MARKOV SWITCHING MODELS IN EMPIRICAL FINANCE

Massimo Guidolin

ABSTRACT

I review the burgeoning literature on applications of Markov regime switching models in empirical finance. In particular, distinct attention is devoted to the ability of Markov Switching models to fit the data, filter unknown regimes and states on the basis of the data, to allow a powerful tool to test hypotheses formulated in light of financial theories, and to their forecasting performance with reference to both point and density predictions. The review covers papers concerning a multiplicity of sub-fields in financial economics, ranging from empirical analyses of stock returns, the term structure of default-free interest rates, the dynamics of exchange rates, as well as the joint process of stock and bond returns.

Keywords: Markov switching; regimes; regime shifts; nonlinearities; predictability; autoregressive conditional heteroskedasticity

JEL classifications: G00; C00

Missing Data Methods: Time-series Methods and Applications
Advances in Econometrics, Volume 27B, 1–86

ISSN: 0731-9053/doi:10.1108/S0731-9053(2011)000027B004

INTRODUCTION

Since the seminal contributions by Hamilton, Nelson, Schwert, Startz, and Turner appeared in *Econometrica*, the *Journal of Economic Dynamics and Control*, and the *Journal of Financial Economics* between 1988 and 1989, the literature has witnessed an explosion in the number and the quality of the academic papers that have applied Markov regime switching econometric methods to model and forecast financial data. To uncover (filter) from financial data the underlying but unobservable "general state" of the economy or more specifically of asset markets through regime switching models – in which such latent state would govern how part or all the parameters of a time series framework may change over time – has in fact become one of the leading methods through which applied financial researchers try and recover important and yet unobservable (missing) data that have turned out to be key to fit and forecast financial phenomena. What have we learned from more than 20 years of applied research based on Markov Switching models (henceforth, MSMs)? What are the distinctive contributions of the hundreds of published papers that have featured MSMs with a prominent role, that is, what chances have they offered us to better understand financial markets phenomena when compared to simpler, single-regime methodologies? How can such missing information – the filtered regime – affect our understanding of the dynamics of financial markets? In this chapter I take stock of the literature, classify a few of the papers that seem to have scored the key contributions, and take the opportunity to collect a number of thoughts concerning either the intimate relationships among different papers and strands of the Markov switching (MS) literature or to suggest directions for further extension and improvement of the way MSMs are currently applied in empirical finance, with special emphasis on modeling and forecasting financial time series.

I do not think that it may be sensible to try and preview the "results" of a literature survey. However, it is useful to list here a number of key questions that I have had in mind when I have approached the task to write this survey. Even though I cannot rule out that readers will find different answers to the same questions as well as different questions to ask the same literature reviewed here, the issues that were on my mind may help better understand the reasons of my choices. I had seven questions on my mind. First, I was curious to assess what fraction of the MS papers had picked MS as a modeling tool on the basis of a statistical reason – for instance, a careful specification search open to nonlinear models of the MS class – and what fraction had arrived at MSMs because of economic motivations. Basically, I had in mind a contrast between papers that use MSMs because "the data ask for them," and papers that use them because general knowledge of a phenomenon or a field "makes them plausible." One interesting implication is that usually the first type of papers tends to report rich comparisons of the performance (fit, prediction, relative pricing errors, etc.) of MSMs versus. other models, especially simpler, linear benchmarks; however, the second type of papers is often more intuitive and deeply grounded in economics. I have found papers in both groups and a rough count gives 50-50 proportions, especially because the literature is increasingly moving toward "rooting" MSMs in primitive asset pricing objects (e.g., the stochastic discount factor, SDF) so that the second approach to MS has become common (see Guidolin, 2011).

Second, given that a paper has elected to rely on MS methods, I have paid attention to whether or not the author(s) have entertained the possibility that the number of regimes (say, K) may exceed two. Basically, it is common to come across papers for which MSMs and two-state nonlinear frameworks are almost equivalent choices, while other papers openly entertain the sensible thought that once a researcher opens up to the possibility that $K \geq 2$, this obviously does not restrict $K=2$. Notice that this second question is partially associated with the first question: many papers that motivate MS on logical grounds, often impose $K=2$, while careful model selection procedures more often than not have led to pick $K>2$. To my surprise, also in this case the relative fractions of papers has turned out to be approximately 50-50, that is, there are many more papers based on K-state MSMs with $K>2$ than I was expecting ex-ante; I also have found cases of empirical papers entertaining as many as 6 or 8 regimes.

Questions 3 and 4 are intimately related. I was curious to record the fraction of papers that had (i) assumed flexible, rich marginal distributions for return innovations (e.g., t-Student shocks) thus going beyond the standard of MSMs as persistent Gaussian mixtures, and/or (ii) used MSMs in which the underlying Markov chain is time-heterogeneous, that is, in which the transition probability matrix is time-varying, as a function of either endogenous or exogenous state variables. Here I found two different answers concerning (i) and (ii). As for (ii), I discovered that the fraction of papers that allow the transition probabilities to change over time has been increasing. This is no random outcome, as modern papers that root MS in the SDF have had solid theoretical reasons to assume that (at least under the objective, physical measure) the transition probabilities may be time-varying (see details in Guidolin, 2011). Interestingly enough, there is a less overwhelming evidence that time-varying transition probabilities (TVTPs) may actually improve the forecasting performance of MSMs. As for (i), it seems that most authors are still finding that traditional Gaussian mixture models are generally sufficient to the task assigned to MSMs. Only in very specific cases, especially because the single-state literature was suggesting to nest alternative distributional assumptions within single-state benchmarks, I have encountered cases of non-Gaussian parameterizations. Obviously, questions 2–4 cannot be asked in isolation: although the "intrinsic" flexibility of a MSM comes from $K \geq 2$, there are other margins that can be exploited by a researcher – the one on TVTPs remains intrinsic to MS and the one on marginal densities is shared with all of time series econometrics – but obvious trade-offs exist, also because the perils of "overfitting" data are always lurking in richly parameterized models.

The fifth concern with which I have approached the task revolved around whether or not papers have relied on the idea that the key challenge of a MSM lies in correctly inferring and predicting regime shifts (sometimes called turning points). The alternative consists of thinking of MSMs not only (or mostly) as devices to anticipate regime switches but as a flexible class of nonlinear tools the usefulness of which cannot be completely pinned to the fact that the Markov chain characterizing them can be easily predicted. Of course, this is a very subtle distinction, because it remains odd to entertain the possibility that a regime switching model may be useful in certain applications (e.g., portfolio choice, risk management) even though it does not really help that much with forecasting the switching dynamics – and yet this is the case.[1] Although most papers I have surveyed remain silent on this aspect, the few that discuss the point seem to have adopted a pragmatic approach. The papers that are more concerned with producing *point* forecasts tend to implicitly discuss the accuracy of regime prediction (see, e.g., Guidolin, Hyde, McMillan, & Ono, 2009). However, a number of papers that instead are more biased toward producing good *density* forecasts usually tend to downplay the regime prediction aspects (see, e.g., Van Dijk & Franses, 2003). A recent paper by Lettau & Van Nieuwerburgh (2007) has surprisingly concluded that the real issue with MSMs is not really locating the switches, but instead precisely estimating the shifts in the conditional mean function, at least from the perspective of the literature on return predictability.

The sixth question concerned the type of markets/assets with respect to which MS tools have been more useful or frequently employed. My prior – that MSMs have been popular with equity market students – has found confirmation, but there have also been a few surprises. In particular, the more interesting, recent literature on MS in asset pricing has mostly originated within the subfield of fixed income pricing. It was also interesting to notice the existence of a cluster of papers devoted to modeling the joint distribution of equity and Treasury bond returns, with special emphasis on measures of their association (e.g., Baele, Bekaert, & Inghelbrecht, 2010; Guidolin & Timmermann, 2005, 2006a, 2006b). However, MSMs have also traditionally found application to currencies, while there are recent papers on MS dynamics in corporate bond, derivatives, and real estate returns and real interest rates. At this point, no asset has been left out.

My final concern was with data frequency: I repeatedly asked whether there was a frequency below/above which it was obvious that MSMs would stop being useful, to be replaced by other dynamic econometric frameworks (such as ARCH or copulas, to name two). In light of the fact that routinely great efforts are spent to interpret regime shifts in the light of or to predict recession/expansion dates – and of the fact that MSMs originated in the empirical macroeconomics literature – the outcome of my survey has been surprising: there is very little literature using MSMs at annual or quarterly frequencies (Engel & Hamilton, 1990, or Liu, 2011, are exceptions). The dominant frequency at which MSMs seem to be most effective is the monthly one, which is adopted by approximately 60% of the papers. However, there is also a substantial body of research that has applied MSMs to daily data. Although it is difficult to net these percentages from the baseline percentages of papers using monthly vs. daily frequencies in all of empirical finance, it seems fair to say that MS tools seem in the end to work well at all possible frequencies, including quarterly ones – even though these are less popular among financial economists.

Let me add the usual caveats and disclaimers. This is just one survey of a (small, alas) portion of existing papers that have used MSMs in applied financial economics. There is no pretense or suggestion that my review may be complete in terms of discussing or citing all the relevant papers. Although I have struggled to avoid that, there is probably a visible bias toward topics or issues of personal interest, which is probably visible in the space I have devoted to density forecasting and risk management. However, this is also where I have found the most obvious examples of new, exciting advances in this subfield of applied time series econometrics. It is also important to openly state that my paper does not aim at becoming a statistical reference on MSMs. My goal is to provide a primer to what MSMs are and to quickly move to review papers that can be taken as examples of what MS tools may yield when applied to finance research questions. Any reader interested in acquiring the basic tools to specify and estimate MSMs is invited to consult the excellent textbooks by – among many others – Franses and van Dijk (2000), Frühwirth-Schnatter (2006), Kim and Nelson (1999), Krolzig (1997), and McLachlan and Peel (2000).

In this survey, I shall keep separate the task of building dynamic asset pricing models in which beliefs are informed by some MSM framework from the (simpler) goal of fitting MSMs to the data and "loosely" using their implications to test one or more hypotheses of interest. Technically, the difference may be seen as stemming from the fact that while in simple fitting exercises, a researcher only cares about modeling (often, predicting) the physical (conditional) density of the data, $\mathbb{P}$, when MSMs are used to build dynamic asset pricing models, she usually cares for the role of regime shifts both under $\mathbb{P}$ as well as under $\mathbb{Q}$, the risk neutral measure. Guidolin (2011) is a companion survey paper devoted to the role of MSMs in asset pricing and portfolio choice research that focuses more on the relationship between MSMs in $\mathbb{P}$ vs. $\mathbb{Q}$ and their implications (see e.g., Bertholon, Monfort, & Pegoraro, 2008; Dai Singleton, & Yang, 2007).

The chapter is structured as follows. The second section provides a short primer to MSMs in their alternative functional forms. I briefly deal with estimation, forecasting, model selection, and diagnostic checks. The goal of this section is not to cover details of the econometrics of MSMs (Hamilton, 1990, 1993, 1994, remain references against which my review cannot compete), but simply to provide a set of definitions and concepts to be referenced later on. Third section starts my survey by devoting attention to a heterogeneous set of papers that have simply modeled financial data using MSMs. This section concerns the most traditional literature that has used MSMs with the objective to show that financial returns contain evidence of multiple, *recurring* regimes or even – which is somewhat questionable given the structure of a MSM – generic breaks in empirical relationships. Thrid section also mostly focuses on univariate applications. Fourth section describes a few genuinely multivariate applications and introduces a powerful set of tools that in my view is destined to play a key role in future developments, MS dynamic factor models (DFMs). Fifth section relates a number of ideas and problems that emerge in third and fourth sections to the debate on the predictability of financial returns. Sixth section examines whether and how we may need to supplement ARCH-type models for variances and covariances with MS components (or vice versa, whether there is a need to augment MSMs with ARCH effects). Seventh section tackles the choice between homogeneous and time-heterogeneous Markov chains in the practice of MSMs. Eighth section asks whether MSMs can forecast financial data. Besides point and density forecasts, financial economists will find in this section elements of risk management addressed in the perspective of MSMs. The ninth section concludes.

A PRIMER ON MARKOV SWITCHING MODELS

Suppose that the $N \times 1$ random vector $\mathbf{y}_t$ follows a K-regime MSVAR(p) (MSVAR) heteroskedastic process, compactly MSIAH(K, p):

$$\mathbf{y}_t = \mu_{S_t} + \sum_{j=1}^{p} \mathbf{A}_{j,S_t} \boldsymbol{y}_{t-j} + \Sigma_{S_t} \varepsilon_t \tag{1}$$

with $\varepsilon_t \sim \text{NID}(\mathbf{0}, \mathbf{I}_N)$.[2] $S_t = 1, 2, \ldots, K$ is a latent state variable driving all the matrices of parameters in Eq. (1): μ_{S_t} is an $N \times 1$ vector that collects N regime-dependent intercepts; the $N \times N$ matrix Σ_{S_t} represents the state S_t factor in a regime-dependent Choleski factorization of the covariance matrix, $\Omega_{S_t} = \Sigma_{S_t} \Sigma'_{S_t}$.[3] A non-diagonal Σ_{S_t} captures *simultaneous* co-movements between asset returns and macro factors, while *dynamic* (lagged) linkages across different variables may be captured by the VAR(p). For instance, in Guidolin and Ono (2006), N is broken down in N_1 asset returns and N_2 macroeconomic predictors, with $N_1 + N_2 = N$. It is typical to assume the absence of roots outside the unit circle, thus making the process stationary.[4] In fact, conditionally on the unobservable state S_t, Eq. (1) defines a standard Gaussian reduced form VAR(p) model. On the contrary, when $K > 1$, alternative hidden states are possible and they will influence both the conditional mean and the volatility/correlation structure characterizing Eq. (1). These unobservable states are generated by a discrete, irreducible, and ergodic first-order Markov chain:[5]

$$\Pr\left(S_t = j | \{S_j\}_{j=1}^{t-1}, \{\mathbf{y}_j\}_{j=1}^{t-1}\right) = \Pr(S_t = j | S_{t-1} = i, \mathscr{F}_t) = p_{ij,t}$$

where $p_{ij,\mathrm{t}}$ is the generic $[i, j]$ element of the $k \times k$ transition matrix $\mathbf{P}_t$. When $\mathbf{P}_t$ is constant over time, I speak of a homogeneous Markov chain. For simplicity, in the rest of this section, I focus on the case of constant transition probabilities, although the seventh section presents frequent examples of economically motivated time-heterogeneous Markov chain models. Ergodicity implies the existence of a stationary vector of probabilities $\bar{\boldsymbol{\xi}}$ satisfying $\bar{\boldsymbol{\xi}} = \mathbf{P}'\bar{\boldsymbol{\xi}}$. Irreducibility implies that $\bar{\boldsymbol{\xi}} > \mathbf{0}$, meaning that all unobservable states are possible. In practice, $\mathbf{P}$ is unknown and hence $\bar{\xi}$ can be at most estimated given knowledge on $\mathbf{P}$ extracted from the information set $\mathcal{F}_t = \{\mathbf{y}_j\}_{j=1}^{t}$.[6]

When N is large, Eq. (1) implies the estimation of a large number of parameters, $K[N+pN^2+N(N+1)/2+(K–1)]$. For instance, for $K=2$, $N=8$, and $p=1$ (the parameters characterizing the application in Guidolin & Ono, 2006), this implies the estimation of 218 parameters! MSMs are known (see Timmermann, 2000) to capture central statistical features of asset returns. For instance, differences in conditional means across regimes enter the higher moments such as variance, skewness, and kurtosis. In particular, the variance is not simply the average of the variances across the two regimes: the difference in means also imparts an effect because the switch to a new regime contributes to volatility; this difference in regime means also generates non-zero conditional skwewness. Finally, differences in means in addition to differences in variances can generate persistence in levels as well as squared values akin to volatility persistence observed in many return series. Again differences in means play an important role in generating autocorrelation in first moments: without such differences, the autocorrelation will be zero. In contrast, volatility persistence can be induced either by differences in means or by differences in variances across regimes. In both cases, the persistence tends to be greater, the stronger the combined persistence, as measured by the diagonal transition probabilities collected in $\mathbf{P}$.[7]

Eq. (1) nests a number of simpler models in which either some of the parameter matrices are not needed or some of these matrices are independent of the regime. These simpler models may greatly reduce the number of parameters to be estimated. Among them, the financial econometrics literature (see, e.g., Ang & Bekaert, 2002a) has devoted special attention to MSIH(K) and MSI(K) models,

$$\mathbf{y}_t = \mu_{S_t} + \Sigma_{S_t}\varepsilon_t$$

in which $p=0$, to MSIA(K, p) homoskedastic models,

$$\mathbf{y}_t = \mu_{S_t} + \sum_{j=1}^{p} \mathbf{A}_{j,S_t}\mathbf{y}_{t-j} + \Sigma\varepsilon_t$$

in which the covariance matrix is constant, and to MSIH(k)-VAR(p) models (see Guidolin & Ono, 2006),

$$\mathbf{y}_t = \mu_{S_t} + \sum_{j=1}^{p} \mathbf{A}_j y_{t-j} + \Sigma_{S_t}\varepsilon_t \tag{2}$$

which are a special case of Eq. (1) in which intercepts and covariance matrices are regime-dependent, and the VAR(p) coefficients are not.[8] For instance, model (2) implies "only" $K[N+N(N+1)/2+(K-1)]+pN^2$ parameters. For the same configuration mentioned above, this means 154 parameters. Of course, a limit case of Eq. (1) is obtained when $K=1$, which yields a multivariate Gaussian VAR(p), a benchmark of a large portion of the empirical finance literature. Certain applications (e.g., the seminal paper by Hamilton, 1989) have also entertained the following variation on Eq. (1), a MSMAH(p, K):

$$(\mathbf{y}_t - \boldsymbol{v}_{S_t}) = \sum_{j=1}^{p} \mathrm{A}_{j,S_t}(\mathbf{y}_{t-j} - \boldsymbol{v}_{S_{t-j}}) + \Sigma_{S_t}\varepsilon_t \tag{3}$$

Krolzig (1997) shows that the dynamic implications of Eqs. (1) and (3) are markedly different. For instance, the definition of Eq. (3) implies that the conditional mean function is governed by a $(p+1)$-th order Markov chain as the terms $(\mathbf{y}_{t-j} - \boldsymbol{v}_{S_{t-j}})$ render the entire sequence $\{S_{t-p}, S_{t-p+1}, \ldots, S_{t-1}, S_t\}$ relevant. With very few exceptions, in the finance literature the more straightforward specification in Eq. (1) dominates.

Estimation and Inference

The first step toward estimation and prediction of a MSIAH model is to put the model in state-space form. Collect the information on the time t realization of the Markov chain in a random vector $x_t=[I(S_t=1)I(S_t=2)\ \ldots\ I(S_t=k)]'$, where $I(S_t=k)$ is a standard indicator variable. In practice the sample realizations of ξ_t will always consist of unit "versors" $\mathbf{e}_k$ characterized by a 1 in the kth position and by zeros everywhere else, assuming $I(S_t=k)=1$. An important property is then $\mathrm{E}[x_t|x_{t-1}]=\mathbf{P}'x_{\mathrm{t}-1}$ (see Hamilton, 1994). Exploiting this property, the state-space form is composed of two equations:

$$\mathbf{y}_t = \mathbf{Y}_t\Psi(\xi_t \otimes \iota_N) + \Sigma^*(\xi_t \otimes \mathrm{I}_N)\varepsilon_t \qquad \text{(measurement equation)}$$

$$\xi_{t+1} = \mathbf{P}'\xi_t + \mathbf{u}_{t+1} \qquad \text{(transition equation)} \tag{4}$$

where $\mathbf{Y}_t$ is a $N\times(Np+1)$ vector of predetermined variables with structure $[1\ \mathbf{y}'_{t-1}\ldots\mathbf{y}'_{t-p}]\otimes\iota_N$, Ψ is a $(Np+1)\times NK$ matrix collecting VAR parameters, both matrices of means and autoregressive coefficients,

$$\Psi = \begin{bmatrix} \mu_1' & \cdots & \mu_k' \\ \mathbf{A}_{11} & \cdots & \mathbf{A}_{1k} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{p1} & \cdots & \mathbf{A}_{pk} \end{bmatrix}$$

Σ^* is a $N \times NK$ matrix collecting all the possible K Choleski factors $[\Sigma_1 \Sigma_2 \ldots \Sigma_k]$ such that $\forall t$, $\Sigma^*(\xi_t \otimes \mathbf{I}_N)(\xi_t \otimes \mathbf{I}_N)'(\Sigma^*)' = \Omega_{S_t}$, the S_t-regime covariance matrix of the innovations ε_t. Moreover, $\varepsilon_{t_{\mathrm{NID}}}(\mathbf{0}, \mathbf{I}_N)$, and in the transition equation $\mathbf{u}_{t+1}$ is a zero-mean discrete martingale difference sequence vector. Also, the elements of $\mathbf{u}_{t+1}$ are uncorrelated with ε_{t+1} as well as x_{t-j}, ε_{t-j}, $\boldsymbol{y}_{t-j}$, and $\mathbf{Y}_{t-j}$ $\forall j \geq 0$. To make the system in Eq. (4) operational, assume that the process in Eq. (1) started with a random draw from the unconditional probability distribution defined by the vector of state probabilities $\overline{\boldsymbol{\xi}}$. The dynamic state-space model in Eq. (4) is neither linear (as the state vector x_t also influences the covariance matrix of the process) nor Gaussian, as the innovations driving the transition equation are Gaussian mixtures.

The state-space representation of Eq. (3) is quite different. As already observed, the conditional mean is now governed by a $(p+1)$-th order Markov chain, so that it is useful to collect the information on the realization of the Markov chain in a $K^{p+1} \times 1$ random vector

$$\xi_t^{(p+1)} = \xi_t \otimes \xi_{t-1} \otimes \ldots \otimes \xi_{t-p} = \begin{bmatrix} I(S_t = 1, S_{t-1} = 1, \ldots, S_{t-p} = 1) \\ I(S_t = 1, S_{t-1} = 1, \ldots, S_{t-p} = 2) \\ \vdots \\ I(S_t = 1, S_{t-1} = k, \ldots, S_{t-p} = k) \\ I(S_t = k, S_{t-1} = k, \ldots, S_{t-p} = k) \end{bmatrix}$$

so that $E[\xi_t^{(p+1)} | \xi_{t-1}^{(p+1)}] = \mathrm{P}'_{(p+1)} \xi_{t-1}^{(p+1)}$ where $\mathbf{P}_{(p+1)} = P \otimes P \otimes \ldots \otimes P$ is the $K^{p+1} \times K^{p+1}$ transition matrix for the transformed set of regimes.[9] Therefore the transition equation will be characterized by a matrix that corresponds to $\mathrm{P}'_{(p+1)}$

$$\xi_{t+1}^{(p+1)} = \mathrm{P}'_{(p+1)} \xi_t^{(p+1)} + \eta_{t+1} \quad \text{or} \quad \xi_{t+1}^{(p+1)} - \overline{\boldsymbol{\xi}} = \mathrm{P}'_{(p+1)}(\xi_t^{(p+1)} - \overline{\boldsymbol{\xi}}) + \eta_{t+1}$$

from the ergodic property that $\mathrm{P}'_{(p+1)}\bar{\boldsymbol{\xi}} = \bar{\boldsymbol{\xi}}$, while the measurement equation becomes:

$$\mathbf{y}_t = \mathbf{Y}_t \mathrm{B}\xi_t^{(p+1)} + \Sigma^*(\xi_t^{(1)} \otimes \mathrm{I}_N)\,\mathbf{u}_t = \mathbf{Y}_t \mathrm{B}\xi_t^{(p+1)} + \Sigma^*(((\mathrm{I}_K \otimes \iota'_{K^p})\xi_t^{(p+1)}) \otimes \mathrm{I}_N)\,\mathbf{u}_t$$

where $\xi_t^{(1)}$ is the standard $K \times 1$ vector collecting state information for period t such that $\xi_t^{(1)} = (\mathbf{I}_K \otimes \iota'_{K^p})\xi_t^{(p+1)}$, and the $(Np+1) \times NK^{p+1}$ coefficient matrix $\mathbf{B}$ has structure:

$$\mathbf{B} = \Psi(\mathbf{I}_K \otimes \iota'_{K^p} \otimes \mathrm{i}_N) + \\ - \begin{bmatrix} \Sigma_{j=1}^{p} \mathbf{A}_{j,S_t^{(1)}} v'_{S_{t-j}^{(1)}} & \cdots & \Sigma_{j=1}^{p} \mathbf{A}'_{j,S_t^{(K^{p+1}-1)}} v'_{S_{t-j}^{(K^{p+1}-1)}} & \Sigma_{j=1}^{p} \mathbf{A}'_{j,S_t^{(K^{p+1})}} v'_{S_{t-j}^{(K^{p+1})}} \\ & & \mathbf{O}_{(N(p-1)+1)\times NK^{p+1}} & \end{bmatrix}$$

Multivariate MSMs are estimated by maximum likelihood, although (E)GMM estimation strategies have recently been proposed (see, e.g., Bansal & Zhou, 2002). Estimation and inference are usually based on the expectation-maximization (EM) algorithm proposed by Dempster, Laird, & Rubin (1977) and Hamilton (1989), a filter that allows the iterative calculation of the one-step ahead forecast of $\xi_{t+1|t}$ given the information set $\mathscr{F}_t$ and the consequent construction of the log-likelihood function.[10] Additional details on the EM algorithm can be found in Hamilton (1993, 1994) or Krolzig (1997). Maximization of the log-likelihood function within the M-step is made faster by the fact that the first-order conditions defining the MLE may often be written down in closed form (see, e.g., Hamilton, 1990). In particular, such first-order conditions can be shown to depend on the smoothed probabilities $\hat{\xi}_{t|T} \equiv \Pr(\xi_t|\mathscr{F}_T;\theta)$ (i.e., the state probabilities estimated on the basis of the full sample of data), and therefore, they all present a high degree of non-linearity in the parameters, collected in the vector θ (that includes the estimable elements of the transition matrix $\mathbf{P}$). As a result, these first-order conditions have to be solved numerically, although convenient iterative methods exist. In fact, the expectation and maximization steps can be used in iterative fashion. Starting with arbitrary initial values $\tilde{\theta}^0$, the expectation step is applied first, thus obtaining a sequence of smoothed probability distributions $\{\hat{\xi}^1_{t|T}\}_{t=1}^{T}$. Given these smoothed probabilities, appropriate first-order conditions are used to derive $\tilde{\theta}^1$. Based on $\tilde{\theta}^1$, the expectation step can be applied again to find a new sequence of smoothed probability distributions $\{\hat{\xi}^2_{t|T}\}_{t=1}^T$. This starts the second iteration of the algorithm. The algorithm keeps being iterated until convergence, that is, until $\tilde{\theta}^l \simeq \tilde{\theta}^{l-1}$. The likelihood function increases at each step and reaches an approximate maximum in correspondence to convergence (see Baum, Petrie, Soules, & Weiss, 1970).

As for the properties of the resulting ML estimators, under standard regularity conditions (such as identifiability, stability and the fact that the true parameter vector does not fall on the boundaries) Hamilton (1989, 1993) and Leroux (1992) have proven consistency and asymptotic normality of the ML estimator $\tilde{\theta}$:

$$\sqrt{T}(\tilde{\theta} - \theta) \xrightarrow{D} N(\mathbf{0}, \mathscr{I}_a(\theta)^{-1})$$

where $\mathscr{I}_a(\theta)$ is the asymptotic information matrix,

$$\mathscr{I}_a(\theta) \equiv \lim_{T\to\infty} -T^{-1}E\left[\frac{\partial^2 \ln \prod_{t=1}^{T} p(\mathbf{y}_t|\theta)}{\partial\theta\partial\theta'}\right]$$

Although other choices exist – that is, the conditional scores or a numerical evaluation of the second partial derivative of the log-likelihood function with respect to $\tilde{\theta}$ – in applied work it is typical to employ a White-style "sandwich" estimator of $\mathscr{I}_a(\theta)$ that yields the estimate

$$\widetilde{\text{Var}}(\tilde{\theta}) = T^{-1}[\mathscr{I}_2(\tilde{\theta})(\mathscr{I}_1(\tilde{\theta}))^{-1}\mathscr{I}_2(\tilde{\theta})]$$

where

$$\mathscr{I}_1(\tilde{\theta}) = T^{-1}\sum_{t=1}^{T}[\boldsymbol{h}_t(\tilde{\theta})][\boldsymbol{h}_t(\tilde{\theta})]' \qquad \boldsymbol{h}_t(\tilde{\theta}) = \frac{\partial \ln p(\mathbf{y}_t|\mathscr{F}_{t-1};\tilde{\theta})}{\partial\theta}$$

$$\mathscr{I}_2(\tilde{\theta}) = -T^{-1}\sum_{t=1}^{T}\left[\frac{\partial^2 \ln p(\mathbf{y}_t|\mathscr{F}_{t-1};\tilde{\theta})}{\partial\theta\partial\theta'}\right]$$

As a consequence, and with one important exception, standard inferential procedures are available to test statistical hypotheses. In particular, call $\phi : \mathbb{R}^q \to \mathbb{R}^r$ a (smooth) function that imposes $q-r$ restrictions on the q-dimensional parameter vector θ. Suppose we want to test $H_0 : \phi(\theta) = \mathbf{0}$ versus. $H_1 : \phi(\theta) \neq \mathbf{0}$ *under the assumption that under both hypotheses the number of regimes K is identical.*[11] Define $\tilde{\theta}_r$ as the restricted estimator, obtained under the null hypothesis. Lagrange multiplier (LM) tests tend to be the preferred tests as they only require the estimation of the restricted model. While the scores of an unrestricted model,

$$\mathrm{s}_t(\tilde{\theta}) \equiv \sum_{\tau=1}^{t} \mathrm{h}_\tau(\tilde{\theta}) = \sum_{\tau=1}^{t} \left\{ \frac{\partial \mathrm{diag}[\eta_\tau(\theta)]\mathrm{P}_\tau(\theta)}{\partial \theta'}\bigg|_{\theta=\tilde{\theta}} \right\}' \hat{\xi}_{\tau|t}$$

have zero mean vector by construction,[12] the scores of the restricted model obtained by MLE, imposing $\phi(\theta) = 0$, can be used to obtain the standard test statistic:

$$LM_T \equiv s_T(\tilde{\theta}_r)' \left[\widetilde{\mathrm{Var}}(\tilde{\theta}_r)\right]^{-1} s_T(\tilde{\theta}_r) \xrightarrow{D} \chi_r^2$$

where $r = \mathrm{rank}[\partial\phi(\theta)/\partial\theta']$ and $\tilde{\theta}_r$ denotes the restricted estimator. For instance, a test of the hypothesis of homoskedasticity $[H_0 : \mathrm{vech}(\Sigma_i) = \mathrm{vech}(\Sigma_K)\ i = 1, 2, \ldots, K]$ implies $r = (K-1)(N(N+1)/2)$ restrictions and can be formulated as a linear restriction on the matrix Σ^*. As an alternative, the likelihood ratio test (LRT) may be employed:

$$LR \equiv 2[\ln L(\tilde{\theta}) - \ln L(\tilde{\theta}_r)] \xrightarrow{d} \chi_r^2$$

Although very simple, this test requires the estimation of both the restricted and the unrestricted models, which for N high enough may be quite cumbersome in the unrestricted case and require a host of diagnostic checks on the performance of the EM algorithm in locating a truly global maximum for the likelihood function. However, my review will show that in practice, the LRT is employed more and more often in applied work, which may be a sign of either cheap computation power being available, or of improving confidence of researchers in their control over the performance of the EM algorithm, or both.

Finally, standard t and F statistics can be calculated in the form of a Wald test. Under asymptotic normality of the unrestricted ML estimator $\tilde{\theta}$, it follows that

$$\sqrt{T}\left[\phi(\tilde{\theta}) - \phi(\theta)\right] \xrightarrow{d} N\left(\mathbf{0}, \frac{\partial\phi(\theta)}{\partial\theta'}\bigg|_{\theta=\tilde{\theta}} \widetilde{\mathrm{Var}}(\tilde{\theta}) \frac{\partial\phi'(\theta)}{\partial\theta'}\bigg|_{\theta=\tilde{\theta}}\right) \text{ and}$$

$$\mathrm{Wald} \equiv T\phi'(\tilde{\theta})\left[\frac{\partial\phi(\theta)}{\partial\theta'}\bigg|_{\theta=\tilde{\theta}} \widetilde{\mathrm{Var}}(\tilde{\theta}) \frac{\partial\phi'(\theta)}{\partial\theta'}\bigg|_{\theta=\tilde{\theta}}\right]^{-1} \phi(\tilde{\theta}) \xrightarrow{D} \chi_r^2$$

The exception to standard inferential procedures mentioned above concerns the *number of non-zero rows* of the transition matrix $\mathbf{P}$, that is, the number of regimes, K. In this case, even under the assumption of asymptotic normality of the estimator $\tilde{\theta}$, standard testing procedures suffer from non-standard asymptotic distributions of the LRT statistic due to the existence of *nuisance* parameters under the null hypothesis. I specifically discuss this problem in the section "Model Selection and Diagnostic Checks."

Although in most applications data are observed only in discrete time, in various problems, it is convenient to use continuous-time models because they allow the derivation of closed-form solutions (see Guidolin, 2011, for a number of such examples). Hahn, Frühwirth-Schnatter, and Sass (2010, HFSS) have recently re-examined a few issues related to the estimation of continuous time MSMs and emphasized the considerable advantages of a Bayesian approach based on Monte Carlo Markov chain (MCMC) methods. Assume, for instance, that the dynamics of a price process $\mathbf{P} = \{\mathbf{P}_t\}_{t\in[0,T]}$ of N stocks are described as

$$d\mathbf{P}_t = \mathrm{diag}(\mathbf{P}_t)(\mu_t dt + \Sigma_t d\mathbf{Z}t)$$

where $\mathbf{Z} = \{\mathbf{z}_t\}_{t\in[0,T]}$ is an N-dimensional Brownian motion, $\mu = \{\mu_t\}_{t\in[0,T]}$ is the drift process, $\Sigma = \{\Sigma_t\}_{t\in[0,T]}$ is the volatility process, and $T>0$ is the time horizon. Suppose that μ and Σ can take K possible values and that switching between these values is governed by a state process S which is a continuous-time Markov chain with state space $\{1,\ldots, K\}$ and rate matrix $\mathbf{\Lambda}$.[13] The Markov chain is again assumed to be time homogeneous, irreducible, and independent of $\mathbf{Z}$. Then the corresponding return process $r = \{r_t\}_{t\in[0,T]}$, defined by $d\mathbf{r}_t = [\mathrm{diag}(\mathbf{P}_t)]^{-1} d\mathbf{P}_t$ satisfies

$$d\mathbf{r}_t = \mu_t dt + \Sigma_t d\mathbf{Z}t$$

and represents a multivariate continuous time MSM. When $\mu_t = \mu$ and $\Sigma_t = \Sigma$, the model simplifies to a standard N-dimensional geometric Brownian motion. Although since Hamilton (1993) the standard of ML estimation of MSMs is represented by the EM approach, HFSS stress that even though EM algorithms for continuous time Markov chain models have been described by, for example, James, Krishnamurthy, and Le Gland (1996) and Jacquier, Johannes, and Polson (2007), they require constant and known volatility Σ: even for constant but unknown Σ, it is impossible to employ the EM algorithm to estimate the volatility jointly with the other parameters, since the change of measure involved in deriving the filters used in the EM algorithm requires known Σ. Furthermore, for a general continuous-time MS model given discrete observations, no finite-dimensional filters are known and, hence, the conditional expectations used in the EM algorithm cannot be computed. HFSS discuss instead the advantages of MCMC methods. In particular, they construct a sampler tailored to a multivariate continuous time MSMs. Furthermore, they adapt a discrete-time sampler to serve as an approximation for the continuous-time model. They compare the proposed discrete and continuous-time methods with simulated data and find that MCMC outperforms ML estimation for difficult cases like high rates of regime switching and considerable noise, the typical situation one faces in financial applications.

Forecasting

Under a mean squared forecast error (MSFE) criterion, the algorithms required to implement standard forecasting are relatively simple despite the nonlinearity of the MSIAH class and naturally derive from (Eq. (4)). Ignoring the issue of parameter uncertainty, that is, the fact that the parameters of a multivariate MSM must be estimated, the function minimizing the MSFE is the standard conditional expectation function. For instance, for a one-step ahead forecast we have:

$$E[\mathbf{y}_{t+1}|\mathscr{F}_t] = \mathbf{Y}_{t+1}\Psi(\hat{\xi}_{t+1|t} \otimes \imath_N)$$

where $\mathbf{Y}_{t+1} = [1\ \mathbf{y}'_t \ldots \mathbf{y}'_{t-p+1}] \otimes \imath_N$, Ψ collects the estimated conditional mean parameters, and $\hat{\xi}_{t+1|t}$ is the one-step ahead, predicted latent state vector filtered out of the information set $\mathscr{F}_t$ according to transition equation $\hat{\xi}_{t+1|t} = \mathrm{P}'\hat{\xi}_{t|t}$. It follows that

$$E[\mathbf{y}_{t+1}|\mathscr{F}_t] = \mathbf{Y}_{t+1}\Psi(\mathrm{P}'\hat{\xi}_{t|t} \otimes \imath_N) \tag{5}$$

For $h>1$ step ahead forecasts the task is much more challenging as: (i) $\mathbf{Y}_{t+h}$ is unknown and must be predicted; (ii) $E[\mathbf{Y}_{t+h}|\mathscr{F}_t]$ involves sequences of predictions $\{E[\mathbf{y}_{t+1}|\mathscr{F}_t], \ldots, E[\mathbf{y}_{t+h-1}|\mathscr{F}_{t+h-2}]\}$ and as such $\{\hat{\xi}_{t+1|t}, \ldots, \hat{\xi}_{t+h-1|t}\}$, which is likely to impress cross-correlation patterns to the predictions used, because of the presence of MS. For instance, for $h=2$, $p=1$, and ignoring the presence of an intercept term, we have

$$\begin{aligned} E[\mathbf{y}_{t+2}|\Im_t] &= E[(\mathbf{y}'_{t+1} \otimes \imath_N)\Psi(\xi_{t+2} \otimes \imath_N)|\mathscr{F}_t] \\ &= E\{[(\mathbf{y}'_t \otimes \imath_N)\Psi(\xi_{t+1} \otimes \imath_N) \otimes \imath'_N]\Psi(\xi_{t+2} \otimes \imath_N)|\mathscr{F}_t\} \end{aligned}$$

which is not the product of the conditional expectations of $[(\mathbf{y}'_t \otimes \imath_N)\Psi(\xi_{t+1} \otimes \imath_N) \otimes i'_N$ and $\Psi(\xi_{t+2} \otimes \imath_N)$ as the future state vectors ξ_{t+1} and ξ_{t+2} are correlated, from $\xi_{t+2} = p'\xi_{t+1} + \eta_{t+2}$. However, in applied work it is customary to follow the suggestion of Doan, Littermann, and Sims (1984) and substitute the sequence of predicted values of $\{\mathbf{y}_{t+1}, \ \mathbf{y}_{t+2}, \ \ldots, \mathbf{y}_{t+h-1}\}$ (as of time t), that is, $\{\hat{E}[\mathbf{y}_{t+1}|\mathscr{F}_t], \ldots, \ \hat{E}[\mathbf{y}_{t+h-1}|\mathscr{F}_t]\}$ for $\{E[\mathbf{y}_{t+1}|\mathscr{F}_t], \ldots, E[\mathbf{y}_{t+h-1}|\mathscr{F}_{t+h-2}]\}$. In this case (Eq. (5)) generalizes to generic $h>2$-step ahead predictions:

$$E[\mathbf{y}_{t+h}|\mathscr{F}_t] = E[\mathbf{Y}_{t+h}|\mathscr{F}_t]\Psi[(\mathbf{P}')^h \hat{\xi}_{t|t} \otimes \iota_N]$$

which gives a recursive formula since $E[\mathbf{Y}_{t+h}|\mathscr{F}_t]$ forces one to forecast a sequence of future $\mathbf{y}_{t+i}$ values, $i = 1, 2, \ldots, \ h-1$. Similar problems apply to multi-step forecasts from the MSMVARH model (3).

Model Selection and Diagnostic Checks

In the absence of MS dynamics in the matrices of autoregressive coefficients and in the covariance matrix of a vector process – that is, for MSI(K,0) and MSI(K)-VAR(p) – it is possible to show that general multivariate MSMs possess a standard VARMA representation that helps define a somewhat precise mapping between nonlinear MS processes and their linear counterparts. In particular, under a few regularity conditions, Eq. (1) possesses a VARMA($K+p-1, \ K-1$) representation, where $K+p-1$ is the autoregressive order and $K-1$ is the moving average order. On the contrary, the MSMVAR(p) process in Eq. (3) has a VARMA($K+Np-1, \ K+Np-2$) representation. In both cases, notice that their VARMA(a,b) representation implies $a \geq b$. These results give a useful starting point in a simple-to-general specification approach:

1. A researcher may start out by conducting a standard Box-Jenkins' style selection applied to the class of VARMA models. The reason is that given the existence of VARMA(a,b) representations for MSMs, it is possible to solve a simple bivariate system of linear equations to recover K and p from the values for a and b. Because in multivariate contexts, VARMA-style model selection remains quite difficult, noting that $K+p-1 \geq p$ and $K+Np-1 \geq p$, this suggests that the autoregressive order in the VARMA is never lower than the autoregressive order in the MSM. Thus a standard VAR lag-selection procedure provides an upper bound to the correct value of p^* in the MSM.
2. Given such a p^*, the focus shifts on the number of regimes K. Krolzig (1997) has suggested the analysis of each component of the vector $\mathbf{y}_t$ in isolation to detect the appropriate number of regimes, say K_i for y_{it}, $i = 1, 2, \ldots, N$. In this case the (V)ARMA equivalence results can be fully

exploited. For each time series, the best fitting ARMA model could be selected using Box-Jenkins or any other ARMA specification criteria. Taking into account that the AR order p^* has been preselected, the optimal number of regimes K_i^* will correspond to the MA order plus one [plus two minus p^* in Eq. (3)]. Call $\{K_i^*\}_{i=1}^N$ the sequence of number of states for each univariate variable under study.

3. Given $\{K_i^*\}_{i=1}^N$, the total number of regimes characterizing the multivariate process might be in principle as high as $\prod_{i=1}^N K_i^*$ if the regimes are not simultaneously perfectly correlated with each other, that is, if it does not occur that at least a subset of variables are governed by the same hidden Markov chain (see the third section for additional comments). This latter hypothesis is usually testable using standard inferential procedures.
4. Once the number of MSIAH (MSMAH) regimes K^* has been selected, it is useful to test for the presence of regime-dependent heteroskedasticity and/or for the presence of regimes in the autoregressive component of the MSM. For instance, an LM test might be employed. Or the MSM might be estimated with and without heteroskedastic component and an LRT used.

As illustrated in a number of papers to be reviewed in the third section, an alternative set of methods to perform data-driven model selection relies on information criteria, such the Schwartz, Hannan-Quinn, and Akaike criteria (see, e.g., Sin & White, 1996, for evidence on information criteria performance in non-linear models). Interestingly, few papers have addressed the issue of the small-sample and asymptotic performance of these information criteria specifically for the case of MSMs. Because these measures rely on the same conditions employed in the asymptotic theory of the LRT, their small and large sample properties are, likewise, largely unknown. However, the literature on mixtures provides some encouraging evidence in the context of unconditional models, suggesting that the BIC may provide a reasonably good indication for the number of components (see, e.g., McLachlan & Peel, 2000, Chapter 6, for a survey and references). As Granger, King, and White (1995) have pointed out, rankings based on information criteria are arguably more appropriate for model selection than procedures based on formal hypothesis testing, partly because, unlike testing, they do not favor unfairly the model chosen to be the null hypothesis. For instance, Psaradakis and Spagnolo (2003) discuss the effectiveness of procedures based on the AIC as a means of selecting the number of regimes in MSVAR models.

Once a restricted set of MSMs has been estimated, the need of further improvements could arise as the result of diagnostic checks.[14] Although the EM algorithm naturally delivers estimates of the parameters $\tilde{\theta}$ and $\tilde{\xi}^1_{1|0}$, besides the smoothed sequence of probability distributions $\{\tilde{\xi}_{t|T}\}_{t=1}^{T}$, it would lead to define the (smoothed) residuals as

$$\tilde{\mathbf{u}}_t \equiv \mathbf{y}_t - \mathbf{Y}_t\hat{\mathbf{B}}\hat{\xi}_{t|T}$$

that are not well suited for diagnostic checks as they are full-sample statistics and hence they structurally overestimate the explanatory power of a MSM. On the contrary the one-step prediction errors

$$\tilde{\mathbf{e}}_{t|t-1} \equiv \mathbf{y}_t - \mathbf{Y}_t\hat{\mathbf{B}}\mathbf{P}\hat{\xi}_{t-1|t-1}$$

are limited information statistics (being based on filtered probabilities) and uncorrelated with the information set $\mathscr{F}_{t-1}$ because $E[\mathbf{y}_t|\mathscr{F}_{t-1}] = \mathbf{Y}_t\hat{\mathbf{B}}\mathbf{P}\hat{\xi}_{t-1|t-1}$ and therefore form a martingale difference sequence $E[\tilde{\mathbf{e}}_{t|t-1}|\mathscr{F}_{t-1}] = \mathbf{0}$. Therefore standard tests of this hypothesis (such as Portmanteau tests of no serial correlation) can be used.[15] In the absence of MS heteroskedastic components (i.e., the covariance matrices of shocks fail to depend on regimes), researchers in empirical finance (e.g., Kim & Nelson, 1999) have also suggested to check whether the smoothed, standardized residuals contain any residual ARCH effects. Standard LM-type as well as Ljung-Box tests can be applied. This is a way to check whether MS variances and covariances may be sufficient to capture most of the dynamics in volatility, else explicit ARCH-type modeling (even of a MS nature, see sixth section) may be required. It is also possible to apply White-style tests that examine the null hypothesis that the score statistics are serially uncorrelated. Hamilton (1996) has shown how White's results may be used to construct tests for possible alternatives to MSM.[16] For instance, by considering the score with respect to the regime-dependent mean coefficients, a White test for autocorrelation can be constructed; an ARCH test can be implemented by examining the serial correlation properties of the scores with respect to regime-dependent variances; the first-order Markov assumption can be tested against the alternatives that it depends on the state at earlier times or that it depends on the realizations of the data by checking whether the scores of transition probabilities can be predicted by the corresponding lagged score or the score of the mean; etc.

Another important type of diagnostic check concerns the number of regimes, K. The problem is that under any number of regimes smaller than K there are a few structural parameters of the unrestricted model – the elements of the transition probability matrix associated with the rows that correspond to "disappearing states" – that can take any values without influencing the resulting likelihood function. We say that these parameters become a *nuisance* to the estimation. The result is that the presence of these nuisance parameters gives the likelihood surface so many degrees of freedom that computationally one can never reject the null that the nonnegative values of those parameters are due to sampling variation. Different alternative ways have been proposed to develop sound inferential procedures concerning the number of regimes in MSMs. Hansen (1992) proposes to see the likelihood as a function of the unknown and nonestimable nuisance parameters so that the asymptotic distribution is generated in each case numerically from a grid of transition and regime-dependent nuisance parameters. The test statistic becomes

$$LW_T \leq \sup_{\rho} LW_T(\rho)$$

where $\rho \equiv \mathrm{vec}(\mathbf{P})$ and the right-hand side converges in distribution to a function of a Brownian bridge. In most of the cases, a closed-form expression cannot be found and the bound must be calculated by simulation and becomes data dependent. Also Davies (1977) bounds the LRT but avoids the problem of estimating the nuisance parameters and derives instead an upper bound for the significance level of the LRT under nuisance parameters:

$$\Pr(LRT > x) \leq \Pr(\chi_1^2 > x) + \sqrt{2x}\exp\left(-\frac{x}{2}\right)\left[\Gamma\left(\frac{1}{2}\right)\right]^{-1}$$

The bound holds if the likelihood has a single peak. A related test is based on another corrected LRT that seems to have been used first by Turner, Startz, and Nelson (1989) but is credited to Wolfe,

$$LR^{\mathrm{Wolfe}} = -\frac{2}{T}(T-3)[\ln L(\tilde{\theta}) - \ln L(\tilde{\theta}_r)] \xrightarrow{D} \chi_r^2$$

where $\tilde{\theta}_r$ is obtained under the null of single-state multivariate normality and $r = K(K-1)$ because in the absence of regime switching there are $K(K-1)$ elements of $\mathbf{P}$ that cannot be estimated. Davidson and MacKinnon's (1981) *J* test for non-nested models can be also applied, because MSMs with K and $K-1$ regimes are logically nested but cannot be treated as such on a statistical basis. To implement a *J* test one has to estimate the model with K and $K-1$ states and calculate their full information fitted values, $\tilde{\mathbf{y}}_t^{(j)} = \mathbf{Y}_t\hat{\mathbf{B}}^{(j)}\hat{\mathbf{x}}_{t|T}^{(j)}$; then estimate the (multivariate) regression

$$\mathbf{y}_t = (\mathbf{I}_N - \Delta)\mathbf{Y}_t\hat{\mathbf{B}}\hat{\xi}_t^{(K-1)} + \Delta\tilde{\mathbf{y}}_t^{(K)} + \eta_t$$

The p-value of an F-test for the matrix of coefficients Δ gives the p-value for the null of K regimes.

Finally, common sense suggests that correct specification of a MSM should give smoothed probabilities $\{\hat{\xi}_{t|T}\}_{t=1}^{T}$ that consistently signal switching among states with only limited periods in which the associated distribution is flatly spread out over the entire support so that uncertainty dominates. Regime classification measures (RCMs) have been popularized as a way to assess whether the number of regimes K is adequate. In simple two-regime frameworks, the early work by Hamilton (1988) offered a rather intuitive RCM:

$$\mathrm{RCM}_1 = 100\frac{K^2}{T}\sum_{t=1}^{T}\prod_{k=1}^{K}\hat{\xi}_{t|T}^{k}$$

where $\hat{\xi}_{t|T}^{k} \equiv \Pr(S_t = k|\mathrm{y}_1,\ \mathrm{y}_2,\ldots,\ \mathrm{y}_T;\ \tilde{\theta})$, that is, the sample average of the products of the smoothed state probabilities. Clearly, when a MSM offers precise indications on the nature of the regime at each time t, the implication is that for at least one value of $k = 1,\ldots,\ K,\ \hat{\xi}_{t|T}^{k} \simeq 1$ so that $\sum_{k=1}^{K}\hat{\xi}_{t|T}^{k} \simeq 0$ because most other smoothed probabilities are zero. Therefore a good MSM will imply $\mathrm{RCM}_1 \simeq 0$.[17] However, when applied to models with $K>2$, RCM_1 has one obvious disadvantage: a model can imply an enormous degree of uncertainty on the current regime, but still have $\sum_{k=1}^{K}\hat{\xi}_{t|T}^{k} \simeq 0$ for most values of t. For instance, when $K=3$, it is easy to see that if $\hat{\xi}_{t|T}^{1} = \hat{\xi}_{t|T}^{2} = 1/2$ and $\hat{\xi}_{t|T}^{3} = 0\ \forall t$, then $\mathrm{RCM}_1 = 0$ even though this remains a rather uninformative switching model to use in practice. As a result, it is common that as K exceeds 2, almost all switching models (good and bad) will automatically imply values of RCM_1 that decline toward 0. Guidolin (2009) proposes a number of alternative measures that may shield against this type of problems, for instance

$$\mathrm{RCM}_2 = 100\left\{1 - \frac{K^{2K}}{(K-1)^2}\frac{1}{T}\sum_{t=1}^{T}\prod_{k=1}^{K}\left[\xi_{t|T}^{k} - \frac{1}{K}\right]^2\right\}$$

Ang and Bekaert (2002b) have also proposed one alternative RCM that has found widespread use:

$$\mathrm{RCM}(K) \equiv 100K^2\frac{1}{T}\sum_{t=1}^{T}\prod\nolimits_{k=1}^{K}\xi_{t|T}^{k}$$

UNIVARIATE AND EARLY APPLICATIONS

A first, traditional application of MSMs in financial economics has consisted in using them as flexible tools to fit the time series dynamics – initially mainly at the univariate level, but more recently also at the multivariate level – of financial data. While historically the 1980s and a portion of the 1990s have been characterized by scores of papers that have shown by examples that MSMs could provide a good fit to popular series, yielding reasonable outcomes in terms of parameter estimates and regime characterizations, starting from the late 1990s MSMs have found increasing application to testing implications and hypotheses derived from finance theories. The "game" in these papers is traditionally simple: given one or more financial time series, the researcher resorts to one or more types of MSMs (more generally, regime switching models) to show that the data contain evidence of different regimes, which often plays an important role in understanding or interpreting the underlying phenomenon. More often than not, it is typical to compare the fitting performance – either using statistical criteria (such as the maximized log-likelihood function and information criteria, see, e.g., Guidolin & Timmermann, 2006a, or Driffill & Sola, 1998) or using the implications for some quantities of interest (e.g., correlations in excess of a Gaussian benchmark, as in Ang & Chen, 2002) – of MSMs with other dynamic time series models that are regarded as close competitors, such as GARCH models. For instance, Guidolin and Nicodano (2009) is an international equity portfolio application of MSMs in which substantial energies are devoted to compare the in-sample fit and the implications for the dynamics of higher-order (co)moments (i.e., co-skewness and co-kurtosis) of a few multivariate MSMs to be compared to DCC-GARCH.

This literature is too vast for any attempt at summarizing it.[18] Therefore I review a few examples of this line of work, noting that these should only be interpreted as examples, and being reminiscent that hundreds of similar papers have been written and published between the late 1980s and today, with additional, related research – especially applications to sub fields of financial economics and types of data that were initially left at the margin of this strand of empirical research – still in the process of appearing. Engel and Hamilton (1990) may be taken as a case study of the first wave of applications of MSMs. It is also interesting that this paper contains an application of MSMs to an asset class – foreign exchange rates – different from equities and Treasury bonds, that instead will play a dominant role in the following. Engel and Hamilton were motivated by a headline policy question: whether one could find a rational framework to explain why the

U.S. dollar had risen (vis-a-vis the Deutsche mark, the French franc, and the British pound) so dramatically in the early 1980s and then fell afterward. They use a simple two-state, time homogeneous MSIH framework to formalize the concept of long swings in the exchange rate:

$$\Delta e_t^{i,j} = \mu_{S_t^{i,j}} + \sigma_{S_t^{i,j}} \varepsilon_t^{i,j} \quad \varepsilon_t^{i,j} \sim \text{NID}(0,1)$$

where $e_t^{i,j}$ is the log-exchange rate, i and j denote the countries/currencies to which the exchange rate refers to, and $S_t^{i,j} = 1,\ 2$. This means that in regime 1, the trend in the exchange rate is μ_1 and in regime 2, it is μ_2. Within the single-regime empirical macroeconomics literature of the period, Engel and Hamilton's quest was well received because the macroeconomic explanations advanced in the 1980s that had focused on effects of U.S. monetary and fiscal policies on real interest rates, on lower capital taxes in the United States than abroad, and on a "safe haven" effects benefiting the U.S. dollar, had not managed to reconcile existing models with empirical facts. Quarterly 1973–1988 data revealed that the U.S. dollar had actually gone through periods of persistent depreciation, followed by appreciation, and then again depreciations – three "long swings." Engel and Hamilton find clean evidence of two regimes: In regime 1 the mark is rising 4% per quarter against the dollar, the franc 3.3%, and the pound 2.6%; regime 2 is associated with quarterly declines in the foreign currencies of −1.2%, −2.7%, and −3.8%, respectively. The point estimates of $p_{11} \equiv \Pr(S_{t+1} = 1 |\ S_t = 1)$ range from 0.82 to 0.93, while the estimates of $p_{22} \equiv \Pr(S_{t+1} = 2 |\ S_t = 2)$ from 0.91 to 0.93. Therefore a given regime is likely to persist for several years, and Engel and Hamilton's inferred turning points that matched the historical record. The expected length of state 1 is seven quarters for Germany, six for France, and fourteen for the United Kingdom; on average state 2 lasts fourteen quarters for the mark, eleven for the franc, and twelve for the pound. States 1 and 2 are differentiated not only by their means but also by the variances of their conditional distributions: the exchange rate would be much more variable when the dollar is appreciating. Formal tests of the null hypothesis that exchange rates follow a martingale reject the null in two cases out of three.[19] In terms of in-sample fit, the MSM reduces the mean forecast error by 9–14% at horizons from two quarters to a year for all three currencies, relative to a random walk (RW).[20]

Engel and Hamilton were among the first researchers who felt the awkwardness of modeling a phenomenon that involves three exchange rates by performing three univariate estimations. They extended their exercise to a joint, trivariate estimation but had little success in using this model when it was assumed to be driven by a common, unique Markov state. Although in their article they speculate that treating the three exchange rates as a group may be inappropriate because country-specific developments played an important role in the evolution of exchange rates in the 1970s, of course we now understand that this result may have come from the data suggesting a need for more than two regimes (or, which is even more complex, from a need to model correlated but country-specific Markov states, see fourth section). In any event, Engel and Hamilton (1990) represents one of the first and highly successful applications of MS methods in empirical finance and, despite its relative simplicity in terms of models estimated and statistical tests applied, it showed how MSMs could concretely help financial economists to formulate and tests hypotheses that would have been impossible to formulate within either regression or single-state ARIMA frameworks.

Another lucid contribution to an exact understanding of the importance of MS (more generally, of Gaussian mixtures) in modeling univariate asset returns came from a paper by Rydèn, Teräsvirta, & Åsbrink (1998, henceforth RTA), who returned to investigate early evidence of nonlinear dynamics in financial returns offered by Granger and Ding (1995).[21] In particular, RTA marks a step forward in the awareness by mainstream financial empiricists that simple, single-regime ARMA models cannot cope with (fit) a wide range of properties commonly found in asset (stock, bond, and currency) returns. Granger and Ding considered long daily stock return series and established a few properties which seem to hold for a large number of such series:

(A) Returns are not autocorrelated in levels (except, possibly, at lag one);
(B) The autocorrelation functions of $|r_t|$ and r_t^2 decay slowly starting from the first autocorrelation and $\mathrm{Corr}(|r_t|, |r_{t-l}|) > \mathrm{Corr}(|r_t|^{\theta}, |r_{t-l}|^{\theta-1}) > 0$ for large l and $\theta \neq 1$ (the *Taylor effect*);
(C) The decay in the autocorrelation functions of squares and absolute values of returns is much slower than the exponential rate of a stationary AR(l) or ARMA(p,q) model;
(D) The autocorrelations of $\mathrm{sign}(r_t)$ are insignificant; moreover, $|r_t|$ and $\mathrm{sign}(r_t)$ are independent, $\hat{E}(r_t) \simeq \hat{\sigma}(r_t)$, and the marginal distribution of $|r_t|$ is exponential (after dropping extreme outliers).

Granger and Ding (1995) and Ding and Granger (1996) proposed to capture these properties of asset returns by assuming they are doubly exponentially distributed with $r_t = \sigma_t \varepsilon_t$, and σ_t following an ARCH(q) process in which the volatility σ_t depends on $|\varepsilon_{t-1}|, \ldots, |\varepsilon_{t-q}|$. RTA present an alternative to the distribution proposed by Granger and Ding based instead on the assumption that the marginal distribution of returns is a mixture of normal distributions. Using daily S&P 500 returns for a long 1928–1991 sample, they show that such a mixture allows DGPs capable of closely reproducing the properties (A) – (D) listed earlier in text. Moreover, RTA stress that postulating a hidden MSM will also generate higher-order temporal dependence in the process.[22] RTA split their long, 63-year long sample in 10 equal consecutive samples, each containing 1,700 observations and perform MS estimation and testing (to find the most appropriate number of regimes, using a parametric bootstrap to pin down the small-sample corrected p-values for their LRTs, and to assess the presence of misspecifications) on each of the 10 sub-series. Their bootstrapped LRTs indicate that $K=2$ is always appropriate against $K=1$, but that in as many as 7 sub-samples out of 10, $K=2$ is rejected in favor of $K=3$, which casts some doubts on the widespread habit in parts of the literature to naively identify MSMs with two-state models. In fact, most estimated two-state MSMs in RTA turn out not fully stable over time, even though some regularities between adjacent models can be singled out. Such instability of parameter estimates in MSMs represents an obvious sign of misspecification.

Interestingly, RTA find considerably higher stability in their 10 daily S&P 500 return sub-samples when the series are "cleaned" of outliers (winsorized), where outliers are defined as all returns falling outside an interval of ± 4 sample standard deviations, to be replaced by the limit of the interval. It remains a bit disconcerting that Gaussian mixtures – which are models proposed in the statistics literature as tools to detect and manage data "contamination" from different regimes – may perform best when they fail to fit the entire empirical distribution of the data. Here the suspicion is that through parameter instability of the MSM, the data may be trying to signal some other type of misspecification. However, although MSMs capture the ratio between implied means and standard deviations of absolute returns, the implied skewness and kurtosis of absolute returns, and the Taylor effect in autocorrelations well, they gave mixed indications with reference to autocorrelations. While the empirical autocorrelation functions of absolute returns decay very slowly, MS autocorrelation functions show a somewhat faster decay.

A recent paper that in no way can be listed among the early applications of MSMs to empirical finance research but that retains the fresh forward push that has characterized the seminal work by Engel, Hamilton, and Schwert is Acharya, Amihud, and Bharath (2010, AAB), who have applied MSMs to investigate the regime switching nature of the exposure of U.S. corporate bond returns to liquidity shocks of stocks and Treasury bonds. The reason why this simple paper retains the innovative impact of the papers mentioned above is due to the fact that AAB simply estimate a number of *MS regressions* (as opposed to the MS ARMA models examined above; see also Guidolin et al., 2009): because regressions play such a key role in the definition and estimation of linear asset pricing models in finance, AAB is an important but recent example of the intuitive (almost naive) idea that in a regression model the slope and intercept coefficients may follow a MS dynamics with important implications in the light of the 2008–2009 Great Financial Crisis. It is well known that liquidity shocks affect asset prices because asset liquidity affects expected returns of both stocks and bonds (Amihud & Mendelson, 1986). Because asset illiquidity is persistent, an unexpected rise in illiquidity raises expected illiquidity. Consequently, investors require higher expected returns, which makes asset prices fall if the rise in illiquidity does not have an appreciable positive effect on assets' cash flows. This generates a negative liquidity beta of assets, that is, a negative relationship between illiquidity shocks and asset realized returns. However, most existing papers in the literature have simply examined the unconditional – econometricians would say, under a false restriction that $K=1$ – effect of liquidity risk not paying enough attention to the casual observation (see, e.g., Acharya & Pedersen, 2005) that the impact of liquidity shocks on asset prices is significantly stronger in bad times. AAB show instead that the response of corporate bond prices to liquidity shocks of stocks and Treasury bonds varies over time in a systematic way, switching between two regimes that they call normal and stress states. In the case of "junk" bonds, the betas of the two illiquidity factors are statistically insignificant in normal times, but they become highly negative and significant in the stress regime. Moreover, the two regimes can be predicted by macroeconomic and financial variables: periods of stress are associated with adverse macroeconomic conditions, such as recessions and adverse financial market conditions. To uncover these properties, AAB regress the probability of being in the stress regime on lagged macroeconomic and financial market variables. AAB provide out-of-sample (OOS) predictions of corporate bond returns for the years 2008–2009: regressions of monthly realized returns on predicted returns produce R^2-s of 74% and 77% for junk and investment grade bonds, respectively. The coefficients on predicted returns are close to one and the intercepts are close to zero (differences are statistically insignificant); the predicted return does a reasonable job at predicting the returns of March 2008 (Bear Stearns' collapse) and September to December 2008 (Lehman Brothers' collapse and the post-Lehman phase). As we shall see in the seventh section, the intuition that regime probabilities could be predictable using macroeconomic factors had already been developed in the nonlinear econometrics literature since the mid-1990s and has recently found considerable play in financial applications.

A similar application of MS regression methods to an important financial question is Alexander and Kaeck (2008) who study the time-varying empirical influence of a wide set of theoretical determinants of daily spread changes in the iTraxx Europe, an equally weighted index that contains 125 single firm investment grade credit default swaps (CDSs), for the period June 2004 to June 2007. They find that most theoretical variables that we would expect to affect spreads (interest rates, stock returns and implied volatility) indeed contribute to the explanation of CDS spread changes but that their influence depends on the prevailing market circumstances. MS regressions reveal that during the volatile CDS regime, credit spreads are highly sensitive to stock market variables; interest rates are instead more significant determinants of credit spreads during the (predominant) regime when CDSs are less volatile; interestingly, their model displays higher explanatory power in the volatile regime. The two market volatility regimes are quite persistent: for example, in the main iTraxx Europe series the probability of remaining in the volatile regime is 0.94; the persistence is even higher for the tranquil regime, with a "stayer" probability of 0.99. Therefore for efficient hedging of CDS exposures traders should adjust equity hedge ratios to the relevant regime. Interestingly, the level of interest rates can influence the probability of entering the turbulent regime – this is captured by a simple logit model applied to filtered regime probabilities and not by a structural modification of the MSM to capture time-varying transitions as it would be possible (see seventh section) – but, once that regime is accessed, interest rates have little effect on CDS spreads.

Recently, the literature has returned to a number of issues that the early papers had raised, such as the importance (or at least, the advantages) of adopting a multivariate approach in MSMs, the opportunity to push the modeling efforts beyond that standard two-state MS case, and the payoff of investigating whether and how Markov chains used to fit the dynamics of individual series may relate to an encompassing, multivariate MSM. Although a number of papers have followed a similar path in recent years, Guidolin & Timmermann (2006a, GT) represents a good example of such efforts. They study various MSMs for the joint distribution of U.S. stock and bond returns, using models in Eq. (1). One of the problems of applications of MSMs to empirical finance is that there are no clear guidelines useful to structure the generalization of univariate nonlinear models to the multivariate case. Naive approaches are known to yield overwhelmingly large models. To see this, suppose that each of N univariate return series is governed by a simple MSIH process, and that the innovations are simultaneously correlated but with zero serial and cross-serial correlations. Under no further restrictions on the relationship between the individual regimes $S_{1t}, S_{2t}, \ldots, S_{Nt}$, the states S_t^* for the joint process of the N return series can be obtained from the (Cartesian) product of the individual states: $S_t^* \equiv S_{1t} \times S_{2t} \times \ldots \times S_{Nt}$. This gives a total of $K = \prod_{n=1}^{N} K_n$ possible states and $K(K-1)$ state transition probabilities. Even assuming $K_i = 2$ for $n = 1, 2, \ldots, N$, for a sufficiently realistic N (such as $N = 8$ in Guidolin & Ono, 2006), this easily would deliver K as large as a few hundred, which clearly turns out to be impossible to handle.[23]

The key empirical finding in GT's (2006a) analysis of monthly data for returns on portfolios of U.S. large and small capitalization stocks and 10-year Treasury bonds is that even though there are well-defined regimes in the marginal distributions of both stock and bond returns, there is very little coherence among them. This complicates the structure of MSMs for the joint dynamics of stock and bond returns and suggests that a richer model with several states is required. Specifically, for all the portfolios they examine, information criteria point to a two-state specification for both stock market portfolios and a three-state specification for bonds. Each of the two regimes identified in the two stock return series has a clear bull and bear-type economic interpretation. However, while there are strong similarities between the smoothed probabilities extracted from large and small stocks, this is not the case for stock and bond returns. GT's specification search on multivariate MSMs including stock and bond data – based on sequential LRTs and on information criteria for the models defined by $K = 1, 2, 3, 4, 5$ and VAR lag orders $p = 1,2$ – confirms that in all cases linearity is strongly rejected no matter how many states and lags are present in the MSM. A Hannan–Quinn information criterion supports four states that display an appealing economic interpretation: Regime 1 is a "crash" state characterized by large, negative mean excess returns and high volatility; regime 2 is a low growth regime characterized by low volatility and small positive mean excess returns on all assets; regime 3 is a sustained bull state in which stock prices – especially small stocks – grow rapidly on average but interest rates surge and excess returns on long-term bonds are negative; regime 4 is a "bounce-back" regime with strong market rallies and high volatility. Correlations between returns also vary substantially across regimes.[24] Interestingly, the transition probability matrix has a very peculiar form. Exits from the crash state are almost always to the recovery state and occur with close to 50% chance suggesting that, during volatile markets, months with large, negative mean returns cluster with months that have high positive returns. The slow growth state is far more persistent with an average duration of seven months. The bull state is the most persistent state with a "stayer" probability of 0.88. On average the market spends eight successive months in this state. Finally, the recovery state is again not very persistent and the market is expected to stay just over three months in this state. The steady-state probabilities, reflecting the average time spent in the various regimes are 9% (state 1), 40% (state 2), 28% (state 3) and 23% (state 4). Hence, although the crash state is clearly not visited as often as the other states, it is by no means an "outlier" state.

These results are typical of the literature but trigger more questions than the answers they provide. For instance, with four regimes, it becomes obvious that even when the individual states are not excessively persistent, these may have non-negligible ergodic probabilities and therefore give an important contribution to characterize the unconditional joint density of the data.[25] It is also interesting that GT end up setting $p = 0$ despite the presence of bond returns in their sample, which are notoriously serially correlated. Once again, this derives from the flexibility (and the ability to generate serial correlation from the persistence of the Markov chain) of a four-state MSM. However, this poses additional issues as to whether VAR components in fixed income returns are as strong as commonly thought, net of regime switching non-linear effects (see, e.g., Aït-Sahalia, 1996).[26] Finally, it may be surprising that GT systematically find that regimes that imply high mean returns tend to display low volatility, and vice versa. However, it should be noted that these are not ex-ante expected returns and ex-ante volatility estimates, since they do not account for the probability of switching across regimes or learning in real time about the regime. Guidolin (2011) discusses the asset pricing literature that has successfully tackled this puzzling evidence using MS.

GT also try an interpretative exercise that is quite common in the literature and that resembles Acharya et al. (2010): they compute correlations between the smoothed state probabilities from their four-state MSIH and NBER recession dates, finding estimates of 0.32 (state 1), 0.13 (state 2), 0.21 (state 3) and 0.18 (state 4). Notice that since the state probabilities sum to one, by construction if some correlations are positive, others must be negative. This suggests that indeed, the high volatility states – states 1 and 4 – occur around official recession periods. Interestingly, and consistent with the idea that the state probabilities backed out from movements in financial asset returns should lead economic recession months (see Ozoguz, 2009), the correlation between the state 1 probability lagged six months and the NBER recession indicator rises to 0.40. Finally, when additional predictor variables are incorporated in the MSM – in this case a typical financial ratio that plays a key role in the recent literature (see fifth section), the dividend yield – GT find that all information criteria as well as sequential LTRs favor a MSVAR(1) model. This is not surprising given the strong persistence of the dividend yield which itself contains multiple regimes. Moreover, although the model is extended by an autoregressive term, a four-state model continues to provide the best trade-off between fit and parsimony. While in a simple single-state VAR model the dividend yield predicts returns on small stocks but does not appear to be significant in the equations for large stocks and bonds, estimates of the four-state MS VAR matrices suggest that the effect of changes in the dividend yield on asset returns is stronger, although regime specific. Inclusion of the dividend yield therefore does not weaken the evidence of multiple states, nor does the presence of such states in a framework that allows for heteroskedasticity remove the predictive power of the dividend yield. This may be interpreted as evidence that MS dynamics is not only (or mostly) a short-cut approach to capture complex predictability patterns in financial data. To the contrary, MS would represent a separate and additional channel of predictability that may be important to exploit in portfolio decisions (see, e.g., Guidolin & Timmermann, 2005, 2007) or risk-management (see, Haas, Mittnik, & Paolella, 2004). However, so far in the literature, we have had few attempts to clearly disentangle the strength and economic value of nonlinear predictability from standard, VAR-type predictability.

MULTIVARIATE MSMs AND DYNAMIC FACTOR MODELS

A renewed and increased attention to multivariate applications is natural in areas of empirical finance – such as international finance – where phenomena such as dynamic correlations and contagion represent key research questions. One leading case study is Baele (2005) who has used MS techniques to investigate to what extent globalization and regional integration may lead to increasing equity market interdependence. The paper by Baele reveals a number of open issues on the best fitting structure of MSMs when the goal of the exercise is to provide a description of the dynamics of the data that is useful to test competing hypotheses. However, this is a literature in which MSMs are destined to become a central econometric tool of empirical investigation. In fact, as it is known since Forbes and Rigobon (2002), what has become known as a *shock spillover model* – when extended to incorporate regime switching dynamics – may lead to more powerful tests of contagion effects: By allowing interdependence among markets to vary with the degree of integration, a gradual increase in market interdependence cannot be mistaken for the occurrence of contagion, a potential pitfall in all single-regime factor model-based tests of contagion. Previous studies had typically used dummy variables to test whether important events had a significant impact on shock spillover intensities. This approach fails in situations where these events are anticipated or need time before becoming effective. On the contrary the MS shock spillover model developed in Baele (2005) has the advantage that the spillover intensities switch endogenously rather than exogenously from one regime to the other, so that probability statements can be formulated about the relative likelihood of the spillovers.

Even though the typical issues with MS GARCH models will be discussed in the sixth section, let me stress that Baele uses a general framework for a range of bivariate processes for U.S. and individual national index returns which nest constant correlation, BEKK GARCH, MSMs, as well as MS GARCH models:[27]

$$\mathbf{r}_t = \mu_t + \mathbf{A}\mathbf{r}_{t-1} + \varepsilon_t \quad \varepsilon_t \sim \text{NID}(\mathbf{0}, \mathbf{\Omega}_t)$$

When μ_t and $\mathbf{\Omega}_t$ are simply switching following a two-state first-order Markov chain, $\mu_t = \mu_{S_t}$ and $\mathbf{\Omega}_t = \mathbf{\Omega}_{S_t}$, we are facing a simple MSIH(2)-VAR(1) model; $\mathbf{\Omega}_{t+1}$ is otherwise allowed to follow a two-state multivariate MS GARCH(1,1) dynamics:

$$\mathbf{\Omega}_t = \mathbf{C}'_{S_t}\mathbf{C}_{S_t} + \mathbf{M}_{S_t}\varepsilon_t\varepsilon'_t\mathbf{M}_{S_t} + \mathbf{B}_{S_t}\mathbf{\Omega}_{t-1}\mathbf{B}_{S_t}$$

These alternative frameworks are estimated to quantify the magnitude and time-varying nature of volatility spillovers from the aggregate EU and U.S. markets to 13 local European equity markets, some of which have followed a path of integration culminated in the introduction of the euro in 1999 and others that – even though they belong to a highly homogeneous area – have elected to stay outside of the European Monetary Union. Weekly data reveal evidence of regime switches to be both statistically and economically important. Statistically, the best fit is provided by the two-state MS GARCH(1,1) and this model outperforms the competing bivariate BEKK GARCH, similarly to the evidence in Guidolin and Nicodano (2009) for a larger set of international stock index returns (and in comparison to DCC GARCH models). Both EU and U.S. shock spillover intensities increased substantially over the 1980s and 1990s, though the rise is more pronounced for EU spillovers. While in the first half of the 1980s common European shocks explained on average about 8% of local variance, this proportion increased to 23% by the end of the 1990s. Similarly, the importance of U.S. shocks increased from 15% to 27%. Finally, Baele uncovers evidence of contagion from the U.S. market to a number of local European markets during periods of high world volatility. There is instead only weak evidence of contagion from the EU market to the German equity market.

In a critical perspective, a number of features of Baele's research design deserve attention. Even though the use of weekly data suggests a need to integrate some ARCH effects in the MS framework, it is noteworthy that Baele's models feature rich bivariate MS GARCH effects and yet entertain the presence of only one and two regimes. Equivalently, even though Baele carefully compares the performance of two-state MSIH(2)-VAR(1) and MS GARCH models, a chance exist that an equally informative comparison would have instead involved MSIH(K)-VAR(1) models, with $K \geq 3$. It would be interesting to check with simulations whether there is a chance that data generated from multi-state processes (e.g., $K=4$ as in Guidolin & Timmermann, 2006a) may lead to a spurious conclusion in favor of two-state MS GARCH models, although the existing evidence for univariate process to be surveyed in the sixth section suggests these substitution effects are likely to manifest themselves. Additionally, Baele draws his economic implications from 26 (13 times two possible sources of contagion, the United States and the EU) different bivariate MS GARCH models, in which the underlying Markov chain has to be re-estimated for every pair, which implies that the regime switching process for the United States and the EU are in principle made depend on the data on the other country included in each pair. Although, even exploiting modern computational resources, it remains doubtful whether MSMs for a system with $N=15$ can be actually estimated using the classical (EM) methods employed by Baele, novel MCMC-based Bayesian frameworks seem to be currently available that may lead to U.S. and EU stock returns being characterized by a unique Markov chain process (see e.g., Kim, Morley, & Nelson, 2005). Finally, it would also be interesting to try and use the structure of the Markov chain process underlying MSMs to test spillover and/or contagion propositions, which seems to be an approach only partially explored in the literature (but see Sola, Spagnolo, & Spagnolo, 2002).

In an effort to explain the cross-sectional and time series variation of the shock spillover intensities, Baele has also used equity market development, trade integration, and price stability in a logit model that explains the filtered probabilities of the spillover regime (transformed into a 0/1 dummy on the basis of whether the filtered probability exceeds 0.5), finding that these factors increase the probability of switching from a low to a high sensitivity to common European shocks. Although the idea of using (transformed) regime probabilities in probit/logit-type regressions is not new and has appeared often in applied work despite its obvious limitations, this approach seems to naively mimic the idea of making regime probabilities a function of some exogenous (as in this case) or endogenous but predetermined variables, that we are about to deal with in a systematic fashion in the seventh section.

A paper that has pushed further the frontier by proposing and estimating multivariate MSMs of relatively large dimensions is Ang and Bekaert (2002b), who apply MSVAR techniques to investigate the joint dynamics of short-term interest rates across the United States, the United Kingdom, and Germany. Although details are provided in the seventh section, let us notice here that similarly to the logic explained with reference to Guidolin and Timmermann (2006a), Ang and Bekaert assume the existence of a two-state MS variable driving the term structure in every country. These country-specific regimes are assumed to be independent across countries. This means that even if S_t^c follows a simple two-state process ($c=$ Ger,UK,US), S_t will in fact follow an eight-state Markov chain, with $S_t = S_t^{\mathrm{US}} \times S_t^{\mathrm{GER}} \times S_t^{\mathrm{UK}}$. However, because it is conceivable that there is a world business cycle driving interest rates in many countries simultaneously, in some of their models they allow for interdependence of various forms across countries to reduce the number of regimes. Another noteworthy feature of Ang and Bekaert's paper is their widespread use of carefully built residual diagnostic tests and statistical tests of sample moment matching adapted to their multivariate MS framework. For instance, they investigate the fit of competing – specifically, VARs versus MSVAR – models to the unconditional moments of the data, including higher-order moments which have often been stressed as an important implication of non-linear dynamics in short-term rates (see, e.g., Ahn & Gao, 2000). To enable comparison across several models, they introduce the statistic $H \equiv (\hat{\mathbf{g}} - \bar{\mathbf{g}})' \sum_g^{-1} (\hat{\mathbf{g}} - \bar{\mathbf{g}})$, where $\bar{\mathbf{g}}$ is a vector collecting the sample estimates of the unconditional moments of interest, $\hat{\mathbf{g}}$ is a vector of collecting the unconditional moments implied by the estimated model, and $\sum_g$ is the covariance matrix of the sample estimates of the unconditional moments, estimated by GMM using a Newey-West estimator.[28]

Using monthly observations on three-month short rates and five-year rates on zero-coupon government bonds from the United States, the United Kingdom, and Germany over a 1972–1996 sample, Ang and Bekaert implement the residual diagnostic tests in the form of a GMM test of the moment conditions on the mean of the scaled residuals, $E[e^c_t e^c_{t-k}] = 0$ $(k = 1, \ldots, 6)$, which they refer to as "mean" residual tests, and a GMM test of the moments of the variance of the scaled residuals, $E[(e^c_t - 1)^2(e^c_{t-k} - 1)^2] = 0$ $(k = 1, \ldots, 6)$ which they refer to as "variance" residual tests. They also test moment-matching of the first four central moments, the autocorrelogram, and cross-correlations. Their evidence is mixed because it strongly depends on the country taken into consideration. The single-state VARs generally outperform the MSVAR models at matching unconditional moments. This warns against the belief that in small samples it may be easy to fit multivariate MSMs deriving precise parameter estimates. In fact, Ang and Bekaert also perform a simulation experiment in which 1,000 short samples (of the same size employed in their paper) are generated from a complex MSVAR(1) with logistic, TVTPs, fitted to short-term rate and term spread data. The competing models they entertain are VAR(1) and an MSVAR(1) with constant transition probabilities.[29] They find that in a striking proportion of the simulations (in excess of 50%), the best fit to population moments as measured by H is given by simple single-state models, *even though the true model is a complex MSVAR*. They also examine the empirical distribution of the moments produced by the models in small samples to find that MSVAR models tend to overestimate the mean and underestimate the variance of the short rate; on the opposite, VAR models produce close to unbiased estimates of the mean and variance. It would be important to benefit for more extensive simulation experiments to verify how widespread the problems with estimating MSVARs may be in empirical finance applications.

Of course, all multivariate MSMs face one common challenge: unless the available time series are long (which usually means that one is to use daily data) and/or computational resource unlimited, it is difficult to seriously entertain the idea of estimating MSMs for large vectors of endogenous variables (think of the example of $N = 15$ given above). Baele et al. (2010, BBI) have recently proposed a smart set of innovative – at least within the MS literature – tools to handle exactly such a challenge: factor models. BBI study the economic sources of U.S. stock-bond return co-movements and their time variation. Even though Baele (2005) has been taken as a key example of work on MS dynamic correlations and contagion in international equity markets, the same logic can find application to modeling the correlation between pairs of asset classes or portfolios within a country, such as the wildly gyrating stock-bond return correlation coefficient: for instance, during the mid-1990s, the (rolling window) stock-bond correlation was as high as 0.6, to drop to levels as low as −0.6 by the early 2000s. BBI develop a DFM that imposes structural restrictions inspired by recent New-Keynesian models and that incorporates MS:

$$\mathbf{r}_{t+1} = E_t[\mathbf{r}_{t+1}] + \mathbf{B}'_{t+1}\mathbf{f}_{t+1} + \varepsilon_{t+1} \quad \varepsilon_{t+1} \sim \mathrm{NID}(\mathbf{0},\ \mathrm{diag}(\sigma_{\varepsilon}^2))$$

$$\mathbf{x}_{t+1} = \mu_{S_{t+1}} + \Psi_{S_{t+1}} E_t[\mathbf{x}_{t+1}] + \mathbf{A}_{S_{t+1}}\mathbf{x}_t + \Gamma_{t+1}\mathbf{f}_{t+1}$$

where $E_t[\mathbf{r}_{t+1}]$ represents the expected excess return vector, $\mathbf{B}_{t+1} \equiv [\beta_{e,t+1}\ \ \beta_{b,t+1}]$ is a $M \times 2$ matrix of stock and bond return factor loadings, $\mathbf{f}_{t+1}$ is an $M \times 1$ vector containing the structural factors, innovations to a set of M MS state variables $\mathbf{x}_{t+1}$ such that $\mathbf{f}_{t+1} \sim N(\mathbf{0},\ \mathrm{diag}(\phi_{t+1}))$, and $\varepsilon_{t+1} \equiv [\varepsilon_{e,t+1}\ \varepsilon_{b,t+1}]'$ represents return shocks not explained by the factors. The MSM is used to accommodate changes in monetary policy and to model heteroskedasticity in the shocks. The time variation in the betas is modeled as $\mathbf{B}_{t+1} = B(\mathscr{F}_t, S_{t+1})$, where S_{t+1} is a discrete variable following an observable (to the investor) homogeneous Markov chain with $K=2$. While a benchmark model forces the betas to be constant, BBI also experiment with a short list of parsimonious models investigating the most likely sources of time variation in betas, for example, with betas that depend on instruments measured at time $t-1$, and with betas that depend on a subset of the MS state variables. In the latter case, BBI preserve the structural interpretation of the implied stock-bond return correlation dynamics by using regime variables exogenously extracted from monetary and business cycle variables without using stock and bond returns.[30] The time variation in the diagonal matrix $\mathrm{diag}(\phi_{t+1})$ is also modeled as depending on a Markov chain, $\mathrm{diag}(\phi_{t+1}) = \mathrm{diag}(\phi_{S_{t+1}})$. Under this model, the conditional correlation between stocks and bonds is:[31]

$$\mathrm{Corr}_t[r_{e,t+1}, r_{b,t+1}] = \frac{\sum_{k=1}^{K} \xi_{t+1|t}^{k} \beta'_{e,t+1} \mathrm{diag}(\phi_{S_{t+1}}) \beta_{b,t+1}}{\sqrt{\sum_{k=1}^{K} \xi_{t+1|t}^{k} \beta'_{e,t+1} \mathrm{diag}(\phi_{S_{t+1}}) \beta_{e,t+1}} \sqrt{\sum_{k=1}^{K} \xi_{t+1|t}^{k} \beta'_{b,t+1} \mathrm{diag}(\phi_{S_{t+1}}) \beta_{b,t+1}}}$$

This decomposition shows the standard effects of a linear factor model: Factors with higher variances have the largest effect on co-movement and if bond and stock betas have the same (opposite) sign, increased factor variances lead to increased (decreased) co-movement. Consequently, to generate substantial variation in correlations, the volatility of the fundamentals must display substantial time variation. Moreover, to generate negative covariances, it must be true that there is at least one factor to which bonds and stocks have opposite exposures, and this factor must display substantial variability compared to other factors.

Using quarterly data 1968–2007, BBI employ a large number of potential economic state variables in their investigation – not only short- and long-term interest rates, inflation, the output gap, and cash flow growth but also a "fundamental" risk aversion measure derived from consumption growth data based on Campbell and Cochrane's (1999) habit formation model and macroeconomic uncertainty measures derived from survey data on inflation and GDP growth expectations, liquidity proxies and the variance premium, a risk-premium proxy representing the difference between the (square of the) VIX (the option-based risk neutral expected conditional variance) and the conditional variance of future stock prices. Given their model, BBI decompose the performance of the factor model in the contributions of the various factors. The key result is that macroeconomic fundamentals contribute little to explaining stock and bond return correlations but that other factors, especially liquidity proxies, play a more important role. The macro factors are still important in fitting bond return volatility, whereas the variance premium is critical in explaining stock return volatility. Interestingly, despite the rich factor structure, BBI report that all models show significant MS in residual volatility. For both inflation and output volatilities, there is a near-permanent switch to the low-volatility regime during the 1990s, which is consistent with the idea of a Great Moderation.

The complexity of BBI's model – characterized by multiple Markov state variables unobservable to the econometrician and by complex latent dynamic processes for state variables and betas – is such that its methodology suffers from a number of limitations. First, BBI follow a two-stage (quasi) ML procedure to estimate their model: In a first stage, they estimate the state variable model; then they estimate the factor model conditional on the economic factor shocks identified in the first step. From an econometric point of view, it would be more efficient to jointly estimate the factor and state variable models. However, an important risk of a one-step estimation procedure is that the parameters of the state variable model may be estimated to help accommodate the conditional stock-bond return correlation, which would make the economic interpretation of the factors problematic. Second, BBI's model is obviously characterized by a problem of parameter proliferation. As a reaction BBI work with "paired-down" versions in which only parameters with t-statistics over 1 are retained. Although this is sensible in the light of a few findings in the forecasting literature (but for simple regression models), more adequate approaches to model simplification may be needed. Third, the baseline specification of BBI's DFM involves five different, two-state first-order Markov state variables, a few of them directly affecting risk exposures and/or risk-factors and others affecting the dynamic process of the factors. However, $2^5 = 32$ and one may argue that – albeit in a highly nonlinear and economically-motivated fashion – BBI eventually work with a MSM with 32 regimes! Here one clearly feels a need of a more systematic treatment (not only within the structure of DFMs) of the issue of how to sensibly go from N MS K-state processes to an encompassing MS process without assuming independence.

MS AND THE PREDICTABILITY DEBATE

In a number of papers reviewed in the third section, the issue of whether MS dynamics may "interfere" with the classical finding (see, e.g., Fama & French, 1989) that simple business-cycle sensitive instruments (such as the dividend yield, the term spread, the default spread, and short-term interest rates) may predict returns or with the result in the exchange rate literature that no time series models can beat a RW in OOS prediction tests, had already surfaced. For instance, while Engel and Hamilton (1990) had suggested that a RW benchmark may be outperformed by MSMs, a number of papers leading up to Guidolin and Timmermann (2006a) had found that MSVAR models imply that the power of standard predictors to forecast subsequent returns may be magnified when compared to simple linear models, although such power may become regime-specific. Moreover, the inclusion of predictors in simple MSMs would not weaken the evidence of multiple states, nor would the presence of such states deny the power of the predictors. This has been interpreted as evidence that MS dynamics is not a short-cut approach to capture complex predictability patterns. Moreover, as recently discussed by Lettau & Van Nieuwerburgh (2007, LvN), MS dynamics (more generally, regimes and/or breaks in the underlying return process) in standard, linear predictive relationships involving financial ratios – such as the price-dividend ratio (PDR) or the earnings-price ratio – may easily rationalize the seemingly incompatible result that linear regressions can accurately forecast returns in-sample, but that they miserably fail in OOS tests (see, e.g., Bossaerts & Hillion, 1999; Goyal & Welch, 2008).[32] In their equity sample, LvN find strong empirical evidence in support of breaks in steady-state relationships and/or of regimes. They then ask how such changes affect the forecasting relationship between returns and lagged price ratios. Standard econometric techniques that assume that the regressor is stationary will lead to biased estimates and incorrect inference. However, since deviations of price ratios from their steady-state values are stationary, it is straightforward to correct for the nonstationarity if the timing and magnitudes of shifts in steady states can be estimated. LvN conduct tests that incorporate such adjustments from the perspective of an econometrician with access to the entire historical sample (in-sample tests), as well as from the perspective of an investor who forecasts returns in real time (OOS tests). For instance, while the raw DPR series is very persistent with the first- and second-order autocorrelations of 0.91 and 0.81 – so that the null hypothesis of a unit root cannot be rejected, according to an augmented Dickey – Fuller test – the adjusted DPR is much less persistent with the first-order autocorrelation drops of 0.77 and 0.61, respectively; the null of a unit root in the adjusted series is thus rejected at the 4% and 1% levels. In-sample results indicate that "adjusted" price ratios have favorable properties compared to unadjusted price ratios. For instance, the slope coefficient in a regression of annual log returns on the lagged log-DPR increases from 0.094 for the unadjusted ratio to 0.455 for the adjusted ratio computed assuming two steady-state shifts; while the statistical significance of the coefficient on the unadjusted DPR is marginal, coefficients on the adjusted DPRs are strongly significant.

In real-time pseudo-OOS experiments, however, the changes in the steady state are not only difficult to detect but also estimated with significant uncertainty, making the in-sample return forecastability hard to exploit. Results from OOS tests in LvN reflect this difficulty. While adjusted price ratios have superior forecasting power relative to their unadjusted counterparts, they do not always outperform the benchmark RW. This is easy to understand because, in real time, an investor faces two challenges: first, she has to estimate the timing of a regime shift; second, if she detects a new regime, she has to estimate the new mean after the shift occurs. Interestingly, LvN report that the estimation of the regime shift dates in real time is not crucial and the resulting prediction errors are smaller than for the RW; however, the estimation of the magnitude of the regime-induced change in the mean DPR entails substantial uncertainty, and is ultimately responsible for the failure of real-time OOS predictions to beat the RW. In particular, using a MSM, LvN show that if the investor did not have to estimate regime-specific means in real time, but only the switching dates, her OOS forecast would improve substantially, and beat the RW. The issues raised by LvN's paper are naturally central to the development of the MS literature. For instance, while the empirical macroeconomics literature has focused on the inability (see, e.g., Clements & Krolzig, 1998; Stock & Watson, 2001) of MSMs to accurately forecast business cycle turning points, LvN surprisingly point out that the biggest obstacle to using MSMs to accurately predict asset returns may lie in the difficulty of estimating shifts in conditional means, and not turning points.[33] Additionally, it remains unclear the exact link between LvN's concept of "adjusting" regressors to take into account regime shifts and the sharper idea that if a parametric model of such shifts – as an MSM is – can be estimated, it should be used in practice to produce such forecasts.

While Lettau & Van Nieuwerburgh (2007) clearly illustrates the role that MS may have in reconciling the predictability "puzzle" – that is, strong in-sample forecastability of asset returns along with low OOS power – a number of other papers have offered examples of how MS may change the common perception that it is very hard to forecast asset returns using the very fundamental variables that in principle should underlie the valuation of assets (see, e.g., James, Koreisha, & Partch, 1985, Rapach & Wohar, 2005). An example of this type of work performed at a multivariate level is Guidolin and Ono (2006) who have investigated the hypothesis of time-varying dynamic linkages across financial markets and the U.S. macroeconomy in a flexible multivariate heteroskedastic MSVAR. Guidolin and Ono's seems to be one of the most ambitious multivariate applications to date, as the MSVAR captures the dynamics of an eight-variable vector that includes stock and bond returns in excess of the T-bill rate, the T-bill yield, typical predictors (such as the default spread between low- and high-grade bond yields and the dividend yield), and three genuine macroeconomic variables, the inflation rate, industrial production growth, and a measure of real money growth. Given their ambitious goal, they use a long monthly data set for the United States, 1926–2004. They find overwhelming evidence of four regimes and of time-varying covariances. The four regimes carry a sensible interpretation as a moderately persistent bull-rebound state, a highly persistent low-volatility state, an expansion high growth state, and a recession-bear state. The last two regimes have low persistence and hence durations limited to 4–5 months. Also in this case, it is possible that a need of four regimes may come from the spurious "amalgamation" of eight variables that may have their own simpler "bull & bear" (or "expansion & recession") regimes. Interestingly, the best in-sample fit to the joint density of the data is provided by a four-state MSIH-VAR(1) model in which the VAR coefficients are restricted to be regime-*independent*. They interpret this as evidence that the dynamic linkages between financial markets and the macroeconomy have been stable over time, which counters a prior of evolving predictability patterns. However, because the intercepts of all the variables in their models are allowed to shift over time, this result is perfectly consistent with LvN's: steady-state relationships are subject to regime shifts, even though this does not directly affect any of the predictive regression coefficients.

MARKOV SWITCHING ARCH

In the fourth section, I have used Baele's (2005) paper to tease the reader into the notion that within the popular family of (G)ARCH models there may be space to model MS dynamics in all or portions of the GARCH coefficients. While in Baele's paper such MS GARCH effects were important to detect potential shifts in dynamic correlation patters, MS GARCH has always been an important general area of research. Although GARCH models driven by normally distributed innovations and their numerous extensions can account for a substantial portion of both the volatility clustering and excess kurtosis found in financial returns, a GARCH-type model has yet to be constructed for which the filtered residuals consistently fail to exhibit clear-cut signs of nonnormality. On the contrary, it appears that the vast majority of GARCH models, when fitted to returns over weekly and shorter horizons, imply heavy-tailed conditional residual distributions. A natural solution has consisted of developing GARCH frameworks that incorporate the original assumption of normal innovations but in which the conditional distribution is a mixture of normals.[34]

As with most good questions, should we ask whether or not MS GARCH models are useful, the most appropriate answer would be that "it depends." Besides the obvious dependence of the usefulness of regime components in ARCH on the specific series under examination, the contribution of MS to the ability of GARCH to fit and forecast the data hinges on two dimensions: First, the frequency of the data and, second, the size N of the data vector. As for the first dimension, it is typical to observe that the higher the frequency, the higher the chances that MS GARCH is needed, with little peril of overfitting the data. As a rule of thumb, all the papers that I have reviewed and that have estimated MSMs on daily data, have specified some form of MS GARCH process, and most papers that have used weekly data have done the same. At a monthly frequency, there is much more uncertainty as to what the right choice may be. For instance, using U.K. equity and bond data, Guidolin and Timmermann (2005) have formally tested for ARCH effects in the residuals of a three-state MSIH model and found that the null of no ARCH cannot be rejected. On U.S. monthly equity data, Guidolin and Timmermann (2007) have reported similar evidence in a four-state model. Their results remind us of our earlier discussion of a potential substitutability between the number of regimes in a MSM and the need of GARCH.

MS GARCH models also have a rich technical dimension that has long attracted the interest of econometricians. Reviewing some of the key papers that have developed this class of MSMs will help me to develop a perspective on these technical issues. Cai (1994) and Hamilton and Susmel (1994) are two contemporaneous papers that have proposed related versions of a simple but already powerful MS ARCH. Cai (1994) develops a MS ARCH model to examine the issue of volatility persistence in monthly excess returns of three-month U.S. T-bills. Cai was concerned that such a high volatility persistence may have been spuriously inflated by the presence of a small number of regimes.[35] Cai proposed to model occasional shifts in the asymptotic, long-run variance of a MS ARCH process. In this case, the conditional variance is no longer determined by an exact linear combination of past squared shocks, as in a standard ARCH. The intercept in the conditional variance is allowed to change in response to occasional discrete shifts. Thus the model is able to retain the volatility-clustering feature of ARCH and, in addition, to capture the discrete shifts in the intercept in the conditional variance that may cause spurious persistence in the process. In the simplest of the two-regime cases explored by Cai (1994), his MS AR(1) ARCH process is:

$$r_t = \mu_0 + \mu_1 S_t + \phi(r_{t-1} - \mu_0 - \mu_1 S_{t-1}) + \varepsilon_t \quad \varepsilon_t = \sqrt{h_t} u_t \quad u_t \sim \text{NID}(0, 1)$$

$$h_t = \omega_0 + \omega_1 S_t + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2, \quad \omega_0, \omega_1, \alpha_i \geq 0$$

where $S_t = 0.1$ follows a first-order, homogeneous, and irreducible two-state Markov chain. The model implies that $S_t = 1$ identifies a high variance state because $\omega_1 \geq 0$. Cai identifies two regime shifts in his 1964–1991 sample, the

1974:2–1974:8 period associated with the oil shocks and the 1979:9–1982:8 period associated with the Federal Reserve's "monetarist" experiment. The variance approached asymptotically in these two episodes is more than 10 times higher than the asymptotic variance of the remainder of the sample. The probability of staying in the low-variance state is 0.99 and the probability of staying in the high-variance state is 0.94; the unconditional probabilities of being in the low-variance and the high-variance states are 0.83 and 0.17. Cai concludes that regime shifts have a great impact on the properties of the data so that earlier empirical results that had adopted an ARCH approach in modeling monthly or low frequency interest rate data may contain severe biases.

A related paper is Hamilton and Susmel (1994) who have proposed a switching-regime ARCH (SWARCH) model in which changes in regimes are captured as changes in the scale of the process

$$r_t = \mu + \sqrt{\delta_0 + \delta_1 S_t}\varepsilon_t \quad \varepsilon_t = \sqrt{h_t}u_t \quad u_t \sim \mathrm{NID}(0,1)$$
$$h_t = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2, \quad \alpha_i \geq 0, S_t = 0, 1, 2$$

so that ε_t follows a standard ARCH(p) process and the MS component concerns the scaling factor $\sqrt{\delta_0 + \delta_1 S_t}$. This is obviously different (and in some sense more powerful) than Cai's MS ARCH where a shift to the volatile regime only affects the unconditional (long-run) variance, while in Hamilton and Susmel's SWARCH also the dynamic process of conditional variance is affected. Hamilton and Susmel's motivation was not directly related to their application but instead more linked to the fact that conventional GARCH models – despite their excellent in-sample fit – often provide worse multiperiod volatility forecasts than constant variance models do. As we have seen, this is a motivation similar to recent work by Lettau and Van Nieuwerburgh (2007) on predictability in regression models, but transposed to explain a differential between in-sample fit and OOS predictive accuracy for variances: in particular, multiperiod GARCH forecasts of volatility tend to be too high (low) just after periods of above (below) normal volatility (see, e.g., Lamoureux & Lastrapes, 1993). With reference to weekly returns of a NYSE value-weighted index, Hamilton and Susmel find that a three-state SWARCH specification offers a better statistical fit to the data, better forecasts, and a better description of the October 1987 crash. Their estimates attribute most of the persistence in the volatility of stock returns to the persistence of the low-, moderate-, and high-volatility regimes, which typically last for several years. The high-volatility regime is associated with economic recessions. In addition, they find that the persistence measure of the ARCH process is much lower that what one would otherwise estimate, similarly to Cai (1994). This contradicts the finding, especially in the aftermath of the October 1987 crash, that the parameter estimates of GARCH models would imply that the conditional variance process were not covariance-stationary.[36]

The reader may have noticed that both Cai (1994) and Hamilton and Susmel (1994) had focused on MS ARCH models. In a way, this is natural because the point of both papers is that the high persistence of volatility often reported in the GARCH literature may have been spuriously inflated by the presence of regimes or breaks. In fact, the reason why Bollerslev (1986) had proposed the GARCH generalization of ARCH was exactly to increase the persistence of the conditional heteroskedastic family within a parsimonious parameterization. However, one may still wonder how we should go about specifying and estimating MS GARCH models. Unfortunately, combining MS with GARCH induces tremendous complications in estimation. As a result of the particular lag structure of a GARCH model – by which all past lags of squared shocks affect conditional variance – the standard equations characterizing the EM algorithm would depend on the entire history of the Markov states through the smoothed probabilities $\Pr(S_T, S_{T-1}, \ldots, S_1|\mathcal{F}_T)$. Because each of the S_t's may take K values, this implies a total of K^T probabilities to be computed and stored. Cai, Hamilton and Susmel concluded that for any data series with a sample size larger than 50, a MS GARCH model would be extremely difficult to estimate. Direct maximum likelihood estimation via a nonlinear filter also turned out to be practically infeasible. Therefore both Cai, Hamilton and Susmel had originally restricted their dynamic lag structure to ARCH models.

A first, important "stab" at the problem of estimating MS GARCH came from Gray (1996), who developed a two-state generalized MS ARCH of the U.S. short-term riskless nominal interest rate (one-month T-bill) that nests many existing interest rate models as special cases. Gray entertains two leading models. The first model is ($S_t = 0.1$)

$$\Delta i_t = (\mu_0 + \mu_1 S_t) + (\phi_0 + \phi_1 S_t) i_{t-1} + (\sigma_0 + \sigma_1 S_t)\sqrt{i_{t-1}}\varepsilon_t \quad \varepsilon_t \sim \text{NID}(0,1) \tag{6}$$

which has finite memory because $Var_{t-1}[\Delta i_t|S_t] = (\sigma_0 + \sigma_1 S_t)^2 i_{t-1}$ depends only on one lag of the short rate. Eq. (6) is a MS version of a standard square root process, see Chan, Karolyi, Longstaff, and Sanders (1992). The second model is a MS GARCH(1,1) in which all GARCH parameters (rather than just an additive or multiplicative scaling parameter) are regime-dependent:

$$\Delta i_t = (\mu_0 + \mu_1 S_t) + (\phi_0 + \phi_1 S_t) i_{t-1} + \sqrt{h_t(S_t)}\varepsilon_t \quad \varepsilon_t \sim \text{NID}(0,1)$$

$$h_t(S_t) = (\omega_0 + \omega_1 S_t) + (\alpha_0 + \alpha_1 S_t)\varepsilon_{t-1}^2 + (\beta_0 + \beta_1 S_t) h_{t-1}(S_{t-1}) \tag{7}$$

$(S_t = 0.1)$ which implies an infinite memory because $\text{Var}_{t-1}[\Delta i_t | S_t] = (\omega_0 + \omega_1 S_t) + (\alpha_0 + \alpha_1 S_t)\varepsilon_{t-1}^2 + (\beta_0 + \beta_1 S_t)\text{Var}_{t-2}[\Delta i_{t-1} | S_{t-1}]$ and this can be solved backwards to show that conditional variance depends on the entire history of shocks to the short-term rate, $\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{t-1}$. Gray tackles the problem of path dependence in MS GARCH adopting an approach that preserves the essential nature of GARCH and yet allows tractable estimation. Under conditional normality, and defining $M(S_{t-1} = i | \mathscr{F}_{t-2})$ to be the conditional mean, the variance of changes in the short rate at time $t-1$ is given by

$$\begin{aligned}\bar{h}_{t-1} = {} & E_{t-2}[(\Delta i_{t-1})^2] - \{E_{t-2}[\Delta i_{t-1}]\}^2 = \Pr(S_{t-1} = 1 | \mathscr{F}_{t-2})[M(S_{t-1} = 1 | \mathscr{F}_{t-2})^2 \\ & + h_{t-1}(S_{t-1} = 1 | \mathscr{F}_{t-2})] + [1 - \Pr(S_{t-1} = 1 | \mathscr{F}_{t-2})][M(S_{t-1} = 0 | \mathscr{F}_{t-2})^2 \\ & + h_{t-1}(S_{t-1} = 0 | \mathscr{F}_{t-2})] + -\{\Pr(S_{t-1} = 1 | \mathscr{F}_{t-2}) M(S_{t-1} = 1 | \mathscr{F}_{t-2}) \\ & - [1 - \Pr(S_{t-1} = 1 | \mathscr{F}_{t-2})] M(S_{t-1} = 0 | \mathscr{F}_{t-2})\}^2\end{aligned}$$

which is not path-dependent and corresponds to a difference of averages across regimes (with probabilities given by filtered probabilities) of the the first and second moments. This value of $\bar{h}_{t-1}$ can be used in the MS GARCH (1,1) specification in Eq. (7) to replace $h_{t-1}(S_{t-1})$.

Using a 1970–1994 weekly sample, Gray shows that his MS GARCH model delivers sensible results, capturing the features of short-term interest rate data better than existing models in terms of both in-sample fit and OOS forecasting performance. Interestingly, a considerable improvement in the statistical fit to the data is already obtained from a simple two-state MSIH with constant variance within regime (i.e., under the restriction that $\sqrt{i_{t-1}}$ does not enter the diffusion coefficient in Eq. (6)) or equivalently setting $\alpha_0 = \alpha_1 = \beta_0 = \beta_1 = 0$ in the GARCH specification. There is persistence in both regimes and the regimes tend to be separated by differential variances, with regime 1 standard deviation being more than four times the standard deviation of regime 2. However the conditional mean parameters are all insignificantly different from zero. In this sense, the short rate essentially follows a RW in all regimes. Yet, the Ljung-Box statistics of the squared standardized residuals are reduced dramatically versus the raw interest rate data: even a simple MSIH can capture much of the time-varying volatility in short-term yields. When a standard, single-state GARCH(1,1) model is estimated, conditional mean parameters remain insignificant and the conditional variance parameters are similar to those typical in the literature, with $\hat{\alpha} + \hat{\beta}$ exceeding one and a LRT unable to reject the hypothesis that the conditional variance follows an IGARCH process. Another LRT rejects the single-regime constant-variance model in favor of the single-regime GARCH model. The single-regime GARCH model, however, does a poor job of modeling the stochastic volatility (SV) in short-term rates, with the Ljung-Box statistics for the squared standardized residuals pointing to the presence of significant serial correlation, and the Jarque-Bera test showing that the standardized residuals are not normally distributed (contrary to model assumptions).

When a MS GARCH(1,1) model is estimated, while none of the conditional mean parameters reaches statistical significance, Gray gets confirmation of the asymmetry across regimes evident in the MSIH. The high-volatility regime is characterized by a higher sensitivity to recent shocks ($\hat{\alpha}_0 > \hat{\beta}_0$) and less persistence ($\hat{\alpha}_0 + \hat{\beta}_0 < \hat{\alpha}_1 + \hat{\beta}_1$) than the low-volatility regime. The effect of individual shocks dies out quickly during periods of very high volatility but has a longer-lasting effect during periods of low volatility. It is this important difference that a single-regime GARCH model is unable to capture. Within each regime, the GARCH processes are stationary ($\hat{\alpha}_k + \hat{\beta}_k < 1$, $k = 0, 1$) and much less persistent than in the single-regime GARCH model. An LRT is constructed to compare the MSIH model with the MS GARCH. The LR statistic is significant at any usual level, indicating that the GARCH effects are important. One final group of models combine both GARCH and square root effects. In such a MS GARCH/square root model, volatility clustering can be caused by three factors. First, the GARCH process in each regime is capable of capturing volatility clustering. Second, if the unconditional variance is higher in one regime than the other, and if regimes are persistent, then periods of high volatility will cluster together during episodes of the high-volatility regime. Third, since volatility depends on the level of interest rates, volatility clustering can emerge in periods of high interest rates, if interest rates are persistent. Gray finds in this case some reversion to a relatively high long-run mean during periods of high interest rates and high volatility, whereas the short rate follows a low variance RW during periods of low and stable interest rates. This separation across regimes of the conditional mean dynamics is similar to Ang and Bekaert's (2002b). The conditional variances appear to separate into a GARCH regime and a "square root" (CIR) regime. In regime 1, both GARCH parameters fail to reach significance, while the CIR (square root) parameter is highly significant. In regime 2, both GARCH parameters are significant and although the CIR parameter is statistically significant its small value renders it economically insignificant. The persistence of an individual shock is much less than is suggested by the single-regime model, with $\hat{\alpha}_k + \hat{\beta}_k < 0.4$ in both regimes.

Dueker (1997) has proposed a different approach to address the issues with the estimation of MS GARCH processes that follows earlier work by Kim (1994) and stresses the importance of allowing the shocks in MS GARCH to follow marginal distributions that are more general than a $N(0,1)$, in particular t-Student shocks. Dueker's work is also important because it investigates the existence of a trade-off between the flexibility allowed by TVTPs taken to be a function of lagged values of the variable(s) of interest in conditionally Gaussian environments, and fat-tailed distributions for the shocks retaining the standard assumption of constant transition probabilities. Dueker's model can be written as

$$
\begin{aligned}
r_t &= (\mu_0 + \mu_1 S_t) + \sqrt{h_t(S_t)}\varepsilon_t \quad \varepsilon_t \sim \text{IID } \mathrm{t}(0, 1, v) \\
h_t(S_t, S_{t-1}) &= (\omega_0 + \omega_1 S_t) + (\alpha_0 + \alpha_1 S_{t-1})\varepsilon_{t-1}^2 + (\beta_0 + \beta_1 S_{t-1})h_{t-1}(S_{t-1}) \qquad (8)
\end{aligned}
$$

where $v > 2$ is the number of t-Student degrees of freedom (an estimable parameter), the GARCH intercept depends on $S_t = 0{,}1$ while α and β depend on $S_{t-1} = 0{,}1$. Kim (1994) had addressed the problem that the t-th observation in a K-regime model implies K^t components of the likelihood function by introducing a collapsing procedure that greatly facilitates computations at the cost of introducing a degree of approximation that – at least based on Dueker's results – does not distort the estimates. The collapsing procedure, when applied to a GARCH process, calls for treating the conditional dispersion, h_t, as a function of the most recent τ values of the state variable S_t. For the filtering to be accurate, Kim noted that, when h is p-order autoregressive, then τ should be at least $p+1$. In the GARCH(1, 1) case, $p = 1$, so we would have to keep track of τ^2 (because the model involves both S_t and S_{t-1}) or four cases, based on the two most recent values of a binary state. Thus, h_t is treated as a function of only S_t and S_{t-1}:

$$h_t^{(i,j)} \equiv h_t(S_t = i, S_{t-1} = j) = \left(\omega_0 + \omega_1 I_{\{S_t=1\}}\right) + \left(\alpha_0 + \alpha_1 I_{\{S_{t-1}=1\}}\right)\varepsilon_{t-1}^2$$
$$+ \left(\beta_0 + \beta_1 I_{\{S_{t-1}=1\}}\right) \times \left\{\Pr(S_{t-1} = 1|S_t = i, \mathscr{F}_{t-1})h_t^{(i,1)}\right.$$
$$\left.+[1 - \Pr(S_{t-1} = 1|S_t = i, \mathscr{F}_{t-1}]h_t^{(i,1)})h_t^{(i,0)}\right\}$$

At this point, $h_{t-1}^{(i,j)}$ simply replaces $h_{t-1}(S_{t-1})$ in Eq. (8). In fact, Dueker also entertains an extension of Hansen's (1994) model in which the Student t degrees-of-freedom parameter, v_t, is allowed to vary over time as a probit-type function of variables dated up to time $t-1$. However, differently from Hansen, Dueker's specification makes v_t follow a first-order, two-state MS process, $v_t = v_0 + v_1 S_t$.[37]

Using daily S&P 500 returns for the period 1982–1991, Dueker reports that a MS GARCH in which v_t follows a MS process produces v_t switching between the values of 2.6 and 8.3. As a result, the fourth moment does not exist in one state, whereas conditional kurtosis is 4.4 in the other state. The weight given to lagged squared residuals in the GARCH process shifts with the state variable between .009 and .027. In this way, shocks drawn from the low degree-of-freedom state do not affect the persistent GARCH dispersion process proportionately. Most importantly, shifts in the degrees-of-freedom parameter bring large discrete shifts in the variance; a shift out of the low degree-of-freedom state causes the variance to decrease by about 68%, holding the dispersion constant. The unconditional probability of being in the low degree-of freedom state is about 10% with a half-life of five trading days. This MS t-GARCH model also suggests that stock returns are negatively skewed because the mean stock return is below normal in the high-volatility state when $S_t = 0$. When a Vlaar-Palm density specification test (see eighth section) is applied, only the MS t-GARCH with switching degrees of freedom is not rejected on an in-sample basis, with a 0.57 p-value; however, it is rejected out of sample. This anticipates to some extent the work on density prediction properties of MSMs that will be reviewed in section "Density Forecasts and Risk Management Applications." As an economic test of MS GARCH models, Dueker uses them to predict the next day's opening level of the VIX volatility index compiled by the CBOE, over the period 1986–1992. In a MSFE metric, he finds that only a MS GARCH in which also v_t follows a MS process predicts the VIX substantially better than a conventional GARCH model, with a notable 14% reduction in MSFE.

Dueker's application to daily stock return data poses a number of interesting questions for subsequent research: the trade-offs between the flexibility in the marginal distribution of return innovations versus the standard flexibility through the selection of K offered by MS mixtures; the choice of optimal approximation/truncation schemes to be applied to the Markov chain dynamics when MS GARCH makes it unfeasible to compute the exact log-likelihood; a desire to push the assessment of OOS performance beyond RMSFE calculations to encompass precise (but necessarily specific) notions of economic value in the perspective of traders and portfolio managers.

Starting in the late 1990s, the debate in the MS GARCH literature has evolved to include a competing family of volatility models, SV. Given their inherent complexity and the substantial payoffs of writing these models in state-space form, which makes them amenable to the application of simulation-based estimation methods, this literature has naturally extended the set of available inferential tools from the classical (ML and EM-based) camp to the Bayesian one. The work of So, Lam, and Li (1998, SLL) is one of the first papers based on Bayesian MCMC methods (see, e.g., Jacquier, Polson, & Rossi, 1994) applied to a MS SV model. Starting from the standard, single-state case (asset returns are assumed to have been demeaned already),

$$r_t = \sqrt{h_t}\varepsilon_t \qquad \ln h_{t+1} = \lambda + \varphi \ln h_t + \eta_t$$

with $|\varphi| < 1$, where ε_t and η_t are IID normal random variables with zero mean and variances 1 and σ^2_η, respectively, r_t is the product of two independent variables, $\sqrt{h_t}$ and ε_t. Asymmetric volatility can be captured by allowing contemporaneous correlation between ε_t and η_t. In terms of state-space representation, the process for $\ln h_{t+1}$ is the transition equation in which $\ln h_{t+1}$ is the state variable, while the measurement equation is simply $\ln r^2_t = \ln h_t + \zeta_t$, where $\zeta_t \equiv \ln \varepsilon^2_t$. Similarly to Hamilton and Susmel's (1994) SWARCH, the parameter determining the *level* of the logarithm of volatility is allowed occasional discrete shifts among K discrete states:

$$r_t = \sqrt{h_t}\varepsilon_t \quad \ln h_{t+1} = \sum_{k=0}^{K-1} \lambda_j I_{\{S_{t+1} \geq k\}} + \varphi \ln h_t + \eta_t \quad |\varphi| < 1, \lambda_j < 0\,(j = 1,\ldots,K-1)$$

which is SLL's MS SV(K) model. Notice that the indicator variable used in the definition of the regime switching intercept has a cumulative nature: for instance, in the case of three regimes (S_t = 0,1,2) we have $\ln h_{t+1} = (\lambda_0 + \lambda_1 + \lambda_2) + \varphi \ln h_t + \eta_t$ in regime 0, $\ln h_{t+1} = (\lambda_0 + \lambda_1) + \varphi \ln h_t + \eta_t$ in regime 1, and $\ln h_{t+1} = \lambda_0 + \varphi \ln h_t + \eta_t$ in regime 2. Because $\lambda_j < 0$ for $j \geq 1$, the higher the index of the state, the lower the value of the intercept in the

SV process. The switching dynamics is governed by a first-order MS process with constant transition probabilities.

SLL is also one of the first papers to show how a rich MSM could be estimated using MCMC. Under a MCMC approach, samples from the joint posterior density of the parameters are obtained through Gibbs sampling. In the MS SV(K) model, let θ be the parameter vector, $\ddot{r}_t$ be the information up to time t, $\ddot{r}_t \equiv [r_1, r_2, \ldots, r_t]'$, and $\mathbf{h}_t = [h_1, \ldots, h_t]'$, $\mathbf{S}_t = [S_1, \ldots, S_t]'$ be the latent variable vectors. As shown by Jacquier et al. (1994), if we augment the parameter vector θ by the latent vectors $\mathbf{h}_T$ and $\mathbf{S}_T$ where T is sample size, the parameter space is enlarged to $[\theta\ \mathbf{h}_T\ \mathbf{S}_T]$. The decomposition of the joint posterior density according to Bayes's theorem is

$$f(\theta, \mathbf{h}_T, \mathbf{S}_T|\ddot{r}_T) \propto f(\ddot{r}_T|\mathbf{h}_T) f(\mathbf{h}_T|\mathbf{S}_T, \theta) f(\mathbf{S}_T|\theta) f(\theta)$$

where

$$f(\ddot{r}_T|\mathbf{h}_T) \propto \prod_{t=1}^{T}\left(\frac{1}{\sqrt{h_t}}\right)\exp\left(-\frac{1}{2}\frac{r_t^2}{h_t}\right)$$

$$\begin{aligned} f(\mathbf{h}_T|\mathbf{S}_T, \theta) &\propto \tfrac{1}{(\sigma_\eta)^T}\sqrt{1-\varphi^2}\left(\textstyle\prod_{t=1}^{T} h_t^{-1}\right) \\ &\exp\left\{-\tfrac{1}{2\sigma_\eta^2}\left[\sum_{t=1}^{T}\left(\ln h_{t+1} - \sum_{k=0}^{K-1}\lambda_k I_{\{S_{t+1}\geq k\}} - \varphi \ln h_t\right)^2 \right.\right. \\ &\left.\left. +(1-\varphi^2)\left(\ln h_1 - \tfrac{1}{1-\varphi}\sum_{k=0}^{K-1}\lambda_k I_{\{S_1 \geq k\}}\right)\right]\right\} \end{aligned}$$

$$f(\mathbf{S}_T|\theta) = \textstyle\prod_{t=2}^{T}\Pr(S_t|S_{t-1})\Pr(S_1 = i) \quad i = 0, 1, \ldots, K-1$$

and $f(\theta)$ is a prior density, assumed to be the product of $(K+1)(K-1)+K+2$ independent priors [i.e., 2 for φ and σ_η, K for the λ_k coefficients, $K(K-1)$ for $\mathbf{P}$, and $K-1$ for the initial state probability, $\Pr(S_1 = i)$]. SLL simulate from $f(\mathbf{S}_T|\theta)$ using the multimove sampler of Carter and Kohn (1994).

Using weekly S&P 500 returns for the sample 1961–1987, SLL find that when $K = 1$, the posterior mean of φ is close to 1 so that the half-life of a shock to volatility has a posterior mean of about 21 weeks; this is the typical conclusion of the GARCH literature. When a three-state MS SV model is considered, the posterior intervals of λ_1 and λ_2 indicate that the two parameters differ from 0. The persistence parameter φ has a posterior mean equal to 0.47, similar to the estimated 0.48 in Hamilton and Susmel (1994). The persistence in volatility is once more considerably lower versus the single-regime model by allowing sudden discrete shifts in the intercept. SLL also uncover an interesting structure for the estimated transition matrix that in some ways echoes the effects later uncovered by other papers on monthly U.S. financial returns. The two diagonal entries p_{11} and p_{22} of the transition matrix display posterior means close to unity, implying that states 1 and 2 tend to be persistent. On average, state 1 and state 2 would last for about 49 and 65 weeks, respectively. Once the high-volatility state is achieved, it has about a 13% chance to switch to the medium-volatility state. Persistence in the high-volatility state is relatively low, the mean duration being less than 18 weeks. Moreover, the posterior means of p_{12} and p_{20} are relatively small. This implies that there is little chance for the switching from state 1 to state 2 and from state 2 to state 0.

A number of empirical researchers (see, e.g., Haas et al., 2004) have argued that despite the advantages of Bayesian inference via MCMC methods such as the Metropolis-Hastings algorithm, given the large sample sizes typically used in high-frequency financial applications and the lack of strong prior information, (quasi) ML estimation may normally be expected to yield results that are very similar to the Bayesian ones. Therefore, we still lack precise information on how much, if anything, MCMC methods based on weak priors may change our understanding of the regime switching dynamics in commonly used series such as S&P 500 daily returns, even though the practical benefits of MCMC in many practical circumstances are hardly debatable (see Hahn et al., 2010). Moreover, despite SLL's compelling application, there are still no papers on multivariate extensions of MS SV methods that could prove very interesting for the debate (see fourth section) on international financial contagion or on dynamic time-varying correlations.

A similar effort, but in a classical estimation framework (quasi-maximum likelihood estimation using the Kalman filter), is illustrated by Hwang, Satchell & Valls-Pereira (2007, HSVP) who propose a family of generalized SV models with MS state equations. HSVP is an interesting paper because it allows us to ask whether and how MS SV may be needed over and beyond more traditional SWARCH and MS GARCH models. HSVP start once more by noting that popular volatility models such as GARCH or SV show extremely high levels of persistence and smooth dynamics in volatility. When volatility is persistent and smooth, estimating or forecasting it becomes easy and estimation errors are small. Such properties of volatility have been assumed in many financial studies, for instance with the implication that a considerable portion of the applied portfolio management literature has focused more on minimizing the estimation errors in expected return than in variances. As already discussed, the possibility remains that high persistence in volatility may actually arise from MS in volatility. Therefore, HSVP generalize So et al.'s (1998) model by allowing regime changes in the level of volatility, the persistence of volatility, and the volatility of volatility. Their key economic insight is that whilst it is an old adage that economists can forecast volatility but not expected returns, their results suggest that economists can forecast neither.

Using S&P500 daily index returns for the period 1994–2004, HSVP show that squared stock returns are better specified with a generalized four-regime MS SV model:

$$\ln r_t^2 = \ln \varepsilon_t^2 + \ln \sigma^2 + h_t \qquad \varepsilon_t \sim \mathrm{NID}(0,1)$$

$$x_t - \mu_{S_t^*} = \varphi_{S_t^*}(x_{t-1} - \mu_{S_t^*}) + \eta_t \qquad \eta_t \sim \mathrm{NID}(0, \sigma^2_{\eta,S_t^*}) \tag{9}$$

where $x_t \equiv h_t + E[\ln \varepsilon_t^2] + \ln \sigma^2 = h_t + \mu$, so that the AR(1) process for the demeaned state x_t is equivalent to $h_t = \varphi h_{t-1} + \eta_t$. Moreover, because of the structure of Eq. (9), S_t^* is defined as a first-order, four-state Markov chain that captures the dynamics of a two-state, second-order Markov chain:[38]

$$S_t^* = \begin{cases} 1 & \text{if } S_t = 1 \text{ and } S_{t-1} = 1 \\ 2 & \text{if } S_t = 1 \text{ and } S_{t-1} = 2 \\ 3 & \text{if } S_t = 2 \text{ and } S_{t-1} = 1 \\ 4 & \text{if } S_t = 2 \text{ and } S_{t-1} = 2 \end{cases}$$

Interestingly, this four-state MS SV nests a number of other models already reviewed. If $\sigma^2_{\eta,1} = \sigma^2_{\eta,2} = 0$, the model becomes a standard MS GARCH; if $\mu_1 = \mu_2$, $\varphi_1 = \varphi_1$, and $\sigma^2_{\eta,1} = \sigma^2_{\eta,2}$, then it is a standard SV; if $\mu_1 = \mu_2$, $\varphi_1 = \varphi_1$, and $\sigma^2_{\eta,1} = \sigma^2_{\eta,2} = 0$, then it is a Gaussian homoskedastic AR(1) model. If either $\sigma^2_{\eta,1}$ or $\sigma^2_{\eta,2}$ are zeros, and the other coefficient is not, then the unobserved variance process consists of a mixture of a SV and GARCH. Finally, when $\varphi_1 = \varphi_1$, and $\sigma^2_{\eta,1} = \sigma^2_{\eta,2}$, we have So et al.'s (1998) model. HSVP confirm that because of the presence of regime shifts in the level of volatility, the latter becomes far less persistent than previously suggested by SV models. More interestingly, the persistence level of volatility also changes over time with high persistence characterizing short spells only. Persistent regimes are more likely to occur when volatility is low, while far less persistence is likely to be observed in high volatility regimes. An LRT comparing SV to MS SV with regime-switching $\mu_{S_t^*}$ gives a statistic of 127, which is highly significant under a $\chi^2_{(3)}$. When persistence is allowed to follow a MS process, one regime shows a low level of persistence whilst the other shows high persistence: they report AR parameters of 0.126 with an unconditional probability of 0.93, and 0.70 with an unconditional probability of 0.07. An LRT comparing a MS SV with regimes in both μ and φ is 119.54, which is again highly significant. The properties of standardized residuals show that allowing for MS in the level of persistence can give a better fit to the data than the SV model does. Finally, HSVP report that changes in regimes do not have memory: regime changes are far more frequent under the generalized MS SV model than those reported by previous studies such as So et al. (1998). So, HSVP's answer to my question on whether or not MS SV may be needed in practice is positive and implies that such MS component may appear not only in the intercept, but also in the persistence and the variance of the SV process, exactly in the same way in which this occurs for the level (mean) of asset returns.

Haas et al. (2004, HMP) have recently returned to the issue of the most efficient approximation to be used in the implementation of univariate MS GARCH. They propose a way to write MS GARCH models that is different from Gray's (1996) and that better fits the standard intuition of volatility persistence in a GARCH framework. Similarly to previous papers, HMP assume that $r_{t+1} = E_t[r_{t+1}] + \varepsilon_{t+1}$, where ε_{t+1} follows a K-component mixture of normals, $\varepsilon_{t+1} \sim NM(\pi_1, \pi_2, \ldots, \pi_K;\ \mu_1, \mu_2, \ldots, \mu_K;\ \sigma^2_{1t+1}, \sigma^2_{2t+1}, \ldots, \sigma^2_{Kt+1})$, with zero unconditional mean,

$$E[\varepsilon_{t+1}] = \sum_{k=1}^{K} \pi_k \mu_k = 0$$

and GARCH(1,1) variances that can be written as[39]

$$\boldsymbol{\sigma}^{(2)}_{t+1} = \omega + \sum_{i=1}^{q} a_i \varepsilon_t^2 + \sum_{j=1}^{p} \mathbf{B}_j \boldsymbol{\sigma}^{(2)}_t$$

where $\boldsymbol{\sigma}^{(2)}_{t+1} \equiv [\sigma^2_{1t+1}\ \sigma^2_{2t+1}\ \ldots\ \sigma^2_{Kt+1}]'$, ω is a $K \times 1$ vector of constants, a_i is a $K \times 1$ vector of coefficients that load the lagged shock ε_t^2 onto the K regime-specific variances $\boldsymbol{\sigma}^{(2)}_{t+1}$, and $\mathbf{B}_j$ is a $K \times K$ matrix that loads past variances in each of the K regimes onto the predicted K regime-specific variances $\boldsymbol{\sigma}^{(2)}_{t+1}$. $E_t[r_{t+1}]$ may simply correspond – especially for daily financial return data – to regime-independent ARMA structures, provided that there is single-state mean dynamics in the mixture components.[40] Differently from the MS GARCH model in Gray (1996), the parameters have a clear interpretation, namely, α_{ik} measures the magnitude of a shock's immediate impact on the next period's σ^2_{kt+1} and the kth row of $\mathbf{B}_j$ reflects the memory in component k's variance in response to the shocks in each of the K components of the mixture. This model is easily generalized to its persistent, Markov chain counterpart in which at each point of time one of the K mixture components generates observation r_{t+1}, where the process that selects the actual component is a (first-order) hidden Markov chain with K-dimensional state space. In this case, the conditional variance of r_{t+1} is:

$$\mathrm{Var}_t[r_{t+1}] = \sum_{k=1}^{K} \Pr(S_{t+1}|S_t, \mathscr{F}_t)(\sigma^2_{kt+1} + \mu^2_{kt+1}) - \left[\sum_{k=1}^{K} \Pr(S_{t+1}|S_t, \mathscr{F}_t)\mu_{kt+1}\right]^2$$

where also the conditional mean has been generalized to follow a MS process.[41] For the empirically relevant GARCH(1,1) case, HMP provide conditions for existence of arbitrary integer moments and analytic expressions of the unconditional skewness, kurtosis, and autocorrelations of the squared process. HMP also entertain the special case of diagonal mixture GARCH models, in which $\mathbf{B}_j$ is diagonal for $j = 1,\ldots, p$ indicating that dynamic conditional variance in each regime only depends on the past of variance within that regime, and partial mixture GARCH models, in which for some $K-G$ regimes out of the K initially specified, all the corresponding elements of a_i ($i = 1,\ldots, q$) and the corresponding rows of $\mathbf{B}_j$ ($j = 1,\ldots, p$) are made of zeroes, indicating that such components of the mixture fail to contain a GARCH component. In this simple case, the conditions for stationarity have standard form but do not apply within each regime but rather to a weighted sum, with the kth weight given by a function of the (ergodic) state probabilities.[42]

HMP use daily NASDAQ returns over a 1972–2001 sample to compare both the in-sample fit and the OOS forecasting performance of mixture GARCH versus simpler MS and MS-GARCH models. On this daily series, they show that even with just two regimes, their mixture GARCH approach can generate a plausible disaggregation of the conditional variance process in which the regime-specific volatility dynamics have a clearly distinct behavior, which is, for example, compatible with the well-known Black's (1976) leverage effect or that generates plausible levels of excess kurtosis and of time-varying skewness without requiring explicit specification of a conditional skewness or kurtosis process (e.g., as in Rockinger & Jondeau, 2002). In particular, within the normal mixture GARCH model class, for a given number of regimes, K, it turns out that the diagonal model is always preferred over the full model when using the BIC criterion. The worst performer is the standard (one-component) normal-GARCH model and this result is robust to extending the single-state GARCH model to include unconditional innovations drawn from either a t-Student or GED. More interestingly, a three-state mixture GARCH(1,1) model with time-varying means minimizes the BIC and passes a wide range of in-sample tests based on density fit.

HOMOGENEOUS VERSUS TIME-VARYING MARKOV CHAIN MODELS

Another issue that has increasingly appeared in the practice of MSMs in empirical finance concerns the modeling Markov chain transition probabilities as depending on (endogenous or exogenous) state variables/factors. On logical grounds, this is no minor enrichment of the baseline MSM presented in the second section: a time-varying transition matrix implies that not only the MSM will capture and potentially predict the instability (regime shifts) in the data, but it will also contain an element of instability, in the form of such regime shifts occurring themselves with probabilities that change as a function of the history of the process (see, e.g., Diebold, Lee, & Weinbach, 1994; Filardo, 1994). Here my reading of the literature is that while MSMs have been originally proposed in their simplest, plain vanilla form in which the transition matrix **P** is constant over time, starting in the mid-1990s it has become increasingly clear that in terms of in-sample fit, the very data may often suggest the opportunity of letting **P** itself be a function of predetermined variables, to capture the idea that the very steady-state of the markets may be subject to shifts. However, to the best of my knowledge, I have been unable to find any systematic comparisons of the predictive (especially, in terms of density forecasts) performance of time homogeneous versus TVTP models. As a matter of fact, it may even sound naive to be surprised by the fact that the data may suggest making **P** itself a time-varying matrix: because the resulting mapping from the history of returns (and/or other state variables) and the density of the data is by construction highly nonlinear, this offers vast opportunities to fit the data. However, this means in no way that the perils of overfitting have been avoided. This is the sense in which systematic OOS tests of predictive power (or better, of economic value) of MSMs with and without TVTPs may be welcome.[43]

As discussed by Guidolin (2011), in many asset pricing applications of MSMs assuming that transition probabilities are time-varying is not only or so much an empirical matter (or a matter of taste in terms of trading off improved in-sample fit for a possible worse OOS performance), but on the contrary this choice has first-order asset pricing implications. It is not clear whether this fact may have created in applied finance researchers an (so far, unfounded) impression that the practice of MSMs implies that transition probabilities must be specified as time-varying as a matter of routine. To try and explain how TVTPs may enter the very fabric of empirical asset pricing research that employs MSMs, let's draw one simple example from Gray (1996). Gray assumes that the switching probabilities may depend on the level of the short rate, for instance, to capture the fact that a switch to the high-volatility regime may be more likely when interest rates are high. Formally:

$$\Pr(S_t = k|\mathscr{F}_{t-1}) = \Phi(c_k + d_k i_{t-1}), \quad k = 0, 1$$

where c_k and d_k are unknown parameters and $\Phi(\cdot)$ is the Normal CDF, which ensures that $0 < \Pr(S_t = k|\mathscr{F}_{t-1}) < 1$. Interestingly, both mean reversion and leptokurtosis in interest rates may then be caused by the switches between regimes, if the TVTPs are correlated with i_{t-1}. To see this, suppose for simplicity that changes in short-term interest rates are parameterized as being $N(\mu_k, \sigma_k^2)$ in regime $k = 0,1$. Mean reversion exists if $\mathrm{Cov}[\Delta i_t, i_{t-1}] < 0$, or

$$\begin{aligned}\mathrm{Cov}[\Delta i_t, i_{t-1}] &= \mathrm{Cov}\left\{\Pr(S_t = 1|\mathscr{F}_{t-1})\mu_1 + [1 - \Pr(S_t = 1|\mathscr{F}_{t-1})]\mu_0, i_{t-1}\right\} \\ &= (\mu_1 - \mu_0)\mathrm{Cov}[\Pr(S_t = 1|\mathscr{F}_{t-1}), i_{t-1}] < 0\end{aligned}$$

Hence, if the regime probability is correlated with the level of interest rates, then switches between regimes may drive the observed mean reversion in the short rate, which is a key-asset pricing phenomenon. A similar argument applies to conditional heteroskedasticity. In his empirical estimates, Gray indeed finds that the probability of staying in the low-variance, near-RW regime decreases as the level of interest rates increases. Conversely, the probability of staying in the high-variance/high mean-reversion regime decreases as the level of interest rates falls. Therefore, when interest rates increase, the probability of staying in or switching to the mean-reverting regime increases. This helps to prevent the interest rate process from wandering off into unreasonable regions, another important payoff of assuming TVTPs: Starting from the RW regime, interest rates are unbounded; however, as interest rates increase, the process is more likely to switch to the regime in which interest rates tend to revert to a long-run mean.

Gray's paper was one of the first contributions to stress that choosing to model transition probabilities as time-varying was not only a statistical choice, but it would and could affect the key financial implications of the exercise. Maheu & McCurdy (2000) have extended Gray's intuition on the importance of the nonlinearities induced by time-varying regime durations to an empirical analysis of high-frequency stock returns and put their emphasis on the fact that MSMs with TVTPs may allow researchers to capture the existence of important nonlinear dynamics in the conditional mean. For example, as a bull market persists, investors could become more optimistic about the future and hence wish to invest more in stocks. This positive feedback means that the probability of switching out of the bull market decreases with its duration, a sort of momentum effect.[44] In addition to duration-dependent hazards, Maheu and McCurdy also use durations as a conditioning variable in both the mean and variance functions; as a result, given persistence in a particular state, the conditional moments change with duration:

$$\begin{aligned} r_t &= (\mu_0 + \mu_1 S_t) + (\psi_0 + \psi_1 S_t) D(S_t) + \sum_{m=1}^{M} \beta_m [r_{t-m} - \mu_0 - \mu_1 S_{t-1} \\ &\quad - (\psi_0 + \psi_1 S_t) D(S_t)] + \varepsilon_t \\ \varepsilon_t &= \sqrt{h_t} u_t \quad u_t \sim \text{NID}(0,1) \quad h_t = (\omega_0 + \omega_1 S_t) + (\delta_0 + \delta_1 S_t) D(S_t) \\ &\quad + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-p}^2 \end{aligned} \tag{10}$$

where $\alpha_i \geq 0$, $S_t = 0, 1$. Conditional on $\{S_{t-1}, S_{t-2}, \ldots, S_1, S_0, \ldots, S_{-\tau+1}\}$, where τ is the memory of the duration dependence process, S_t is assumed to be independent of $\{r_{t-1}, r_{t-2}, \ldots, r_1\}$. While in plain-vanilla MSMs the evolution of S_t is governed by a first-order Markov chain, introducing duration dependence in the model results in a higher-order Markov chain. Define duration as

$$D(S_t) \equiv \begin{cases} D(S_{t-1}) + 1 & S_{t+1} = S_t \\ 1 & S_{t+1} \neq S_t \end{cases}$$

In words, duration is the length of a run of consecutive states. Theoretically, $D(S_t)$ could grow very large. To make estimation feasible, Maheu and MacCurdy keep track of duration up to and including τ, which becomes an estimable parameter subject to $\tau \geq \max\{M, p\} + 1$. At this point, the TVTPs driving the two-state Markov chain are parameterized using a logistic function to ensure that the probabilities are in (0,1). Using i and d to index realizations of states and duration, the transition probabilities are, for $k = 0,1$ and parameters γ_{1k}, γ_{2k}:[45]

$$\Pr(S_t = k|S_{t-1} = k, D(S_{t-1}) = d) = \begin{cases} \frac{\exp(\gamma_{1k}+\gamma_{2k}d)}{1+\exp(\gamma_{1i}+\gamma_{2i}d)} & d \leq \tau \\ \frac{\exp(\gamma_{1k}+\gamma_{2k}\tau)}{1+\exp(\gamma_{1k}+\gamma_{2k}\tau)} & d > \tau \end{cases}$$

That is, duration is allowed to affect the transition probabilities up to τ periods, after which the transition probabilities are constant. Given state k, $\gamma_{2k} < 0$ (> 0) implies that $\Pr(S_t = k|S_{t-1} = k, D(S_{t-1}) = d)$ declines (increases) or, equivalently, that the hazard function $1 - \Pr(S_t = k|S_{t-1} = k, D(S_{t-1}) = d)$ – the conditional probability of switching off state $k = 0,1$, given the duration d – increases (decreases): the longer the market stays in state k, the lower (higher) is the probability of an additional period of stay in that regime. $\gamma_{2k} = 0$ implies no duration dependence, that is, the hazard function is independent of past duration. One key payoff is then that one is allowed to investigate the dynamic behavior of conditional mean and variance *within* each state and not only as determined by switching among states. This may be taken as an appealing alternative to the need to specify and work with multi-state models in which K often needs to be set to equal 3 or even 4 (see, e.g., Guidolin & Timmermann, 2006a, 2006b; Rydèn et al., 1998; Kim, Nelson, & Startz, 1998). This also implies that there is a clear trade-off between specifying simple first-order Markov chain dynamics but resorting to a relatively large value of K and modeling duration dependence for simpler, two-state Markov processes.[46] Because in Eq. (10) duration effects in the conditional mean and variance are measured by the coefficients $(\psi_0 + \psi_1 S_t)$ and $(\delta_0 + \delta_1 S_t)$, this model is also capable of capturing complicated correlation patterns between the conditional mean and variance. Another crucial result in Maheu and McCurdy's paper – that at least to some extent echoes the intuition of some insights in Dueker (1997) – is that their model captures ARCH effects, in the sense that their specification tests reveal that any conditional heteroskedasticity left by a baseline Durland and McCurdy's (1994) logistic parameterization of the duration model is fully explained by the endogenous duration variable in the conditional variance function. Obviously, this further complicates the complex task of selecting the appropriate components to be appended to a simple, plain vanilla two-state MSM – basically, among MS ARCH or GARCH, specifying a relatively large number of regimes, selecting a non-normal, fat-tailed parametric density for return innovations, and a TVTP matrix, possibly with complex duration-dependence effects – to forecast the data, because it shows that a sufficiently flexible TVTP model that interacts with the conditional variance function may relieve the modeler of a need to worry about MS GARCH effects altogether.

Estimates of Maheu and McCurdy's two-state duration-dependent models on monthly, 1834–1995 U.S. stock returns classify the states into bull and bear markets. The bull market displays high returns coupled with low volatility, but the bear market has a low return and high volatility. The memory parameter, τ, of the Markov chain is estimated between 12 and 20, depending on the specific version of Eq. (10) (e.g., with duration effects in the conditional mean only, and with ARCH effects included or not). That is, duration is significant in affecting the transition probabilities for a little over a year. The empirical hazard functions are declining (which implies a negative duration dependence) in both the bull and bear market states. For example, when the economy is in a bull market, the probability of staying in the bull market actually increases with duration. Although the probability of staying in the bear market also increases with duration, the probability of staying in the state is still less than 1/2 until after four consecutive occurrences of the bear regime. That is, the low return, high-volatility state is not persistent until after we have stayed in it for several months. On average, the stock market spends 90% of the time in a bull market and only 10% in a bear market. Conditional mean estimates stress that the best market gains come at the start of a bull market. That is, returns in the bull market state are a decreasing function of duration. Volatility in the bear market state, however, is an increasing function of duration. For instance, during the first period in the high-return state, the conditional return is 2.8%, but if the bull state persists for 16 periods, the conditional return drops to 0.7%. Thus, the bull market delivers decreasing positive returns.[47]

A paper that has extended Gray's (1996) seminal work on MS GARCH time-heterogeneous models for short-term interest rates to a multivariate dimension is Ang and Bekaert (2002b). They provide an analysis of the econometric properties of MSVAR(1) models, both with constant and TVTPs, for short-term interest rates in the United States, Germany, and the

United Kingdom. For instance, in the special case in which only short-term rates are modeled, their baseline (restricted) MSVAR(1) is:

$$\mathbf{y}_{t+1}=\mu_{S_{t+1}}+\mathbf{A}_{S_{t+1}}\mathbf{y}_t+\Sigma_{S_{t+1}}\varepsilon_{t+1}\ \ \varepsilon_{t+1}\sim\text{NID}(\mathbf{0},\mathbf{I}_3),\Sigma_{S_{t+1}}\Sigma'_{S_{t+1}}=\Omega_{S_{t+1}}\Omega_{S_{t+1}}$$

$$=\begin{bmatrix} \left(\sigma^{\text{US}}_{S_{t+1}}+\delta^{\text{US}}_{S_{t+1}}\sqrt{i^{\text{US}}_t}\right)^2 & \rho^{\text{GER}}_{S_{t+1}}\left(\sigma^{\text{US}}_{S_{t+1}}+\delta^{\text{US}}_{S_{t+1}}\sqrt{i^{\text{US}}_t}\right) & \rho^{\text{UK}}_{S_{t+1}}\left(\sigma^{\text{US}}_{S_{t+1}}+\delta^{\text{US}}_{S_{t+1}}\sqrt{i^{\text{US}}_t}\right) \\ \rho^{\text{GER}}_{S_{t+1}}\left(\sigma^{\text{US}}_{S_{t+1}}+\delta^{\text{US}}_{S_{t+1}}\sqrt{i^{\text{US}}_t}\right) & \left(\sigma^{\text{GER}}_{S_{t+1}}+\delta^{\text{GER}}_{S_{t+1}}\sqrt{i^{\text{GER}}_t}\right)^2+\left(\rho^{\text{GER}}_{S_{t+1}}\right)^2 & \rho^{\text{GER}}_{S_{t+1}}\rho^{\text{UK}}_{S_{t+1}} \\ \rho^{\text{UK}}_{S_{t+1}}\left(\sigma^{\text{US}}_{S_{t+1}}+\delta^{\text{US}}_{S_{t+1}}\sqrt{i^{\text{US}}_t}\right) & \rho^{\text{GER}}_{S_{t+1}}\rho^{\text{UK}}_{S_{t+1}} & \left(\sigma^{\text{UK}}_{S_{t+1}}+\delta^{\text{UK}}_{S_{t+1}}\sqrt{i^{\text{UK}}_t}\right)^2+\left(\rho^{\text{UK}}_{S_{t+1}}\right)^2 \end{bmatrix} \tag{11}$$

where $\mathbf{y}_t \equiv i^{\text{US}}_t\, i^{\text{GER}}_t\, i^{\text{UK}}_t]'$ and i^c_t is the short-term rate for country c. When $\delta^{\text{US}}_{S_{t+1}}=\delta^{\text{GER}}_{S_{t+1}}=\delta^{\text{UK}}_{S_{t+1}}=0$, the model becomes homoskedastic within regime, otherwise this is a version of a square root interest rate process with regime shifts, that extends Gray's model.[48] The process S_t follows a Markov chain with K regimes and with transition probabilities that may be logistic functions of lagged endogenous variables,

$$\Pr(S_{t+1}=k|S_t=i,\ddot{\mathbf{y}}_t)=\frac{\exp(\gamma^0_{i,k}+\gamma'_{i,k}\mathbf{y}_t)}{1+\exp(\gamma^0_{i,k}+\gamma'_{i,k}\mathbf{y}_t)}$$

with $\ddot{\mathbf{y}}_t \equiv \mathbf{y}_0\, \mathbf{y}_1, \ldots, \mathbf{y}_t]'$. As already seen, Ang and Bekaert assume the existence of a two-state MS variable in every country driving the entire term structure. Using monthly observations on three-month short rates and five-year rates over a 1972–1996 sample, Ang and Bekaert find that the MSMs all produce one regime with a unit root and lower conditional volatility and a second regime that is stationary with higher conditional volatility. This type of estimation is found in univariate, multicountry, and term spread models. Economically, the first regime corresponds to normal periods where monetary policy smoothing makes interest rates behave like a RW (see Mankiw & Miron, 1986). When extraordinary shocks occur, interest rates are driven up, volatility becomes higher, and interest rates become "more mean reverting," which is captured by the second regime. For some of their models, the null hypothesis of constant probabilities cannot be rejected, which casts some doubts on the true need to specify TVTPs in rich MSMs that also include conditional heteroskedasticity components. In the normal RW regime, U.S. shocks propagate to Germany and the United Kingdom, whereas in the second regime U.S. short rates Granger-cause only U.K. short rates. The regime classification implied by the MSVAR models turns out to be closely related to economic business cycles and the filtered regime probabilities are good short-horizon predictors of the business cycle in the United States. Ang and Bekaert also use their RCM presented in section "Model Selection and Diagnostic Checks" to find that multicountry models produce sharper regime classification for the United Kingdom and Germany at the expense of the United States, for which the RCM declines versus the univariate exercise. This means that a multivariate MSM in some way "sacrifices" its power to fit U.S. data to instrumentally use such U.S. data to improve the fit of British and German interest rates. In particular, there is a large improvement in regime classification for the United Kingdom by adding U.S. information. Including term spread information leads to lower RCM statistics for all countries.

CAN MSMS FORECAST FINANCIAL TIME SERIES?

In applied work, all dynamic time series models are as good as their forecasting performance is. This is especially true in the case of nonlinear models, such as MSMs. In many financial applications, such as the pricing and hedging of complex portfolios of securities and risk management, what matters the most is the evolution of asset prices in the future, not in the past. In general, there is no guarantee that a model that fits historical data well will also perform well OOS due to at least three reasons. First, the extensive search for more complicated models using the same (or similar) data set(s) may suffer from a so-called data-snooping bias, as pointed out by Lo and MacKinlay (1989) and White (2000). A more complicated model can always fit a given data set better than simpler models, but it may overfit some idiosyncratic features of the data without capturing the true DGP. OOS evaluation will alleviate, if not eliminate completely, such data-snooping bias. Second, large, possibly overparameterized models contain an excessive number of parameters and inevitably exhibit excessive sampling variation in parameter estimates, which in turn may adversely affect their OOS performance. Third, a model that fits a historical data set well may not forecast the future well because of unforeseen structural changes in the DGP (e.g., Boero & Marrocu, 2002; Dacco & Satchell, 1999). In the case of MSMs, this may derive from the number of regimes being misspecified.[49] Therefore, from both a practical and a theoretical standpoint, in-sample analysis alone is not adequate, and it is necessary to examine the OOS predictive ability of nonlinear models.

As a matter of fact, the OOS predictive performance – in several dimensions, as we shall see below – of MSMs has been the subject of intense scrutiny. Unfortunately, and especially in financial economics, there does not seem to exist a clear consensus on whether or not – and in the negative, why – MSMs may produce satisfactory in-sample fits and at the same time poor OOS forecasts. It has sometimes been reported that nonlinear models provide a richer understanding of the in-sample dynamics, but that they could also be much less useful for prediction purposes (see, e.g., Brooks, 1997; Clements & Hendry, 1998). Yet, the literature abounds of cases in which MSMs outperform single-state benchmarks in OOS tests. For instance, Engel and Hamilton (1990) report a strong OOS forecasting performance (in terms of MSFE) for their MSIH models, even when compared with RWs that have traditionally held their ground in the empirical exchange rate literature. Another paper that reports encouraging results for MSMs – in a much more complex multivariate framework – is Ang and Bekaert (2002b) who stress that when it comes to MSVAR models, performing extensive forecasting exercises is very important because the estimation may suffer from a peso problem, in that the fraction of observations drawn from one particular regime in a sample may not correspond to the population frequency of that regime. In that case, the estimation is biased. In fact, simple VARMA models may generally constitute good approximations to any covariance stationary process and hence may outperform MSVAR models in small samples if the parameter estimates of the MS models are severely biased and inefficient. In this sense, even if an MSVAR were to correctly capture the unknown DGP, peso problems always make it possible for simpler and misspecified VARMA models to forecast more accurately than MSVARs over short samples, as in Ang and Bekaert (2002b). As we have seen, to overcome these problems, Ang and Bekaert extend their effective sample size through two channels. First, they investigate multi-country systems of interest rates, including U.S., U.K., and German interest rates in their MSVAR and therefore allowing the possibility that short rates in the U.S. Granger-cause rates in other countries (or vice versa) and that Granger causality may be regime dependent. Second, they exploit the information in the term structure by adding term spreads to their models.[50] Ang and Bekaert's discussion of MSVARs is also important because they entertain the idea that multivariate MS modeling exercises may provide more accurate predictions univariate ones. However, as I have illustrated above, deploying multivariate MSMs does open a range of problems on its own, so that the net benefit from such a modeling strategy remains not obvious, at least in general.

Using monthly data on three-month and five-year rates of zero-coupon bonds from the United States, the United Kingdom, and Germany, Ang and

Bekaert find that whereas MSMs do not always outperform single-regime models in terms of in-sample diagnostics, they forecast very well OOS. Moreover, *multivariate* MSMs perform systematically better than univariate models in terms of both RCMs and forecasting: the best forecasting model is invariably a multivariate MS model, incorporating information from the term structure in other countries. This is an important aspect that also explains the relatively disappointing univariate results in Gray (1996). Ang and Bekaert's forecast methodology consists of estimating only using the in-sample period and forecasting without updating the parameters in the OOS. They employ two point statistics for comparison of unconditional forecast errors, the root MSFE, and the mean absolute deviation (MAD). The predicted moments are the level of interest rates, their squares (to mimic volatility forecasting), and the cross-moments involving short-term rates and spreads. Interestingly, MSMs under TVTPs forecast better than constant probability counterparts, even though they perform very poorly at matching sample moments. Unfortunately, Ang and Bekaert's simulations stop short from clarifying the sources of the superior forecasting performance from TVTPs. Finally, their finding that better RCMs tend to correlate with superior OOS accuracy is also interesting as it would seem to imply that better forecasting scores may come from superior ability to predict turning points (regime shifts), even though it must be noted that a low RCM does not necessarily imply that the switches are correctly predicted – only that they are sharp, with a small frequency of periods of uncertainty on the nature of regimes.

Also Guidolin and Ono (2006) have performed systematic multivariate comparisons of single-state VARs versus MSVARs with reference to U.S. equity and bond returns, when standard macroeconomic aggregates are admitted as predictors.[51] They report that a relatively parsimonious four-state MSIH-VAR(1) model can be helpful in forecasting, in the sense that for many relevant variables (especially equity-related ones, stock returns and dividend yields) its recursive OOS performance is superior to a simpler (and nested) VAR(1), as well as competing MSMs and simpler benchmarks, such as a multivariate RW with drift. Interestingly, the performance differential is statistically significant when formal tests of superior predictive accuracy are applied, especially at intermediate and long horizons. This is important because many papers have warned that while in-sample tests generally show that a large number of macroeconomic variables appear to predict future stock returns in linear models, OOS tests of return predictability that protect against data mining typically return disappointing results, (see, e.g., Rapach & Wohar, 2005).

A related paper is Henkel, Martin, and Nardari (2011, HMN) who use Bayesian MSVAR methods to capture time-variation in stock market return predictability from the dividend yield and commonly used term structure variables (the short rate, the slope of the term structure, and the default premium) in the G7 countries. Their key finding is that standard predictors in the finance literature are effective almost exclusively during recessions. For instance, the cumulative proportion of recession months in the U.S. data and the adjusted R^2 from one-month-ahead predictive OLS regression using the contemporaneously available sample, reveal that the adjusted R^2 rises and falls with the proportion of recession months in the sample that had been available to investors in real time. In the United States, over the 1953–2007 period, the average R^2 is about 15% during recessions and less than 1% in expansions; in the G7 no country has R^2 significantly different from zero during expansions, and no individual predictor is more important in expansions than in recessions. The underlying economic mechanism is that risk premia are countercyclical and that the time series behavior of risk premia lets some structure sip into realized excess return predictability. If investors demand higher risk premia in bad times, and volatility is higher in bad times as well, then overall adjustments to discount rates per unit of change in economic state are larger in bad times. HMN also draw intriguing (but ex-post) implications of the cumulative fraction of recessions characterizing the data and the development of the predictability debate in the literature, in which researchers would have discovered and argued in favor of predictability after recession periods and would vice versa emphasized the validity of the RW theory of efficient prices after protracted economic expansions.

Guidolin et al. (2009, GHMO) have recently returned to these issues, but in a simpler, univariate predictive regression framework. They report mixed results for the performance of (simple) MSMs, which depend on the data used in their experiment. GHMO perform a systematic evaluation of whether, when, and where a wide class of nonlinear econometric models that also includes a few plain-vanilla MSMs may provide accurate forecasts of monthly financial returns, in particular stock and bond returns in G7 countries.[52] Although their framework is based on predictive regressions (as in Acharya et al., 2010, as opposed to VARs in which all variables are endogenous and forecastable), similarly to Guidolin and Ono (2006) they also employ as predictors a standard set of macroeconomic variables: changes in short-term interest rates, the term spread, the dividend yield, the inflation rate, the rate of growth of industrial production, the change in the unemployment rate, the rate of growth in oil prices, and the change in a weighted log-effective exchange rate versus the dollar. First, GHMO report that U.S. and U.K. (and, to a lesser extent, Canadian) asset returns appear to be "special" in the sense that good predictive performances for returns in these markets can be obtained only from nonlinear models, especially (but not exclusively) MSMs. Although occasionally also prediction of stock and bond returns from other G7 countries are improved by nonlinear effects (but not of the MSM type), data from France, Germany, and Italy tend to yield good predictions based on simple linear benchmarks, including a homoskedastic RW. Second, the United States and the United Kingdom are the only two countries for which GHMO detect support of statistically significant differences in the recursive OOS performance of different models.[53] Third, although a few patterns could be found, the role of nonlinear models does not depend on any particular part of their sample (1979–2006). Of course, it is tempting to observe that it would be interesting to recheck their empirical findings using data from the 2008–2009 Great Financial Crisis.

GHMO's intuition for the country-specific nature of their findings lies in the heterogeneous pricing frameworks that may generate international stock and bond returns in the presence of international market segmentation: If a researcher estimates simple two-state MSMs, their forecasting performance is likely to get worse as one moves away from the prediction of returns on portfolios that are mostly driven by global factors and toward portfolios that are driven by both global and multiple local factors. The intuition is that a nonlinear model helps forecasting asset returns if it helps identifying and predicting turning points and regime shifts in the process followed by the factors that are compounded into realized asset returns. However, when many alternative factors are priced and most factors are characterized by a different dynamics of regime shifts, if a nonlinear framework is too simple in the sense of falsely imposing $K=2$, then the performance of such model will get increasingly poor as the number of independent, priced factors grows. In GHMO's case, the presumption is that while U.S. returns are driven by one global latent factor, in many other countries a number of regional or local factors are at work that quickly weaken the forecasting performance of a simple two-regime model.[54] To verify their intuition, GHMO perform a small-scale simulation experiment in which three asset markets are described by a factor model. The first market is exclusively driven by a global factor f_t^W which follows a two-state model in which both mean and variance are regime-dependent and driven by the Markov state S_t^W. A second market is driven not only by f_t^W but also by a regional factor f_t^R; also f_t^R follows a two-state MS process with regime-dependent mean and variance; the corresponding Markov chain variable is S_t^R. Finally, the third market is affected both by global and regional factors, and also by a local factor f_t^L driven by another two-state variable, S_t^L. The three Markov state variables S_t^W, S_t^R, and S_t^L are assumed to be independent. Using U.S., U.K., and Italian stock return data to estimate parameters, worldwide and regional bear and bull states are identified; Italian returns are instead mostly driven by a local factor. GHMO simulate 1,000 336-observation long times series of returns from the estimated model and in correspondence to each simulation, they recursively estimate and predict subsequent simulated returns using a two-state MSIH to be compared to the RW and to an AR model. The implied distribution over the 1,000 simulation trials of ratios of forecasting performance measures (e.g., RMSFE) of the RW and the AR model over the MSIH measure, reveal that results for simulated U.S. returns are overwhelmingly favorable to the two-state MSIH, they are mixed in the case of simulated U.K. returns, and they reject the usefulness of the MS in the case of the simulated Italian returns. This perfectly fits the overall intuition: Italian stock returns are generated by a very complicated nonlinear model in which in principle eight different regimes ought to be specified and estimated; as a result a basic two-regime model ends up losing to even the naivest of the prediction benchmarks. Although their intuition may be valuable, GHMO's results would imply a larger scope for MS factor models in applied forecasting in finance, which – as we have seen – has been recently pursued by Baele et al. (2010). More generally, the fact remains that for important stock and bond markets such as the U.S., the British, and to some extent the Canadian one, GHMO overturn a number of presumptions and conjectures in the applied finance literature, showing that nonlinear time series models can produce accurate OOS performances.

A different but equally interesting strand of the literature has instead examined the performance of MSMs at predicting volatilities, which are key inputs in a number of risk management and derivative pricing applications. Although its general findings are not easy to summarize as they seem to strongly depend on the specific asset or sample period under examination, many of the papers reviewed above on MS GARCH generally report OOS prediction results that tend to be favorable to MSMs. For instance, Haas et al. (2004) have reexamined the forecasting performance of their MS-GARCH(1,1) model against a range of alternative benchmarks popular in the literature, including both standard GARCH(1,1) models as well as GARCH models with nonstandard innovations. On their daily NASDAQ return data, and similarly to Klaassen (2002), HMP report that the OOS performance of the mixture GARCH is superior to both simpler MSMs (like a plain MSIH) and MS-GARCH models. The worst performer is the standard (one-component) normal-GARCH model and this result is robust to extending this model to display unconditional innovations drawn from a *t*-Student.

Also Hwang, Satchell, & Valls-Pereira (2007) test forecast accuracy of their MS SV models, after approximating actual variance with alternative nonparametric models. One commonly used method to test the performance of a volatility predictor is to regress realized volatility (RZV) computed using high-frequency (30-minute) squared returns on the daily volatility prediction function that is being tested,

$$\mathrm{RZV}_{t+1} = \beta_0 + \beta_1 V_{t+1} + \zeta_t$$

where V_{t+1} is the volatility forecast. As explained in the realized volatility literature, the choice of RZV as the dependent variable is justified as it provides a robust estimator of latent volatility. If a volatility predictor is unbiased, we expect $\beta_0 = 0$ and $\beta_1 = 1$. HSVP report that none of the volatility predictions tested – which include SV, several forms of MS SV, a simple GARCH(1,1), and S&P 500 index option implied volatilities – satisfy these restrictions, in the sense that chi-square tests of the restrictions lead to rejections. However, the MS SV model yields the smallest χ^2 statistic while SV and GARCH yield the lowest estimates of β_1 and their R^2 coefficients are smaller than those of other models. Moreover, a nonparametric sign test shows that MS SV volatilities are not significantly different from the RZV, while all other volatility predictions are significantly different from RZV in both sign and the Wilcoxon signed rank tests. Unfortunately, HSVP refrain from conducting genuine OOS tests of the ability of MS SV models to forecast subsequent realized volatility.

A topic that has recently drawn the interest of applied researchers is the interaction between forecasting with MSMs and optimal combination schemes of different forecasts. An expanding literature on forecast combinations has found empirically that combining forecasts from different models tends to improve upon the performance of the best individual models (see, e.g., the review in Timmermann, 2006). It is simple to understand why this may occur: In most applications, forecasting models provide at best simple approximations to DGPs that are likely to be far more complicated than assumed by the econometric specification at hand. In this context it is unlikely that an individual model forecast encompasses all other models at all points in time. One type of misspecification that is likely to be particularly relevant empirically is related to time-variations in the conditional relationship between the target variable and the underlying predictor variables. It is quite possible that models that on average (i.e., unconditionally) generate superior predictions in some states of the world may be slower to adapt than other models that generate higher expected loss on average. Similarly, if the parameters of the underlying DGP are subject to transitory shifts, it is natural to expect that there can be gains from combining forecasts from different econometric specifications and that the optimal weights may be both horizon-specific and time-varying as the underlying state changes. Elliott and Timmermann (2004) have argued that a combination scheme that puts more weight on adaptive (e.g., MSMs) models around "breaks" and loads more heavily on stable models away from periods of instability may be expected to give better results *than either of the two alone*. This means that the existence of structural shifts of unknown form/timing makes it likely *not* that some nonlinear framework able to capture such features would manage to produce the best OOS performance, but that on the contrary pooling forecasts may be the winning strategy. For instance, Deutsch, Granger, and Terasvirta (1994) consider switching regressions where the regime is determined by some function of lagged forecast errors. They use rolling regressions to estimate the parameters underlying the combination equations and find that using time-varying combinations leads to MSFE reductions. Elliott and Timmermann (2005) have used a MSM framework where the predicted variable is driven by factors that also affect the prediction signals observed by the decision maker. Both the mean and variance of the factors can vary across different regimes. They find evidence of time-variations in the optimal forecast combination weights under the most common loss functions adopted in the literature, namely quadratic (symmetric), linex and linlin loss. Guidolin and Na (2008) have shown that in a large multivariate forecasting problem similar to Guidolin and Ono's (2006), modeling MS dynamics may lead to superior forecasting performance through forecast combinations and that these gains in prediction accuracy are statistically significant.

With reference to monthly U.S. short-term yields, Guidolin and Timmermann (2009) use MSMs to capture the existence of common, latent factors driving both the stochastic process of a variable of interest (the one-month T-bill spot rate) and of a related market variable (the one-month forward rate implied by the term structure of T-bills) that can be construed as a predictor of the target variable. Even though they admit that a MSM only provides a reduced form for the underlying joint process, its flexibility allows a researcher to capture the two features most likely to explain the success of forecast combinations: the presence of time-variations in the underlying model parameters and the potential difference between the conditional (short-term) and unconditional (long-term) process driving the variables of interest. In particular, in finance it is well known that the EH, when coupled with the standard assumption that markets form rational expectations (RE), imposes tight restrictions on the relationship between spot and forward interest rate in the sense that the forward ought to represent the optimal forecast for future spot rates (apart from terms reflecting a risk premium). In forecasting language, this is a strikingly stringent condition: EH and RE implies that no feasible forecast combination ought to systematically outperform forward rates as predictors of future spot rates. Guidolin and Timmermann examine whether imposing theory-driven restrictions may improve the forecasting performance, when such restrictions are allowed to hold in a time-varying fashion.

Using data on one-month U.S. interest rates over the period 1950–2003, GT find that four regimes are required to provide an adequate description of the joint density of spot and forward rates. The regimes can be interpreted in terms of a low stable interest rate state, one intermediate-rate stable state, one state with high interest rates, and a turbulent regime of high volatility. The choice of the optimal combination weights solves a problem where both the presence of regime shifts and the difference between conditional and unconditional distributions plays a major role. Under a few alternative loss functions, GT report evidence that the optimal combination weights should strongly depend on the underlying regime. The expected loss decline from taking advantage of the opportunity to combine forecasts is high across regimes and reaches levels in excess of 50% (of the highest expected loss) for long forecast horizons, which is a remarkable finding. Moreover, because in their framework the combination weights are time-varying and the optimality of forward rates as predictors of future spot rates implies they should receive a unit weight (with zero weights attached to other variables in the current information set), they conclude that the EH appears to hold only in certain periods (regimes), when forecasts should be entirely based on this benchmark. However, most of the time, deviations from the EH appear and these are exploitable to build better forecasts. Interestingly, this type of findings seems to contradict the classical result that simple averaging outperforms estimated combination weights that reflect the covariance structure of forecast errors and hence efficiently incorporate the available information.

Density Forecasts and Risk Management Applications

As discussed by van Dijk and Franses (2003), most nonlinear models – among them, certainly MSMs – are sufficiently stylized to offer a clean, easy-to-understand trade-off: they are often good frameworks to forecast the predictive density from which future observations will be drawn; however, unless the nonlinear model is unusually good at pinning down the exact (functional) "shape" of the nonlinearity – in the case of MSMs, at filtering and predicting regime shifts – it is often naive to expect much accuracy in point predictions. Because MSMs are simply (Gaussian) mixture distributions with finite memory (as captured by the Markov chain), most of their power may actually reside in the ability to forecast the shape and the dynamics of the tails of the predictive densities, not really the location (mean or median) of such density. In fact, van Dijk and Franses conjecture that as researchers move their focus away from point forecasts and toward emphasizing the importance of accurately predicting the tails, nonlinear models are bound to offer increasingly appreciable performances. This fairly intuitive point has had two implications in the applied finance literature. First, starting from the late 1990s, the focus on testing the predictive performance of MSMs has moved away from point forecasts and toward density forecasts. Second, because most economic-based loss functions depend on a range of features (e.g., moments) of the predictive density – usually not only on its mean (think of simple mean-variance portfolio choice problems to visualize this claim) – the literature has increasingly translated the assessment of the performance of MSMs from a purely statistical domain to an economic one, in which specific (yet, stylized and manageable) decision problems are solved both under single-state and MSMs, to compare in OOS experiments their average payoffs (realized loss). In this subsection I survey some of the work on forecasting density functions. Guidolin (2011) is instead dedicated to understanding the performance of MSMs in asset allocation problems.

In recent developments of time series econometrics, there is a growing interest in OOS probability distribution forecasts and their evaluation, motivated in the context of decision-making under uncertainty. Density forecasts are important not only for statistical evaluation, but also directly relevant to many financial applications. For example, the booming industry of financial risk management is dedicated to providing density forecasts of portfolio returns, and to tracking certain aspects of distributions, such as value at risk (VaR), that quantify the risk exposure of a portfolio. One of the most important issues in density forecasting is how to evaluate the quality of a forecast: Suboptimal density forecasts will have real adverse impact in practice. For example, an excessive forecast of VaR would force risk managers and financial institutions to hold too much capital, imposing an unnecessary cost. Among many others, Granger and Pesaran (2000) have shown that accurate density forecasts are essential for decision making under uncertainty when the forecaster's objective function is asymmetric and the underlying process is non-Gaussian. In particular, if a density forecast coincides with the true conditional density of the DGP, then it will be preferred by all forecast users regardless of their loss functions (e.g., risk preferences). Thus testing the optimality of a forecast boils down to checking whether the density forecast model can capture the true DGP. To my knowledge, one of first papers to provide a comprehensive empirical analysis of the OOS density prediction performance of MSMs in finance is Hong, Li, and Zhao (2004, HLZ). HLZ consider a wide variety of popular short-term (spot) interest rate models – including single-factor diffusions, GARCH, MSMs, and jump-diffusion models – whose density forecast performance was previously largely unknown. Importantly, although some existing studies (e.g., Bali & Wu, 2006; Duffee, 2002; Gray, 1996) have conducted OOS analysis of interest rate models, what distinguishes HLZ's contribution is that they focus on forecasting the conditional density of future interest rates, rather than just their conditional mean.

Evaluating density forecasts, however, is nontrivial because the density function is not observable, even ex post. Relative to point forecasts, there are fewer statistical tools for evaluating density predictions. In a pioneering contribution, Diebold, Gunther, and Tay (1998) have shown how to evaluate density forecasts by examining the probability integral transforms of the residuals with respect to the empirical conditional density, defined below. Such a transformed series is often called the "generalized residuals" of the forecast model. Diebold et al. (1998) prove that these generalized residuals ought to be IID $U[0,1]$ if the density forecast model correctly captures the dynamics of the process: Any departure from IID $U[0,1]$ is evidence of misspecification.[55] Intuitively, the $U[0,1]$ property characterizes the correct specification of the stationary distribution of the data implied by a given model, and the IID property characterizes correct dynamic specification. When these tests are applied to MSMs, the problem is compounded by the fact that, as we have already discussed in section "Model Selection and Diagnostic Checks," in a MS framework it is impossible to directly evaluate the distributional properties of the residuals because, even if the MSM were correctly specified, standardized residuals would not be identically distributed. To circumvent this, it is common (see, e.g., Guidolin & Ono, 2006; Haas et al., 2004) to transform the residuals by computing the corresponding value of the empirical conditional CDF, that is, $\hat{u}_{t+1} \equiv F(\hat{\mathbf{u}}_{t+1}|\mathscr{F}_t)$, where $\hat{\mathbf{u}}_{t+1} \equiv \varepsilon^1_{t+1}\ \varepsilon^2_{t+1}\ \ldots\ \varepsilon^K_{t+1}]'$ collects the K regime-specific residuals from a MSM. To test the IID uniform properties that should hold under the null of a correctly specified model, it has been common to follow Vlaar and Palm (1993) and use the Pearson goodness-of-fit test statistic,

$$X^2 = \sum_{i=1}^{G} \frac{(n_i - n_i^*)^2}{n_i^*} \xrightarrow{D} \chi^2_{G-1}$$

where G is the number of (equally spaced) intervals (or bins) into which the researcher groups the transformed residuals, n_i is the number of observations in interval i, and n_i^* is the expected number of observations under the null hypothesis of uniformity.[56] A drawback of this test is the arbitrariness of the choice of the number of classes, G. In addition, one may wish to test whether the specified distribution captures some specific characteristics of the data such as (conditional) skewness and kurtosis. This can be accomplished by a further transformation of the generalized residuals, namely

$$z_{t+1} = \Phi^{-1}(\hat{u}_{t+1})$$

where Φ is the standard normal CDF, such that IID uniform $\hat{u}$'s should imply that the z's are IID $N(0,1)$. Berkowitz (2001) and Vlaar and Palm (1993) show that inaccuracies in the specified density will be preserved by the transformed z-scores. Thus this transformation allows the use of moment-based normality tests for checking features such as correct specification of conditional skewness and kurtosis. Berkowitz (2001) also uses the z-scores to test the dynamic properties of the conditional distribution. For instance, a LM test can be used to test whether the conditional volatility is successfully captured by a MSM (see, e.g., Haas et al., 2004): the relevant test statistic, $\mathrm{LM}_{\mathrm{ARCH}} = \mathrm{TR}^2$ is approximately χ^2_q distributed, with R^2 denoting the coefficient of determination obtained for a regression $z^2_{t+1} = \alpha_0 + \sum_{i=1}^{q} \alpha_i z^2_{t+1-i} + \varepsilon_{t+1}$. Rather naturally, also LRTs have been employed.

HLZ go beyond the methods in Diebold et al. (1998) and Berkowitz (2001) and use instead Hong's (1999) omnibus evaluation procedure for density forecasts by measuring the departure of the generalized residuals from IID $U[0,1]$. The evaluation statistic provides a metric of the distance between the density forecast model and the true DGP and explicitly addresses the impact of parameter estimation uncertainty. Specifically, Hong's generalized spectral density function becomes a known "flat" function under the null hypothesis of IID $U[0,1]$, whereas if the generalized residuals depend on their past (in any possible form) or they are not $U[0,1]$, the modified generalized spectrum will be nonflat. The evaluation statistic for OOS density forecasts compares a kernel estimator of the modified generalized spectrum with the flat spectrum. HLZ find that although previous studies have shown that simpler models, such as RW, tend to provide better forecasts for the conditional mean of many financial time series, including interest rates (e.g., Duffee, 2002), more sophisticated spot rate models that capture volatility clustering, excess kurtosis, and heavy tails of interest rates yield superior density forecasts. For instance, even a simple GARCH improves the modeling of the dynamics of the conditional variance and kurtosis of the generalized residuals, whereas MS and jumps significantly improve the modeling of the marginal density of interest rates. In particular, using daily 1-month T-bill rates for the period 1961–2000, HLZ report that a two-state first-order MSM with TVTP in which the *constant elasticity of variance* (CEV) coefficient λ_{S_t} is[57]

$$\Delta i_t = \frac{\alpha_{0,S_t}}{i_{t-1}} + \alpha_{1,S_t} + \alpha_{2,S_t} i_{t-1} + \alpha_{3,S_t} i_{t-1}^2 + \sigma_{S_t} i_{t-1}^{\lambda_{S_t}} \sqrt{h_t}\varepsilon_t \quad \varepsilon_t \sim \mathrm{NID}(0,1) \quad (12)$$

depends on the latent Markov state, provides a remarkably good in-sample fit. The parameter estimates of the various MSMs show that the spot rate behaves quite differently between the two regimes. As typical, in models with a linear drift (i.e., $\alpha_{0,s} = \alpha_{2,s} = 0$ for $s = 1,2$), in the first regime the spot rate has a high long-run mean (about 10%) and exhibits strong mean reversion. The spot rate in the second regime behaves almost like a RW, because most drift parameter estimates are close to 0 and insignificant. For models with a nonlinear drift, all drift parameters are insignificant. Volatility in the first regime is much higher, about three times of that in the second regime. The estimates also show that level effect, although significant in both regimes, is much stronger in the second regime in the absence of GARCH. After including GARCH, the elasticity parameter λ_{S_t} estimate becomes insignificant in the second regime, but remains the same in the first regime. Moreover, MS helps capture volatility clustering, as the sum of GARCH parameters measuring persistence strongly declines in MS models versus pure GARCH models, a finding that we have seen to be ubiquitous in the MS GARCH literature. The estimated TVTPs show that the low-volatility regime is much more persistent than the high-volatility regime.

In OOS density tests, HLZ find that their omnibus evaluation statistic overwhelmingly rejects all single-factor diffusion models. Moreover, models that include a drift term (either linear or nonlinear) have much worse OOS performance than those without a drift; this despite the finding that Chan et al. (1992) and Aït-Sahalia's (1996) nonlinear drift models have the highest in-sample likelihood values. This seems to suggest that for the purpose of density prediction, it is more important to model the diffusion function than the drift function. Similar to single-factor diffusion models, GARCH models with a zero drift have much better density forecasts than models with a linear or nonlinear drift. However MS models have better density forecasts than single-factor diffusions and GARCH models. Although modeling the drift does not improve density forecasts for diffusions and GARCH, the best MSMs contain a linear drift in each regime. Even though an MS GARCH model with a linear drift in each regime provides the best OOS forecasts among all of the MSMs, the advantages of MSMs seem to come from better modeling of the marginal density, rather than from the dynamics of generalized residuals. Overall, HLZ OOS analysis demonstrates that more complicated models that incorporate conditional heteroskedasticity and heavy tails of interest rates tend to yield better density forecast. These results are simple to rationalize: As widely documented in the literature, the predictable component in the conditional mean of the interest rate appears insignificant. As a result, RW models tend to outperform more sophisticated models in terms of mean forecast. However, density forecasts include all conditional moments, and as a result, those models that can capture the dynamics of higher-order moments tend to perform better. MSMs appear at the forefront of the wide class of nonlinear models that can produce accurate density predictions. It would be of course interesting to generalize these results beyond short-term interest rates, although my conjecture is that these will likely extend to many asset classes.

A few papers – among others, Guidolin and Timmermann (2006b) and Haas et al. (2004) – have investigated the forecasting performance of MSMs with respect to value-at-risk (VaR). The VaR for period t with shortfall probability ϑ, denoted by $\text{VaR}_t(\vartheta)$, associated with model $\mathscr{M}$ is defined by $\hat{F}^{\mathscr{M}}_{t+1}(\text{VaR}_t(\vartheta)) = \vartheta$, where $\hat{F}^{\mathscr{M}}_{t+1}(\cdot)$ is the return distribution function at time $t+1$ predicted by model $\mathscr{M}$ using information up to time t. For a correctly specified model, we expect $100\vartheta\%$ of the observed return values not to exceed the respective VaR forecast. Thus, it is typical to report and examine the quantity $U_\vartheta = 100 \times \hat{\vartheta}$, where $\hat{\vartheta} \equiv x/T$ is the empirical shortfall probability, T denotes the number of forecasts evaluated, and x is the observed shortfall frequency. If $\hat{\vartheta}$ is less (higher) than ϑ, then model $\mathscr{M}$ tends to overestimate (underestimate) the risk of the position. To formally test whether a model correctly estimates the risk inherent in a given financial position, that is, whether the empirical shortfall probability coincides with the specified shortfall probability, ϑ, one can use the LRT statistic proposed by Kupiec (1995):

$$LR_{VaR} = -2\left\{\ln[\vartheta^{x}(1-\vartheta)^{T-x}] - \ln[\hat{\vartheta}^{x}(1-\hat{\vartheta})^{T-x}]\right\} \sim \chi^2_1$$

Guidolin and Timmermann (2006b) is one example of VaR assessment of MSMs. They consider the term structures – that is, the measures as a function of the forecast horizon – of commonly used risk measures under a range of econometric specifications including multivariate MS, multivariate GARCH-in-mean models with fat tails, a nonparametric historical simulation method, and a component GARCH model fitted to univariate portfolio return series. Because these are all highly nonlinear dynamic specifications that account for time-varying mean, variance and higher order moments, they use simulation methods to compute VaR. GT apply their research design to a strategic asset allocation (SAA) problem, portfolios composed of broad asset classes such as U.S. T-bills, bonds, and stocks. They find evidence of large variations both in levels and shapes of term structures of risk measures across econometric specifications. In a recursive, OOS experiment in which model parameters are re-estimated at monthly intervals, they report that the GARCH(1,1)-in-mean model overestimates VaR for bond portfolios. The GARCH component model is much better at short horizons but suffers from the opposite problem at longer horizons where it underestimates the risk of bond portfolios. An IID Gaussian model tends to underestimate tail risk especially at the longer horizons and for stock portfolios. VaR estimates from a four-state MSM similar to Guidolin and Timmermann's (2006a) are generally good for $\vartheta = 1\%$ but are too low when $\vartheta = 5\%$, particularly at the longest horizons. There is strong evidence of predictability in the hit sequence generated by the component GARCH models, for all horizons and portfolios. At the 1% VaR level, the MSM does not produce any significant rejections of the null of no predictability of the hit indicator, while both the IID Gaussian and bootstrap methods do so in a number of cases.

CONCLUSIONS

What have we learned from this trip through 20 years of advances of applications of MSMs in empirical finance? There are two lessons that one may take away. First, I have accumulated evidence based on several dozen papers that modeling MS dynamics in asset returns makes a difference for our understanding of key financial phenomena, for our ability to forecast them, and for the accuracy of our work in risk management applications. Second, although I have surveyed too many papers to summarize here all of their economic implications, there are a few empirical findings that tend to emerge repeatedly and that represent now a new body of generally available knowledge that all empiricists in finance ought to take into account. For instance, MSMs of short-term interest rates typically isolate one regime in which rates are high, volatile, but quickly mean-reverting toward some unconditional mean level, and a second regime in which rates are instead low, stable, but highly persistent – so that it is occasionally impossible to reject the null of a within-regime unit root. This makes short-term rates very persistent, even though the existence of a stationary regime is often sufficient for them to be globally stationary. Another widespread finding concerns stock returns that are generally found to display one regime in which their risk premia are high but volatility is low, and another regime in which the opposite occurs.[58] Another important lesson is that although a number of empirical studies have found that simpler models, such as RW, tend to provide better forecasts for the conditional mean of many financial time series, more-sophisticated models – among them MSMs – that capture volatility clustering, excess kurtosis, and heavy tails of interest rates yield superior density forecasts. Therefore, and even in a pure forecasting perspective, the selection of the econometric model to be used should also depend on the specific objective (loss function) of the empiricist, as we now understand that heterogeneous objective functions may require methods of different statistical complexity.

NOTES

1. If MSMs are interpreted as tools to optimally recover information on missing regimes, this may appear less surprising because errors in timing the dynamics of the states may still lead to good quality h-step ahead density predictions.
2. NID stands for "normal and identically distributed."
3. It is immaterial whether regimes are labeled starting from 0 and going up to $K-1$ or from 1 to K. Most two-state MS papers prefer the convention $S_t=0,1$ when S_t plays a direct role in writing down the econometric model. The reason is that for $K\geq 3$, S_t cannot, in general, be interpreted as a dummy variable, although it is possible to interpret it as such in special cases.
4. Formally, it is just sufficient for such a condition to be verified in at least one of the K alternative regimes, for covariance stationarity to hold.
5. The assumption of a first-order Markov process is not restrictive, since a higher order Markov chain can always be reparameterized as a higher dimensional first-order Markov chain, that is, substitutability exists between the order of the Markov chain driving S_t and the number of regimes K.

6. An alternative to the assumption of recurring regimes is change point processes, as considered by Chib (1998) and applied in empirical finance by Pastor and Stambaugh (2001) and Pettenuzzo and Timmermann (in press). In this model, the set of regimes expands over time, each regime is unique, and previous regimes are not visited again. This type of model is likely to be a good representation of regime shifts related to technological change and certain types of legislative or political changes that are irreversible. Of course, a combination of recurrent regimes and new regimes is possible. In this survey we will mostly be concerned with recurrent regime models only.

7. This is the sense in which Marron and Wand (1992) emphasize that mixtures of normal distributions provide a flexible family that can be used to approximate many distributions. Mixtures of normals can also be viewed as a nonparametric approach to modeling the return distribution if the number of states, K, is allowed to grow with sample size.

8. MSH(K) models in which $p=0$ and intercepts (often interpreted as expected return vectors) are constant over time – so that only the covariance matrix depends on S_t – have been popular in theoretical asset pricing research, where interesting effects may derive simply from MS in second moments.

9. Krolzig (1997, pp. 38–39) shows that $\mathbf{P}_{(p+1)}$ has structure: $\mathrm{P}_{(p+1)} = (\mathrm{P}' \otimes \mathrm{i}_{K^{p-1}}\mathrm{i}'_{K^{p+1}}) \odot (\mathrm{i}_K \otimes \mathrm{I}_{K^p} \otimes \mathrm{i}'_K)$.

10. Some assumptions have to be imposed to guarantee the (local) identifiability of the parameters under estimation. One possibility relies on the results in Leroux (1992) to show that under the assumption of multivariate Gaussian shocks to the measurement equation, MSIAH models are identifiable up to any arbitrary re-labeling of unobservable states.

11. Notice though that some hypotheses involving elements of $\mathbf{P}$ set to zero cannot be entertained as they fall on the boundaries of the parameter space and may imply a change in the number of actual regimes. However, other hypotheses involving $\mathbf{P}$ can be tested without restrictions, for instance the important statistical hypothesis of independent MS, when $\Pr(S_t = j|S_{t-1} = i) = \Pr(S_t = j) \forall j = 1, \ldots, K$ (i.e., $\mathbf{P}$ has rank 1).

12. This is because $\mathbf{s}_T(\tilde{\theta}) \equiv \sum_{t=1}^{T} \mathbf{h}_t(\tilde{\theta}) = \sum_{t=1}^{T} \left\{ \frac{\partial \operatorname{diag}[\eta_\tau(\theta)]P_\tau(\gamma)}{\partial \theta'} \Big|_{\theta=\tilde{\theta}} \right\}' \hat{\xi}_{t|T} = 0'$, from the MLE first-order conditions.

13. The state process S is characterized by the rate matrix $\Lambda \in \mathbb{R}^{K \times K}$ as follows: setting $\lambda_j = -Q_{jj} = \sum_{l=1, l \neq j}^{K} Q_{jl}$, in state j the waiting time for the next jump is λ_j exponentially distributed and $\Pr(Y_t = l | Y_{t-} = j, Y_t = Y_{t-})$, the probability of jumping to state $l \neq j$ when leaving j, is given by Q_{ji}/λ_j.

14. In what follows we focus for simplicity on MSMAH models because they are logically and computationally more complicated than MSIAH models. However, all of our remarks apply once one replaces $\tilde{\mathbf{u}}_t \equiv \mathbf{y}_t - \mathbf{Y}_t \hat{\mathbf{B}} \hat{\xi}_{t|T}$ with $\tilde{\mathbf{u}}_t \equiv \mathbf{y}_t - \mathbf{Y}_t \hat{\Psi} \hat{\xi}_{t|T}$.

15. With the caveat that the one-step prediction errors do not have a Gaussian distribution and hence the approximate validity of standard tests can only be guessed. For instance, Turner et al. (1989) devise tests in which the filtered probabilities are used as predictors of future variance and test the absence of serial correlation in the resulting regression residuals.

16. Hamilton (1990) has also noted that in small samples the size of these tests may be significantly under-stated.

17. On the opposite, the worst possible MSM has $\hat{\xi}^1_{t|T} = \ldots = \hat{\xi}^k_{t|T} = 1/K$ so that $\sum_{k=1}^{K} \hat{\xi}^k_{t|T} = 1/K^2$ and $\mathrm{RCM}_1 = 100$. Therefore, $\mathrm{RCM}_1 \in [0, 100]$ and lower values are to be preferred to higher ones.

18. An incomplete set of early references that contain univariate MS work, includes Flood and Garber (1983), Driffill (1992), Hamilton (1988), Kaminsky (1993), Kandel and Stambaugh (1990), Schaller and van Norden (1997), Pagan and Schwert (1990), Schwert (1989), and Tucker and Pond (1988).

19. The nuisance parameter problems discussed in section "Model Selection and Diagnostic Checks" prevented Engel and Hamilton to test their two-state MSIH against a single-state random walk model. Therefore, they entertained the null hypothesis of an IID Gaussian mixture (i.e., $p_{11} = 1 - p_{22}$) in which transition probabilities are independent of the current regime. Under this assumption, changes in log-exchange rates are martingale difference sequences.

20. Engel and Hamilton also proceed to test whether markets perceived these swings in real time and therefore investigate the hypothesis of uncovered interest parity, by which the nominal interest differential between two countries forecasts future exchange rate changes or, equivalently, that the three-month forward exchange rate is a rational forecast of the future spot exchange rate. They find no evidence to support this hypothesis in the data.

21. Although the points in RTA were originally put forth with explicit reference to high-frequency (daily) returns – more generally in cases when modeling and forecasting of conditional means cannot be the only or main goal of the empirical researcher – their validity is considerably more general.

22. A hidden Markov chain model (HMCM) is a restricted MSM in which all columns of the transition matrix $\mathbf{P}$ contain identical values for the transition probabilities, such that the probability of switching to a state j at time $t+1$ fails to depend on the state occupied by the system at time t.

23. Only if a researcher is ready to assume independence between the individual states, the transition probability matrix defined on the joint outcome space would simply be the Kronecker product of the individual transition matrices and $K^{\mathrm{ind}} = \sum_{i=1}^{N} K_i(K_i - 1)$, which can be considerably smaller than $\left(\prod_{n=1}^{N} K_n\right)\left(\prod_{n=1}^{N} K_n - 1\right)$.

24. The correlation between large and small firms varies from a high of 0.82 in the crash state to a low of 0.50 in the recovery state. The correlation between large cap and bond returns even changes signs across different regimes and varies from −0.37 to 0.40. The correlation between small stock and bond returns goes from 0.26 to 0.12.

25. This is obviously impossible in two-state models, unless all regimes are weakly persistent, which would cast strong doubts on the appropriateness of the MS framework.

26. Lanne and Saikkonen (2003) show that a mixture of autoregressions with two regimes improves forecasts of weekly U.S. 3-month Treasury bill rates relative to standard AR models.

27. Sola et al. (2002) have considered spillover effects in international equity markets by estimating MSMs to compare effects during tranquil and turbulent periods. Their tests are based on restrictions imposed on a MS constant transition matrix.

28. This GMM-style statistic assigns weights to the deviations between sample versus model-implied unconditional moments that are inversely proportional to the error with which sample moments are estimated.

29. The true model with TVTPs cannot be used because of numerical problems that prevent estimation on short samples.

30. BBI also estimate a structural model for the subset of $\mathbf{x}_{t+1}$ (composed of macroeconomic variables) that extends a standard New-Keynesian three-equation model (e.g., Bekaert, Cho, & Moreno 2010). The monetary policy rule is the typical forward-looking Taylor rule with a positive interest rate smoothing parameter.

31. When betas vary through time, they can also generate sign changes over time. When they are constant, however, the sole driver of time variation in the covariance between stock and bond returns is the heteroskedasticity in the structural factors. The betas simply determine the sign of the covariance.

32. A number of papers (e.g., Nelson & Kim, 1993) have also stressed that correct inferences for regression models based on financial ratios are often problematic because financial ratios are extremely persistent.

33. Guidolin (2011) links this insight to research on MS-based optimal asset allocation strategies. He notes that in many financial applications it may be more crucial to accurately estimate the parameters that enter the conditional mean and variance functions than to predict future regimes.

34. Of course, other competing approaches exist such as the specification of GARCH structures driven by IID innovations from a fat-tailed and, possibly, asymmetric distribution as in Hansen (1994). Harvey and Siddique (1999) and Rockinger and Jondeau (2002) employ autoregressive structures to allow for time variation in skewness and kurtosis.

35. The idea that ignoring either regimes or breaks in the conditional volatility process may spuriously inflate ARCH persistence had already been popularized by Lamoureux and Lastrapes (1990) and Diebold (1986). Lamoureux and Lastrapes (1990) found that allowing for the presence of deterministic shifts in the conditional variance intercept in GARCH produces substantially lowered estimates of the persistence parameters for 30 randomly selected stocks. Diebold (1986) also indicated that although interest rate equations appear to have integrated-variance disturbances, this may be due to the failure to include a monetary regime dummy in the variance intercept.

36. Frequently, statistical tests have been unable to reject the hypothesis that the conditional variance of the short rate follows an integrated GARCH (IGARCH) process (i.e., the null of $\alpha + \beta = 1$ cannot be rejected at standard significance levels). For instance, Engle, Ng, and Rothschild (1990) report a sum of the GARCH coefficients of 1.01 for a portfolio of U.S. Treasuries.

37. Haas et al. (2004) also show how their mixture component GARCH(q, p) model may be generalized to different parametric distributions for their innovations, for instance, K t-densities with degrees-of-freedom parameters that may differ across regimes.

38. As typical of MSMIA models (see section "Estimation and Inference"), this derives from the fact that the process $x_t - \mu_1 = \varphi_1(x_{t-1} - \mu_1) + \eta_t$ ($\eta_t \sim N(0, \sigma^2_{\eta,1})$) differs from $x_t - \mu_1 = \varphi_1(x_{t-1} - \mu_2) + \eta_t(\eta_t \sim N(0, \sigma^2_{\eta,1}))$, etc. Because the four-state

chain derives from expanding the state space for a second-order two-state chain, this implies restrictions on the transition matrix.

39. Vlaar and Palm (1993) have proposed a restricted version of this model in which there are only two regimes, $\sigma^2_{2t+1} = \sigma^2_{1t+1} + \Delta\omega$ and σ^2_{1t+1} follows a standard Gaussian GARCH(1,1).

40. Mixture autoregressive models with different AR structures in each component have also been employed in the literature (e.g., see, Guidolin & Timmermann, 2006a; Lanne & Saikkonen, 2003 Wong & Li, 2000). However, theoretical results on the mixture autoregressive model with GARCH errors are not available and their dynamic properties need to be evaluated by simulation, as in Lanne and Saikkonen (2003).

41. With reference to daily data, HMP argue in favor of using simpler, multi-state normal independent mixtures over MS processes because by allowing for skewness in MS-GARCH models through different regimes, dependent means are inevitably associated with autocorrelation in the raw returns (see Timmermann, 2000). Therefore the zero autocorrelation in daily raw returns paired with dependencies in higher moments – which is one of the most pronounced theoretical properties of GARCH – would be lost under MS-GARCH. Of course, such superiority of simple mixtures over MS-GARCH may be lost at lower frequencies (such as monthly), when returns may be predictable and in fact portions of systematic conditional skewness may even be priced.

42. Consequently, and similarly to Francq & Zakoan (2001) results for Hamilton and Susmel's SWARCH, the process can have finite variance even though some components are not covariance stationary, as long as the corresponding component weights are sufficiently small. This may help to isolate regimes and sample periods over which the persistence in volatility may produce properties consistent with IGARCH, even within the variance process that is globally stationary.

43. Ang and Bekaert (2002b) report results that suggest that TVTPs are useful in forecasting. However, their application is sufficiently complex and rich of important caveats that it is difficult to assess how general this finding may be.

44. Duration dependence has also been related to stochastic, rational bubbles, see, for example, McQueen and Thorley (1994) who have formally proven that a testable implication of rational bubbles is that high returns will exhibit negative duration dependence.

45. Maheu and McCurdy have also investigated a few alternative specifications of these functional forms mapping durations into transition probabilities, for instance

$$\frac{(\gamma_{1k} + \gamma_{2k}d)^2}{1 + (\gamma_{1k} + \gamma_{2k}d)^2}, \quad \frac{\exp(\gamma_{1k} + \gamma_{2k}\ln d)}{1 + \exp(\gamma_{1k} + \gamma_{2k}\ln d)}, \quad \sin(\gamma_{1k} + \gamma_{2k}d)^2$$

In most cases the logistic function gave the best log likelihood value. Moreover, all functional forms suggest negative duration dependence in the transition matrix.

46. Interestingly, filtering and smoothing can be performed by conveniently re-writing the duration dependent MSM as a standard MSM driven by a first-order Markov chain, and expanding the number of regimes as required, see the Appendix in Maheu and McCurdy (2000). This shows that what matters in fitting and forecasting the data is the structure imposed on the regimes, not only (or even primarily) their number.

47. Maheu and McCurdy have also estimated a duration-dependent MSM in which the latent process for the conditional mean is decoupled from the latent process for the variance, that is, in which there are two independent Markov chains, one for the conditional mean and the other for the conditional variance. Both Markov chains allow for duration dependence. They find a similar negative duration-dependence structure in the conditional variance and its associated transition matrix. Differently from earlier papers (McQueen & Thorley, 1994), Maheu and McCurdy's findings do not support a bubble explanation for duration dependence, because there is a negative relationship between duration and conditional mean in the bull market.

48. When the multivariate model is expanded to fit data on both short-term rates and terms spreads, then $y_t \equiv i_t^{US}\ i_t^{GER}\ i_t^{UK}\ ts_t^{US}\ ts_t^{GER}\ ts_t^{UK}]'$ where $ts_t^i \equiv i_t^{5Y,c} - i_t^c$ is the long-short rate spread for country c.

49. Clements, Franses, and Swanson (2004) evaluate these arguments against and in favor of nonlinear models and conclude that, even though the evidence in favor of forecasting using nonlinear models is sparse, there is reason to be optimistic.

50. Under the null of the expectations hypothesis (EH), spreads should forecast future short rates, so the potential for improved estimation and prediction is obvious. Moreover, the spread may be informative about the underlying regimes because it is well known that the spread increases during expansions and correlations between the spread and the short rate are generally higher (less negative) during recessions than in expansions (see, e.g., Estrella & Mishkin, 1997).

51. Using linear regression models, numerous studies have found that selected macroeconomic variables can predict US asset returns, although only in-sample: see, for example, Balvers, Cosimano, and Mcdonald (1990, industrial production and real GNP for stock returns), Campbell (1987, term structure variables for excess stock and bond returns), Cutler, Poterba, and Summers (1989, industrial production for stock returns), Fama and French (1989, 1993, term structure, dividend yield and default spread for stock and bond returns), Fama and Schwert (1977, inflation for real stock returns).

52. Besides MS regressions, GHMO consider threshold predictive regressions, smooth transition predictive regressions, and GARCH-in-mean predictive regressions.

53. This is done using various methods, from naive Diebold and Mariano (1995) tests to more sophisticated van Dijk and Franses (2003) tests that overweight the importance of accurately predicting in the tails, to the new conditional testing framework proposed by Giacomini and White (2006). For most of all these tests, they find that many nonlinear models – among them MS predicting regressions – outperform most other models in OOS experiments.

54. For instance, German, French, and Italian asset returns may be influenced by European (e.g., driven by common monetary policy influences) factors, besides global ones. In the case of Japan, one may think of a geopolitical, regional Asian factor.

55. Diebold et al. (1998) used an intuitive graphical method to separately examine the IID and uniform properties of generalized residuals: the autocorrelations of the generalized residuals to check the IID property and their histograms to check the $U[0,1]$ property. This method is simple and informative about possible sources of suboptimality in density forecasts. Unfortunately, to test the joint hypothesis of IID $U[0,1]$ on the generalized residuals is nontrivial.

56. If the statistic is used to test the in-sample fit, that is, the same observations are used both to estimate the parameter vector and to test the goodness-of-fit, the asymptotic distribution is actually unknown, but is bounded between the $\chi^2_{G-\dim(\theta)-1}$ and χ^2_{G-1} distributions, where $\dim(\theta)$ is the number of estimated parameters.

57. The MS model in Eq. (12) contains a MS non-linearity term in the conditional mean, through the terms $\alpha_{0,S_t}/i_{t-1}$ and $\alpha_{3,S_t}i^2_{t-1}$; moreover, h_t is allowed to follow a standard GARCH(1,1) process in which dependence on lagged variance is established only with reference to unconditional variance, as in Gray (1996). For identification purposes, one needs to set the diffusion constant $\sigma_{S_t} = 1$ in all regimes when h_t follows a GARCH process.

58. However, we now understand that, especially in a MS framework, this does not necessarily contradict the basic tenet that more risk should be compensated by higher risk premia. To the contrary, many papers have used MS dynamics to estimate economically strong and statistically significant volatility feedback effects that are perfectly consistent with a positive relationship between risk and rewards in financial markets, see Guidolin (2011).

REFERENCES

Acharya, V., Amihud, Y., & Bharath, S. (2010). *Liquidity risk of corporate bond returns*. NBER Working Paper No. 16394.

Acharya, V., & Pedersen, L. (2005). Asset pricing with liquidity risk. *Journal of Financial Economics*, *77*, 375–410.

Ahn, D., & Gao, B. (2000). A parametric nonlinear model of term structure dynamics. *Review of Financial Studies*, *12*, 721–762.

Alexander, C., & Kaeck, A. (2008). Regime dependent determinants of credit default swap spreads. *Journal of Banking and Finance*, *32*, 1008–1021.

Amihud, Y., & Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics*, *17*, 223–249.

Ang, A., & Bekaert, G. (2002a). International asset allocation with regime shifts. *Review of Financial Studies*, *15*, 1137–1187.

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

Ang, A., & Chen, J. (2002). Asymmetric correlations of equity portfolios. *Journal of Financial Economics*, *63*, 443–494.

Aït-Sahalia, Y. (1996). Testing continuous-time models of the spot interest rate. *Review of Financial Studies*, *9*, 385–426.

Baele, L. (2005). Volatility spillover effects in European equity markets. *Journal of Financial and Quantitative Analysis*, *40*, 373–401.

Baele, L., Bekaert, G., & Inghelbrecht, K. (2010). The determinants of stock and bond return comovements. *Review of Financial Studies*, *23*, 2374–2428.

Bali, T., & Wu, L. (2006). A comprehensive analysis of the short-term interest-rate dynamics. *Journal of Banking and Finance*, *30*, 1269–1290.

Balvers, R., Cosimano, T., & Mcdonald, B. (1990). Predicting stock returns in an efficient market. *Journal of Finance*, *45*, 1109–1128.

Bansal, R., & Zhou, H. (2002). Term structure of interest rates with regime shifts. *Journal of Finance*, *57*, 1997–2043.

Baum, L., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, *41*, 164–171.

Bekaert, G., Cho, S., & Moreno, A. (2010). New-Keynesian macroeconomics and the term structure. *Journal of Money, Credit, and Banking*, *42*, 33–62.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, *19*, 465–474.

Bertholon, H., Monfort, A., & Pegoraro, F. (2008). Econometric asset pricing modelling. *Journal of Financial Econometrics*, *6*, 407–458.

Black, F. (1976). Studies of stock price volatility changes. *Proceedings of the Meetings of the American Statistical Association*, Business and Economics Statistics Division, pp. 177–181.

Boero, G., & Marrocu, E. (2002). The performance of non-linear exchange rate models: A forecasting comparison. *Journal of Forecasting*, *21*, 513–542.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*, 307–327.

Bossaerts, P., & Hillion, P. (1999). Implementing statistical criteria to select return forecasting models: What do we learn? *Review of Financial Studies*, *12*, 405–428.

Brooks, C. (1997). Linear and non-linear (non-)forecastability of high frequency exchange rates. *Journal of Forecasting*, *16*, 125–145.

Cai, J. (1994). A Markov model of switching-regime ARCH. *Journal of Business and Economic Statistics*, *12*, 309–316.

Campbell, J. (1987). Stock returns and the term structure. *Journal of Financial Economics*, *18*, 373–399.

Campbell, J., & Cochrane, J. (1999). By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy*, *107*, 205–251.

Carter, C., & Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, *81*, 541–553.

Chan, K., Karolyi, G., Longstaff, F., & Sanders, A. (1992). An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance*, *47*, 1209–1227.

Chib, S. (1998). Estimation and comparison of multiple change point models. *Journal of Econometrics*, *86*, 221–241.

Clements, M., Franses, P.-H., & Swanson, N. (2004). Forecasting economic and financial time-series with non-linear models. *International Journal of Forecasting*, *20*, 169–183.

Clements, M., & Hendry, D. (1998). *Forecasting economic time series*. Cambridge: Cambridge University Press.

Clements, M., & Krolzig, H. (1998). A comparison of the forecast performance of Markov-switching and threshold autoregressive models of US GNP. *Econometrics Journal*, *1*, C47–C75.

Cutler, D., Poterba, J., & Summers, L. (1989). What moves stock prices? *Journal of Portfolio Management*, *15*, 4–12.

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

Dai, Q., Singleton, K., & Yang, W. (2007). Are regime shifts priced in U.S. treasury markets? *Review of Financial Studies*, *20*, 1669–1706.

Davidson, R., & MacKinnon, J. (1981). Several tests for model specification in the presence of alternative hypothesis. *Econometrica*, *49*, 781–793.

Davies, R. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, *64*, 247–254.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, *39*, 1–38.

Deutsch, M., Granger, C., & Terasvirta, T. (1994). The combination of forecasts using changing weights. *International Journal of Forecasting*, *10*, 47–57.

Diebold, F. (1986). Modelling the persistence of conditional variances: A comment. *Econometric Reviews*, *5*, 51–56.

Diebold, F., Gunther, T., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, *39*, 863–883.

Diebold, F., Lee, J.-H., & Weinbach, G. (1994). Regime switching with time-varying transition probabilities. In: C. Hargreaves (Ed.), *Nonstationary time series analysis and cointegration* (pp. 283–302). Oxford University Press: Oxford.

Diebold, F., & Mariano, R. (1995). Computing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–263.

Ding, Z., & Granger, C. (1996). Modeling volatility persistence of speculative returns: A new approach. *Journal of Econometrics*, *73*, 135–215.

Doan, T., Littermann, R., & Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, *3*, 1–14.

Driffill, J. (1992). Changes in regime and the term structure: A note. *Journal of Economic Dynamics and Control*, *16*, 165–173.

Driffill, J., & Sola, M. (1998). Intrinsic bubbles and regime-switching. *Journal of Monetary Economics*, *42*, 353–373.

Dueker, M. (1997). Markov switching in GARCH processes and mean-reverting stock market volatility. *Journal of Business and Economic Statistics*, *15*, 26–34.

Duffee, G. (2002). Term premia and interest rate forecasts in affine models. *Journal of Finance*, *57*, 405–443.

Durland, J., & McCurdy, T. (1994). Duration-dependent transitions in a Markov Models of US. GNP growth. *Journal of Business & Economic Statistics*, *12*, 279–288.

Elliott, G., & Timmermann, A. (2004). Optimal forecast combinations under general loss functions and forecast error distributions. *Journal of Econometrics*, *122*, 47–79.

Elliott, G., & Timmermann, A. (2005). Optimal forecast combination weights under regime switching. *International Economic Review*, *46*, 1081–1102.

Engel, C., & Hamilton, J. (1990). Long swings in the dollar: Are they in the data and do markets know it? *American Economic Review*, *80*, 689–713.

Estrella, A., & Mishkin, F. (1997). The predictive power of the term structure of interest rates in Europe and the United States: Implications for the European Central Bank. *European Economic Review*, *41*, 1375–1401.

Fama, E., & French, K. (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, *25*, 23–49.

Fama, E., & French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*, 3–36.

Fama, E., & Schwert, W. (1977). Asset returns and inflation. *Journal of Financial Economics*, *5*, 115–146.

Filardo, A. (1994). Business cycle phases and their transitional dynamics. *Journal of Business and Economic Statistics*, *12*, 299–308.

Flood, R., Garber, P., & Peter. (1983). A Model of stochastic process switching. *Econometrica*, *51*, 537–551.

Forbes, K., & Rigobon, R. (2002). No contagion. Only interdependence: Measuring stock market CoMovement. *Journal of Finance*, *57*, 2223–2261.

Francq, C., & Zakoan, J.-M. (2001). Stationarity of multivariate Markov-switching ARMA models. *Journal of Econometrics*, *102*, 339–364.

Franses, P. H., & van Dijk, D. (2000). *Nonlinear time series models in empirical finance*. Cambridge: Cambridge University Press.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York, NY: Springer.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*, 1545–1578.

Goyal, A., & Welch, I. (2008). A comprehensive look at the empirical performance of the equity premium prediction. *Review of Financial Studies*, *21*, 1455–1508.

Granger, C., & Ding, Z. (1995). Some properties of absolute returns. An alternative measure of risk. *Annales d'Economie et de Statistique*, *40*, 67–91.

Granger, C., King, M.-L., & White, H. (1995). Comments on testing economic theories and the use of model selection criteria. *Journal of Econometrics*, *67*, 173–187.

Granger, C., & Pesaran, H. M. (2000). A decision theoretic approach to forecasting evaluation. In: W. S. Chan, W. K. Li & H. Tong (Eds.), *Statistics and finance: an interface* (pp. 261–278). London: Imperial College.

Gray, S. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, *42*, 27–62.

Guidolin, M. (2011). Markov switching in portfolio choice and asset pricing models: A survey. in missing data methods: time-series methods and applications. *Advances in Econometrics*, *27B*, 87–178.

Guidolin, M., & Gregoriou, G. (Eds.). (2009). *Detecting and exploiting regime switching ARCH dynamics in US stock and bond returns*. Stock market volatility, London:Chapman Hall.

Guidolin, M., Hyde, S., McMillan, D., & Ono, S. (2009). Non-linear predictability in stock and bond returns: When and where is it exploitable? *International Journal of Forecasting*, *25*, 373–399.

Guidolin, M., & Na, F.-Z. (2008). The economic and statistical value of forecast combinations: Regime switching: An application to predictable US returns. In: M. Wohar & D. Rapach (Eds.), *Forecasting in the presence of structural breaks and model uncertainty* (pp. 595–657). Emerald Publishing Ltd & Elsevier Press.

Guidolin, M., & Nicodano, G. (2009). Small caps in international equity portfolios: The effects of variance risk. *Annals of Finance*, *5*, 15–48.

Guidolin, M., & Ono, S. (2006). Are the dynamic linkages between the macroeconomy and asset prices time-varying? *Journal of Economics and Business*, *58*, 480–518.

Guidolin, M., & Timmermann, A. (2005). Economic implications of bull and bear regimes in UK stock and bond returns. *Economic Journal*, *115*, 111–143.

Guidolin, M., & Timmermann, A. (2006a). An econometric model of nonlinear dynamics in the joint distribution of stock and bond returns. *Journal of Applied Econometrics*, *21*, 1–22.

Guidolin, M., & Timmermann, A. (2006b). Term structure of risk under alternative econometric specifications. *Journal of Econometrics*, *131*, 285–308.

Guidolin, M., & Timmermann, A. (2007). Asset allocation under multivariate regime switching. *Journal of Economic Dynamics and Control*, *31*, 3503–3544.

Guidolin, M., & Timmermann, A. (2009). Forecasts of US short-term interest rates: A flexible forecast combination approach. *Journal of Econometrics*, *150*, 297–311.

Haas, M., Mittnik, S., & Paolella, M. (2004). A new approach to Markov-switching GARCH models. *Journal of Financial Econometrics*, *2*, 493–530.

Hahn, M., Früuhwirth-Schnatter, S., & Sass, J. (2010). Markov Chain Monte Carlo methods for parameter estimation in multidimensional continuous time Markov switching models. *Journal of Financial Econometrics*, *8*, 88–121.

Hamilton, J. (1988). Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates. *Journal of Economic Dynamics and Control*, *12*, 385–423.

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

Hamilton, J. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, *45*, 39–70.

Hamilton, J. (1993). Estimation, inference, and forecasting of time series subject to changes in regime. In: G. Maddala, C. Rao & H. Vinod (Eds.), *Handbook of statistics* (vol. 11). Amsterdam: North Holland.

Hamilton, J. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Hamilton, J. (1996). Specification testing in Markov-switching time-series models. *Journal of Econometrics*, *70*, 127–157.

Hamilton, J., & Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, *64*, 307–333.

Hansen, B. (1992). The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP. *Journal of Applied Econometrics*, *7*, S61–S82.

Hansen, B. (1994). Autoregressive conditional density estimation. *International Economic Review*, *35*, 705–730.

Harvey, C., & Siddique, A. (1999). Autoregressive conditional skewness. *Journal of Financial and Quantitative Analysis*, *34*, 465–487.

Henkel, S., Martin, J., & Nardari, F. (2011). Time-varying short-horizon predictability. *Journal of Financial Economics*, *99*, 560–580.

Hong, Y. (1999). Hypothesis testing in time series via the empirical characteristic function: A generalized spectral density approach. *Journal of the American Statistical Association*, *94*, 1201–1220.

Hong, Y., Li, H., & Zhao, F. (2004). Out-of-sample performance of discrete-time spot interest rate models. *Journal of Business and Economic Statistics*, *22*, 457–473.

Hwang, S., Satchell, S., & Valls-Pereira, P. (2007). How persistent is stock return volatility? An answer with Markov regime switching stochastic volatility models. *Journal of Business Finance and Accounting*, *34*, 1002–1024.

Jacquier, E., Johannes, M., & Polson, N. (2007). MCMC maximum likelihood for latent state models. *Journal of Econometrics*, *137*, 615–640.

Jacquier, E., Polson, N., & Rossi, P. (1994). Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics*, *12*, 371–389.

James, C., Koreisha, S., & Partch, M. (1985). A VARMA analysis of casual relations among stock returns, real output, and nominal interest rates. *Journal of Finance*, *40*, 1375–1384.

James, M., Krishnamurthy, V., & Le Gland, F. (1996). Time discretization of continuous-time filters and smoothers for HMM parameter estimation. *IEEE Transactions on Information Theory*, *42*, 593–605.

Kaminsky, G. (1993). Is there a peso problem? Evidence from the dollar/pound exchange rate, 1976–1987. *American Economic Review*, *83*, 450–472.

Kandel, S., & Stambaugh, R. (1990). Expectations and volatility of consumption and asset returns. *Review of Financial Studies*, *3*, 207–232.

Kim, C.-J. (1994). Dynamic linear models with Markov switching. *Journal of Econometrics*, *64*, 1–22.

Kim, C.-J., Morley, J., & Nelson, C. (2005). The structural break in the equity premium. *Journal of Business and Economic Statistics*, *23*, 181–191.

Kim, C. J., & Nelson, C. (1999). *State-space models with regime switching: Classical and gibbs-sampling approaches with applications*. Cambridge, MA: MIT Press.

Kim, C.-J., Nelson, C., & Startz, R. (1998). Testing for mean reversion in heteroskedastic data based on Gibbs-sampling-augmented randomization. *Journal of Empirical Finance*, *5*, 131–154.

Klaassen, F. (2002). Improving GARCH volatility forecasts with regime-switching GARCH. *Empirical Economics*, *27*, 363–394.

Krolzig, H. M. (1997). *Markov-switching vector autoregressions: Modeling, statistical inference, and application to business cycle analysis*. Berlin: Springer-Verlag.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, *3*, 73–84.

Lamoureux, C., & Lastrapes, W. (1990). Persistence in variance, structural change, and the GARCH model. *Journal of Business and Economic Statistics*, *8*, 225–234.

Lamoureux, C., & Lastrapes, W. (1993). Forecasting stock return variance: Toward an understanding of stochastic implied volatilities. *Review of Financial Studies*, *5*, 293–326.

Lanne, M., & Saikkonen, P. (2003). Modeling the U.S. short-term interest rate by mixture autoregressive processes. *Journal of Financial Econometrics*, *1*, 96–125.

Leroux, B. (1992). Maximum likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, *40*, 127–143.

Lettau, M., & Van Nieuwerburgh, S. (2007). Reconciling the return predictability evidence. *Review of Financial Studies*, *21*, 1607–1652.

Liu, H. (2011). Dynamic portfolio choice under ambiguity and regime switching mean returns. *Journal of Economic Dynamics and Control, 35,* 623–640.

Lo, A., & MacKinlay, A. C. (1989). Data-snooping biases in tests of financial asset pricing models. *Review of Financial Studies*, *3*, 175–208.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

Maheu, J., & McCurdy, T. (2000). Identifying bull and bear markets in stock returns. *Journal of Business and Economic Statistics*, *18*, 100–112.

Mankiw, N., & Miron, J. (1986). The changing behavior of the term structure of interest rates. *Quarterly Journal of Economics*, *101*, 211–228.

Marron, J., & Wand, M. (1992). Exact mean integrated squared error. *Annals of Statistics*, *20*, 712–736.

McQueen, G., & Thorley, S. (1994). Bubbles, stock returns, and duration dependence. *Journal of Financial and Quantitative Analysis*, *29*, 379–401.

Nelson, C., & Kim, M. J. (1993). Predictable stock returns: The role of small sample bias. *Journal of Finance*, *43*, 641–661.

Ozoguz, A. (2009). Good times or bad times? Investors' uncertainty and stock returns. *Review of Financial Studies*, *22*, 4377–4422.

Pagan, A., & Schwert, W. (1990). Alternative models for conditional stock volatility. *Journal of Econometrics*, *45*, 267–290.

Pastor, L., & Stambaugh, R. (2001). The equity premium and structural breaks. *Journal of Finance*, *56*, 1207–1245.

Pettenuzzo, D., & Timmermann, A. (in press). Predictability of stock returns and asset allocation under structural breaks. *Journal of Econometrics*.

Psaradakis, Z., & Spagnolo, N. (2003). On the determination of the number of regimes in Markov-switching autoregressive models. *Journal of Time Series Analysis*, *24*, 237–252.

Rapach, D., & Wohar, M. (2005). In-sample vs. out-of-sample tests of stock return predictability in the context of data mining. *Journal of Empirical Finance*, *13*, 231–247.

Rockinger, M., & Jondeau, E. (2002). Entropy densities with an application to autoregressive conditional skewness and kurtosis. *Journal of Econometrics*, *106*, 119–142.

Rydèn, T., Teräsvirta, T., & Åsbrink, S. (1998). Stylized facts of daily returns series and the hidden Markov model. *Journal of Applied Econometrics*, *13*, 217–244.

Schaller, H., & van Norden, S. (1997). Regime switching in stock market returns. *Applied Financial Economics*, *7*, 177–191.

Schwert, G. (1989). Why does stock market volatility change over time? *Journal of Finance*, *44*, 1115–1153.

Sin, C. Y., & White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics*, *71*, 207–225.

So, M., Lam, K. P., & Li, W. K. (1998). A stochastic volatility model with Markov switching. *Journal of Business and Economic Statistics*, *16*, 244–253.

Sola, M., Spagnolo, F., & Spagnolo, N. (2002). A test for volatility spillovers. *Economics Letters*, *76*, 77–84.

Stock, J., & Watson, M. (2001). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In: R. Engle & H. White (Eds.), *Cointegration, causality, and forecasting: A festschrift in honour of clive granger* (pp. 1–44). Oxford: Oxford University Press.

Timmermann, A. (2000). Moments of Markov switching models. *Journal of Econometrics*, *96*, 75–111.

Timmermann, A. (2006). Forecast combinations. In: G. Elliott, C. Granger & A. Timmermann (Eds.), *Handbook of economic forecasting* (pp. 135–196). Elsevier Press: Amsterdam.

Tucker, A., & Pond, L. (1988). The probability distribution of foreign exchange price changes: Tests of candidate processes. *Review of Economics and Statistics*, *70*, 638–647.

Turner, C., Startz, R., & Nelson, C. (1989). A Markov model of heteroskedasticity, risk, and learning in the stock market. *Journal of Financial Economics*, *25*, 3–22.

van Dijk, D., & Franses, P. H. (2003). Selecting a nonlinear time series model using weighted tests of equal forecast accuracy. *Oxford Bulletin of Economics and Statistics*, *65*, 727–744.

Vlaar, P., & Palm, F. (1993). The message in weekly exchange rates in the European monetary system: Mean reversion, conditional heteroscedasticity, and jumps. *Journal of Business and Economic Statistics*, *11*, 351–360.

White, H. (2000). A reality check for data snooping. *Econometrica*, *69*, 1097–1127.

Wong, C., & Li, W. K. (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society*, *62*, 95–115.