017 |
| \{3,1\} | 0.0-0.1 | 0.998 | 1.016 | 1.020 | 1.014 | 0.999 | 1.010 | 1.012 | 1.020 |
| | 0.1-0.2 | 1.000 | 1.008 | 1.011 | 1.013 | 1.001 | 1.001 | 1.002 | 1.023 |
| | 0.2-0.3 | 1.002 | 0.990 | 0.990 | 0.993 | 1.002 | 0.971 | 0.970 | 0.988 |
| | 0.3-0.4 | 1.004 | 0.969 | 0.965 | 0.966 |
| \{4,2\} | 0.0-0.1 | 0.999 | 1.011 | 1.013 | 1.011 | 1.000 | 1.004 | 1.004 | 1.014 |
| | 0.1-0.2 | 1.001 | 0.997 | 0.998 | 0.999 | 1.001 | 0.985 | 0.984 | 1.011 |
| | 0.2-0.3 | 1.003 | 0.973 | 0.971 | 0.971 | 1.002 | 0.947 | 0.946 | 0.970 |
| | 0.3-0.4 | 1.003 | 0.954 | 0.950 | 0.951 |

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.

The performance of \( w_\hat{\delta} \) is very similar to the performance of the Markov switching weights. Overall, however, the conclusions from the experiments with two state models carry over to the case of three state models.

6 Application to US GNP

Models of business cycles in US GNP have been the original application of Markov switching model by Hamilton (1989) and they have remained one of the most important applications. Many variants of Markov switching models have been used to analyze business cycles, and we will use the classification scheme by Krolzig (1997) to label the different models. Applications of different models to US GNP can be found in Clements and Krolzig (1998), Krolzig (1997) and Krolzig (2000).

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 \( p = 4 \):

\[
y_t = \beta_{s_t} + \sum_{i=1}^{p} \phi_i (y_{t-i} - \beta_{s_{t-i}}) + \sigma \varepsilon_t
\]
where \( s_t = \{1, 2\} \). Here, \( y_t \) depends on the current state but also on the previous \( p = 4 \) states. We can therefore rewrite this model as

\[
y_t = \beta_{s_t} + \sum_{i=1}^{p} \phi_i y_{t-i} + \sigma \varepsilon_t
\]

where \( s_t^* \) can take on \( m^{p+1} = 2^5 = 32 \) states for an MSM(2)-AR(4) model, and \( 4^5 = 1024 \) states for an MSM(4)-AR(4) model. In practice, this means that only MSM(m)-AR(p) models with low \( m \) or low \( p \) are feasible to estimate. If the model has a state dependent variance 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 \( p = 4 \):

\[
y_t = \beta_{s_t} + \sum_{i=1}^{p} \phi_i y_{t-i} + \sigma \varepsilon_t
\]

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 again 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 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. We select the Markov switching model that delivers the best forecast, where the selection is based on the historic forecast performance of the standard Markov switching model as measured by the MSFE. Using this model, we then compare the 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 2013Q3 obtained from the Federal Reserve Economic Data (FRED). The data is seasonality adjusted. In total the series consists of 266 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 1948Q1 so that the models that use lagged dependent variables can all be initialized based on the available data.
Table 7: GNP forecasts: relative forecasting performance

|                | $w_{MS}$ | $w_{\hat{s}}$ | $w_{\xi}$ | $w_{\xi|asy}$ | $w_{M}$ |
|----------------|----------|----------------|-----------|----------------|---------|
| 1983-1993      | 0.207    | 0.208          | 0.193     | 0.192          | 0.188   |
| 1993-2003      | 0.286    | 0.286          | 0.288     | 0.288          | 0.295   |
| 2003-2013      | 0.544    | 0.544          | 0.527     | 0.527          | 0.515   |
| 1983-2013      | 0.346    | 0.346          | 0.336     | 0.336          | 0.333   |

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1983-1993</td>
<td>1.002</td>
<td>0.929*</td>
<td>0.927*</td>
<td>0.905*</td>
<td></td>
</tr>
<tr>
<td>1993-2003</td>
<td>0.999</td>
<td>1.005</td>
<td>1.005</td>
<td>1.031</td>
<td></td>
</tr>
<tr>
<td>2003-2013</td>
<td>1.000</td>
<td>0.969**</td>
<td>0.969**</td>
<td>0.947**</td>
<td></td>
</tr>
<tr>
<td>1983-2013</td>
<td>1.001</td>
<td>0.971**</td>
<td>0.970**</td>
<td>0.962**</td>
<td></td>
</tr>
</tbody>
</table>

Note: The top half of the table reports the MSFE based on the best Markov switching model as indicated by historical forecast performance. Historical forecast performance is calculated using an expanding window that starts 40 quarters prior to the first actual forecast. The bottom half 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.

We use the following set of candidate 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 $p = 0, 1, 2, 3, 4$. In total we thus estimate 32 models (as MSM$(m)$-AR$(0)$ and MSIH$(m)$-AR$(0)$ are identical). At each step in the forecasting exercise all the models are re-estimated to include all the available data at that point in time.

The out-of-sample forecast period is 1983Q4-2013Q3, which amounts to 120 observations. We start evaluating forecasts for model selection purposes based on a training period 1973Q4-1983Q3 (40 observations) and calculating the MSFE over this period. Then the model that is used to forecast 1983Q4 is selected. In this way no information is used that is not available to a researcher at that point in time. In the next period we evaluate the forecast error and determine the model we will use in 1984Q1 by recalculating the MSFE in expanding window from 1973Q4. Based on this procedure the following models are selected: MSM$(3)$-AR$(1)$ for 1-67, MSI$(3)$-AR$(0)$ for 68-73, 95, 100-101, MSI$(3)$-AR$(2)$ for 74-94, 96-99,102-103,107-120 and
Table 8: GNP forecasts: comparison to linear models

<table>
<thead>
<tr>
<th></th>
<th>AR dyn</th>
<th>$w_{MS}$</th>
<th>$w_s^\hat{}$</th>
<th>$w^\hat{}_\xi$</th>
<th>$w^\hat{}<em>{\xi</em>{asy}}$</th>
<th>$w^\hat{}_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1983-1993</td>
<td>0.224</td>
<td>0.928*</td>
<td>0.929*</td>
<td>0.862**</td>
<td>0.860**</td>
<td>0.840**</td>
</tr>
<tr>
<td>1993-2003</td>
<td>0.284</td>
<td>1.008</td>
<td>1.013</td>
<td>1.013</td>
<td>1.031</td>
<td>1.039</td>
</tr>
<tr>
<td>2003-2013</td>
<td>0.511</td>
<td>1.065</td>
<td>1.065</td>
<td>1.032</td>
<td>1.031</td>
<td>1.008</td>
</tr>
<tr>
<td>1983-2013</td>
<td>0.339</td>
<td>1.019</td>
<td>1.019</td>
<td>0.989</td>
<td>0.989</td>
<td>0.980</td>
</tr>
</tbody>
</table>

Note: The table reports in the first column the MSFE of a forecasting exercise where linear alternative to the MS model is selected based on historical performance of the various linear alternatives. This selects the AR(1) model for the first 68 forecasts and AR(2) for the final 52 forecasts. The other columns give the relative performance of the selected MS model under the various weights.

MSM(2)-AR(2) for 104-106.

The forecasting performances are reported in Table 7. The results show that, over the entire forecast period, the optimal weighting schemes conditional on the state probabilities improve the forecast performance over the standard weights and that the improvements are significant. The largest improvement is achieved by $w^\hat{}_M$. The three state models have an average estimated break size $\hat{\lambda}_{21} = 1.99$ and $\hat{\lambda}_{31} = 3.78$. The average minimum normalized variance of the smoothed probability vector $\hat{\sigma}^2_{\xi|T} = 0.26$. The two state model that occurs for three periods has $\hat{\lambda} = 2.70$. 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.

We also compare forecast performance in subperiods. In the first subperiod, 1983Q4–1993Q3, forecasts based on the optimal weights conditional on the state probabilities improve significantly over the standard weights with the largest improvement coming from $w^\hat{}_M$. Forecasts based on the plug-in weights, $w_s$, in contrast, cannot improve on the standard MS forecasts. In the second subperiod, 1993Q4–2002Q3, which largely covers the great moderation, only $w_s$ offers a marginal improvement, which is, however, not statistically significant. The forecasts that use weights based on state probabilities fail to improve on the standard MS forecasts, and the worst forecasts are from $w^\hat{}_M$. In the last subperiod, 2003Q4–2013Q3, again all optimal weights conditional on the state probabilities lead to more precise forecasts than the standard weights and these improvements are again significant.

In order to understand the differences in forecast performance over the subperiods, we also compare the forecasting performance with a selection
of linear alternatives. These include a random walk forecast and AR($p$) models with $p = 1, 2, 3, 4$. We compare the Markov switching forecasts to the best AR($p$) model based on historic forecast performance in line with out model selection for the Markov switching model. The AR(1) model is selected for the first 68 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 linear models beat the Markov switching model with standard weights, which is a standard result in the literature. The same is true for the weights conditional on the states. This contrasts with the forecast based on optimal weights conditional on state probabilities, which 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, the optimal weights conditional on the state probabilities improve over the linear models but in the middle and last subperiod the linear model cannot be beaten. Comparing these results to those in Table 7, suggests that the optimal weights do well when the data exhibit strong switching behavior but do worse than the standard weights when the data do not exhibit strong switching behavior and Markov switching models are beaten by a linear model. This ties in with the results from our theory that suggests that the weights conditional on the states are tending more towards equal weighting, that is in the direction of the linear models, whereas the optimal weights emphasize the Markov switching nature of the data.

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 influence of uncertainty around the timing of the breaks is shown by first deriving optimal forecasts when the states of the Markov chain are assumed to be known. The optimal weights 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. Optimal weights are then derived under the assumption that the states of the Markov chain are uncertain. In this case the optimal weights 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 theory shows that the optimal weights in this case aim to correct for
the uncertainty around the states. The performance of the forecasts constructed using optimal weights is tested through Monte Carlo simulations. These simulations show that conditional on the states, improvements can be made if both sample size and the size of the break are small. Taking into account that the states are uncertain, improvements can be made when the sample size and break size are large. We apply the theoretical results in an empirical application on forecasting quarterly US GNP. The results show that the optimally weighted forecasts outperform both, the standard Markov switching forecast and the best linear alternative.

The theoretical results on the importance of uncertainty of states carry over the case of discrete structural breaks. In this case, however, finite sample results for the probability distribution of break dates are difficult to obtain. Practical implementations of our theoretical results to the case of structural breaks are the subject of ongoing research.

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 $\eta = \frac{\sigma_1}{\sigma_2}$, $\pi_1 = \frac{1}{T} \sum_{t=1}^{T} s_{1t}$, and $\pi_2 = \frac{1}{T} \sum_{t=1}^{T} s_{2t}$. Then we have

$$M = Q + \tilde{S}' \lambda \lambda \tilde{S}$$

$$= q_1^2 S_1 + S_2 + \lambda^2 s_2 s_2'$$

where $S_t$ is a $T \times T$ diagonal matrix with typical $t,t$-element $s_{i,t}$. Of which the inverse is

$$M^{-1} = (q_1^2 S_1 + S_2)^{-1} - \frac{\lambda^2 (q_1^2 S_1 + S_2)^{-1} s_2 s_2' (q_1^2 S_1 + S_2)^{-1}}{1 + \lambda^2 s_2' (q_1^2 S_1 + S_2) s_2}$$

$$= \frac{1}{q^2} S_1 + S_2 - \frac{\lambda^2 s_2 s_2'}{1 + \lambda^2 s_2' (q_1^2 S_1 + S_2) s_2}$$

The weights are given by

$$w = \lambda^2 M^{-1} s_2 s_{2,T+1} + \frac{M^{-1} t}{t' M^{-1} t} (1 - \lambda^2 t' M^{-1} s_2 s_{2,T+1})$$

11
So we need

\[ M^{-1}s_2 = s_2 - \frac{\lambda^2 T \pi_2}{1 + \lambda^2 T \pi_2} s_2 \]

\[ = \frac{1}{1 + \lambda^2 T \pi_2} s_2 \]

\[ M^{-1}l = \frac{1}{q^2} s_1 + s_2 - \frac{\lambda^2 T \pi_2}{1 + \lambda^2 T \pi_2} s_2 \]

\[ = \frac{s_1(1 + \lambda^2 T \pi_2) + q^2 s_2}{q^2(1 + \lambda^2 T \pi_2)} \]

and

\[ u' M^{-1} s_2 = \frac{T \pi_2}{1 + \lambda^2 T \pi_2}, \quad u' M^{-1} l = \frac{T \pi_1 + \lambda^2 T \pi_1 \pi_2 + q^2 \pi_2}{q^2(1 + \lambda^2 T \pi_2)} \]

This yields the weights

\[ w = \frac{\lambda^2}{1 + \lambda^2 T \pi_2} s_{2,s_{2,T+1}} + \frac{1}{T \pi_1 + q^2 \pi_2 + \lambda^2 T \pi_1 \pi_2} \left( s_{2,s_{2,T+1}} + \frac{1}{1 - \lambda^2 T \pi_2 (s_{2,T+1})} \right) \]

Suppose \( s_{2,T+1} = s_{2,t} = 1 \), then

\[ w_{22} = \frac{1}{1 + \lambda^2 T \pi_2} \left( \lambda^2 + \frac{1}{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T} \right) \]

\[ = \frac{1}{1 + \lambda^2 T \pi_2} \left( \frac{1}{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T} \right) \left( q^2(1 + \lambda^2 T \pi_2) + \lambda^2 T \pi_1 (1 + \lambda^2 T \pi_2) \right) \]

\[ = \frac{1}{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T} \]

when \( s_{2,T+1} = s_{2,t} = 0 \), then

\[ w_{21} = \frac{1}{1 + \lambda^2 T \pi_2} \left( \frac{1}{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T} \right) \]

\[ = \frac{1}{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T} \]

when \( s_{2,T+1} = 0, s_{2,t} = 1 \), then

\[ w_{12} = \frac{1}{1 + \lambda^2 T \pi_2} \left( \frac{1}{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T} \right) \]

\[ = \frac{1}{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T} \]

32
finally, when $s_{2,T+1} = 0$, $s_{2,t} = 0$, then

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

$$= \frac{1}{1 + \lambda^2 T \pi_2} \frac{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T}{T^2 \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T}$$

In order to show the symmetry of the weights, consider the definition of $\lambda$ and $q$ conditional on the regime $s_{i,T}$. If $s_{2,T+1} = 1$, define $\lambda = \frac{\beta_2 - \beta_1}{\sigma_2}$ and $q = \frac{q_2}{\sigma_2}$, but if $s_{1,T+1} = 1$, define $\lambda = \frac{\beta_1 - \beta_2}{\sigma_1}$ and $q = \frac{q_2}{\sigma_1}$. Then, $\lambda^2 = \frac{\lambda_2^2}{q_2^2}$ and we have for $w_{12}$ and $w_{11}$

$$w_{12} = \frac{1}{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T}$$

$$= \frac{1}{1 + \lambda^2 T \pi_2} \frac{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T}{T^2 \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T}$$

$$w_{11} = \frac{1 + \lambda^2 T \pi_2}{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T}$$

$$= \frac{1 + \lambda^2 T \pi_2}{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T} \frac{T \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T}{T^2 \pi_1 + q^2 \pi_2 + \lambda^2 \pi_1 \pi_2 T}$$

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 $s_{T+1,k} = 1$ as we have shown above that the weights are symmetric. In this case, define $\lambda = \frac{\beta_i - \beta_k}{\sigma_k}$ and $q = \frac{\sigma_i}{\sigma_k}$. The model is given by

$$y_t = \sum_{i=1}^{m} \beta_i s_{it} + \sum_{i=1}^{m} \sigma_i s_{it} \varepsilon_t$$

$$= \beta_k + \sum_{i=1}^{m} \left( \frac{\beta_i - \beta_k}{\sigma_k} \right) s_{it} + \sum_{i=1}^{m} \sigma_i s_{it} \varepsilon_t$$

$$= \sigma_k \left( \frac{\beta_k}{\sigma_k} + \sum_{i=1}^{m} \lambda_i s_{it} + \sum_{i=1}^{m} q_i s_{it} \varepsilon_t \right)$$

33
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

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

Squaring and taking expectations gives

$$E[\sigma_k^{-2} (y_{T+1} - w'y)^2] = 1 + \sum_{i,j} \lambda_i \lambda_j w'_i s'_j w + \sum_i q_i^2 w_i s_i$$

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

$$w = \left( \sum_{i,j} \lambda_i \lambda_j w'_i s'_j + \sum_i q_i^2 w_i s_i \right)^{-1} (-\theta \lambda)$$

$$= -\theta M^{-1} \lambda$$

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

$$M^{-1} = \sum_{i=1}^{m} \frac{1}{q_i^2} s_i - \frac{\left( \sum_{i=1}^{m} \frac{1}{q_i^2} s_i \right) \left( \sum_{i,j} \frac{1}{q_i^2} \pi_i s_i \right)}{1 + \left( \sum_{j=1}^{m} \lambda_j \pi_j \right) \left( \sum_{i=1}^{m} \frac{1}{q_i^2} s_i \right)}$$

$$= \sum_{i=1}^{m} \frac{1}{q_i^2} s_i - \frac{\sum_{i,j} \frac{\lambda_i}{q_i^2} \lambda_j s'_j}{1 + T \sum_{i=1}^{m} \frac{\lambda_i^2}{q_i^2} \pi_i}$$

Multiplying with $\lambda$ as in equation (37) gives

$$M^{-1} \lambda = \sum_{i=1}^{m} \frac{1}{q_i^2} \pi_i \frac{T \sum_{i,j} \frac{\lambda_i}{q_i^2} s_i}{1 + T \sum_{i=1}^{m} \frac{\lambda_i^2}{q_i^2} \pi_i}$$

Since the weights should sum up to one, we have

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

$$= 1$$
which gives

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

The weights are then given by

\[
w = \frac{1}{T} \sum_{i=1}^{m} \frac{1}{q_i^2} s_i + T \sum_{i,j} \frac{1}{q_i q_j} \frac{1}{2} \pi_j \lambda_j (\lambda_j - \lambda_i) s_i
\]

So that if \( s_{tt} = 1 \) the weight at time \( t \) is

\[
w_t = \frac{1}{T} \sum_{i=1}^{m} \frac{1}{q_i^2} + T \sum_{i,j} \frac{1}{q_i q_j} \frac{1}{2} \pi_j \lambda_j (\lambda_j - \lambda_i)
\]

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

\[
E \left[ \sigma_k^{-2} (y_{T+1} - w^T y)^2 \right] = 1 + \sum_{i,j} \lambda_i \lambda_j w_i^T s_i s_j w + \sum_i q_i^2 w_i^T S_i w
\]

\[
= 1 - \theta
\]

\[
= 1 + w_{kk}
\]

where \( w_{kk} \) is the weight when \( s_{T+1,k} = s_{tk} = 1 \).

A.2 Derivations conditional on state probabilities

A.2.1 Large \( T \) approximation for optimal weights

Rewrite (21) as

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

Furthermore,

\[
d_t = \left[ \lambda^2 \xi_{2t} (1 - \xi_{2t}) + q_1^2 (1 - q_1^2) \xi_{2t} \right]^{-1}
\]

Define

\[
a_t = \lambda^2 \xi_{2t} (1 - \xi_{2t}) + q_1^2 (1 - q_1^2) \xi_{2t}
\]
so that \( d_t = \frac{1}{a_t} \). Note that \( a_t > 0 \) and

\[
\frac{1}{T} \sum_{t=1}^{T} d_t = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{a_t} \leq \frac{1}{T} T \frac{1}{a_{\text{min}}} = \frac{1}{a_{\text{min}}}
\]

where \( a_{\text{min}} \) is the minimum value of \( a_t \) over \( t = 1, 2, \ldots, T \). Similarly, define

\[
\tilde{a}_t = \frac{\xi_{2t}}{a_t} = \frac{1}{\lambda^2(1 - \xi_{2t}) + q^2(1 - \xi_{2t})/\xi_{2t} + 1}
\]

Then \( \tilde{a}_t > 0 \) and

\[
\frac{1}{T} \sum_{t=1}^{T} \xi_{2t} d_t = \frac{1}{T} \sum_{t=1}^{T} \xi_{2t} a_t = \frac{1}{T} \sum_{t=1}^{T} \tilde{a}_t \leq \frac{1}{\tilde{a}_{\text{min}}}
\]

where \( \tilde{a}_{\text{min}} \) is the minimum value of \( \tilde{a}_t \) over \( t = 1, 2, \ldots, T \). Finally, define

\[
\bar{a}_t = \xi_{2t}^{\frac{2}{a_t}} = \frac{1}{\lambda^2(1 - \xi_{2t})/\xi_{2t} + q^2(1 - \xi_{2t})/\xi_{2t}^2 + 1/\xi_{2t}}
\]

Then \( \bar{a}_t > 0 \) and

\[
\frac{1}{T} \sum_{t=1}^{T} \bar{a}_t \xi_{2t} d_t = \frac{1}{T} \sum_{t=1}^{T} \bar{a}_t \xi_{2t} a_t = \frac{1}{T} \sum_{t=1}^{T} \bar{a}_t \tilde{a}_t \leq \frac{1}{\bar{a}_{\text{min}}}
\]

where \( \bar{a}_{\text{min}} \) is the minimum value of \( \bar{a}_t \) over \( t = 1, 2, \ldots, T \).

Denote \( \bar{d} = \frac{1}{T} \sum_{t=1}^{T} d_t \), \( \bar{d}_\xi = \frac{1}{T} \sum_{t=1}^{T} d_t \xi_{2t} \), and \( \bar{d}_\xi^2 = \frac{1}{T} \sum_{t=1}^{T} d_t \xi_{2t}^2 \), then (38) can be written as

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

where \( \theta = \frac{\bar{d}}{\lambda^2 \left[ \bar{d} \xi^2 - d \xi^2 \right]} \). Dividing \( w_t \) by \( \sum_{t=1}^{T} d_t \) yields (23).
A.2.2 Weights and MSFE for standard Markov switching model

The Markov switching weights can be written as

$$w_{\text{MS}} = \frac{\xi_{1,T+1}\xi_1 + \xi_{2,T+1}\xi_2}{\sum_t \xi_{1t}} + \frac{1}{\xi_2} \left(1 - \xi_{2,T+1}\xi_2\right) \left(1 - \xi_2\right)$$

For a general vector of weights $w$, subject to $\sum_{t=1}^{T} w_t = 1$, and assuming a constant error variance, we have the following MSFE

$$E[\sigma^2 e_{T+1}^2] = 1 + \lambda^2 \xi_{2,T+1} + w'\mathbf{M}w - 2\lambda^2 w'\xi_{2,T+1}$$

For a general vector of weights $w$, subject to $\sum_{t=1}^{T} w_t = 1$, and assuming a constant error variance, we have the following MSFE

$$E[\sigma^2 e_{T+1}^2] = 1 + \lambda^2 \xi_{2,T+1} + w'\mathbf{M}w - 2\lambda^2 w'\xi_{2,T+1}$$

where $D = (1 + \lambda^2 \sigma^2)I$.

Using (39) we have that

$$w'_{\text{MS}} \xi = \bar{\xi}_2 + \frac{\xi_{2,T+1} - \bar{\xi}_2}{(1 - \bar{\xi}_2)} \left(\frac{1}{T} \sum_t \xi_t^2 - T\bar{\xi}_2^2\right)$$

$$= \bar{\xi}_2 + \frac{\xi_{2,T+1} - \bar{\xi}_2}{(1 - \bar{\xi}_2)} (\bar{\xi}_2(1 - \bar{\xi}_2) - \sigma^2)$$

$$= \xi_{2,T+1} - \frac{\xi_{2,T+1} - \bar{\xi}_2}{\xi_2(1 - \bar{\xi}_2)} \sigma^2$$

where we have used (24), and

$$w'_{\text{MS}} \mathbf{D}w_{\text{MS}} = (1 + \lambda^2 \sigma^2) \left(1 + \frac{(\xi_{2,T+1} - \bar{\xi}_2)^2}{T\xi_2^2(1 - \bar{\xi}_2)^2}(\bar{\xi}_2(1 - \bar{\xi}_2) - \sigma^2)\right)$$
So that the MSFE is

\[ E[\sigma^{-2}e_{T+1}^2]_{MS} = 1 + \lambda^2 \xi_{2,T+1} + \lambda^2 \left[ \frac{\xi_{2,T+1}^2}{\xi_2(1 - \xi_2)} - 2 \frac{\xi_{2,T+1}(\xi_{2,T+1} - \bar{\xi}_2)\sigma_\xi^2}{\xi_2(1 - \xi_2)} + \frac{(\xi_{2,T+1} - \bar{\xi}_2)^2\sigma_\xi^4}{\xi_2^2(1 - \xi_2)^2} \right] \]

\[ - \lambda^2 \left[ 2\xi_{2,T+1} - 2 \frac{\xi_{2,T+1}(\xi_{2,T+1} - \bar{\xi}_2)\sigma_\xi}{\xi_2(1 - \xi_2)} \right] + (1 + \lambda^2\sigma_\xi^2) \frac{1}{T} \left[ 1 + \frac{(\xi_{2,T+1} - \bar{\xi}_2)^2}{\xi_2^2(1 - \xi_2)^2} (\xi_2(1 - \bar{\xi}_2) - \sigma_\xi^2) \right] \]

\[ = 1 + \lambda^2 \xi_{2,T+1}(1 - \xi_{2,T+1}) + \lambda^2 \frac{(\xi_{2,T+1} - \bar{\xi}_2)^2\sigma_\xi^4}{\xi_2^2(1 - \xi_2)^2} + (1 + \lambda^2\sigma_\xi^2) \frac{1}{T} \left[ 1 + \frac{(\xi_{2,T+1} - \bar{\xi}_2)^2}{\xi_2^2(1 - \xi_2)^2} (\xi_2(1 - \bar{\xi}_2) - \sigma_\xi^2) \right] \]

\[ = 1 + \lambda^2 \xi_{2,T+1}(1 - \xi_{2,T+1}) + \lambda^2 \frac{(\xi_{2,T+1} - \bar{\xi}_2)^2}{\xi_2^2(1 - \xi_2)^2} \left[ \lambda^2\sigma_\xi^4 + (1 + \lambda^2\sigma_\xi^2) \frac{1}{T}(\xi_2(1 - \bar{\xi}_2) - \sigma_\xi^2) \right] \]

### A.2.3 MSFE for Markov switching model using optimal weights

Equation (22) for an arbitrary number of states is derived as follows

\[ E[\sigma^{-2}e_{T+1}^2] = (\iota'M^{-1}\iota)^{-1}(1 - \iota'M^{-1}\bar{\xi}\xi_{T+1})^2 + \sum_{j=2}^{m} \lambda_j^2 \xi_j,T+1 - \bar{\xi}_T^2 \xi'_{T+1}M^{-1}\xi + \sum_{j=1}^{m} q^2_j \xi_{j,T+1} \]

\[ = \frac{1 + \bar{\xi}'D^{-1}\xi}{\iota'D^{-1}\iota(1 + \bar{\xi}'D^{-1}\xi) - (\iota'D^{-1}\xi)^2} \left[ 1 + \frac{\bar{\xi}_{T+1}(\iota'D^{-1}\xi)^2}{(1 + \bar{\xi}'D^{-1}\xi)^2} \right] \]

\[ - \frac{2\bar{\xi}_{T+1}\iota'D^{-1}\xi}{1 + \bar{\xi}'D^{-1}\xi} + \bar{\xi}_{T+1} = \frac{\bar{\xi}_{T+1}\iota'D^{-1}\xi}{1 + \bar{\xi}'D^{-1}\xi} + \frac{1}{d_{T+1}} \]

\[ = \frac{\sum_s \xi_s^2 - 2\bar{\xi}_{T+1} \sum_s d_s \xi_s + \bar{\xi}_{T+1}^2 \sum_s d_s^2}{\sum_s d_s(1 + \sum_s d_s \xi_s^2) - (\sum_s d_s \xi_s)^2} + \frac{1}{d_{T+1}} \]

\[ = \frac{wt_{T+1}}{d_{T+1}} + \frac{1}{d_{T+1}} \]

\[ = \frac{1}{d_{T+1}(1 + wt_{T+1})} \]

### A.2.4 Derivation of (32)

To save on notation, in the following we use \( p(s_{j,t}|s_{i,t+3}, \Omega_T) \) to write \( p(s_{j,t} = 1|s_{i,t+3} = 1, \Omega_T) \). To derive (32), take for example a three state model and
calculate

\[ p(s_{jt}|s_{i,t+3}, \Omega_T) = \sum_{k=0}^{2} p(s_{jt}|s_{k,t+1}, s_{i,t+3}, \Omega_T) p(s_{k,t+1}|s_{i,t+3}, \Omega_T) \]

\[ = \sum_{k=0}^{2} p(s_{jt}|s_{k,t+1}, \Omega_T) \sum_{l=0}^{2} p(s_{k,t+1}|s_{l,t+2}, \Omega_{t+1}) p(s_{l,t+2}|s_{i,t+3}, \Omega_{t+2}) \]

\[ = \sum_{k=0}^{2} \frac{p(s_{jt}|\Omega_t)}{p(s_{i,t+3}|\Omega_{t+2})} \sum_{l=0}^{2} \frac{p_{jk} p(s_{jt}|\Omega_t)}{p(s_{k,t+1}|\Omega_{t+1})} \frac{p_{kl} p(s_{k,t+1}|\Omega_{t+1})}{p(s_{l,t+2}|\Omega_{t+2})} = \frac{p(s_{jt}|\Omega_t)}{p(s_{i,t+3}|\Omega_{t+2})} \sum_{k=0}^{2} \sum_{l=0}^{2} p_{jk} a_{l+1}^{k} p_{kl} a_{l+2} p_{li} \]

\[ = \frac{p(s_{jt}|\Omega_t)}{p(s_{i,t+3}|\Omega_{t+2})} \left( P' A_{t+1} P' A_{t+2} P' \right)_{jj} \]

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

## B Monte Carlo results for MSI and MSM models

Table 9 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 9 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


Table 9: Monte Carlo results: MSI and MSM models

| λ     |  $\hat{\sigma}_{\xi|T}^2$ | $w_\delta$ | $w_\xi$ | $w_{\xi|\text{asy}}$ | $w_M$ | $w_\delta$ | $w_\xi$ | $w_{\xi|\text{asy}}$ | $w_M$ |
|-------|-----------------|-----------|-------|-------------------|------|-----------|-------|-------------------|------|
| **MSI** |                |           |       |                   |      |            |       |                   |      |
| 1     | 0.0-0.1         | 0.988    | 1.008 | 1.024             | 1.002| 0.994     | 1.006 | 1.015             | 1.006| 0.994     | 1.006 |
|       | 0.1-0.2         | 0.994    | 1.019 | 1.032             | 1.016| 0.997     | 1.016 | 1.022             | 1.020| 0.997     | 1.020 |
|       | 0.2-0.3         | 0.997    | 1.018 | 1.030             | 1.018| 0.999     | 1.017 | 1.023             | 1.026| 0.999     | 1.026 |
| 2     | 0.0-0.1         | 0.997    | 1.005 | 1.009             | 1.006| 0.999     | 1.003 | 1.004             | 1.020| 0.999     | 1.004 |
|       | 0.1-0.2         | 1.000    | 1.005 | 1.008             | 1.017| 1.002     | 0.994 | 0.994             | 1.030| 1.002     | 0.994 |
|       | 0.2-0.3         | 1.003    | 0.993 | 0.994             | 1.007| 1.003     | 0.985 | 0.985             | 1.018| 1.003     | 0.985 |
| 3     | 0.0-0.1         | 1.000    | 0.999 | 1.000             | 1.004| 1.000     | 0.999 | 0.999             | 1.012| 1.000     | 0.999 |
|       | 0.1-0.2         | 1.004    | 0.983 | 0.981             | 1.026| 1.004     | 0.972 | 0.969             | 1.020| 1.004     | 0.972 |
|       | 0.2-0.3         | 1.005    | 0.970 | 0.964             | 0.986| 1.005     | 0.944 | 0.939             | 0.981| 1.005     | 0.944 |
| **MSM** |                |           |       |                   |      |            |       |                   |      |
| 1     | 0.0-0.1         | 0.991    | 1.010 | 1.021             | 1.008| 0.994     | 1.019 | 1.028             | 1.020| 0.994     | 1.019 |
|       | 0.1-0.2         | 0.994    | 1.023 | 1.032             | 1.017| 0.996     | 1.033 | 1.042             | 1.042| 0.996     | 1.033 |
|       | 0.2-0.3         | 0.995    | 1.029 | 1.039             | 1.037| 0.998     | 1.033 | 1.039             | 1.043| 0.998     | 1.033 |
| 2     | 0.0-0.1         | 0.996    | 1.011 | 1.017             | 1.009| 0.999     | 1.012 | 1.015             | 1.028| 0.999     | 1.012 |
|       | 0.1-0.2         | 0.998    | 1.015 | 1.019             | 1.019| 1.000     | 1.010 | 1.012             | 1.034| 1.000     | 1.010 |
|       | 0.2-0.3         | 0.999    | 1.015 | 1.020             | 1.022| 1.001     | 1.007 | 1.008             | 1.024| 1.001     | 1.007 |
| 3     | 0.0-0.1         | 0.999    | 1.004 | 1.007             | 1.004| 1.000     | 1.002 | 1.003             | 1.015| 1.000     | 1.002 |
|       | 0.1-0.2         | 1.000    | 1.002 | 1.003             | 1.013| 1.002     | 0.991 | 0.990             | 1.012| 1.002     | 0.991 |
|       | 0.2-0.3         | 1.000    | 1.006 | 1.008             | 1.007| 1.003     | 0.974 | 0.972             | 0.983| 1.003     | 0.974 |

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


Deschamps, P. J. (2008). Comparing smooth transition and markov switch-


