MARKOV SWITCHING IN PORTFOLIO CHOICE AND ASSET PRICING MODELS: A SURVEY

Massimo Guidolin

ABSTRACT

I survey applications of Markov switching models to the asset pricing and portfolio choice literatures. In particular, I discuss the potential that Markov switching models have to fit financial time series and at the same time provide powerful tools to test hypotheses formulated in the light of financial theories, and to generate positive economic value, as measured by risk-adjusted performances, in dynamic asset allocation applications. The chapter also reviews the role of Markov switching dynamics in modern asset pricing models in which the no-arbitrage principle is used to characterize the properties of the fundamental pricing measure in the presence of regimes.

Keywords: Markov switching; regimes; risk-return trade-off; volatility feedback; no arbitrage pricing; price of regime risk

JEL classification: G00; C00

Missing Data Methods: Time-series Methods and Applications
Advances in Econometrics, Volume 27B, 87–178

ISSN: 0731-9053/doi:10.1108/S0731-9053(2011)000027B005

INTRODUCTION

In the late 1980s, the landscape of time series econometric models has witnessed the appearance of a new and powerful set of tools to model the presence of instability in dynamic relationships: Markov switching models (MSMs). In this survey, I review a number of chapters that have applied MSMs to two key subfields of modern financial economics, asset allocation models and asset pricing. In a companion chapter, Guidolin (2011), I have surveyed how MSMs have changed our ability to model and forecast financial phenomena relative to simpler, single-state benchmarks, such as simple regression models or vector autoregressions (VARs). This chapter is instead devoted to understanding whether and how MSMs may have affected the ability of financial economists to advise – at the normative level – and better understand – at a positive level – how optimal portfolio choice is and should be dealt with in the presence of regimes in econometric predictive relationships. Moreover, I critically discuss a relatively recent body of research work that has either used MSMs to capture key asset pricing phenomena, for instance the complex and unstable relationships between average excess returns and risk (variance), or that has directly captured the existence of regimes in the fundamental pricing measure, the stochastic discount factor (SDF).

Let me immediately emphasize that there are way to many interesting papers for them to be summarized in one simple, encompassing framework. Therefore I will proceed by organizing my survey of a few selected contributions around five key questions: whether, why, and how MSMs may affect our ability to optimally diversify wealth across alternative portfolios and asset classes, when our goal is to maximize expected future wealth or utility from real consumption streams; whether MSMs may be consistent with the existence of an inverse (or at least nonmonotonic) risk-return trade-off; whether MSMs may change our assessment of the presence of (rational) bubbles in asset prices; whether Markov switching (MS) in fundamentals is responsible for the evidence of MS in asset returns; how can we incorporate MS dynamics in the SDF and what are the effects of such a choice for no-arbitrage asset prices. Indirectly, these questions relate to a more general concern: whether MS papers in the financial economics literature have been concerned more with the economic value – that is, the ability to support financial decisions – or with the statistical predictive performance of MSMs, even conditioning on the fact that the econometric methods are applied to financial data. Interestingly, I have noticed that over time the focus seems to have been shifting toward the first type of issues, because while in the 1990s it had been typical to simply ask whether and how MSMs could fit the data

and forecast them, a rising literature on MS portfolio choice that has sprang off the seminal paper by Ang and Bekaert (2002a) has asked whether (and by how much) MSMs can be the basis for value-enhancing portfolio management. Additionally, an increasing number of papers on the risk management applications of MSMs has pushed through the pages of journals. As third section will show, starting in mid-1990s it is possible to find an impressive number of papers that have rooted MSMs deep in otherwise standard asset pricing frameworks.

The usual caveats and disclaimers apply to this survey. This is just one review of a portion of existing papers that have used MS methods in the portfolio choice and asset pricing literatures (see, e.g., Ang & Timmermann, 2011). There is no assumption that my review may be complete in terms of discussing (or even just 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 the role that MS may play in the SDF and to the issue of whether MS risk should or could be priced in the absence of arbitrage. However, this is also where I have found the most obvious samples 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 providing an introduction to the econometrics or statistical theory of MSMs. Even though the companion paper, Guidolin (2011), provides a primer to what MSMs are and how they can be used, any reader interested in acquiring the basic tools to specify and estimate MSMs are invited to consult the excellent textbooks by – also here, among many others – Franses and van Dijk (2000), Frühwirth-Schnatter (2006), Kim and Nelson (1999), and Krolzig (1997).

Finally, I have made a conscious choice to limit my efforts to MSMs only and leave aside the equally important family of *threshold models*. To see the difference, let me borrow the general representation of a univariate (for simplicity) two-state smooth transition autoregressive (STAR) model from van Dick, Terasvirta, and Franses (2000):

$$r_t = \left(\mu_1 + \sum_{j=1}^{p} \phi_{j1} r_{t-j}\right)[1 - G(S_t; \theta, c)] + \left(\mu_2 + \sum_{j=1}^{p} \phi_{j2} r_{t-j}\right) G(S_t; \theta, c) + \sigma\varepsilon_t \quad \varepsilon_t \sim \text{NID}(0, 1)$$

where the transition function $G(S_t;\theta,c)$ is bounded between 0 to 1 and S_t may either correspond to a lagged value of returns, to some exogenous variable (z_t), or a combination of the two ($S_t = h(r_{t-1},z_t)$). In the STAR literature one of the typical choices of the transition function $G(S_t;\theta,c)$ is the

logistic function, $G(S_t;\theta,c) = [1+\exp(-\theta(S_t-c))]^{-1}$. This important case illustrates the fact that when $\theta \to 0$ the STAR model becomes a linear model because $G(S_t;0,c) = 1/2$; when $\theta \to \infty$ the change of $G(S_t;\theta,c)$ from 0 to 1 becomes almost instantaneous at $S_t = c$ and, consequently, the logistic function approaches the indicator function $I(S_t > c)$ so that the STAR model nests a two-regime threshold autoregressive model; when $\theta \to \infty$ and S_t is an exogenous two-state Markov chain that takes one value below c and another above c, then the logistic STAR becomes a two-regime MSM. Therefore this chapter may seem to focus on the rather special case of logistic STARs in which θ is large and S_t follows a Markov chain. Yet, as I shall endeavor to show, even such a peculiar combination has drawn the attention of hundreds of papers.

MARKOV SWITCHING PORTFOLIO SELECTION

As discussed in Guidolin (2011), 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 because there is no guarantee that a model that fits historical data well will also perform well out-of-sample (OOS) due to at least three reasons. First, the extensive search for more complicated models using the same (or similar) dataset(s) may suffer from a so-called "data-snooping bias," as pointed out by Lo and MacKinlay (1989). A more complicated model can always fit a given dataset 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 forecast performance. Third, a model that fits a historical dataset well may not forecast the future well because of unforeseen structural changes in the DGP (e.g., see Boero & Marrocu, 2002). 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 any nonlinear model.

In particular, as recently emphasized by van Dijk and Franses (2003), because MSMs are simply (Gaussian) mixture distributions with finite memory, 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 overweighting the importance of accurately predicting in the tails, then nonlinear models are bound to offer increasingly beneficial 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 (see Guidolin, 2011, for a survey of relevant papers and methods). Second, because most economic-based loss functions depend on a range of features (e.g., moments) of the predictive density – usually not only from its mean – 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). One area of empirical finance in which researchers have been routinely computing and assessing the (in- as well as out-of-sample) density prediction performance of alternative models of financial returns is the portfolio choice literature, based on the specification and estimation of dynamic econometric models.

The structure of chapters in this literature is rather simple: if a given time series model yields a forecast $f(\mathrm{r}_{t+1}|\mathscr{F}_t;\hat{\theta}_t)$ for the conditional density of asset returns, the implications and performance of the model may be evaluated by endowing some fictional asset manager (or household) with $f(\mathrm{r}_{t+1}|\mathscr{F}_t;\hat{\theta}_t)$ and allowing her to maximize (minimize) the conditional expectation of some objective (loss) function $E_t[L_{t+1}(\omega_t, C_t)]$ – where ω_t is a vector of portfolio weights and C_t represents consumption – where the conditional expectation is computed with respect to $f(\mathrm{r}_{t+1}|\mathscr{F}_t;\hat{\theta}_t)$. From basic principles, unless $L_{t+1}(\omega_t, C_t)$ is a linear function of wealth and hence returns – that is, $L_{t+1}(\omega_t, C_t) = (W_t - C_t)\omega_t'\mathrm{r}_{t+1}$ – the expectation $E_t[L_{t+1}(\omega_t, C_t)]$ will be a nonlinear function of r_{t+1} and such it may involve *all* the moments of r_{t+1}, hence the entire conditional density $f(\mathrm{r}_{t+1}|\mathscr{F}_t;\hat{\theta}_t)$, and in particular not only $E_t[\mathrm{r}_{t+1}]$. This shows that dynamic asset allocation applications are density forecasting exercises performed under a very specific (realistic) loss function that in general will not have the properties necessary and sufficient for the exercise to boil down to a simple point (conditional mean) prediction. For instance, in a simple mean-variance objective function,

$$L_{t+1}(\omega_t, C_t) = E_t[W_{t+1}] - \gamma \mathrm{Var}_t[W_{t+1}]$$

which may help understanding that $E_t[L_{t+1}(\omega_t, C_t)] = (W_t - C_t)\omega_t' E_t[\mathrm{r}_{t+1}] - \gamma^2(W_t - C_t)^2\omega_t'\mathrm{Var}_t[r_{t+1}]\omega_t$, so that correctly forecasting r_{t+1} is not equivalent to only forecasting r_{t+1} or W_{t+1}, because $\mathrm{Var}_t[W_{t+1}]$ and

therefore $\text{Var}_t[r_{t+1}]$ do play a crucial role. On the one hand, this way of assessing OOS performance is less general than the literature on testing comparative predictive accuracy, as all evaluations are mediated by the role played by the loss function $L_{t+1}(\omega_t, C_t)$, which is of course not only or simply of a mean-variance type. On the other hand, portfolio choice models are based on economically grounded loss functions and it is important that the evaluation of an econometric model may be related to exact structure and meaning of $L_{t+1}(\omega_t, C_t)$. In what follows, we shall examine a range of choices for such a loss function.

The application of MS portfolio methods has also marked a step forward in the empirical finance literature. Optimal asset allocation decisions can only be made in the context of a model for the joint distribution of asset returns and, when they can be found, predictors. Although most studies have assumed that asset returns are generated by a linear process with stable coefficients so the predictive power of state variables such as dividend yields, default, and term spreads does not vary over time (see, e.g., Campbell & Viceira, 2002), since the late 1990s mounting empirical evidence has shown that asset returns follow more complicated, nonlinear processes (see, e.g., Detemple, Garcia, & Rindisbacher, 2003). This new literature based on nonlinear dynamic processes has often shown that findings in standard linear VAR frameworks – for example, that the equity allocation should be higher the longer the investment horizon – may be overturned in the presence of nonlinearities. Moreover, because MS implies that all conditional moments of the asset return distribution are time-varying, the MS portfolio literature has marked an extension of the classical literature to the case where the dynamics in all conditional moments may need appropriate hedging and hence affect portfolio choices.

A few papers in this literature have simply computed and reported mean-variance portfolio weights. For instance, Ramchand and Susmel (1998) examine the relationship between correlation and variance in an MS ARCH model estimated on weekly international stock return data. They find that correlations between the United States and other world markers are 2–3.5 times higher when the U.S. market is in a high-variance state. They also calculate mean-variance portfolios and find that their switching framework leads to higher Sharpe ratios than under linear models. Guidolin and Na (2008) assess the improvement in (risk-adjusted) portfolio performance achievable by a portfolio manager who – confronted with the evidence of regimes and the availability of a multiplicity of forecasting models – implements a number of alternative forecast combination schemes. They report large and statistically significant economic gains from

combining forecasts for portfolio management purposes, for example, an increase in the realized, recursive OOS Sharpe ratio from 0.20 to 0.36. Guidolin and Ria (2011) have recently extended results and intuitions in Ang and Bekaert (2004) to characterize in closed-form the dynamic process of the minimum variance efficient frontier in the presence of MS dynamics in returns, and have shown that simply tracking such simple dynamics might have improved considerably realized performances during the 2008–2009 financial crisis, relative to textbook, naive static mean-variance strategies.

On the one hand, resorting to a mean-variance framework in the presence of MS may seem natural because the structure of a simple model such as MSIH seems to suggest that MS simply impresses MS dynamics in expected returns, variances, and covariances. However, upon further thinking, this turns out to be misleading: it is well known (see, e.g., Timmermann, 2000) that Gaussian mixtures and MSMs impress strong patterns of time-variation in departures from a multivariate Gaussian benchmark, for instance in the form of nonzero conditional skewness and excess kurtosis. As a result, unless restrictive assumptions on preferences are enforced, simple mean-variance asset allocation may be logically inconsistent with the very nature and implications of MS, even though we shall see that Ang and Bekaert (2004) have derived very crisp insights even under these slightly contradictory assumptions on preferences.

When it comes to the role and OOS performance of MSMs in the asset allocation literature, Ang and Bekaert (2002a) is the seminal paper on the effects of MS on optimal dynamic portfolio choice, a contribution that has spurred a substantial volume of additional refinements and applications with important general insights on optimal portfolios under predictable (time-varying) investment opportunities (see Campbell & Viceira, 2002). Crucially, their paper assumes a power utility function that – by depending on the entire joint predictive density of asset returns – allows optimal portfolios to depend on all the conditional moments implied by any given econometric framework. Ang and Bekaert's starting point is that while in standard textbook international portfolio choice models agents optimally hold the world market portfolio and a series of portfolios to hedge against real exchange rate risk, a substantial body of empirical research (e.g., French & Poterba, 1991; Tesar & Werner, 1995) has concluded that in practice investors allocate their wealth differently, with the result that most investors/portfolios would display strongly home-biased choices that, for example, overweight domestic equities (securities). One popular argument often heard to rationalize this *home bias puzzle* relies on the asymmetric correlation behavior of international equity returns: correlations between

international equity returns are higher during bear markets than during bull markets; if the diversification benefits from international investing are not forthcoming at the time that investors need them the most (when their home market experiences a downturn), the strong case for international investing may be attenuated and substantial degrees of home bias might be rational (see, e.g., Butler & Joaquim, 2002). Ang and Bekaert set out to formally evaluate these informal claims in a partial equilibrium setting with an exogenous return-generating process. They develop (numerical) methods of dynamic portfolio optimization under time-varying investment opportunities characterized by a two-state MSM that accommodates time-varying correlations and volatilities *when regimes are observable to the investor*:

$$\mathrm{r}_{t+1} = \mu_{S_{t+1}} + \mathbf{A}_{S_{t+1}}\mathrm{r}_t + \sum_{S_{t+1}} \varepsilon_{t+1} \quad \varepsilon_{t+1} \sim \mathrm{NID}(\mathbf{0}, \mathbf{I}_{N-1})$$

where r_{t+1} is a vector of $N-1$ excess returns. Although in all the earlier literature, time variation in expected returns characterizes the changes in the investment opportunity set and the time variation is captured by a linear function of the state variables – normally expressed in the form of a single-state VAR – in their paper linear and MS (state-dependent) predictability are allowed to coexist. In particular, Ang and Bekaert quantify and compare the constant relative risk aversion (CRRA, as captured by the coefficient γ) utility cost, using the certainty equivalent notion, $\varkappa$,[1]

$$E_t\left[\frac{W_T^{1-\gamma}(\hat{\omega}_t, \ldots, \hat{\omega}_{T-1})}{1-\gamma}\right] = E_t\left[\frac{((1+\varkappa)W_T)^{1-\gamma}(\check{\omega}_t, \ldots, \check{\omega}_{T-1})}{1-\gamma}\right] \Rightarrow \varkappa$$
$$= \left(\frac{\hat{Q}_{t,T}}{\check{Q}_{t,T}}\right)^{1/1-\gamma} - 1$$

of (i) not being internationally diversified and (ii) ignoring MS.

Let's take an in-depth look at Ang and Bekaert's (2002a) framework because it has been the baseline reference for a number of subsequent papers. They consider a U.S. investor with CRRA preferences, that is, a standard power utility function parameterized by γ, with a T-month horizon, and who rebalances her portfolio over N assets every month to maximize expected end-of-period utility,

$$\max_{\omega_t, \omega_{t+1}, \ldots, \omega_{T-1}} E_t\left[\frac{W_T^{1-\gamma}}{1-\gamma}\right] \quad (\gamma \neq 1)$$

subject to the constraint that the portfolio weights at time t must sum to 1, where $\{\omega_{t+j}\}_{j=0}^{T-t-1}$ are the $N \times 1$ vectors of portfolio weights. There are no costs for short-selling or rebalancing. Next period wealth, W_{t+1}, is given by

$$W_{t+1} = \left[\sum_{n=1}^{N-1} \exp(r_{t+1}^n)\omega_t^n + \exp(r_{t+1}^N)\omega_t^N\right] W_t$$

Ang and Bekaert focus on the investment problem of a U.S. investor and ignore intermediate consumption. Using dynamic programming, they obtain the portfolio weights at each time t, for horizon $T-t$, by maximizing the (scaled) indirect utility:

$$\hat{\omega}_t = \arg\max_{\omega_t} E_t\left[Q_{t+1,T} W_{t+1}^{1-\gamma}\right]$$

$$Q_{t+1,T} = E_{t+1}\left[(R_{t+2}^p(\hat{\omega}_{t+1})R_{t+3}^p(\hat{\omega}_{t+3}), \ldots, R_T^p(\hat{\omega}_{T-1}))^{1-\gamma}\right]$$

where $R_{t+\tau}^p(\hat{\omega}_{t+\tau-1})$ denotes realized portfolio returns at time $t+\tau$ as a function of the weights $\hat{\omega}_{t+\tau-1}$. The $(N-1)\times 1$ vector of first-order conditions (FOCs) of the investor's problem can be summarized as

$$E_t\left[Q_{t+1,T}(R_{t+2}^p(\hat{\omega}_{t+1})R_{t+3}^p(\hat{\omega}_{t+3}), \ldots, R_T^p(\hat{\omega}_{T-1}))^{-\gamma} \mathrm{r}_{t+1}\right] = 0 \qquad (1)$$

Under the assumption that regimes are known to the agent at time t, the random variable $Q_{t+1,T}(S_{t+1})$ may take on one of the K values, one for each regime $S_{t+1} = 1, \ldots, K$. Therefore the optimal portfolio weights become functions of the regime at time t, $\hat{\omega}_t = \hat{\omega}_t(S_t)$. Moreover, the investor wants to hedge herself against future regime switches. These intertemporal hedging demands cause portfolio weights for different horizons, $\hat{\omega}_\tau(S_t)$, to differ from current portfolio weights, $\hat{\omega}_t(S_t)$, for $t < \tau \leq T-1$.[2] Under MS, the FOCs do not have a closed-form solution. Ang and Bekaert obtain a numerical solution to Eq. (1) by quadrature.[3] For instance, consider the one-period problem at $T-1$. For $S_{T-1} = k$, the FOCs are approximated by

$$E_t\left[(R_T^p(\hat{\omega}_{T-1}(k)))^{-\gamma} \mathrm{r}_{t+1} | S_{T-1} = k\right] = \sum_{i=1}^{K} \Pr(S_{t+1} = i | S_t = k)$$
$$E_t\left[(R_T^p(\hat{\omega}_{T-1}(k)))^{-\gamma} \mathrm{r}_{t+1} | S_T = i\right]$$
$$\simeq \sum_{i=1}^{K} \Pr(S_{t+1} = i | S_t = k)\left\{\sum_{g=1}^{G_i}\left[\exp(\mathrm{r}_{ig,t+1})'\hat{\omega}_{T-1}(k)\right]^{-\gamma} \mathrm{r}_{ig,t+1} W(\mathrm{r}_{ig,t+1})\right\}$$

where the $\{\mathrm{r}_{ig,t+1}\}_{g=1}^{G_i}$ are valid quadrature points assuming $S_T = i$. The optimal portfolio weights $\hat{\omega}_{T-1}(k)$ $(k = 1, \ldots, K)$ are the solution to this system of nonlinear equations that can be obtained numerically. At this

point, to start the dynamic programming algorithm, define $Q_{T-1,T}(S_{T-1}=k)$ as $Q_{T-1,T}(k) \equiv \sum_{i=1}^{K} \Pr(S_{t+1}=i|S_t=j)\{\sum_{g=1}^{G_k}[\exp(\mathrm{r}_{g,t+1})'\hat{\omega}_{T-1}(k)]^{-\gamma}\mathrm{r}_{g,t+1} W(\mathrm{r}_{g,t+1})\}$. The $T-2$ problem can then be solved for each regime $S_{T-2}=j$ $(j=1,\ldots,K)$ by finding the roots of

$$\sum_{k=1}^{K} p_{kj}\left\{\sum_{g=1}^{G_i} Q_{T-1,T}(k)\left[\exp(\mathrm{r}_{kg,t+1})'\hat{\omega}_{T-1}(j)\right]^{-\gamma}\mathrm{r}_{kg,t+1} W(\mathrm{r}_{kg,t+1})\right\}=0$$

This process is to be continued for $t=T-3$ to t.[4] Ang and Bekaert also characterize uncertainty in the portfolio choices from a classical econometric perspective, using the delta method (applied to the asymptotic normal distribution of the MS parameters), to test for the presence of intertemporal hedging demands (the difference between the investor's one-period-ahead and long-horizon portfolio choice under dynamic rebalancing), for the presence of regime-dependent asset allocation for investors with different horizons, and for the statistical significance of international diversification. These tests are more than interesting empirical exercises: if the asset allocations are similar across regimes, then in practice investors may not go to the trouble of rebalancing, especially if transactions costs are high. If intertemporal hedging demands are small, then investors may lose very little by solving a simple one-period problem at all horizons rather than solving the rather more complex dynamic problem.

Using Morgan Stanley Capital International (MSCI) monthly excess equity total unhedged (i.e., dollar-denominated) returns indices for the United States, the United Kingdom, and Germany (over a short rate represented by the 1-month eurodollar rate) for a sample 1972–1997, Ang and Bekaert find that in one regime equity returns have a lower conditional mean, much higher volatility, and are more highly correlated than in the other; volatilities and correlations increase together simultaneously. The strongest differentiating effect across the regimes for both systems is volatility. They reject the equality of volatilities across regimes at a 1% significance level. The evidence of different correlations across the regimes is not as strong as expected as they fail to reject that correlations for the United Kingdom and Germany are constant across regimes.[5] Ang and Bekaert also obtain mixed evidence on the usefulness of specifying regime-dependent means and fail to reject the constraint of equal means across regimes (p-value of 0.23). However, regime classification measures (RCMs, see Guidolin, 2011) improve only slightly when this restriction is imposed. Regime-dependent pure-equity portfolio calculations with continuous (monthly) rebalancing reveal the existence of strong (but not always

statistically significant) differences in regime-dependent weights and that in the first, high-volatility (bear) regime, the null of the optimality of no international portfolio diversification can sometimes not be rejected. Even in the basic model with large standard errors around the conditional means, they reject that a pure U.S. portfolio is optimal in the bull regime with $\gamma = 5$. Interestingly, the Gaussian IID portfolio weights – which imply no predictability of asset returns – lie in-between the regime-dependent weights and give a reasonable approximation of the optimal weights in each regime. Portfolio holdings of U.S. equity increase as the horizon increases, although the increase is small. After 3 years the portfolio weights converge to a constant. Finally, formal statistical tests of nonzero intertemporal hedging demands turn out to produce large p-values.

Ang and Bekaert report relatively large benefits to international diversification, although on statistical grounds it is not always possible to reject the optimality of home-biased portfolios. In the case of restricted means ($m_1 = m_2$), the costs of not diversifying internationally are substantial: an investor with a horizon of 1 year and risk aversion of 5 needs to be compensated with $\varkappa = 0.97\%$ in regime 1 to hold no United Kingdom or German equity under the benchmark model; for $\gamma = 10$, this compensation roughly doubles. At long horizons, $\varkappa$ exceeds 10% for $\gamma = 10$. Even though correlations are higher in regime 1, the costs of no international diversification in that regime are often not less than they are in regime 2, especially for modest levels of γ. This demonstrates that increasing correlations a priori does not make international diversification less valuable. The results are qualitatively the same for the case $\mu_1 \neq \mu_2$, but the costs of not diversifying internationally are even larger. When a conditionally riskless asset (with a constant 5% return) is added to the asset menu, Ang and Bekaert observe that leveraging occurs in the bull regime whereas a dramatic shift back to cash occurs in the bear regime. They now systematically fail to reject the hypothesis that a 100% home-biased position in only U.S. cash or equity is optimal and also that portfolio weights are equal across regimes. The costs of ignoring MS may instead be small or large depending on the presence of a conditionally risk-free asset. When investors ignore regimes, the IID weights they hold are reasonable approximations to the optimal weights, especially the weights in regime 2, the longest duration regime. The cost of ignoring regimes is higher in regime 1 than in regime 2. This is in accordance with intuition, since in the bull regime, conditional means and variances are closer to their unconditional counterparts than they are in regime 1. When a conditionally risk-free asset is introduced, ignoring MS becomes much more costly and of a similar

order of magnitude as ignoring the investment opportunities in overseas equities.[6] Intertemporal hedging demands under MS are always economically negligible and statistically insignificant. Investors have little to lose by acting myopically instead of solving a more complex dynamic programming problem for horizons greater than one period. The p-values of tests of intertemporal hedging remain small and below what is typical of the predictability literature (see, e.g., Lynch & Balduzzi, 2000). This is because the MSM does not rely on a highly correlated predictor like the dividend yield.

Ang and Bekaert have also analyzed the effect of time-varying short rates by incorporating the U.S. short rate as an additional state variable in their MSVAR. In this model, i_t is the driving variable predicting asset returns. Excess stock returns now follow

$$\mathrm{r}_{t+1} = \mu_{S_{t+1}} + \theta_{S_{t+1}} i_t + \Sigma_{S_{t+1}} \varepsilon_{t+1} \quad \varepsilon_{t+1} \sim \mathrm{NID}(0, \mathrm{I}_{N-1}),\ \Sigma_{S_{t+1}} \Sigma'_{S_{t+1}} = \Omega_{S_{t+1}}$$

which implies regime-dependent predictability in the conditional mean. They use an MS discretized square root process to model i_t:

$$i_{t+1} = \mu^i_{S_{t+1}} + \phi^i_{S_{t+1}} i_t + \sigma^i_{S_{t+1}} \sqrt{i_t} \varepsilon^i_{t+1} \quad \varepsilon^i_{t+1} \sim \mathrm{NID}(0, 1)$$

where the normally distributed error terms ε^i_{t+1} and ε_{t+1} are correlated in each regime. The transition probabilities for $S_{t+1} = 1,2$ are assumed to be logistic functions of the short rate. This more complex MSM incorporates a number of features discussed in Guidolin (2011) such as the coexistence of nonlinear and linear (VAR-style) predictability patterns and the potential role of time varying transition probabilities (TVTPs). A likelihood ratio test (LRT) reveals that in regime 2, as the short rate increases, a transition to the first regime becomes increasingly likely. Moreover, a test fails to reject the null of $\phi^i_{S_{t+1}} = 0$ with a p-value of 0.92 showing the absence of predictability in the conditional mean. Moreover, Ang and Bekaert cannot once more reject the hypothesis that $\mu_1 = \mu_2$. The resulting model exhibits nonlinear predictability through TVTPs rather than linear predictability through the conditional mean. In regime 1, equity returns are much more volatile and more highly correlated across countries. However, in this regime, short rates and equity returns are more negatively correlated than in regime 2. This means that two effects increase the attractiveness of cash for investors in this regime. First, interest rates are higher in this regime; second, shocks to equity and short rates are more negatively correlated in bear markets. Portfolio weight calculations reveal that hedging demands are again small,

although they are now nonnegligible in the first regime. In regime 2, as the short rate increases, investors hold less equity, but in regime 1 there is almost no effect of the short rate on the portfolio allocations. This is driven by the nonlinear predictability via TVTPs. The portfolio holdings in regime 1 are flat because the excess returns are constant and no significant short-rate predictability drives the transitions from this regime. In the second regime, as the short rate increases, a transition to regime 1 becomes increasingly likely. As the first regime has much higher equity volatility, investors seek to hold less equity to mitigate the higher risk. The costs of no international diversification – where investors are permitted to hold only cash and U.S. equity – are not small: at a 12-month horizon, this cost is 3.39% at $i_t = 5.1\%$ in the bull regime. In the bear regime, at a 12-month horizon and $i_t = 5.1\%$, the cost is 3.33%. Similarly to the case of the constant risk-free asset, the costs of ignoring regimes are substantial. Finally, like the all-equity portfolios and the constant risk-free asset case, the cost of myopia remains negligible.

Ang and Bekaert (2004) have extended their seminal 2002 paper to an MS Black's (1972)-style (i.e., in which the existence of a riskless asset need not to be assumed) intertemporal CAPM-like framework in which the excess world market return follows a simple two-state MSIH

$$r_t^W = \mu_{S_t}^W + \sigma_{S_t}^W \varepsilon_t^W \quad \varepsilon_t^W \sim \text{NID}(0,1)$$

and the expected excess return on any other international portfolio or security (indexed by j) is linear in its beta with respect to the world market:

$$r_t^j = \mu^z + \beta^j(\mu_{S_t}^W - \mu^z) + \beta^j \sigma_{S_t}^W \varepsilon_t^W + \sigma^j \varepsilon_t^j \quad \varepsilon_t^j \sim N(0,1)$$

where μ^z is the zero-beta excess return. The unexpected return on security or portfolio j is determined by the security's sensitivity to the world market return and by an idiosyncratic term, which has volatility σ^j. Of course, this is also a rather stylized factor, CAPM-style model for stock returns, that has been later extended by Baele, Bekaert and Ingelbrecht (2010). Because the mean of the world excess return switches between regimes, the 1-month ahead expected excess return of country j is given by

$$E[r_{t+1}^j - \mu^z | S_t = i] = p_{i1}\beta^j(\mu_1^W - \mu^z) + p_{i2}\beta^j(\mu_2^W - \mu^z)$$

Expected returns of individual equity markets differ only through their different betas with respect to the world market. The conditional variance for the individual assets is more complex:

$$\mathrm{Var}[r^j_{t+1} - \mu^z | S_t = i] = p_{i1}(\beta^j \sigma_1^W)^2 + p_{i2}(\beta^j \sigma_2^W)^2 + p_{i1}p_{i2}(\beta^j)^2(\mu_1^W - \mu_2^W)^2 + (\sigma^j)^2$$

Intuitively, the conditional variance depends on three components. First, as in a standard CAPM, an asset's conditional variance depends on the asset's exposure to systematic risk through the asset's beta. In the world CAPM, however, the world market return switches regimes, so the market conditional variance now also depends on the regime prevailing at time t. Second, also as in a standard CAPM, each asset has an idiosyncratic volatility term unrelated to its systematic (beta) exposure. Finally, the variance of an individual asset depends not only on the realization of the current regime but also on a "jump"-type component, $p_{i1}p_{i2}(\beta^j)^2(\mu_1^W - \mu_2^W)^2$, which arises because the conditional means differ across regimes.

Ang and Bekaert observe that although the model structure is parsimonious, the model generates rich patterns of stochastic volatility and time-varying correlations. In particular, the model captures the asymmetric correlation structure in international equity returns already discussed in Ang and Bekaert (2002a). The first regime is a normal, quiet regime, where world excess returns are expected to yield 0.90% a month, with volatility of 2.81% a month. The other regime is a volatile regime, with standard deviation 5.04% a month and with a lower but imprecisely estimated mean of 0.13% a month. The estimate of μ^z is larger than the expected excess equity return in the low-volatility regime. The country betas are all estimated precisely, and their magnitudes seem economically appealing. Because the betas are close to 1, expected returns are close to each other in the normal regime. In the bear regime, expected excess returns are dramatically lower and more dispersed, with the United Kingdom and Japan giving the lowest expected excess returns. In this regime, the zero-beta excess return, μ^z, is higher than the excess return of the world market, which causes the high beta countries to have lower expected returns. North America and the smaller countries in Europe have the lowest idiosyncratic volatility implied by the model. Correlations in the bear state are, on average, some 20% higher than those in the bull state.

Based on these parameter estimates, Ang and Bekaert perform a simple, recursive mean-variance asset allocation exercise in which there are two distinct optimal tangency portfolios, one for each regime, and compare them with the unconditional tangency portfolio that will be obtained when

the unconditional moments are used and regime switches are ignored. Because the mean-variance frontiers also become time-varying, the world market portfolio (based on average market-cap weights) will be inefficient and inside the unconditional frontier. Theoretically, the presence of two regimes and two frontiers means that the MS investment opportunity set dominates the investment opportunity set offered by a single unconditional frontier. In particular, in the bull state the unconditional tangency portfolio yields a Sharpe ratio of 0.62. The investor could improve this trade-off to 0.87 by holding the optimal tangency portfolio for this low-variance regime. In the bear state, the unconditional tangency portfolio yields a Sharpe ratio of only 0.13, which could be doubled to 0.27 by holding the optimal tangency portfolio for the high-variance regime. Finally, in a recursive OOS exercise based on a crude reduction of the filtered state probabilities to a 0/1 regime classification – which is consistent with their assumption that regimes be observable – Ang and Bekaert report a realized portfolio Sharpe ratio that is more than double the Sharpe ratio for a simple OOS buy-and-hold strategy on the world market portfolio and also higher than the Sharpe ratio for the nonregime dependent, unconditional mean-variance frontier portfolio. Besides its interesting use of an MS CAPM to support optimal portfolio decisions, Ang and Bekaert's (2004) is a key paper because it first opens the stage to OOS tests of the realized performance of MS-based portfolios (i.e., back-testing).[7]

However, Ang and Bekaert (2004) also assume that a portfolio manager knows which regime is being realized at each point of time, but does not know which regime will be realized in the future. Ang and Bekaert (2002a) had indeed conjectured that under the alternative assumption where investors are uncertain about the regimes, the effects of MS should be uniformly weaker since the regime-dependent solutions would deviate less from the IID solution, and that their assumption of observable regimes would be a worst-case scenario: if there are weak effects when the agents perfectly observe the regimes, the effects will be even smaller when learning about the regimes is introduced. Yet, this logic hides a key aspect of optimal asset allocation problems under MS. When regimes are not observable, similarly to what econometricians normally do, and an investor may only filter the nature of regimes out of the data, the conditional, 1-month ahead expected excess return of country j is

$$E[r^j_{t+1} - \mu^z | \mathscr{F}_t] = \xi_{1t}[p_{11}\beta^j(\mu_1^W - \mu^z) + p_{12}\beta^j(\mu_2^W - \mu^z)] + (1 - \xi_{1t})[p_{21}\beta^j(\mu_1^W - \mu^z) + p_{22}\beta^j(\mu_2^W - \mu^z)]$$

which is different from the formula shown above for the case of observable regimes as the filtered probabilities ξ_{1t} also get to play a key role. Moreover, the conditional variance for the individual assets is even more complex:

$$\begin{aligned}\mathrm{Var}[r^j_{t+1} - \mu^z | S_t = i] = \xi_{1t}[p_{11}(\beta^j\sigma_1^W)^2 + p_{12}(\beta^j\sigma_2^W)^2 + p_{11}p_{12}(\beta^j)^2(\mu_1^W - \mu_2^W)^2] \\ + (1-\xi_{1t})[p_{21}(\beta^j\sigma_1^W)^2 + p_{22}(\beta^j\sigma_2^W)^2 \\ + p_{21}p_{22}(\beta^j)^2(\mu_1^W - \mu_2^W)^2] + (\sigma^j)^2\end{aligned}$$

These differences imply that in practice a considerably richer dynamics may characterize the time variation in both predicted risk premia and variances, and this could be reflected by long-run optimal portfolio shares. Guidolin and Timmermann (2005, 2007) have analyzed exactly this problem and have extended Ang and Bekaert's (2002a) seminal contribution to the case in which regimes are not observable to the investor, in line with the information set that belongs to an econometrician who makes inferences on an MSM. Guidolin and Timmermann (2005, 2007) shift their applied focus away from international equity portfolio diversification problems and apply MS-driven asset allocation tools to a typical strategic asset allocation (SAA) decision, that is, how much to invest in major asset classes such as cash, stocks, and bonds. Guidolin and Timmermann (2005, 2007, GT) characterize an investor's SAA (and consumption decisions, in GT 2007) under a heteroskedastic MSM with four states. As far as solution methods are concerned, GT use Monte Carlo methods for integral (expected utility) approximation which are advised by the difficulty of applying Ang and Bekaert's (2002a) quadrature approximation methods when the Markov state is unobservable. For instance, in a simple buy-and-hold case with no interim consumption, this means that GT simply draw a large number L of simulated paths for asset returns under an estimated MSM and solve

$$\max_{\omega_t} L^{-1}\sum_{l=1}^{L}\left\{\frac{\left[(1-\omega_t'\iota_3)\exp(Tr^f) + \omega_t'\exp\left(\sum_{i=1}^{T}(r^f\iota_3 + \mathrm{r}_{t+i,l})\right)\right]^{1-\gamma}}{1-\gamma}\right\}$$

where $\omega_t'\exp(\sum_{i=1}^{T}(r^f\iota_3 + \mathrm{r}_{t+i,l}))$ is the portfolio return in the lth Monte Carlo simulation.[8] Interestingly, GT's methods also allow to fully capture learning on the regimes along simulated paths when rebalancing is allowed, in the form of the investor simply applying filtering algorithms on simulated data, analogously to what an econometrician would do in recursive, real time estimations (see e.g., Veronesi, 1999, for a characterization of the resulting hedging demands).

GT (2007) extend the results in GT (2006a) to find that in monthly U.S. data on stock (large and small caps) and bond excess returns there is evidence that four separate regimes – characterized as crash, slow growth, bull, and recovery states – are required to capture their joint conditional distribution. However, none of the states can be perfectly anticipated, that is, starting from any of the states the investor always attaches positive (and never negligible, the minimum value being 0.12) probability to the event of transitioning to a different, unknown state. GT show that optimal asset allocation differs strongly across regimes. For instance, stocks are attractive to short-to-medium term investors in the bull state since the probability of staying in such a state is high. Stocks are far less attractive in the crash state even though this state is not very persistent. Even if, as seems plausible, investors never know with certainty which regime the economy is currently in, beliefs about state probabilities become important. For instance, with reference to the effect of the investment horizon on optimal SAA, GT report that while in simple, single-state VARs Barberis (2000) and others have found that the weight on stocks should increase as a function of the investor's horizon, even in the absence of predictor variables MSMs imply that horizon effects strongly vary across states. Since stocks are not very attractive in the crash state, investors with a short horizon hold very little in stocks in this state. At longer investment horizons, there is a high chance that the economy will switch to a better state and so investors allocate more toward stocks. In the crash state the allocation to stocks is therefore an increasing function of the investment horizon. In the more persistent slow growth and bull states, investors with a short horizon hold large positions in stocks. At longer horizons investment opportunities will almost certainly worsen so investors hold less in stocks, thereby creating a downward sloping relation between stock holdings and the investment horizon. In addition to these horizon effects, GT find interesting substitution effects among small and large stocks. As the horizon expands, the allocation of small stocks as a proportion of the total equity portfolio typically declines, whereas the allocation of large stocks increases. This extends earlier findings that predictability of returns on small and large stocks can lead to important shifts in the composition of equity portfolios.[9]

Finally, similarly to Guidolin and Timmermann (2006a), GT extend their MSIH model for asset returns to include predictability from state variables such as the dividend yield. Compared to a benchmark with constant expected returns, predictability from the dividend yield in a linear VAR is known to reduce risk at longer horizon and to lead to an increased demand for stocks, the longer the investment horizon (see Barberis, 2000;

Campbell & Viceira, 2002). In contrast, in an MSVAR GT uncover a positive correlation between return innovations and shocks to future expected returns, thereby increasing risk and lowering the long-term demand for stocks compared to the benchmark model with no predictability. As a result, within an MSVAR a nonmonotonic relationship between the allocation to stocks and the horizon appears: At short horizons the effect of regimes tends to dominate while at longer horizons the mean reverting component in returns tracked by the yield leads to an increasing demand for stocks. GT also compute the welfare (annual percentage certainty equivalent return, $\varkappa$, CER) costs of ignoring MS and find that the CER is as high as 3% at short horizons – when investors can exploit market timing more aggressively – whereas at the longest horizons the loss is around 1.3% per annum.[10]

GT have also followed Ang and Bekaert (2004) and proposed to evaluate the economic significance of their SAA results by examining real time OOS performance of asset allocation rules based on both standard VARs that use the dividend yield as a predictor variable, simple MSIH, and MSVAR models for the joint dynamics of stock and bond returns. They find that the recursively updated portfolio weights vary significantly over time as a result of changing investment opportunities and that optimal asset holdings are sensitive to how predictability is modeled. Interestingly, the turnover in optimal portfolios is found to be smaller under MSIH than under the single-state VAR model. Once the dividend yield is included in an MSVAR, the volatility of the equity weights increases and becomes comparable to that under the VAR(1) benchmark. MS increases the overall demand for stocks (to approximately 60%) relative to the benchmark VAR(1) model (40%). Bonds receive a substantial weight under MS, between 35% and 60% depending on T and irrespective of whether the dividend yield is included as a predictor. Conversely, the VAR(1) model puts a large weight on cash investments (in excess of 50%). This suggests that the presence of MS is important in understanding the demand for (nominal) long-term bonds. Furthermore, GT generally find that the average realized utility is highest for models that account for regime switching. The VAR(1) model performs best over the shortest investment horizon ($T=1$) although 10% confidence intervals for the realized utility overlap under the VAR(1) and MSMs, suggesting that their performances are statistically indistinguishable. For the longer horizons, $T=12$, 120 months, an MSIH model produces the highest mean realized utility. At a 12-month horizon the OOS performance of MSIH is sufficiently good to be statistically significant against 3 of the 5 alternative models.

Ang and Bekaert (2004) have also examined an SAA problem in which a CRRA investor diversifies among U.S. stocks, bonds, and one-month *T*-bills. Differently from the rich MSIH and MSVAR models in Guidolin and Timmermann (2005, 2007), Ang and Bekaert write a multivariate, two-state MS model in which stock and bond returns follow a simple process with MS-expected returns, and variances, correlations, and the short rate follow an autoregressive process,

$$i_t^{\mathrm{TB}} = \mu_{S_t}^{\mathrm{TB}} + \phi_{S_t} i_{t-1}^{\mathrm{TB}} + \sigma_{S_t}^{\mathrm{TB}} \varepsilon_t^{\mathrm{TB}} \quad \varepsilon_t^{\mathrm{TB}} \sim N(0,1)$$

where both the constant term and the autoregressive parameter depend on the regime, similarly to Ang and Bekaert (2002b, 2002c). Consistent with the literature on MSMs for interest rates, this is useful to capture the existence of one regime in which interest rates are highly persistent and another regime that captures periods of higher, volatile interest rates that revert quickly to lower rates.[11] Ang and Bekaert report that the portfolio strategy that times regime switches is more volatile but delivers higher average returns than a nonregime-dependent strategy. The MS strategy's positions are more leveraged, however, and although they yield higher realized returns, they also lead to higher realized portfolio volatility. The MS market-timing model is the best performing model in terms of realized, recursive Sharpe ratios. However, it remains questionable whether Sharpe ratios may be the most appropriate performance indicators when optimal diversification strategies are derived under CRRA preferences.

MS Portfolio Choice in Continuous Time

A literature exists that has investigated similar – usually stylized, essentially univariate (in the sense that there is a single risky asset) – portfolio problems in continuous time. For instance, Honda (2003) solves a portfolio and consumption problem in which only the expected return (drift) of a single risky asset depends on an unobservable regime governed by a continuous Markov chain with two states.[12] Also in Honda, optimal policies are computed using Monte Carlo methods. Similarly to Guidolin and Timmermann (2005, 2007), he also finds that the shape of the function connecting optimal portfolio weights to the investment horizon may depend on the perception of the current regime. Unfortunately, Honda (2003) has not computed real-time, pseudo-OOS portfolio weights and performances: it would be interesting to assess what portion of any improved performance

from MSMs may come from the incorporation of MS in second moments, as opposed to homoskedastic models.

Liu (2011) is a related paper in which investment opportunities are MS but in which investors are modeled as being subject to ambiguity aversion (when an investor with concave preferences also integrates over the space of possible probability measures, which can be incorporated by specific modified functional forms for preferences, see Guidolin & Rinaldi, 2010, for a recent review). In Liu's paper, the investor treats the model of filtered probabilities as ambiguous and has multiple, uncertain beliefs with respect to the states resulting from continuous Bayesian updating. The investor obtains the conditional estimates of the unobservable state by observing past and current asset prices and she employs a nonlinear recursive filter to extract regime probabilities that are updated according to the Bayes rule. As in Honda (2003), the asset menu is simply composed of a riskless short-term bond paying an instantaneous return r^f and a risky asset with Ito dynamics, $\mathrm{d}P_t = \mu_t P_t t + \sigma_t P_t \mathrm{d}Z_t$. The drift process μ_t follows a continuous-time Markov chain with two states, $\mu_H > \mu_L$, and infinitesimal rate matrix with generic element λ_{ij} ($i, j = 0, 1$). The infinitesimal generating matrix governs the dynamics of MS between states of high and low drift. The investor can observe neither the expected return μ_t, nor the Brownian motion Z. Instead, she can only observe the stock price P. Given an initial prior π_0 over the two regimes, the investor estimates the unobservable state, that is, the probability of the current state being in the high-mean-return regime, based on the observed asset prices, $\pi_t \equiv \Pr(\mu_t = \mu_H | \mathscr{F}_t)$. It can be shown that the law of motion of the estimate π_t can be explicitly written as:

$$\mathrm{d}\pi_t = [\lambda_0 - (\lambda_0 + \lambda_1)\pi_t]\mathrm{d}t + \pi_t(1 - \pi_t)\frac{\mu_H - \mu_L}{\sigma_t}(\mathrm{d}Z_t - \theta \mathrm{d}t)$$

Using Malliavin calculus (see, e.g., Detemple, Garcia, & Rindisbacher, 2005), Liu (2011) explicitly characterizes the optimal consumption and portfolio rules. Ignoring ambiguity, under power utility the optimal stock demand may be decomposed as[13]

$$\omega_t = \frac{(\pi_t \mu_H + (1 - \pi_t)\mu_L) - r^f}{\gamma \sigma_t^2} + \mathrm{hedge}^{\mathrm{IIR}}$$

where γ is a constant risk aversion coefficient. The first term is the standard myopic demand for the risky asset under MS, which is instantaneously mean-variance efficient and depends on the current estimate of the unobservable state. The second term quantifies the intertemporal hedging demand, which practically insures the investor's portfolio against future

time variation of the conditional estimates of the unobservable state. Liu has computed the (discounted) CERs under MS vs. a simpler IID as a function of the investment horizon ranging from 1 to 20 years when the current state probability is, starting from its steady-state value, updated after each period's return is realized. The economic value of regimes is nonnegligible and the welfare loss of ignoring regimes is increasing in the horizon. The utility cost measured by the difference in CERs reaches 1.5% per year at long horizons. In a real time OOS exercise over a quarterly sample 1996–2009 for three investment horizons (1, 5, and 10 years), Liu finds that MS under ambiguity produces the highest mean realized utility for all the three investment horizons. Among all the horizons considered, the difference in mean realized utilities between the MS/ambiguity strategy and other strategies is the highest for a 10-year horizon. An IID strategy that ignores both MS and ambiguity, generates a slightly lower mean realized utility for all horizons. Of course, it would be very interesting to extend both Liu's closed-form solutions and OOS performance comparisons to richer asset menus and more realistic applications, which seems to remain rather challenging in technical terms.

The Effect of Transaction Costs

Many practitioners and asset management scholars are in principle ready to admit that MS strategies may be useful to time investment opportunity regimes; however the find in the very regime shifting nature of the econometric model the seeds for a number of alleged difficulties in its implementation, especially for the most active traders. As the reader may have noticed, the few real time OOS portfolio backtesting assessments that we have so far available with reference to MS portfolio strategies yield heterogeneous implications for the variability of optimal weights. Guidolin and Timmermann (2005, 2007) have observed that for intermediate and high values of γ, an investor will optimally trade *less* than the continuous adjustments implied by a highly persistent predictor such as the dividend yield or the term spread in single-state VAR models à la Barberis (2000) and Campbell and Viceira (2002). They therefore emphasize that MS strategies are likely to be considerably robust to transaction costs because they are designed to exploit low-frequency changes in expected returns and volatilities: when the probability of staying within the same regime is relatively high, portfolio turnover will be low. However, Ang and Bekaert (2004) have reported that in practice the volatility of MS portfolios may be

higher than what linear predictability implies. This is a crucial aspect because a standard objection against MS strategies in applied portfolio management maintains that these may be "prohibitively" expensive, because of the implied need to frequently trade (re-balance) that would be costly in the presence of fixed and variable transaction fees. The logic is that because a well-specified MSM implies a number of regime shifts across states that – again assuming correct specification – will usually correspond to rather different investment opportunities, as a result MS strategies will force an investor to massive portfolio re-shuffling often enough to lead her into ruin (or at least, mediocre performances) just because of the trading costs. Because, as we have mentioned, in an MSM the predicted state probabilities converge to the ergodic frequencies as the prediction horizon grows, this caveat to the benefits of MSMs in portfolio management would strongly apply to aggressive investors characterized by low (or no) risk aversion and with short investment horizons, who will want to boldly time the state of financial markets. Therefore, comprehensive investigations of the net-of-transaction costs realized performance from MSMs are certainly needed.

A recent paper by Jang, Koo, Liu, and Loewenstein (2007, JKLL) has specifically examined this problem in a normative, continuous time framework. Even though JKLL's main goal is to re-examine the odd conclusion in the earlier constant investment opportunities literature that the utility loss due to the presence of transaction costs may be small, contrary to standard intuition, JKLL results hold the promise to soon allow an effective examination of the claim that MS portfolio strategies may be less appealing than commonly thought, once transaction costs are taken into account.[14] Empirical research has gathered a great deal of evidence that is inconsistent with the constant investment opportunity set hypothesis, because both volatilities and expectations of stock returns vary substantially over time: JKLL's conjecture is that by taking into account the stochastic nature of the investment opportunity set may qualitatively change the conclusion in the standard literature that transaction costs only have second-order asset pricing effects because as market conditions change over time, an investor should rebalance more often to avoid being too far away from the target if the transaction cost rate is small.

To quantify this intuition, JKLL build a model with MS in fundamental parameters, in the form of two regimes (bull and bear) with different fundamental parameters such as expected return, volatility, and liquidity. One regime switches to the other regime at the first jump time of a market-independent but possibly regime-dependent Poisson process with intensity

λ_s, for $s =$ bull, bear. In regime s, the risk-free interest rate is r_i^f, and the investor can buy the stock at the ask $P_t^a = (1 + \theta_s)P_t$ or sell the stock at the bid $P_t^b = (1 - \alpha_s)P_t$, where $\theta_s \geq 0$ and $0 \leq \alpha_s \leq 1$ represent the proportional transaction cost rates and P_t satisfies a standard MS Geometric Brownian motion with drift μ_s and diffusion coefficient σ ($s =$ bull, bear), where all parameters are positive constants and $\mu_s > r_s^f$ (i.e, the stock will never be shorted). JKLL characterize the optimal consumption and investment problem for a small investor (i.e., with no price impact) who derives CRRA utility from intertemporal consumption and a terminal bequest at death, with a time discount rate of ρ.[15] The investor can invest in one risky asset and one risk-free asset. Under no transaction costs, $\theta_s = \alpha_s = 0$ (for $s =$ bear, bull), JKLL prove that both the optimal consumption and dollar amount invested in the stock are constant fractions of the investor's wealth in each regime. Even though the investor smooths consumption across regimes, the optimal investment policy is "within-regime myopic" in the sense that the optimal weights only depend on the current regime parameters. This follows from the fact that the risk of regime switching is unhedgeable using the existing securities and that the investor can rebalance at regime-switching time without any transaction costs.

When transaction costs appear in the problem, the solvency region of the investor splits into three regions: A "no-trading" region, a (stock) "buy" region, and a (stock) "sell" region. The homogeneity of the regime-specific value function implies that the transaction boundaries are straight lines. In addition, there exists an interval for portfolio weights such that in regime s, the investor trades only the minimum amount to keep the bond to stock ratio inside the interval. JKLL's extensive numerical analysis demonstrates that in contrast to the standard conclusion that transaction costs only have a second-order effect, transaction costs can have a first-order effect if the investment opportunity set varies over time.[16] Additionally, unlike the no transaction cost case, smoothing of trading strategies *across regimes* is optimal in the presence of transaction costs. Without transaction costs, the optimal investment policy in one regime is independent of parameters in the other regime. In contrast, in the presence of transaction costs, an investor optimally responds to changes in one regime by altering investment behavior in both regimes. Finally, JKLL notice that under MS the expected holding period for stocks is significantly shorter: whereas the expected time to sale after a purchase is monotonically decreasing in the risk aversion in a model with constant investment opportunities, it is nonmonotonic under MS. Intuitively, as risk aversion decreases, on average the investor keeps more invested in the stock. In the single-regime model, lower risk aversion

reduces the frequency of trade in part through a higher consumption to wealth ratio and in part through a modification to the relative position of the boundaries to reduce rebalancing costs. However, if there are two regimes, the investor needs to also take into account the transaction costs to be paid at future regime switching times. Less frequent rebalancing before the regime shift can lead to the possibility of larger revisions when the regime changes. Thus, loosely speaking, in the MS model the investor optimally trades off the large lump-sum transaction costs incurred at future regime switching times against more frequent but smaller rebalancing costs within a regime. JKLL report that for their parameters and low levels of risk aversion, the expected discounted lifetime transaction costs as a fraction of wealth can amount to 40% versus only about 5% under constant investment opportunities. This would corroborate the cautions on the actual, ex-trading costs realized performance of MS portfolio strategies illustrated earlier, even though JKLL do not perform any recursive backtesting of their strategies that may be useful to assess this conjecture.

A related effort has been recently made by Guidolin and Hyde (2010) who adopt a discrete time MS portfolio choice framework in which transaction costs are numerically incorporated by constraining an investor to trade at time t only if the increase in expected utility from timing a regime shift in (perceived, as in Guidolin & Timmermann, 2005, 2007) investment opportunities exceeds the transaction costs to be paid to re-shuffle the portfolio. In fact, Guidolin and Hyde (2010) motivate their work in the light of a number of claims related to Ang and Bekaert's (2002b) finding that – in small-sample Monte Carlo simulation from MSVARs – misspecified but simple linear VAR models may often fit simulated data that do contain regime shifts better than MSMs do. Therefore they ask whether it is possible to find VAR models that approximate the optimal portfolio weights implied by simple MS strategies or that may even outperform MS strategies in real time, recursive OOS tests as in Guidolin and Timmermann (2005, 2007) and Liu (2011). Although linear models are key benchmarks in empirical finance and their simplicity makes them obvious choices in many applications, their use in asset allocation exercises has relied on two often-implicit premises. First, that although most normative papers have to be taken only as examples of how practical portfolio choice ought to proceed, when the scope of the investigation is extended beyond the class of small-scale (i.e., with 3–4 predictors at most) VAR(1) models typical in the literature, *some* more complicated VAR must exist that is of practical use in terms of consistently improving portfolio performances. This means that some VARs can be found that can efficiently summarize the overall balance of predictability in

asset returns and that make the modeling of any residual nonlinear effects of second-order importance. Second, that although more complicated, large-scale VAR(p) models may yield complex portfolio strategies, simple, small-scale VAR(1) models must be illustrative already of the first-order effects of predictability on dynamic portfolio selection. Guidolin and Hyde tackle both these conjectures at their roots and provide a systematic examination of whether, when, and how small- and medium-scale VAR(p) models may deliver dynamic portfolio choices that are: (i) able to approximate the portfolio choices of an investor that exploits both linear and nonlinear predictability patterns in the data and (ii) competitive in terms of realized portfolio performance.

Interestingly, these results are obtained with reference to the long-horizon portfolio problem (up to a 5-year horizon) of an investor who derives utility from real consumption when rebalancing is allowed at the same frequency as the data. Moreover, Guidolin and Hyde's investor selects optimal portfolio weights taking into account the presence of both fixed and variable transaction costs using the discretization methods applied to both time and states to obtain a numerical approximation, similarly to Lynch and Balduzzi (2000). Using monthly U.S. stock and bond return data for a 1953–2009 sample and an SAA application similar to earlier papers, Guidolin and Hyde find that a relatively large set of small- and medium-scale (with up to 7 predictors) VAR(p) models (with $p = 1, 2, 4$, and 12) fails to imply portfolio choices that approximate those from a three-state MSIH.[17] This of course gives only an ex ante perspective on the problem: "different" does not mean worse in the view of a portfolio manager and what could be misspecified is not the VAR family, but the MSIH benchmark. However, VARs also systematically fail to perform better than MSMs in recursive (pseudo) OOS tests, in the sense that VARs generally produce lower realized ex post CERs than MSMs do. This means that VARs provide no approximation tool for more complicated, nonlinear dynamics either ex ante or ex post.

MS Portfolio Choice Driven by Asset Pricing Models

At least in my view, two problems remain unresolved in the empirical MS portfolio choice literature. First, applications need to be based on numerical solution of complex dynamic optimization programs. This has severely constrained the actual range of applications to rather small-scale problems, mostly concerning either SAA or international portfolio diversification with 3–4 asset classes at most. Second, especially MSVAR models applied to

$N \geqslant 3$ and $K \geqslant 3$ easily imply the need to estimate hundreds of free parameters. Although data availability – also because portions of the literature have not hesitated to employ weekly data (see, e.g., Baele, 2005) – is not a key factor, computational resources are, and one always wonders how reliable portfolio strategies can be, when they are built on the basis of large sets of estimated parameters.[18] A few recent papers have also moved steps in a direction appropriate to tackle these issues. Guidolin and Timmermann (2008) is a paper that has proposed approximations that may prove important to enlarging the size of the asset menus to which MS portfolio choice methods are applicable. GT modify the standard international CAPM (ICAPM) specification that assumes mean-variance preferences over a time-invariant distribution of local stock returns in two ways. First, they allow investor preferences to depend not only on the first two moments of returns but also on third and fourth moments such as skewness and kurtosis, similar to related approaches such as Harvey and Siddique (2000) and Dittmar (2002). Like in these papers, GT's approach approximates the unknown marginal utility function by means of a Taylor series expansion. Second, they model international equity returns in the context of a four-moment ICAPM with regimes that track time-variations in the means, volatility, correlations, skew and kurtosis (as well as co-skewness and co-kurtosis) of all equity return indices and the world market portfolio; also the world price of covariance, co-skewness, and co-kurtosis risk are allowed to vary across regimes. This MS ICAPM accurately approximates the return distribution and captures volatility clustering, return correlations that strengthen in down markets, outliers that occur simultaneously in several markets, fat tails, and skewness. This is because MSMs are known to capture central statistical features of asset returns (see, e.g., Guidolin, 2011). GT consider a relatively large number of assets and as such need an efficient method to compute optimal portfolio choices. Therefore they propose a new tractable approach that, when coupled with a utility specification that is based on moments, reduces an otherwise complicated numerical problem to finding the roots of a low-order polynomial.

GT return to Ang and Bekaert's (2002a) application of MS portfolio selection to explain the home bias in the portfolios of U.S. investors. GT find evidence of two regimes in the joint distribution of MSCI international stock return data: A bear state with high volatility and low mean returns and a bull state with high mean returns and low volatility. Differently from Ang and Bekaert (2002a), GT find strong evidence of regime switching equity premia which strongly contribute to the ability of their econometric framework to fit time-varying patterns in Sharpe ratios, co-skewness, and

co-kurtosis. Both modifications to the standard ICAPM model are needed to explain the home country bias. Regimes in the distribution of international equity returns generate skew and kurtosis and therefore affect the asset allocation of a mean-variance investor differently from that of an investor whose objectives depend on higher moments of returns. This is significant since the single state model is severely misspecified and fails to capture basic features of international stock market returns. Their sample estimates suggest that a U.S. mean-variance investor with access to T-bills, U.S., U.K., European, Japanese and Pacific stock markets should hold only 30% in domestic stocks. The presence of bull and bear states raises this investor's weight on U.S. stocks to 50%. Introducing both skew and kurtosis preferences and bull and bear states further increases the weight of U.S. stocks to 70% of the equity portfolio. Therefore the contribution of the dynamics in (priced) conditional skewness and kurtosis brought forth by the MS process is crucial in biasing the optimal choices of a U.S. investor toward domestic securities. Interestingly, GT's paper is based on a strategy that incorporates MS in an asset pricing framework, in their case the ICAPM, and then derives asset allocation implications from it. However, it is clear that their model also represents an example of how MS dynamics may be incorporated in the SDF, a topic that will be developed in third section.[19]

An approach that allows to entertain large asset menus and to integrate large numbers of MS parameters into predictive densities that would account for all the estimation uncertainty characteristic of a portfolio problem is the Bayesian one, as popularized by Barberis (2000) with reference to simple linear VAR models. Incidentally, simulation-based Bayesian methods also appear to be potentially fit to deal with the calculations required by relatively large asset menus. One example of such a line of work is Tu (2010) who proposes a Bayesian framework for analyzing portfolio decisions under a simple two-state MS. In particular, Tu incorporates MS pricing model uncertainty into portfolio decision making, in the sense that MS concerns also the alphas (to be interpreted as measure of abnormal portfolio performance within a standard three-factor linear framework) that characterize portfolios in the asset menu vs. benchmark portfolios. In this sense, also Tu's paper an interesting mix of MS asset pricing and portfolio choice, similarly to GT (2008). Tu considers an asset menu composed of a large number of assets identified with a set of 28 risky portfolios (the 25 size- and book-to-market sorted U.S. equity portfolios, besides 1-month *T*-bills, the value-weighted market portfolio, the size factor portfolio, and the value factor portfolio) which is still amenable to

performing mean-variance calculations. With reference to monthly data for the sample 1963–2006, he finds that asset returns generally have higher means and lower standard deviations in the bull regime than in the bear regime. The bull regime appears to be more persistent and therefore to have a higher ergodic probability, as it accounts for approximately two-thirds of the entire sample period. Also correlations and betas between the 25 nonbenchmark portfolios and the three factor portfolios are regime-dependent, with sizable cross-regime differences. There is evidence of mispricing in both bull and bear regimes. The economic value of accounting for regimes is substantial, independently of whether model uncertainty and parameter uncertainty are incorporated. In particular, the CER losses associated with ignoring MS are generally above 2% per year, and can be as high as 10%. However, Tu's paper is based on a mean-variance assumption for preferences which makes his results completely static, and it is not clear how these could generalize to preferences that imply nonzero hedging demands. Yet, it represents a first example of a type of application of MS methods that is likely to gain increasingly popularity.

DYNAMIC ASSET PRICING MODELS UNDER MARKOV SWITCHING

The most recent strand of the financial economics literature that has resorted to MSMs to capture the presence of instability and regime shifts in the data is the asset pricing literature. In fact, even the early literature from the late 1980s and the 1990s was rich of cases in which empirical papers, besides estimating MSMs, also went further using their very MS framework to test asset pricing restrictions. For instance, Engel and Hamilton (1990) relate their MS estimates to the test of the uncovered interest parity model of exchange rate determination. In fact, their seminal work was to be later extended by contributions with a more distinct rational asset pricing slant. For instance, Evans and Lewis (1995a) have taken steps from Engel and Hamilton's evidence to argue that their results made it likely that traders in the market could anticipate shifts between regimes. If so, traders' expectations in turn affect the behavior of forward rates relative to observed spot rates. In fact, when Evans and Lewis estimate an MSM (with observable states) for the dollar exchange rate against the German mark, the British pound, and the Japanese yen, the restriction that the actual forward premium equals the foreign exchange risk premium plus the expected change

in the exchange rate implied by a simple model ($e_t = \phi E_t[e_{t+1}] + y_t(S_t)$, where e_t is the logarithm of the exchange rate and y_t the logarithm of the composite effects of variables that affect the exchange rate) cannot be rejected.

A related paper is Bekaert, Hodrick, and Marshall (2001, BHM), who tackle the failure of the Expectations Hypothesis of the term structure of (riskless) interest rates focussing on *peso problems* – determined by the possibility that the ex post frequencies of states within a sample may differ substantially from their ex ante probabilities, so that sample moments may not coincide with population moments (see, e.g., Barro, 2006) in small-sample inference. BHM assume a three-state MSM which is a simplification of Gray's MS generalized square root model with TVTPs driven by a logistic functional form restricted to capture the fact that observed short rates move gradually. Suppose that short interest rates can evolve through three different regimes, with the mean and volatility of interest rates all increasing as we move across regimes. Further, suppose that any shock that increases (decreases) the short rate also increases the probability of switching to a higher rate (lower rate) regime. Then, as short rates rise, the term spread may rise as agents rationally forecast transitions into a higher rate regime. But, if in a particular sample the higher rate regimes are observed less frequently than their unconditional probabilities, this increase in the spread will appear unjustified ex post. In such a sample, a typical Campbell–Shiller regression of long-term rates on yield curve spreads will fail to deliver an estimated slope of unity and could produce negative coefficients if increases in the spread are subsequently followed by surprising transitions to lower rate regimes. Although the single-state version of this model is well-known to be unable to generate population distributions for the term premiums that resolve the expectations hypothesis (EH) puzzles (see, e.g., Backus, Gregory, & Zin, 1989), BHM report that even their model of the peso problem cannot salvage the EH: whereas evidence against the EH is considerably weakened when peso issues are taken into consideration, the hypothesis is still rejected in joint tests of the Campbell–Shiller statistics. However, introducing small variation in term premiums dramatically increases the dispersion of the small-sample distributions for all statistics. As a result, the evidence against the model is weak for the 12- and 36-month horizons.

Recent years have certainly marked an acceleration in the number and quality of asset pricing papers that have placed MSMs at the heart of their efforts. In this section, I start by reviewing a few early papers that had speculated on the nature and implications of MS for the theory of financial

prices and belief formation and then proceed to describe a few recent papers that – especially in the subfield of default risk-free term structure modeling – have developed a consistent and structural approach to the role of MS dynamics in the process of asset price formation.

Explaining the Risk-Return Trade-Off

Despite its key role in many applications, estimating and understanding the market risk premium has proven very difficult. For instance, even though Merton (1980) has suggested estimating the risk premium based on the theoretical relationship between expected returns and the contemporaneous variance of returns, for a long-time empirical research has failed to document a significant positive relationship between average returns and the filtered/predicted levels of market volatility. In fact, a number of researchers have instead unveiled a negative relationship between volatility and market prices, the so-called *volatility feedback* effect. (see, Campbell & Hentschel, 1992; French et al., 1987), the intuitive idea that an exogenous change in the level of market volatility initially generates additional return volatility as stock prices adjust in response to new information about future discounted expected returns. For instance, in a modified GARCH-in-mean framework using post–World War II monthly data, Glosten, Jagannathan, and Runkle (1993) find that the estimated coefficient on volatility in a return/volatility regression is negative. Using similar data, when both conditional moments are estimated as functions of predetermined financial variables, Whitelaw (1994) finds that the long-run correlation between fitted moments is negative. Harrison and Zhang (1999) show that at longer horizons (i.e., 1–2 years) there is a significantly positive relation between expected returns and conditional volatility. The shorter horizon phenomenon is also present in international data. For example, De Santis and Imrohoroglu (1997) find a significant positive relation in only 2 countries out of a sample of 14 emerging and 3 developed markets.

Recently, Lettau and Ludvigson (2001) have provided evidence suggesting that the failure to find a positive relationship between excess returns and market volatility may result from the failure to control for shifts in investment opportunities. However, within applications of MSMs in financial economics, this idea dates back at least to a seminal paper by Turner, Startz, and Nelson (1989, TSN) that had traced a connection between MS as a time series technique and asset pricing theory. TSN introduce a model of the aggregate market portfolio in which the *excess*

return is drawn from a mixture of two normal densities because market portfolio returns are assumed to switch between two states. The states are characterized by the variances of their densities as a high-variance state and a low-variance state. The state is assumed to be generated by a first-order Markov process,

$$r_t = \mu_t + \varepsilon_t \quad \varepsilon_t \text{NID}(0, \sigma^2_{S_t})$$

where $\sigma^2_1 \geq \sigma^2_0$ and μ_t is discussed below. TSN develop two models that incorporate alternative assumptions about agents' information sets. In the first model, economic agents know the realization of the Markov state process, even though the econometrician does not observe it. There are two risk premiums in this specification. The first is the difference between the mean of the distribution in the low-variance state and the riskless return. Agents require an increase in return to hold an asset with a random return. The second premium is the added return necessary to compensate for increased risk in the high-variance state:

$$E[r_t|S_t] = \begin{cases} \mu_0 & \text{if} \quad S_t = 0 \\ \mu_1 & \text{if} \quad S_t = 1 \end{cases}$$

The parameter estimates from this model suggest that the first risk premium is positive and the second is negative, $\hat{\mu}_0 > 0$ and $\hat{\mu}_1 < 0$. Monthly data on S&P 500 index returns for 1946–1987 reveal that the two regimes identified by $\sigma^2_1 \geq \sigma^2_0$ and $\hat{\mu}_1 \neq \hat{\mu}_0$ are highly persistent, with median durations of 3 months for the high-variance regime and of 43 months for the low-variance one. Estimates of this simple MSIH model, in which agents are assumed to know the state, do not support a risk premium that increases with risk, which is puzzling, as commented earlier. Furthermore, not only is $\hat{\mu}_1$ significantly less than $\hat{\mu}_0$, it is also significantly negative. Therefore, TSN reject the hypothesis of a risk premium increasing in the variance. Misspecification is a likely explanation for this result. If agents are uncertain of the state, so that they are basing their decisions on forecasts of the state in the following period, estimates assuming they know the state with certainty will be inconsistent.

Accordingly, in their second model TSN assume that neither economic agents nor the econometrician observe the regime. In each period, agents form probabilities of each possible state in the following period conditional on current and past excess returns, and update their prior beliefs about that period's state using Bayes' rule. The parameter of interest is then the

increase in return necessary to compensate the agents for a given percentage increase in the prior probability of the high-variance state, θ:

$$\mu_t = \alpha + \theta \Pr(S_t = 1 | \mathscr{F}_{t-1})$$

where the constant, α, represents agents' required excess return for holding an asset in the low-variance state. TSN generalize slightly this model to

$$\mu_t = (1 - S_t)\alpha_0 + S_t\alpha_1 + \theta \Pr(S_t = 1 | \mathscr{F}_{t-1})$$

TSN are able to sign all the parameters in this simple empirical model. The stock price at time t should reflect all available information. This requires that the price at t should fall below its value at $t-1$ if some new unfavorable information about fundamentals, such as an increase in variance, arrives between $t-1$ and t. This fall is necessary to ensure that the return from time t to $t+1$ is expected to be higher than usual so as to compensate stockholders. According to this scenario, the return between $t-1$ and t will be negative on average for those periods in which adverse information is newly acquired, and positive on average when favorable information is acquired. This means that the coefficient θ attached to $\Pr(S_t = 1 | \mathscr{F}_{t-1})$ represents the effect when agents anticipate as of time $t-1$ that the return at time t will be drawn from the high-variance distribution. According to standard mean-variance theory, foreknowledge of a high-variance should be compensated by a higher expected return. The expected variance in this model is simply

$$E[\sigma_t^2 | \mathscr{F}_{t-1}] = [1 - \Pr(S_t = 1 | \mathscr{F}_{t-1})]\sigma_0^2 + \Pr(S_t = 1 | \mathscr{F}_{t-1})\sigma_1^2 + [1 - \Pr(S_t = 1 | \mathscr{F}_{t-1})]\Pr(S_t = 1 | \mathscr{F}_{t-1})(\alpha_1 - \alpha_0)^2$$

Thus when $\Pr(S_t = 1 | \mathscr{F}_{t-1}) \in (0, 1/2)$ is high, because

$$\frac{\partial E[\sigma_t^2 | \mathscr{F}_{t-1}]}{\partial \Pr(S_t = 1 | \mathscr{F}_{t-1})} = (\sigma_1^2 - \sigma_0^2) + [1 - 2\Pr(S_t = 1 | \mathscr{F}_{t-1})](\alpha_1 - \alpha_0)^2$$

is positive when $\Pr(S_t = 1 | \mathscr{F}_{t-1}) < 0.5$, the expected excess return should be positive so that the parameter θ is positive. On the contrary, it could be that today's high-variance state, $S_t = 1$, was not anticipated in the previous period. In this case $\Pr(S_t = 1 | \mathscr{F}_{t-1})$ is small so that the average return between $t-1$ and t is dominated by α_1. During a period in which agents are surprised by $S_t = 1$, the stock price must fall below what would have been seen had $S_t = 0$ occurred instead. This will make the return between $t-1$ and t lower and will show up as a negative value for α_1. Similar reasoning

suggests that if the variance unexpectedly decreases, the return between $t-1$ and t will turn out to be higher than usual, suggesting that α_0 is positive. TSN can also sign a linear combination of the parameters. The risk premium is

$$\mu_t = [1 - \Pr(S_t = 1|\mathscr{F}_{t-1})]\alpha_0 + (\alpha_1 + \theta)\Pr(S_t = 1|\mathscr{F}_{t-1})$$

If agents are risk-averse, this equation should always be positive and increase with $\Pr(S_t = 1|\mathscr{F}_{t-1})$. The expectation will always be positive as long as $\alpha_0 \geq 0$ and $\alpha_1 + \theta \geq 0$. Finally, if both these conditions hold with inequality and $\alpha_1 + \theta > \alpha_0$ then

$$\frac{\partial E[r_t|\mathscr{F}_{t-1}]}{\partial Pr(S_t = 1|\mathscr{F}_{t-1})} = \alpha_1 + \theta - \alpha_0 > 0$$

that is, the risk premium will increase with agents' prior probability of the high-variance state.

When estimated on S&P 500 monthly data, this model yields parameter estimates that are largely consistent with asset pricing theory. The estimates ($\hat{\alpha}_0 = 0.70\%$, $\hat{\alpha}_1 = -3.36\%$, and $\hat{\theta} = 2.88$) provide support for a risk premium rising as the anticipated level of risk rises. If the agents are certain that next period's return will be drawn from the low-variance density, agents anticipate a monthly return of 5%. Likewise, if agents are certain next period's return will be drawn from the high-variance density, they will require a return of 180% per year. These estimates suggest that agents perceive stocks to be a very risky asset during high-variance periods. The unconditional probability of the high-variance state is only 0.035, however. This suggests the risk premium will average approximately 9% on an annual basis. This number is close to the average excess return observed in the data, 7.5%. However, one problem remains: because $\hat{\alpha}_1 + \hat{\theta} - \hat{\alpha}_0 = -1.18 < 0$, the risk premium does not increase with the anticipated variance; the variance of the linear combination is large in relation to the point estimate, the t-statistic is -0.21, so that the model provides no evidence for a risk premium changing with or against the variance. This result is consistent with French, Schwert, and Stambaugh's (1987) who also find little evidence of a relation between the risk premium and volatility.

Mayfield (2004) has refined TSN's seminal intuition on the role played by a potential MS risk premium, a unit compensation for the additional risk caused by the presence of regime shifts in volatility, and developed a method for estimating the market risk premium based on the equilibrium

relationship between volatility and expected returns when there are discrete shifts in investment opportunities, as governed by a simple two-state MSM for volatility, where $\sigma_H^2 \geq \sigma_L^2$, $\Pr(S_{t+1} = L|S_t = L) = 1 - \pi_L$, $\Pr(S_{t+1} = H|S_t = H) = 1 - \pi_H$, and the Markov state $S_t = L,H$ is observable.[20] Mayfield solves the utility maximization problem of a representative investor in an infinite horizon, continuous-time model. Preferences are described by a power utility function. There are only two assets in which the investor can invest: a risk-free asset yielding a certain rate of return equal to r_t^f and a risky asset yielding a rate of return equal to $\mathrm{d}S_t/S_t$. The investor chooses an amount of consumption C_t and a fraction ω_t of her wealth W_t to invest in the risky asset:

$$\max_{C_t,\omega_t} E_v \int_v^\infty e^{-\rho s} \frac{C_t^{1-\gamma}}{1-\gamma} \mathrm{d}t \quad C_t > \bar{C}$$

$$\text{s.t. } \mathrm{d}W_t = \omega_t W_t \frac{\mathrm{d}S_t}{S_t} + (1-\omega_t) r_t W_t \mathrm{d}t - C_t \mathrm{d}t$$

$$\mathrm{d}S_t = \mu_t S_t \mathrm{d}t + \sigma_t S_t \mathrm{d}Z + J_t S_t \mathrm{d}N(\pi_t)$$

where $\mathrm{d}Z$ is a standard Wiener process, $\mathrm{d}N(l)$ is a Poisson process that equals either zero or one with intensity l, and the process followed by the drift, volatility, and the state probabilities are $\mathrm{d}\mu_t = 2(\bar{\mu} - \mu_t)\mathrm{d}N(\pi_t)$, $\mathrm{d}\sigma_t = 2(\bar{\sigma} - \sigma_t)\mathrm{d}N(\pi_t)$, and $\mathrm{d}\pi_t = 2(\bar{\pi} - \pi_t)\mathrm{d}N(\pi_t)$. Barred variables are simple averages of the state-dependent variables, for instance $\bar{\pi} = 0.5\pi_L + 0.5\pi_H$. The process of the drift μ_t is determined in equilibrium. When $\mathrm{d}N = 1$, this causes the drift, volatility, and transition parameters to jump to the alternative state. The parameter J_t is the magnitude of the jump in the stock price that occurs when the economy switches state. The value of the jump parameter J_t takes on two values, J_L and J_H such that $\mathrm{d}J_t = 2(\bar{J} - J_t)\mathrm{d}N(\pi_t)$. Using the principle of optimality, Mayfield proves that the optimal consumption-wealth ratio is

$$\frac{\hat{C}_t}{W_t} = \frac{\rho + (\gamma - 1)\mu_t - 0.5\gamma(\gamma - 1)\sigma_t^2}{\gamma(1 - \lambda_t)} + \frac{\pi_t}{\gamma}\left[1 - \frac{(1 + J_t)^\gamma}{(1 + K_t)^\gamma}\right]$$

where K_t is the jump in consumption that is expected conditional on the switching state, while the expression for the conditional equilibrium risk premium (imposing $\pi_t J_t = \pi_t' \ln(1 + J_t)$, that is, that over the expected duration of each volatility state, the continuously compounded expected

change in wealth is equal to the actual change in wealth associated with a change in state) is:

$$E_t\left[\frac{\mathrm{d}S_t}{S_t}\right] - r_t^f = E_t[r_{t+1}] - r_t^f = \gamma\sigma_t^2 + \hat{\pi}_t \ln(1+J_t)[1-(1+K_t)^{-\gamma}] \quad (2)$$

where $\hat{\pi}_t$ is a discrete time estimate of the probability of a switch. This result shows that the equilibrium risk premium in each state can be decomposed into two state-dependent risk premia, the sum of an intrastate risk premium and an interstate risk premium. The first term, $\gamma\sigma_t^2$, describes the required intrastate premium required to compensate for diffusion risk within the current state. The second term, $\hat{\pi}_t \ln(1+J_t)[1-(1+K_t)^{-\gamma}]$, describes the required interstate premium that compensates for potential jump risk arising from a change in volatility state. Notice that $E_t[r_{t+1}] - r_t^f < \gamma\sigma_t^2$ in the low-volatility state (i.e., $K_L < 0$) because investors expect a reduction in wealth when the economy enters the high-volatility state because of a positive precautionary savings motive; similarly, when the economy is in the high-volatility state $E_t[r_{t+1}] - r_t^f > \gamma\sigma_t^2$ (i.e., $K_H > 0$) because when the economy re-enters the low-volatility state, consumption jumps up.[21]

The expression for the equilibrium risk premium in Mayfield's model is a special case of Merton's (1973) intertemporal CAPM. Because individuals anticipate future changes in the volatility state and the corresponding changes in the level of stock prices, ex post measured returns are not equal to the ex ante expected returns. When individuals place a nonzero probability on the likelihood of a future change in the volatility state, the expected returns include the expected change in stock prices associated with a change in volatility state. While the economy remains in the low-volatility state, actual ex post returns are higher on average than the expected returns. Conversely, while the economy remains in the high-volatility state, actual ex post returns will be lower on average than the expected returns. Within each state, the difference between ex post returns and the expected returns is similar to the peso-type problem discussed in Rietz (1988). Therefore Mayfield's model generates periods of low-volatility and high ex post returns alternating with periods of high-volatility and low ex post returns, reconciling the empirical finding that returns may be lower in periods of high volatility with the theoretical intuition that the expected returns should be positively related to the level of market volatility.

Empirical estimation on value-weighted CRSP index data (in excess of one-month *T*-bills) of the MSM for volatility shows that market returns can be described as having been drawn from two significantly different distributions: a low-volatility/high-return distribution, from which about

88% of the returns are drawn, and a high-volatility/low-return distribution, from which about 12% of the returns are drawn. In the low-volatility state, the annual standard deviation of returns is 13.0% and the mean annualized excess return is 12.4%. In contrast, the annual standard deviation of returns in the high-volatility state is 38.2% and the mean annualized excess return is −17.9%. After having estimated the state-dependent moments and transition probabilities using standard techniques, Mayfield uses Eq. (2) and the optimal consumption process to find the corresponding values of γ, J_t, and K_t that are consistent with sample moments. Next, he uses the expression for the risk premium together with the estimated model parameters to calculate the intrastate and interstate components of the risk premium in each volatility state. The point estimate for the jump parameter J_L equals 29.6% and is significantly different from zero; the corresponding value of J_H is 42.1%. The implied values for the optimal percent change in consumption K_t in the low- and high-volatility states are 28.8% and 40.4%, respectively. The resulting estimates of the annualized state-dependent risk premia in the low- and high-volatility states are 5.2% and 32.5%, respectively. Based on the estimated preference parameters, about 50% of the unconditional risk premium is related to the risk of future changes in the level of market volatility.

Kim, Morley, and Nelson (2004, KMN) have tackled the volatility feedback puzzle of TSN using Campbell and Shiller's (1988) log-linear present value framework to derive an estimable model of stock returns under the assumption of MS volatility. The log-linear present value framework is used to derive an analytical expression for volatility feedback in terms of the parameters of the model:

$$r_t = E_{t-1}[r_t] \underbrace{- \left\{ E_t^* \left[\sum_{j=1}^{\infty} \kappa^j r_{t+j} \right] - E_{t-1} \left[\sum_{j=1}^{\infty} \kappa^j r_{t+j} \right] \right\}}_{\text{volatility feedback } (-f_t)} + \underbrace{\left\{ E_t^* \left[\sum_{j=0}^{\infty} \kappa^j \Delta d_{t+j} \right] - E_{t-1} \left[\sum_{j=0}^{\infty} \kappa^j \Delta d_{t+j} \right] \right\}}_{\text{news about dividends } (\varepsilon_t)}$$

or $r_t = E_{t-1}[r_t] - f_t + \varepsilon_t$, where d_{t+j} the log dividend at time $t+j$, r_{t+j} the log-return on a portfolio held from $t+j-1$ to $t+j$, and $\kappa \lessapprox 1$ the average ratio of the stock price to the sum of the stock price and the dividend. Notice that the starred time t expectations, $E_t^*[\cdot]$, denote expectations conditional on all information at time t except the final realized return r_t, which makes them different from the standard conditional operator $E_t[\cdot]$.

Differently from Campbell and Hentschel (1992), a time-homogeneous MSM for dividend news (ε_t) is deemed to be more plausible than GARCH specifications in the light of Hamilton and Susmel (1994) finding that most of the ARCH dynamics disappear at monthly frequencies when structural variance shifts are allowed. Implicitly, KMN use this evidence as an indication that an alternative theory for the negative empirical relationships between realized stock returns, the *leverage hypothesis*, is rejected by the data.[22] Under the leverage hypothesis a large movement in stock prices alters the debt/equity ratios of firms, changing their risk profiles and therefore, the volatility of future returns. The log-linear present value model implies that the expected return for a given period $t+j$ is a linear function of the market expectation about the volatility of news. Given this assumption and MS volatility, the expected return is assumed to be a linear function of the conditional probability of the high-volatility regime, $E_t[r_{t+j}] = \mu_0 + \mu_1 \Pr(S_{t+j} = 1|\mathscr{F}_t)$, where μ_0 is the expected return in a perfectly anticipated low-variance regime and μ_1 reflects the marginal effect on the expected return of a perfectly anticipated high-variance regime. Because this implies that

$$E_t[r_{t+j}] = \mu_0 + \mu_1 \Pr(S_t = 1) + \mu_1[\Pr(S_t = 1|\mathscr{F}_t)(p+q-1)^j - \Pr(S_t = 1)]$$

and

$$E_t^*\left[\sum_{j=1}^{\infty} \kappa^j r_{t+j}\right] = \frac{\mu_0}{1-\kappa} + \frac{\mu_1}{1-\kappa}\Pr(S_t = 1) + \frac{\mu_1}{1-\kappa(p+q-1)}[\Pr(S_t = 1|\mathscr{F}_t^*) - \Pr(S_t = 1)]$$

it is easy to show that

$$\begin{aligned} -f_t = \sum_{j=1}^{\infty} \kappa^j E_t^*[r_{t+j}] - \sum_{j=1}^{\infty} \kappa^j E_{t-1}[r_{t+j}] &= \frac{\mu_0}{1-\kappa} + \frac{\mu_1}{1-\kappa}\Pr(S_t = 1) \\ &+ \frac{\mu_1}{1-\kappa(p+q-1)}[\Pr(S_t = 1|\mathscr{F}_t^*) - \Pr(S_t = 1)] \\ &- \frac{\mu_0}{1-\kappa} - \frac{\mu_1}{1-\kappa}\Pr(S_t = 1) + \frac{\mu_1}{1-\kappa(p+q-1)}[\Pr(S_t = 1|\mathscr{F}_{t-1}) \\ &- \Pr(S_t = 1)] = \varphi\left[\Pr(S_t = 1|\mathscr{F}_t^*) - \Pr(S_t = 1|\mathscr{F}_{t-1})\right] \end{aligned}$$

so that

$$r_t = \mu_0 + \mu_1 \Pr(S_t = 1|\mathscr{F}_{t-1}) + \varphi[\Pr(S_t = 1|\mathscr{F}_t^*) - \Pr(S_t = 1|\mathscr{F}_{t-1})] + \varepsilon_t$$

where $\varepsilon_t \sim N(0, \sigma^2_{S_t})$ and $\varphi \equiv \mu_1/[1-\kappa(p+q-1)]$. Under a volatility feedback effect, if market volatility is persistent and positively related to

the equity premium, then stock prices should immediately move in the opposite direction to the level of market volatility, $\varphi<0$. Notice that a necessary and sufficient condition for a volatility feedback to occur is that $\mathscr{F}_t^* \supseteq \mathscr{F}_{t-1}$; when $\mathscr{F}_t^*$ and $\mathscr{F}_{t-1}$ are identical, $\Pr(S_t = 1|\mathscr{F}_t^*) = \Pr(S_t = 1|\mathscr{F}_{t-1})$ so that $r_t = \mu_0 + \mu_1 \Pr(S_t = 1|\mathscr{F}_{t-1})$ holds, as in Turner et al. (1989).[23]

Using monthly excess returns for the value-weighted NYSE equity portfolio to estimate the model under different assumption about the information available to economic agents, KMN find that the evidence of time-varying risk premia is mixed. When agents are assumed to observe only past returns (i.e., $\mathscr{F}_t^*$ and $\mathscr{F}_{t-1}$ are identical but do not include S_t), an LRT for the null hypothesis of a constant mean has a p-value of 0.38 for a 1926–1951 sample but it is 0.02 for a 1952–2000 sample. When agents are assumed to observe the true volatility regime (i.e., $\mathscr{F}_t^*$ and $\mathscr{F}_{t-1}$ are identical and include S_t), there is strong evidence of a time-varying mean, with p-values of 0.03 and less than 0.01 for the two samples. However, in this case, the estimated trade-off between the mean and variance is significantly negative, that is, the high-volatility regime is the one with a negative risk premium. When investors are assumed to know the prevailing volatility regime only by the end of each month as in Mayfield (2004), that is, $\mathscr{F}_t^* \supseteq \mathscr{F}_{t-1}$, which is plausible given the availability of data at frequency higher than daily – there is statistically significant evidence of a positive relationship between market volatility and the equity premium. The volatility feedback effect is significantly negative for both sample periods when the feedback parameter is unrestricted, $\hat{\varphi}<0$. The estimated partial effect, $\hat{\mu}_1$, is even positive, though not significant, for the 1952–2000 sample. Since volatility regimes appear to be very persistent (i.e., $\hat{p} + \hat{q} - 1 > 0$), these results provide strong support for a positive relationship between market volatility and the equity premium. Similarly, when the restriction $\varphi = \mu_1/[1 - \kappa(p + q - 1)]$ is imposed, the estimated relationship is always positive. The estimated partial effect, $\hat{\mu}_1$, is positive when φ is restricted. KMN's conclusion is therefore that accounting for volatility feedback is important to avoid confusing a negative relationship between return volatility and realized returns with the underlying feedback relationship between market volatility and the equity premium.

Bae, Kim, and Nelson (2007, BKN) have used an MS GARCH framework to explicitly test two competing explanations for the negative relationship between average market returns and their conditional volatility, that is, (i) Black's (1976) leverage story and (ii) volatility feedback, by which *if volatility is persistent and priced*, an increase in volatility raises the expected future volatility and thus the required return on stocks. BKN argue

that the test results in Turner et al. (1989), Kim et al. (2004), and Mayfield (2004) – who, as we have seen, all find evidence in support of a volatility feedback effect – are unsatisfactory because they fail to distinguish between leverage and feedback effects. Moreover, the monthly estimates typical of the literature are puzzling in the light of the evidence (see, e.g., Hamilton & Susmel, 1994) that ARCH effects in equity returns die out within a month when volatility is MS, so that the usefulness of ARCH may be doubtful. However, MS volatility models in which volatility is constant within each regime may also ignore a potential source of asymmetric volatility stemming from changing conditional volatility within a regime. As a result, BKN propose a model of asymmetric volatility that identifies leverage and volatility feedback effects by controlling for the actual change in the leverage ratio and that is based on MS GARCH, a framework that nests both MS ARCH and MSIH. BKN take steps from an accounting identity by which the sum of equity and debt equals total assets: The return on total assets is then identical to a value-weighted sum of returns on equity and debt. An assumption of risk-free debt and some manipulations of the identity yields the following relationship between excess returns on total assets and excess returns on equity, $r_t \equiv l_t r_t^a$, where r_t is returns on equity in excess of the risk-free rate, r_t^a the returns on total assets in excess of the risk-free rate, and l_t the one plus the leverage ratio (the value of debt divided by the value of equity). Since the leverage effect matters only for equity, BKN first formulate asymmetric volatility at the level of total assets and interpret the asymmetric volatility as a volatility feedback:

$$r_{t+1}^a = \mu_1 E_t[\sigma_{t+1}^2] + \mu_2(E_t[\sigma_{t+1}^2] - E_t[\sigma_t^2]) + \varepsilon_{t+1} \quad \varepsilon_{t+1} \sim N(0, \sigma_{t+1}^2)$$

$$\sigma_{t+1}^2 = \omega_{S_{t+1}} + \alpha\varepsilon_t^2 + \delta I_{\{\varepsilon_t \leq 0\}}\varepsilon_t^2 + \beta\sigma_t^2 \quad S_{t+1} = 1,\ 2 \text{ and } \omega_1 > \omega_2$$

If volatility is priced as implied by volatility feedback, market participants require compensation for the risk associated with volatility, which can be expressed as $\mu_1 E_t[\sigma_{t+1}^2]$. Evidence of a significant and positive μ_1 reflects that risk, measured by volatility, and return are positively related, and support volatility feedback. Similarly to Mayfield (2004), market participants face another risk: a possible change in the volatility regime. Suppose that the market is in the low-volatility regime at time t. There is a nonzero possibility of switching to the high-volatility regime at time $t+1$. Higher volatility increases the risk premium associated with volatility by decreasing stock prices. Facing a possible capital loss at time $t+1$, market participants require at time t compensation for risk, $\mu_2(E_t[\sigma_{t+1}^2] - E_t[\sigma_t^2])$, the expected

returns associated with the market participants' revision of the volatility regime, where μ_2 is expected to be positive.

In BKN's model, MS effects simply involve the GJR GARCH(1,1) constant, $\omega_{S_{t+1}}$, and therefore only affect implied unconditional variance, $\omega_{S_{t+1}}/(1-\alpha-\delta/2-\beta)$. If $\omega_1=\omega_2$, then the volatility process collapses to a typical asymmetric GARCH process with only one volatility regime; if $\alpha=\delta=\beta=0$, the model reduces to the MS variance model considered by Turner et al. (1989) and Kim et al. (2004). $\delta>0$ suggests the existence of a negative correlation between returns and subsequent volatility. The negative correlation can be interpreted as volatility feedback if intraregime conditional volatility is persistent and priced. BKN call this volatility feedback channel of asymmetric volatility as *intraregime* volatility feedback. Although changes in conditional volatility within regimes may induce volatility feedback, shifts in the volatility regime provide an additional source which they call *interregime* volatility feedback. In order to capture the negative relationship between returns and subsequent volatility regimes, BKN also make the transition probabilities governing the dynamics of S_{t+1} dependent on past return innovations and assume that the TVTPs have a logistic functional form.[24] Finally, BKN combine the volatility feedback model formulated at the level of total assets with the financial leverage ratio to produce a model for equity returns, which incorporates interactions between leverage and volatility feedback effects:

$$r_{t+1}=l_t\mu_1 E_t[\sigma^2_{t+1}]+l_t\mu_2(E_t[\sigma^2_{t+1}]-E_t[\sigma^2_t])+u_{t+1}\quad u_{t+1}\sim N(0,l_t^2\sigma^2_{t+1})$$

where the expression for σ^2_{t+1} is identical to the one reported earlier. This model enables BKN to assess the relative contribution of each type of volatility to asymmetric volatility. Because volatility feedbacks can account for the response of stock prices to a change in volatility that persists but not to those that do not persist, one may speculate that MS may be able to account for persisting changes in volatility while GARCH picks up less persistent changes within regimes. Under this assumption, BKN's framework represents a natural test for volatility feedback against the alternative that there is some other, not well-understood source of asymmetric volatility, having controlled for leverage.

Using monthly 1952–1999 returns on the value-weighted NYSE index in excess of 1-month U.S. *T*-bill yields and data on the leverage ratio for this portfolio constructed as in Schwert (1989), BKN assess the importance of each potential source of a negative relationship between expected returns and volatility, by first estimating by ML the model under the restriction

$\omega_1 = \omega_2$. The resulting model is an asymmetric GARCH model with no MS volatility. The model allows for volatility feedback stemming from changing conditional volatility but no interregime volatility feedback. The point estimate of δ is 0.256 with a standard error of 0.126, suggesting that there is a significant negative correlation between returns and subsequent volatility. The negative correlation could be interpreted as volatility feedback, if the volatility process is persistent and the price of risk is positive. The estimate of the price of risk parameter μ_1 is significantly positive. With a 1987 crash dummy variable included, the volatility process is also reasonably persistent, as measured by $\hat{\alpha} + \hat{\beta} + \hat{\delta}/2 = 0.79$. Thus, in the absence of MS, the volatility feedback arising from changing conditional volatility may play an important role in explaining asymmetric volatility in returns, which is consistent with the single-regime results in Campbell and Hentschel (1992) and Bekaert and Wu (2000). To examine the implications of MS volatility, the full model with $\omega_1 \neq \omega_2$ is then estimated (using a Gray-type approximation in the MS GARCH case). The α estimate is zero as above, but MS volatility significantly affects the δ and β estimates: the point estimate of δ is 0.189 in the absence of MS but almost zero in the presence of MS; β is estimated to be 0.694 in its absence but almost zero in its presence. This has two important implications. First, as already discussed, in the presence of MS, conditional volatility within regimes is neither persistent nor correlated with returns at all, suggesting that GARCH type volatility is not an important source of asymmetric volatility. That is, once MS is accounted for, persistence of volatility within regimes almost disappears. Second, the potential of MS to act as a source of asymmetric volatility is not affected by the presence of varying conditional volatility within regimes. Although intraregime volatility feedback does not exist, BKN results support interregime volatility feedback; there is evidence of negative correlation between return innovations and the subsequent volatility regime, with negative shocks increasing the probability of switching from the low-volatility regime to the high-volatility regime.[25] After controlling for the leverage effect, interregime volatility feedback weakens intraregime volatility feedback, indicating that recurrent regime shifts are the main source of the negative correlation between returns and subsequent volatility. Market participants also require additional compensation associated with expected changes in the volatility regime, consistent with Mayfield's (2004) results: the point estimate of the revision to the volatility regime parameter μ_2 equals 0.610 and is significantly different from zero at the 5% level.

Kim, Morley, and Nelson (2005, KMN) have performed related tests but in a Bayesian framework that admits structural breaks in the process of

value-weighted NYSE excess stock returns. Although in a classical framework model comparisons and tests for structural breaks with unknown breakpoints are complicated because of the presence of nuisance parameters under alternative hypotheses (see Guidolin, 2011, for a review of the issues involved), in a Bayesian framework, nuisance parameters do not pose any special problem as long as they, along with the other parameters, can be integrated out of the likelihood function to solve for the marginal likelihood. More generally, a Bayesian approach based on the Gibbs sampler provides a computationally feasible approach to the estimation of models that feature both Markov regimes and structural breaks with unknown breakpoints. KMN consider four models of excess stock returns. First, a model that assumes a constant equity premium and a constant level of market volatility within each structural regime (i.e., excess returns are IID normal within each subsample between structural breaks), $r_t = \mu + \varepsilon_t$, $\varepsilon_t \sim \text{NID}(0, \sigma^2)$. Second, a model that assumes a constant equity premium within each structural regime, but that allows a two-state MS market volatility process, $r_t = \mu + \varepsilon_t$, $\varepsilon_t \sim \text{NID}(0, \sigma^2_{S_t})$, $S_t = L, H$. Third, a model that assumes that the equity premium changes in response to MS volatility within each structural regime:

$$r_t = \alpha + \beta E_{t-1}[\sigma_t^2] + \varepsilon_t \quad \varepsilon_t \sim \text{NID}(0, \sigma^2_{S_t})$$

where the intercept $\alpha \geq 0$ allows the marginal price of risk, measured by the slope $\beta \geq 0$, to differ from the average price of risk. Fourth and final, a model in which the equity premium responds to volatility within each structural regime and also accommodates volatility feedback effects:

$$r_t = \alpha + \beta E_{t-1}[\sigma_t^2] + \varphi(\sigma_t^2 - E_{t-1}[\sigma_t^2]) + \varepsilon_t \quad \varepsilon_t \sim \text{NID}(0, \sigma^2_{S_t})$$

where $(\sigma_t^2 - E_{t-1}[\sigma_t^2])$ is the volatility feedback term and φ the volatility feedback coefficient. Given a persistent volatility process ($p + q > 1$), stock prices should initially move in the opposite direction to a change in the equity premium. Because the price movement generated by the change in volatility reflects the effect on all future discounted expected returns, rather than just the partial effect on the current period equity premium, KMN constrain the sign of the volatility feedback effect to be opposite to the change in the equity premium (i.e., $\beta\varphi \leq 0$). Breaks in some (all) the parameters are modeled to occur at unknown and estimable breakpoints. The timing of a breakpoint τ_i, $i = 1, \ldots, n$, where n is the total number of

breakpoints, is determined by a latent two-point variable, D_t ($=0, 1$), which follows a $(n+1)$-state MS process with constrained transition probabilities as in Chib (1998). The transition probabilities for D_t are given by $\Pr(D_t = j|D_{t-1} = j) = \kappa_j$, $\Pr(D_t = j+1|D_{t-1} = j) = 1 - \kappa_j$, and $\Pr(D_t = j'|D_{t-1} = j) = 0$ for $j' \neq j$, $j+1$ with $\kappa_j \in (0, 1)$ and $\kappa_{n+1} = 1$ (i.e., the last structural regime is "absorbing").[26]

The empirical Bayes factors – ratios of marginal posteriors under different model configurations – strongly favor the three models that incorporate MS over the simple IID model. KMN's most preferred model relates the equity premium to MS changes in the level of market volatility and also accommodates volatility feedback effects. The evidence in favor of a time-varying equity premium depends crucially on whether volatility feedback is taken into account. For all four models, there is strong evidence of only one break that leads to a permanent reduction in the general level of stock market volatility in the 1940s. However, because the equity premium depends on the level of volatility in the most preferred model, the structural break in the volatility process also leads to a break in the equity premium. KMN's estimated annualized equity premium drops from around 12% before the structural break to about 9% after the break.

Also Whitelaw (2000) has investigated the issue of whether it is sensible to find a negative relationship between market expected returns and conditional variance. Whitelaw re-examines this questions using a representative agent, Lucas-type exchange economy in which consumption growth is modeled as an MS autoregressive process with two regimes and in which the probability of a regime shift is modeled as a function of the level of consumption growth, under TVTPs. When the parameters are estimated by ML using monthly consumption data over the period 1959–1996, the two-regime MSM is able to identify the expansionary and contractionary phases of the business cycle consistently with NBER business cycle dating. Moreover, the model generates results that are broadly consistent with the empirical evidence: Expected returns and conditional volatility exhibit a complex, nonlinear relation; they are negatively related in the short run and this relation varies widely over time. In marked contrast, a single-regime consumption CAPM (CCAPM) calibrated to the same data generates a strong positive, and essentially linear, relation between expected returns and volatility. Therefore Whitelaw concludes that the empirical evidence is consistent with reasonable parameterizations of a relatively simple MS version of the CCAPM. The intuition for such a finding is that under CRRA preferences, the equity risk premium is a function of the correlation between

equity returns and the SDF, here the intertemporal marginal rate of substitution (IMRS):

$$E_t[r_{t+1} - r_t^f] = -r_t^f \sigma_t[r_{t+1}] \sqrt{\mathrm{Var}_t[\mathscr{M}_{t+1}]} \mathrm{Corr}_t[r_{t+1}, \mathscr{M}_{t+1}]$$
$$= -r_t^f \sigma_t[r_{t+1}] \sqrt{\mathrm{Var}_t[\mathscr{M}_{t+1}]} \mathrm{Corr}_t \left[\frac{c_{t+1}}{c_t} \frac{\psi_{t+1} + 1}{\psi_t}, \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} \right]$$

where ψ_{t+1} is the time-varying price-dividend ratio (PDR). The conditional moments of returns will be positively related (period by period) as long as the correlation between $\mathscr{M}_{t+1} \equiv (c_{t+1}/c_t)^{-\gamma}$ and equity returns is negative. Holding the PDR constant (for the sake of argument), this condition holds (for $\gamma > 0$) since $\mathrm{Corr}_t[c_{t+1}/c_t, (c_{t+1}/c_t)^{-\gamma}] < 0$ However, the IMRS depends only on next period's consumption growth, whereas the equity return depends on the infinite future via its dependence on the stock price next period (via ψ_{t+1}). The only way to duplicate the salient features of the data (i.e., weak or negative short-run and positive long-run relations between expected returns and volatility) is to formulate a model in which variation in the PDR partially offsets the variation in the dividend growth component of the equity return in some states of the world. Whitelaw observes that these requirements are difficult, if not impossible, to overcome if consumption growth follows a simple ARMA process. The correlation will vary little over time because the PDR, which is an expectation of future consumption growth, will be less variable than consumption growth itself. Moreover, correlations will be relatively stable because both the immediate and distant future depend on a limited number of past values of consumption growth.[27] Whitelaw notices that on the opposite, one appealing way to overcome these problems is to consider an MSM with TVTPs. For regimes that are sufficiently far apart in terms of the time-series behavior of consumption growth, the regime switching probability will control the conditional volatility of returns. That is, states with a high probability of switching to a new regime will have high volatility. At the same time, however, increasing the probability of a regime switch may decrease the correlation between equity returns and the SDF, thus reducing the risk premium. This second effect will occur because the PDR, which depend on expected future consumption growth, will be related to the regime not to short-run consumption growth. Put differently, regime shifts introduce large movements in the investment opportunity set, and therefore induce a desire among investor's to hedge adverse changes.

In papers such as Whitelaw (2000), in order to preserve tractability, the MS specification is kept simple. As a consequence, the reduced form model,

while providing insights into the relation between risk and return, fails to match other features of the equity return data. For example, the magnitude of the equity premium in Whitelaw (2000) is low. Recently, a number of papers have extended these intuitions in richer and more realistic frameworks. For instance, David and Veronesi (2009) consider a three-state MSI model transitory "good" and "bad" regimes and a more persistent "normal" regime in fundamentals. Asset prices are dominated by directional information and so are lowest in the bad regime and highest in the good regime. However, uncertainty is also the highest in the good and bad regimes, due to the low probability of remaining in these regimes, and lowest in the normal regime. This creates a V-shaped relation between return volatility and valuation measures which in turn can give rise to an inverse V-shaped relation between volatility and expected returns. Calvet and Fisher (2007) develop a parsimonious CCAPM with shocks of heterogeneous durations to capture the fact that equity prices are driven by news with heterogeneous degrees of persistence, ranging from intraday intervals to several decades. They present an MS multifractal model, which is a stochastic volatility model characterized by a small number of parameters but an arbitrarily large number of frequencies, as in Calvet and Fisher (2001, 2002). Under this specification, volatility is hit by exogenous shocks with heterogeneous durations, which range from one day to more than a decade in empirical applications. That is, given a state vector $\mathbf{S}_{t+1}$ that contains K elements, $S_{j,t+1} = 1, 2, \ldots, K$, volatility follows the process

$$\sigma^d_{S_{t+1}} = \bar{\sigma}^d \left(\prod_{j=1}^{K} S_{j,t+1} \right)^{1/2}$$

where $\bar{\sigma}^d > 0$. This specification permits each component to evolve at a separate frequency. For example, the first component may have transitions measured in years or even decades, corresponding to low-frequency shocks to technology or demographics, medium-run components might represent business cycle fluctuations, and high-frequency components may capture liquidity or other transient effects. The number of components and their frequencies can be inferred directly from returns data. To maintain parsimony, Calvet and Fisher assume that the components of S_{t+1} evolve independently; further, the $S_{j,t+1}$ are first-order Markov, and for each component $S_{j,t+1}$, a single parameter ρ_j controls the persistence of innovations. The components are constructed recursively as follows. Given a value $S_{j,t}$ for the jth component at date t, the next-period multiplier $S_{j,t+1}$ is either (i) drawn from a fixed distribution with probability ρ_j, or (ii) left unchanged. The volatility components therefore differ in their transition

probabilities but not in their common marginal distribution.[28] MS multifrequency models are sufficiently flexible to account for market conditions that change considerably over a long-time span and captures the outliers, volatility persistence, and power variation of financial series.

In Calvet and Fisher (2007), an Epstein–Zin consumer receives an exogenous consumption stream and prices a flow of dividends under multifrequency MS. The agent receives an exogenous consumption stream C_t, with the continuously compounded growth rate of consumption following a random walk with constant drift and volatility (called μ^c and σ^c), where the shocks to consumption growth are IID Normal. The model separates stock dividends from aggregate consumption growth and the SDF; this assumption is consistent with the imperfect correlation between real consumption growth and real dividend growth. The log-dividend follows a random walk with state-dependent drift and volatility:

$$\ln D_{t+1} - \ln D_t = \mu^d_{S_{t+1}} - \frac{1}{2}(\sigma^d_{S_{t+1}})^2 + \sigma^d_{S_{t+1}}\varepsilon^d_{t+1} \quad \varepsilon^d_{t+1} \sim \text{NID}(0,1)$$

where ε^d_{t+1} is correlated with shocks to consumption growth (with coefficient $\rho_{c,d}$), and S_{t+1} a first-order Markov state that may take K different values. The Ito's term $-0.5(\sigma^d_{S_{t+1}})^2$ guarantees that expected dividend growth $E_t[D_{t+1}/D_t]$ is controlled only by $\mu^d_{S_{t+1}}$. This tight parameterization induces tractable expressions for equilibrium prices, return dynamics, and filtered probabilities. When the representative agent directly observes the true state of the economy, the SDF is

$$\mathcal{M}_{t+1} = \beta\{(E_t[C_{t+1}/C_t])^{1-\gamma}\}^{1/(\alpha-1)}\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}$$

an expression that is proportional to the SDF obtained under expected utility ($\alpha \equiv (1-\gamma)/(1-1/\psi) = 1$) suggesting that the elasticity of intertemporal substitution (EIS) affects the interest rate but not the price of risk. Given this SDF, the real interest rate $r_f = -\ln E_t[\text{SDF}_{t+1}]$ is constant and obeys the familiar relationship

$$\begin{aligned} r_f &= -\ln\beta + (\alpha-1)\ln E_t\left[\left(\frac{C_{t+1}}{C_t}\right)^{1-\gamma}\right] + \ln E_t\left[\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}\right] \\ &= -\ln\beta + \frac{1}{\psi}\mu^c - \left(\gamma + \frac{\gamma-1}{\psi}\right)\frac{(\sigma^c)^2}{2} \end{aligned}$$

Similarly to Hung (1994), the equilibrium PDR is controlled by the Markov state, $P_t/D_t = \Psi_{S_t}$ so that the gross return on the stock is

$$\frac{P_{t+1}+D_{t+1}}{P_t} = \frac{D_{t+1}}{D_t}\frac{(P_{t+1}+D_{t+1})/D_{t+1}}{P_t/D_t} = \frac{D_{t+1}}{D_t}\frac{1+\Psi_{S_{t+1}}}{\Psi_{S_t}}$$

and satisfies the standard Euler condition

$$\left\{\left(E_t\left[\frac{C_{t+1}}{C_t}\right]\right)^{1-\gamma}\right\}^{1/(\alpha-1)} E_t\left[\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}\frac{D_{t+1}}{D_t}\frac{1+\Psi_{S_{t+1}}}{\Psi_{S_t}}\right]$$
$$= e^{-r_f}E_t\left[\frac{D_{t+1}}{D_t}\frac{1+\Psi_{S_{t+1}}}{\Psi_{S_t}}\right] = 1$$

The equilibrium price-dividend ratio therefore solves the fixed-point

$$\Psi_{S_t} = E_t\left[(1+\Psi_{S_{t+1}})\frac{D_{t+1}}{D_t}\right] = E_t[(1+\Psi_{S_{t+1}})e^{-r_f+\mu^d_{S_{t+1}}-\gamma\rho_{c,d}\sigma^c\sigma^d_{S_{t+1}}}]$$

Interestingly, when the volatility process $\sigma^d_{S_{t+1}}$ is persistent, a large standard deviation of dividend growth at a given date t implies a low contemporaneous PDR. High volatility therefore feeds into low asset prices for any choices of γ and of the EIS. By the fixed-point condition, the equilibrium PDR can be computed numerically for every possible state.

Calvet and Fisher conduct structural estimation by MLE on an index of U.S. equities over the period 1926–2003. Their model is partially calibrated, setting μ_c, ψ, and δ to match the long-run mean of the one-month real T-bill rate, and $\gamma\rho_{c,d}\sigma^c$ to match a plausible long-run value for the PDR of 25 (this implicitly pins down γ given estimates of $\rho_{c,d}$ and σ^c).[29] As far as the number of MS multifrequency components are concerned, Calvet and Fisher observe that the likelihood function appears monotonically increasing and concave in K, and tends to flatten markedly by the time they reach $K=8$. Because MS multifrequency specifications with different Ks are nonnested, Calvet and Fisher assess significance in the log-likelihood differences using a heteroskedasticity and autocorrelation consistent (HAC) version specification test proposed in Calvet and Fisher (2004). They find that the test is significant at the 1% level for $K\leq 6$, and at the 5% level for all specifications. The model generates volatility feedback that, under sensible parameterizations, is substantially higher than in previous research. In particular, the strength of the feedback rises with the number of components and the likelihood function, increasing by 20–40% for the preferred specifications. As a result, the multifrequency equilibrium model generates

an unconditional feedback that is 10–40 times larger than in previous literature. Calvet and Fisher also compute an ex post decomposition of equity returns into conditional expectation, feedback innovation, and dividend news when the investor, consistently with econometrics practice, cannot observe the regimes. They report that while smoothed versions of the conditional expected return $E[r_{t+1}|\mathscr{F}_t]$ show small persistent variations, the smoothed feedback $E[r_{t+1}|\mathscr{F}_{t+1}] - E[r_{t+1}|\mathscr{F}_t]$ deriving from changes of regimes in the dividend volatility process appears in intermittent bursts. On most days it is small, but its occurrences coincide with the most substantial variations in the series, and on these days it contributes a large portion of realized returns. These features are consistent with the intuition that low-frequency volatility changes are infrequent but have a large price impact. For instance, this ex post analysis attributes over half of the 1987 crash to volatility feedback.

Markov Regimes or Bubbles?

Another subfield of modern asset pricing in which MS has made a considerable impact is in understanding the potential that simple present value models (PVMs), augmented by the presence of bubbles, may have for observed stock prices. In a PVM current stock prices are the present discounted value of expected future cash flows (net dividends) paid by the stock:

$$P_t = e^{-r} E_t[D_{t+1} + P_{t+1}]$$

where the real rate of discount r is assumed to be constant. Since the seminal papers by Le Roy and Porter (1981), Shiller (1981), and Grossman and Shiller (1981), finance students have become aware of the existence of large and persistent deviations of U.S. stock prices from those predicted by PVMs. Several explanations for this divergence have been proposed, including time varying discount factors, noise traders, and fads (see, e.g., Shiller, 1989). Although some of these explanations have been considered to be appealing, they fail to account successfully for the bulk of the divergence between fundamentals and stock prices. Froot and Obstfeld (1991) were more successful in using an *intrinsic bubble* component:

$$\mathscr{B}(D_t) = cD_t^{\lambda} \tag{3}$$

where the bubble is *rational* if and only if $\mathscr{B}(D_t) = e^{-r} E_t[\mathscr{B}(D_{t+1})]$, λ is the positive root of the quadratic equation $0.5\sigma^2\lambda^2 + \mu\lambda - r = 0$ and c an arbitrary constant. μ and σ are parameters from a simple Gaussian IID

process for real dividend growth assumed by Froot and Obstfeld, $\ln D_{t+1} - \ln D_t = \mu + \sigma\varepsilon_t$, $\varepsilon_t \sim \text{NID}(0,1)$. This bubble is a nonlinear function of current dividends, and thus it captures the apparent overreaction of stock prices to dividend changes. In fact, the no-arbitrage stock price under Eq. (3) is

$$P_t = \frac{1}{e^r - e^{\mu+0.5\sigma^2}} D_t + cD_t^{\lambda}$$

Although the fundamental stock price implies that the PDR is constant, the bubble allows it to be a function of the current dividend

$$\frac{P_t}{D_t} = \frac{1}{e^r - e^{\mu+0.5\sigma^2}} + cD_t^{\lambda-1} + \eta_t \tag{4}$$

where η_t is interpreted as a random measurement error. Given the strong evidence of MS in dividends, Driffill and Sola (1998) have asked whether Froot and Obstfeld might have incorrectly labeled as an intrinsic bubble some rational pricing components which in fact may derive from the presence of MS in the underlying dividend process. Of course, Driffill and Sola's point also relates to the issue of whether and how MS in fundamentals may lead to MS in real stock returns, but the punch-line of their paper was a set of conclusions strongly in support of the efficient market hypothesis (EMH) and against bubbles.

Using 1900–1987 annual U.S. data on real dividends and stock prices, Driffill and Sola find that a two-state MSIH for real dividend growth

$$\ln D_{t+1} - \ln D_t = \mu_0 + \mu_1 S_{t+1} + (\sigma_0 + \sigma_1 S_{t+1})\varepsilon_t \quad \varepsilon_t \sim \text{NID}(0,1)$$

accounts for most of the differences between stock prices and fundamentals, and also appears to provide a better characterization of the evolution of stock prices than the simple intrinsic bubble model without MS.[30] ML estimates reveal that the first regime is a low-growth/high-variance state, and the other regime is a high-growth/low-variance state. Score-based tests fail to reject the model while this occurs for the single-state model. Under a two-state MSIH for real dividends, the no-arbitrage stock price is given by

$$P_t = \kappa_s D_t + c_s D_t^{\lambda} \quad s = 0,1 \quad \kappa_j = e^{-r}[1 + q\kappa_0 e^{\mu_0+0.5\sigma_0^2} + (1 - q\kappa_1 e^{\mu_1+0.5\sigma_1^2})]$$

where c_0, c_1, and λ solve the system of equations:

$$\frac{c_1}{c_0} = \frac{e^r - qe^{\lambda\mu_0+0.5\lambda^2\sigma_0^2}}{(1-q)e^{\lambda\mu_1+0.5\lambda^2\sigma_1^2}} \quad \frac{c_1}{c_0} = \frac{(1-p)e^{\lambda\mu_0+0.5\lambda^2\sigma_0^2}}{e^r - pe^{\lambda\mu_1+0.5\lambda^2\sigma_1^2}}$$

This system has a unique positive solution for c_1/c_0 and λ. Driffill and Sola estimate a system for the PDR and the log of dividends:

$$\frac{P_t}{D_t} = \kappa_s + c_s D_t^{\lambda - 1} + \sigma_\eta^j \eta_t \quad s = 0,1 \quad \ln D_{t+1} - \ln D_t = \mu_s + \sigma_s \varepsilon_t \quad \varepsilon_t \sim \text{NID}(0,1)$$

in which the state S_{t+1} is treated as latent. This system is estimated by QML subject to the restrictions concerning κ_s, c_1/c_0, and λ, whereas r is assumed to equal to average sample real stock return, 8.16% per year, as in Froot and Obstfeld (1991). The estimated means and variances of real dividend growth rate are still separated into a state 0 of a low-mean and high-variance state and a state 1 of high-mean and low-variance; $\hat{\kappa}_0 \simeq 15$ and $\hat{\kappa}_1 \simeq 18$, respectively. The bubble coefficient c is higher in state 1 ($\hat{c}_1 \simeq 0.29$) than in state 0 ($\hat{c}_0 \simeq 0.15$). The elasticity of the bubble with respect to the dividend is $\hat{\lambda} \simeq 2.1$. The model shows no sign of ARCH errors or misspecifications. The bubble term accounts for most of the large deviations between fundamentals and stock prices, especially those in the second part of the sample. The hypothesis that there is no bubble in stock prices is rejected by an LRT; however, the improvement in fit obtained by including a bubble, once MS has been allowed for, appears to occur in periods in which the bubble element of the estimated stock price is small. This suggests that the bubble may simply be acting as a proxy for any additional omitted nonlinearities in the data. More interesting is a comparison between an MSM without the bubble and the bubble model without MS: both models account successfully for the boom in stock prices of the 1950s and 1960s and the collapse in the early 1970s. However, they offer two very different interpretations of these events. The intrinsic bubble diagnoses them as an over-reaction to the current dividend, whilst the MSM interprets the boom as a response of the present-value stock price to a change of regime into an era of rapidly growing dividends. The Akaike, Schwarz, and Hannan-Quinn information criteria all clearly favor the MSM over the bubble. Because intrinsic bubbles have the unappealing feature that they are permanent and their expected value is explosive, Driffill and Sola conclude that it is encouraging – and consistent with the tenets of the EMH – that MS provides a better explanation of the data than intrinsic bubbles do. It is also interesting to notice that the estimation of Driffill and Sola's two-state MSIH subject to the restrictions imposed by the PVM may also be taken as an early example of estimation of an MSM implicit in asset prices subject to no-arbitrage conditions, an empirical approach that has recently been greatly developed (see section "Structural Models of Markov Regimes in the SDF").

Psaradakis, Sola, and Spagnolo (2004, PSS) have later extended Driffill and Sola's (1998) results to the case of MS vector models of error correction of stock prices toward dividends. An assumption frequently made in the empirical finance literature is that the dynamic adjustment of prices toward their long-run steady state is linear. In fact, Campbell and Shiller (1988) show that simple PVMs imply that real stock prices and dividends ought to be cointegrated in the long-run, so that some sort of error correction mechanism should be at work. There are, however, good reasons to expect dynamic adjustment processes to be asymmetric and nonlinear. For instance, for many variables the speed of adjustment is unlikely to be the same during a recession and an expansion. MS vector error-correction models (MSVECMs) are flexible enough to accommodate situations where both regimes are error-correcting but the speed of disequilibrium adjustment is different, as well as situations where deviations from the long-run equilibrium temporarily follow nonstationary paths so that the system behaves as if cointegration has been "switched off." PSS have proposed to use a two-state MSVECM which allows for different rates of adjustment to long-run equilibrium, and in particular in which deviations from equilibrium tend to decay to the mean level of zero as long as $S_t = 1$; otherwise stock prices behave like a nonstationary process, and there is no tendency for the system to move toward equilibrium.[31] Using annual U.S. S&P real stock prices and dividends for the period 1900–1992, PSS find evidence in favor of MSVECM adjustment in which the estimated MS adjustment coefficients of -0.561 in regime 1 and -0.163 in regime 2 imply that the first regime is associated with very fast disequilibrium adjustment while the second regime is associated with very slow adjustment. In fact, the hypothesis that the second adjustment coefficient is zero cannot be rejected on the basis of a standard Wald test. The filtered probabilities reveal that the unstable regime is clearly associated with the periods 1959–1978 and 1987–1992. The filter correctly associates periods of low (or no) adjustment to deviations from long-run equilibrium with periods characterized by a high price-dividend ratio. This implies that the data would contain little evidence of rational bubbles, but only of MS dynamics in the adjustment of PDRs to their long-run steady-state.

Markov Regimes in Dynamic General Equilibrium Pricing Models

A number of papers have shown that – under a variety of asset pricing frameworks and assumptions concerning preferences and market

completeness – simple MSMs of fundamentals (such as real dividends, real consumption, short-term interest rates, output, etc.) underlying asset prices may generate the ubiquitous finding that most asset classes would be characterized by MS dynamics involving both conditional means and conditional variances (see Guidolin, 2011). On the one hand, this literature seems removed from the main bulk of empirical finance papers that have implemented MS estimation. In the chapters that I review in this subsection, the framework is extremely simple: to goal is simply to derive equilibrium asset prices and return processes from MS assumptions on fundamentals. Often the papers have only performed rough calibrations of the resulting models, in the form of estimation of simple MSMs for fundamentals and then simulated the model to re-produce key stylized facts concerning asset returns (in fact, this was already the scheme followed by Whitelaw, 2000). On the other hand, there is no doubt that even though these early papers proceeded in a mechanical way to assemble MS assumptions for the forcing variables to deliver models – often containing itself MS dynamics – of the SDF, this is the literature that has provided an essential background to the development of models of MS in SDFs, to be reviewed in section "Structural Models of Markov Regimes in the SDF."

The context for a number of the papers discussed later is a quest to propose models of preferences and/or the dynamics of beliefs (hence, models of the SDF) that could explain a number of asset pricing puzzles (below, I shall refer to them simply as "the puzzles") that had been discussed in the literature starting in the 1980s.[32] For instance, Mehra and Prescott (1985) had examined whether Lucas' (1978) asset pricing model could reasonably account for the average rates of return on stocks and short-term bills in the United States. They found that over the period 1889–1978 the average real rate of return was 6.98% per year on stocks but only 0.80% per year on short-term bills. The 6.18% per year excess return on stocks relative to bills – the equity premium – is much larger than Mehra and Prescott could account for using a simple version of Lucas model in which stocks are priced off the consumption process (i.e., assuming that the aggregate stock market pays out consumption flows directly, instead of dividends, the CCAPM). The inability of the CCAPM to account for the average equity premium has been dubbed the "equity premium puzzle." Subsequently, Weil (1989) had emphasized a "risk-free rate puzzle," the fact that models that can produced a large equity premium produced a riskless rate that was much higher than the historically observed average riskless rate.

Let us mention first, keeping details to a very minimum, a number of papers that since the late 1980s had ventured into an analysis of the effects of MS fundamentals for equilibrium asset returns. Abel (1994) has addressed the puzzles assuming that the conditional growth rates of consumption and dividends – assumed to be different – may depend on an underlying random state that evolves according to an MSM and derived simple closed-form solutions for the conditional and unconditional expected rates of return on stocks and short-term riskless bills.[33] However, Abel concludes that the unconditional riskless rate is higher under MS than under conditional IID shocks, which further exacerbates the risk-free rate puzzle. In addition, under conditional lognormality the added stochastic richness of MS fundamentals will reduce the equity premium predicted by a general equilibrium asset pricing model (GEAPM), and thus will exacerbate rather than resolve the equity premium puzzle. More precisely, Abel proves that if the conditional expected growth rates of consumption and dividends are positively correlated (as they must be in models that cannot distinguish between consumption and dividends), then introducing MS will *reduce* the unconditional equity premium under conditional lognormality and CRRA preferences, which makes the puzzles worse.

Another paper with negative results for the ability of MSMs to solve the puzzles is Bonomo and Garcia (1994). In the face of the apparent success of earlier papers such as Cecchetti, Lam, and Mark (1990, CLM) and Kandel and Stambaugh's (1990) at calibrating simple GEAPMs in which consumption equals dividends and follows an MS process to reproduce the predictability patterns in (aggregate) stock returns – for instance, the evidence of negative serial correlation at long horizons and the finding that ratios of variances of long- and short-horizon returns would decline as the horizon grows – Bonomo and Garcia show that one needs to be careful in not misspecifying the MS, least the effects on the dynamic asset pricing implications may be disruptive. In particular, they stress that if instead of imposing one particular MSM, one specifies a larger class of MSMs and lets each historical series of endowment decide on the best model according to various testing procedures, the chosen specification turns out to be a two-state MSM with constant mean and regime-dependent variances. Therefore they reject CLM's (1990) specification, an MS with regime-dependent means and constant variance, as well as Kandel and Stambaugh's (1990) specification with four regimes (two state for the mean and two states for the variance) for consumption growth. In a standard version of the

CCAPM, Bonomo and Garcia report that their GEAPM is unable to produce negative serial correlation in stock returns in the magnitude observed in the data, similarly to single-state models.

A paper that in my view deserves careful analysis is Hung (1994), because it pursues goals and solution methods common to many of the papers cited earlier, extending them to encompass a class of preferences (hence, SDFs) richer than power utility.[34] In particular, Hung works with nonexpected utility preferences à la Kreps and Porteus (1978) and develops a GEAPM in which market fundamentals follow a bivariate MSM.[35] In practice, the representative investor is assumed not to be indifferent to the timing of the resolution of uncertainty over temporal lotteries and her preferences are represented recursively by $V_t = U(C_t, E_t[V_{t+1}])$, where $U(\cdot,\cdot)$ is an aggregator function, and C_t the consumption. As shown by Kreps and Porteus, agents prefer early resolution of uncertainty over temporal lotteries if $U(\cdot,\cdot)$ is convex in its second argument; if $U(\cdot,\cdot)$ is concave in its second argument, agents will prefer late resolution of uncertainty over temporal lotteries. Hung specializes to the parametrization

$$\left[(1-\beta)C_t^{1-\rho} + \beta E_t[V_{t+1}]^{(1-\rho)/(1-\gamma)}\right]^{(1-\gamma)/(1-\rho)}$$

where β is the agent's subjective time discount factor, γ can still be interpreted as the Arrow-Pratt CRRA coefficient, and $1/\rho$ measures EIS. If the agent's CRRA coefficient γ is greater than the reciprocal of his EIS (ρ), then she prefers early resolution of uncertainty; if $\gamma = \rho$, the agent's utility becomes an isoelastic, von Neumann-Morgenstern utility function and she is indifferent to the timing of the resolution of uncertainty. In this case, the basic Euler condition for the value-weighted portfolio *of all assets* simplifies to

$$1 = \beta E_t\left\{\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} R_{t+1}^W\right\}$$

where R_{t+1}^W is the gross return on the wealth portfolio, which is equivalent to

$$P_t^W u'(C_t) = \beta E_t[(P_{t+1}^W + C_{t+1})u'(C_{t+1})] \tag{5}$$

when $u(C_t)$ is a standard power utility function. Eq. (5) simply states that a consumer can buy a share of stock in period t by giving up $P_t{}^W$ units of consumption in period t, thereby reducing utility by $P_t^W u'(C_t)$ in period t. In

the following period, the consumer receives the dividend C_{t+1} and can sell the share of stock at a price P^W_{t+1}; therefore, the consumer can increase consumption in period $t+1$ by $(P^W_{t+1} + C_{t+1})$ units, thereby increasing utility by $(P^W_{t+1} + C_{t+1})u'(C_{t+1})$, of which the consumer considers the expected value discounted one period. At an optimum, these two effects have to be identical. Epstein and Zin (1989) have shown that when in general $\gamma \neq \rho$, because the SDF is given by

$$\left[\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\rho}\right]^{(1-\gamma)/(1-\rho)} (R^W_{t+1})^{(1-\gamma)/(1-\rho)-1}$$

the Euler condition that pins down the price of any asset or portfolio i is:

$$1 = E_t\left\{\left[\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\rho}\right]^{(1-\gamma)/(1-\rho)} (R^W_{t+1})^{(1-\gamma)/(1-\rho)-1} R^i_{t+1}\right\} \tag{6}$$

I briefly summarize one method of solution of CCAPMs of this type, warning that although equivalent solution methods have been proposed (see, e.g., Abel, 1994), Hung's represents a very common approach. To derive equilibrium asset prices and returns under Epstein and Zin's preferences, one needs to start off with the pricing function for the wealth portfolio, which is assumed to pay off C_t. Denote its price and return by P^W_t and R^W_t, respectively. As in Mehra and Prescott (1985) and Weil (1989), a stationary equilibrium is assumed such that $P^W_t = \psi_{S_t} C_t$ where S_t is the state prevailing at time t and ψ_{S_t} an undetermined, regime-dependent PDR (in consumption) for the wealth portfolio (as in Cecchetti et al., 1990). The one-period rate of return for the wealth portfolio, if the current state is $S_t = i$ and the next period state is $S_{t+1} = j$ $(i,j = 0,1)$, is $R^W(i,j) = [(\psi_j + 1)/\psi_i]g_j$, where g_j is the consumption growth rate in state $S_{t+1} = j$. Substituting this definition into Eq. (6), the following equation is obtained:

$$\begin{aligned}
1 &= E_t\left\{\beta^{(1-\gamma)/(1-\rho)}[g^{-\rho}_{t+1}]^{(1-\gamma)/(1-\rho)}(g_{t+1})^{(1-\gamma)/(1-\rho)}\left(\frac{1+\psi_j}{\psi_i}\right)^{(1-\gamma)/(1-\rho)}\right\} \\
&= \Pr(S_{t+1} = 1|S_t = i)\left[\beta^{(1-\gamma)/(1-\rho)} g_1^{1-\gamma}\left(\frac{1+\psi_1}{\psi_i}\right)^{(1-\gamma)/(1-\rho)}\right] \\
&\quad + \Pr(S_{t+1} = 0|S_t = i)\left[\beta^{(1-\gamma)/(1-\rho)} g_0^{1-\gamma}\left(\frac{1+\psi_0}{\psi_i}\right)^{(1-\gamma)/(1-\rho)}\right]
\end{aligned}$$

By using the lognormal moment generator, this is equivalent to:

$$1=\beta^{(1-\gamma)/(1-\rho)}\exp[0.5(1-\gamma)^2\sigma_c^2]$$
$$\left\{p_{i1}\exp[(1-\gamma)g_1]\left(\frac{1+\psi_1}{\psi_i}\right)^{(1-\gamma)/(1-\rho)}+p_{i0}\exp[(1-\gamma)g_0]\left(\frac{1+\psi_0}{\psi_0}\right)^{(1-\gamma)/(1-\rho)}\right\}$$

These equations form a nonlinear system in which the ψ_is ($i=0$ and 1) are positive solutions if the equilibrium exists. It is possible to solve ψ_i as a function of the estimated parameters in the bivariate MS process, and the preference parameters γ, β, and ρ. After obtaining the solutions for ψ_i, the expression for returns on the market portfolio follows easily. The stock market pays D_t in period t. Denote its price and return by P^M and R^M, respectively. In a stationary equilibrium, the price for the stock market can be expressed as $P_t^M(i)=\kappa_i D_t$, where D_t is the dividend level at time t, i the state of nature at time t, and κ_i an undetermined PDR, $i=0,1$. Let $d_{t+1}=D_{t+1}/D_{t+1}$ the dividend growth rate at time t. Then the one-period rate of return for the stock market, if the current state is i and the next period state is j, can be written as $R^M(i,j)=[(\kappa_j+1)/\kappa_i]d_j$. Substituting this expression into Eq. (6), and following steps similar to what shown above the wealth process, implies

$$\kappa_i=\beta^{(1-\gamma)/(1-\rho)}\exp[0.5\gamma^2\sigma_c^2+0.5\sigma_d^2-\gamma\sigma_{cd}]$$
$$\left\{\Pr(S_{t+1}=1|S_t=i)\exp[-\gamma g_1+d_1]\left[\left(\frac{1+\psi_1}{\psi_i}\right)^{(1-\gamma)/(1-\rho)-1}(1+\kappa_1)\right]\right.$$
$$\left.+\Pr(S_{t+1}=0|S_t=i)\exp[-\gamma g_0+d_0]\left[\left(\frac{1+\psi_0}{\psi_0}\right)^{(1-\gamma)/(1-\rho)-1}(1+\kappa_0)\right]\right\}$$

These equations form a system of linear equations. The κ_i ($i=0$, 1) are positive solutions to these equations if the equilibrium exists. If the current state is $i=0$, 1, the regime-switching risk-free rate $R^f(i)$, can be calculated as follows:

$$\frac{1}{R^f(i)}=\beta^{(1-\gamma)/(1-\rho)}\exp[0.5\gamma^2\sigma_c^2]$$
$$\left\{\Pr(S_{t+1}=1|S_t=i)\exp[-\gamma g_1]\left[\left(\frac{1+\psi_1}{\psi_i}\right)^{(1-\gamma)/(1-\rho)-1}\right]\right.$$
$$\left.+\Pr(S_{t+1}=0|S_t=i)\exp[-\gamma g_0]\left[\left(\frac{1+\psi_0}{\psi_i}\right)^{(1-\gamma)/(1-\rho)-1}\right]\right\}$$

Finally, if we call $\bar{\xi}$ be the vector of stationary (ergodic) probabilities, the unconditional expected rate of return on equity is

$$E[R^M] = \bar{\xi}_0 \left[\Pr(S_{t+1}=1|S_t=0)\left(\frac{1+\kappa_1}{\kappa_0}\right)d_1 + \Pr(S_{t+1}=0|S_t=0)\left(\frac{1+\kappa_0}{\kappa_0}\right)d_0 \right]$$
$$+ (1-\bar{\xi}_0)\left[\Pr(S_{t+1}=1|S_t=1)\left(\frac{1+\kappa_1}{\kappa_1}\right)d_1 \right.$$
$$\left. + \Pr(S_{t+1}=0|S_t=1)\left(\frac{1+\kappa_0}{\kappa_1}\right)d_0 \right]$$

while the unconditional riskless rate is: $E[R^f] = \bar{\xi}_0 R^f(0) + (1-\bar{\xi}_0)R^f(1)$ and the unconditional equity premium is $E[R^M] - E[R^f]$.

After estimating his bivariate MS model on U.S. annual (real) data for the period 1889–1985, Hung finds that the combination of nonexpected utility preferences and of MS fundamentals over recessions and expansions are able to resolve the equity premium and risk-free rate puzzles. The ML estimates illustrate the presence of strong asymmetries in the process of fundamentals: the probability that consumption and dividends will be both in a high-growth state next period, given that they are in a high-growth state this period, is estimated to be 0.96. If dividends and consumption are presently depressed, there is instead only a probability of 0.30 that both will continue to be depressed. This asymmetric transition matrix implies that the unconditional probabilities of both consumption and dividends being in either a boom state or a depression state are 0.94 and 0.06, respectively. The resolution of the puzzles is then due to the following forces: First, using aggregate dividends rather than aggregate consumption as the payoff to the stock market decreases the covariance between the equity return and the IMRS and hence increases equity returns. Second, nonexpected utility allows a separation between EIS and the inverse of the CRRA coefficient and this is used to lower the real risk-free rate by decreasing the EIS. Hung also investigates the ability of his GEAPM to match the variances and covariance of the risk premium and the risk-free rate. Asymmetric market fundamentals are capable of matching the negative sample correlation between the risk premium and the risk-free rate. Although Hung's model is not capable of matching all first and second moments of the risk premium and the risk-free rate, the MS-driven model cannot be rejected using CLM's (1993) GMM-style tests. The parameter value for the EIS coefficient for which the model is not rejected is around 0.1, whereas the parameter value for γ that prevents rejections ranges from 7 to 20.

The importance of results such as Hung's and the role of MS in capturing the asymmetric dynamics of consumption and dividends over the business cycles, as well as their time-varying correlations is difficult to understate. Cecchetti et al. (1990) and Kocherlakota (l990) had already argued that consumption and dividend growth rates are less than perfectly correlated, and thus the return to the market portfolio and the return to the stock market portfolio are different. Consequently, instead of using a univariate degenerate process to govern consumption, they had adopted a bivariate, single-state VAR process for consumption and dividends. However, they all found that a GEAPM governed by a bivariate linear process is not capable of resolving the puzzles. Moreover, Cecchetti et al. (1990, CLM) had also experienced with simple, bivariate MSVAR models with regime switching mean growth rates, when preferences are described by simple, time-separable power utility functions in which CRRA is the inverse of EIS. Similarly to Abel (1994), they had reported that although MSVAR processes are nonlinear and hence able to capture asymmetries over the business cycle, the solution to the equity premium puzzle could not be found by employing asymmetric market fundamentals only. Hence the importance of Hung's results that by putting together both regime switching consumption/dividends and Epstein–Zin preferences, a solution to extant asset pricing puzzles could be found.

Similarly to the subsequent work by Calvet and Fisher (2007), MS pricing models such as Cecchetti et al. (1990) with reference to power utility and Hung (1994) with reference to Epstein–Zin preferences, are significantly different from single-state frameworks. Because $R^M(S_{t+1}, S_t) = [(\kappa_{S_{t+1}} + 1)/\kappa_{S_t}]d_{S_{t+1}}$, return variations arise from two sources. First, there is the usual variation due to uncertainty about future dividend growth, which in this case becomes compounded by the dependence of such growth on the unknown future regime. Second, there is variation over time in the PDR. This second source is induced by the presence of regimes and arises because realized returns depend on both current and next-period regimes. If the parameters of the dividend process are sufficiently different across the two regimes and preferences are different from log-utility, PDRs can be highly regime-dependent and regime switches will have large effects on returns. Similar expressions can be derived for the variance of returns conditional on the current regime. In fact, the conventional finding of a monotonic and linear relation between the equity premium and the conditional variance of returns need not hold in this model, similarly to the argument made by Whitelaw (2000).

Cecchetti, Lam, and Mark (2000, CLM) have recently re-examined the issue of whether an MSM for fundamentals may be nested within a standard

endowment-type CCAPM with isoelastic preferences to produce realistic implications. Their main conclusion is that the representative investor's beliefs need to be distorted in order for a calibrated model under MS to produce realistic implications for the mean equity premium and risk-free rate, volatility, persistence, and long-horizon predictability of asset returns and its relationship to the business cycle. These belief distortions enter along two empirically plausible dimensions. First, they introduce systematic deviations of the subjective probabilities of transitions between high- and low-growth states, that are motivated by showing how agents who use simple rules of thumb to estimate the transition probabilities may form subjective probabilities that deviate from ML estimates obtained from U.S. per capita consumption growth data, assumed to be identical to market dividends. Rule-of-thumb estimates that imply relatively pessimistic beliefs about the persistence of the expansion state and relatively optimistic beliefs about the persistence of the contraction state, allow CLM to match the mean equity premium and risk-free rate found in the data. Unfortunately, systematic distortions alone are not sufficient to explain the volatility of asset returns or the patterns of serial correlation and predictability exhibited in the data. To go beyond an explanation of the first moments of asset returns, CLM introduce a second distortion in which beliefs about the transition probabilities fluctuate randomly about their subjective (distorted) mean values. However, fluctuations of a particular type is required to match the data, not randomization per se: it is random fluctuations in beliefs about the persistence of the low endowment growth state that generate volatility and predictability in asset returns; furthermore, it is necessary for the subjective transition probabilities themselves to be persistent. When the subjective discount rate is instead selected to match the risk-free rate and the CRRA representative agent is fully rational with beliefs that match the ML estimates of their two-state MSMs, as expected from Abel (1994) and Bonomo and Garcia (1994), the model fails badly: there is virtually no equity premium, the volatility of equity returns is far below its sample value, and excess returns have neither the persistence nor the predictability found in the data. Clearly, CLM's use of MSMs to develop realistic asset pricing implications acquires a distinct behavioral flavor and it may be interpreted not as a description of a rational expectations equilibrium framework, but of the deviations that would be required of an MSM for it to correctly price the aggregate U.S. market. Moreover, it may seem questionable the fact that CLM end up specifying a rich eight-state MSM, when almost none of the empirical finance literature based on MSMs has ever resorted to model of this extreme complexity (see Guidolin, 2011).

Another successful attempt at tackling complex asset pricing issues using an MSM for fundamentals has been recently made by Lettau, Ludvigson, and Wachter (2008, LLW). LLW's motivation comes from the fact that the relationship between stock prices and fundamentals in the 1990s appears to have changed. A growing body of literature has been devoted to the evidence of structurally higher PDRs in the 1990s, and explanations have ranged from declining costs of equity market participation and diversification to irrational exuberance to changes in technology and demography. LLW propose a CCAPM with Epstein–Zin preferences with a IES greater than 1 to study the influence of a decline in macroeconomic risk on aggregate stock prices. They estimate a four-state MSIH model with unobservable regimes for the volatility and mean of consumption growth and find evidence of a shift to substantially lower consumption volatility at the beginning of the 1990s.[36] LLW's MSM is built on two different Markov states driving the mean and the variance of consumption growth (assumed to be equal to dividend growth), respectively,

$$\Delta \ln C_{t+1} = \mu_{S^{\mu}_{t+1}} + \sigma_{S^{\sigma}_{t+1}} \varepsilon_{t+1} \quad \varepsilon_{t+1} \sim \text{NID}(0,1)$$

similarly to Kandel and Stambaugh (1990). LLW assume that the probability of changing mean states is independent of the probability of changing volatility states, and vice versa, so that $S^{\mu}_{t+1} = 1, 2$ and S^{σ}_{t+1} may be equivalently summarized by a four-state Markov chain, $S_{t+1} \equiv S^{\mu}_{t+1} \times S^{\sigma}_{t+1} = 1, 2, 3, 4$, under adequate restrictions on the transition matrix. Over a period of roughly 6 years, the smoothed probability of being in a low-volatility state switches from essentially zero, where it resided for most of the post-war period before 1991, to unity, where it remains for the rest of the decade. In contrast to CLM (1990, 2000) and Bonomo and Garcia (1994, 1996), LLW assume that agents cannot observe the regime but must infer it from consumption data. This is important because in this case, the equilibrium stock price is a convex function of their posterior estimate of the regime probability; this model leads to higher asset price volatility during times with high uncertainty about the underlying regime which typically occurs around recessions, thus matching the stylized finding that stock return volatility is countercyclical. The (posterior) filtered probabilities summarize the information upon which conditional expectations are based; the PDR may be then computed by summing the discounted value of future expected dividends across states, weighted by the posterior probabilities of being in each state. Feeding to the model historical

quarterly filtered probabilities from the MSIH, LLW find that plausible parameterizations of the model that can account for an important fraction of the run-up in PDRs observed in the late 1990s. Moreover, the increase in valuation ratios predicted by the model is not well described as a sudden jump upward, but instead occurs gradually over several years, as in the data. This is a result of the learning built into the model by the assumption that agents cannot observe the underlying state, as in Veronesi (1999) and Guidolin and Timmermann (2007). Interestingly, LLW predictions for the risk-free rate are also reasonable: under their baseline parameters, the short-term rate has a mean of 1.44% per annum and a standard deviation of 0.35% per annum, in line with sample values. However, it must be recognized that when LLW do not use a common volatility regime persistence parameter of 0.9999 (or even of 1, an absorbing state) for both Markov states, but instead calibrate the consumption volatility Markov chain to the exact ML point estimates obtained in the chapter, the model explains only about 20% of the total run-up in the price-dividend ratio, with the equilibrium PDR rising from 35 to 40 only. These findings illustrate the importance of the perceived permanence of the volatility decline in determining the magnitude of the rise in the equilibrium PDR: only assuming that the volatility moderation has been perceived to be very persistent – lasting many decades – a large fraction of the run-up in stock prices can be explained. This is consistent with the results in CLM (2000) or peso problem implications for asset prices in Barro (2006).[37]

Structural Models of Markov Regimes in the SDF

Recently the asset pricing literature has been proposing increasingly sophisticated models in which MS does not simply occur in the data under investigation – one can say, in the physical, objective measure $\mathbb{P}$ – but it also (or better, they emerge in $\mathbb{P}$ because they) occurs in the risk neutral pricing measure $\mathbb{Q}$. As we shall see, this is equivalent to modeling the SDF as containing MS effects. A few papers have assumed not only MS in the $\mathbb{Q}$ measure, but also that MS itself may represent a priced risk factor that is reflected in the SDF. This has a rather intuitive meaning: if the state variables vary over time not only as a result of continuous ("diffusive") shocks, but also because of discrete shifts in parameters, it is natural to ask whether such shifts may represent a source of risk that may be priced in equilibrium. Interestingly, these papers have initially appeared in a rather

specific subfield of asset pricing research, that is, in research on (default risk-free) yield curve models. In a way, this is not surprising because the role of MSMs is to capture the important feature that the aggregate economy is subject to discrete and persistent business cycle fluctuations. Such cycles, together with the monetary policy response to them, have significant impact not only on short-term interest rates, but also on the entire term structure. Because business cycles and monetary policy reactions primarily affect the yield curve, and there is now an extensive empirical literature on bond yields that suggests that MSMs may describe interest rate data better than single-regime models (see, e.g., Gray, 1996; Ang & Bekaert, 2002b, 2002c), it is sensible that fixed income scholars were the ones who have first felt a need to develop structural models of the effects of MS on equilibrium asset prices. In fact, MS term structure models (MSTSMs) represent a parsimonious way of introducing interactions between business cycles, the term structure, and risk premia on bonds and may represent the key to understanding the widespread rejection of classical term structure models, such as Cox, Ingersoll, and Ross' (1985) and linear affine models.[38]

Before reviewing the key lessons from a range of leading papers in this field, it may be useful to stress that similar to the body of work surveyed in section "Markov Regimes in Dynamic General Equilibrium Pricing Models", modern MS empirical yield curve models find their underpinning in theoretical papers such as Veronesi and Yared (1999), who have presented a GEAPM in which a representative investor prices nominal and real bonds when fundamentals (consumption and the inflation rate) contain MS components in their drift functions and the investor cannot observe the current drift rates. These papers usually derive closed form solutions for nominal and index-linked (real) bond prices. It turns out that the no-arbitrage price of (real) bonds is simply a filtered probability weighted sum of expected discounted values of one unit of (real) consumption good in the future. Veronesi and Yared (1999) report a number of interesting calibration results. For instance, a number of heterogeneous shapes – including humped and inverted hump yield curves – for both the nominal and real term structure may be generated using three regimes for the deflator process and two regimes for the real consumption one (independent of inflation), for a total of six. In addition, real rates can be high or low independently on whether nominal rates are high or low. These theoretical insights have been subsequently reflected in the empirical results reported by a number of papers that we are about to review.

Bansal and Zhou (2002) is the seminal paper on dynamic term structure models (DTSMs) with regime shifts. In the simplest, single-factor case, the model can be written as

$$\Delta x_{t+1} = \kappa_{S_{t+1}}(\theta_{S_{t+1}} - x_t) + \sigma_{S_{t+1}}\sqrt{x_t}\varepsilon_{t+1} \quad \varepsilon_{t+1} \sim \text{NID}(0,1)$$

where S_{t+1} follows a two-state, time homogeneous Markov chain, x_{t+1} is a continuous variable that describes the state of the economy, and $\kappa_{S_{t+1}}$, $\theta_{S_{t+1}}$, and $\sigma_{S_{t+1}}$ the regime-dependent mean reversion, long-run mean, and volatility parameters, respectively. All of these parameters are subject to discrete regime shifts. For analytical tractability, the Markov chain S_{t+1} is assumed to be independent of x_{t+1-l}, $l = 0, 1, \ldots, \infty$. The agents in the economy observe the regimes, although the econometrician may possibly not observe them. The SDF for this economy is similar to that of standard models, except for incorporating regime shifts:[39]

$$\mathcal{M}_{t+1} = \exp\left[-i_t - \left(\frac{\lambda_{S_{t+1}}}{\sigma_{S_{t+1}}}\right)^2 \frac{x_t}{2} - \frac{\lambda_{S_{t+1}}}{\sigma_{S_{t+1}}}\sqrt{x_t}\varepsilon_{t+1}\right]$$

The $\lambda_{S_{t+1}}$ parameter that affects the risk premia on bonds is also subject to regime shifts.[40] Bansal and Zhou develop a GEAPM of a standard Lucas-type (with log-preferences and a CIR-type process for consumption growth) that leads to this SDF. Importantly, this MSTSM does not entertain the possibility of separate risk compensation for regime shifts. In other words, the risk premium for a security that pays 1 dollar contingent on a regime shift at date $t+1$ is 0. Under MS, Bansal and Zhou conjecture that the bond price with n periods to maturity at date t depends on the regime $S_t = 1, 2$ as well as x_t according to a simple linear affine function, $P_{S_t}(t,n) = \exp[-A_{S_t}(n) - B_{S_t}(n)x_t]$, so that the bond return is

$$\begin{aligned} r_{t+1}(n) &= \ln P_{S_{t+1}}(t+1,\ n-1) - \ln P_{S_t}(t,n) \\ &= [A_{S_t}(n) - A_{S_{t+1}}(n-1) + B_{S_t}(n)x_t - B_{S_{t+1}}(n-1)\kappa_{S_{t+1}}\theta_{S_{t+1}}] \\ &\quad + B_{S_{t+1}}(n-1)\sigma_{S_{t+1}}\sqrt{x_t}\varepsilon_{t+1} \end{aligned}$$

and the volatility is $B_{S_{t+1}}(n-1)\sigma_{S_{t+1}}\sqrt{x_t}$. Under standard boundary conditions $A_{S_t}(0) = B_{S_t}(0) = 0$ and the normalization $A_{S_t}(1) = 0$, $B_{S_t}(1) = 1$, for $S_t = 0,1$ (so that $P_{S_t}(t,1) = \exp(-x_t)$, which means that the continuous state is just the short-term rate, $x_t = i_t$) and exploiting a log-linear approximation

$(\exp(z) - 1 \simeq z)$ of the key asset pricing condition $E[r_{t+1}(n)\mathscr{M}_{t+1}|x_t, S_t] = 1$, they show that the bond risk premium is

$$E\left[\mu_{n,S_{t+1}} + \frac{\sigma^2_{n,S_{t+1}}}{2} - i_t \middle| x_t, S_t\right] \simeq -x_t \sum_{k=1}^{K} \Pr(S_{t+1} = k|S_t) B_k(n-1)\lambda_k$$
$$= -x_t \sum_{k=1}^{K} p_{ik}\xi_{it} B_k(n-1)\lambda_k \qquad (7)$$

where p_{ik} is the probability of a transition from state i to state k ($k = 1,2$). In the absence of MS, the risk premium would simply be $-x_t B(n-1)\lambda$. Hence incorporating regime shifts makes the beta of the asset (i.e., the coefficient on x_t) time-varying and dependent on the regime. Given Eq. (7), the solution for bond prices can be derived by recursively solving for the unknown coefficients $A_{S_t}(n)$ and $B_{S_t}(n)$ with initial conditions $A_1(0) = A_2(0) = B_1(0) = B_2(0) = 0$. The bond yield ($y_{S_t}(t,n)$) of a K-regime MSTSM can be then simply derived as $y_{S_t}(t,n) \equiv -(1/n)\ln P_{S_t}(t,n) = A_{S_t}(n) + B_{S_t}(n)x_t$.

Using monthly 6-month and 5-year U.S. Treasury yield data from 1964 to 1995, Bansal and Zhou estimate three classes of models – one- and two-factor MSTSMs, one-, two-, and three-factor CIR models, and three-factor linear affine models – by the efficient method of moments (EMM) by Gallant and Tauchen (1996). Bansal and Zhou (2002) find that the by far preferred model has two regimes and two factors. In both regimes, the first factor has a far greater mean reversion than the second factor. Regime 1 volatility is larger for both factors. Additionally, the risk premium parameters for both factors are larger in the more volatile regime 1. Both regimes are highly persistent, with $\hat{p}_{11} = 0.91$ and $\hat{p}_{22} = 0.94$. All the parameters of the model are estimated fairly accurately. The recovered first factor tracks the short yield very well, while the second tracks the long yield. The model-implied regime changes usually lead or coincide with economic recessions: Regimes with low-yield spreads occur before or during business contractions. For instance, the correlation between the NBER business cycle the recovered recession (regime 1) indicator is 0.15. During NBER business cycle expansions, the yield spread is 86 basis points, whereas in contractions it is -8 basis points. The other key finding is that the benchmark CIR and affine model specifications with up to three factors are sharply rejected with p-values of zero. The only model specification that finds support in the data (with p-value of 0.14) is the preferred two-factor MSTSM where the market price of risks depends on regime shifts. The two-state MSTSM can duplicate the violations of the expectations hypothesis as documented in Campbell

and Shiller (1991) while the affine and CIR specifications have difficulty in matching these violations.[41] Using simulated data from different model specifications, Bansal and Zhou estimate the typical EH regression of long-term yields on yield spreads: Only the preferred two-factor MSM can duplicate the violations of the EH documented in the data.

Unfortunately, Bansal and Zhou (2002) assume that MS risk is not priced, which is equivalent to assume that MS is not an aggregate risk, and therefore that regime switching shocks do not affect the SDF. Since most empirical MSMs are motivated by business cycle fluctuations or shifts in monetary policies, it seems important to treat regime shifts as an aggregate risk. As discussed by Wu and Zeng (2008), MS risk premia in the yield curve may appear not only because regime shifts have a direct impact on the bond price, but also because regime shifts have a direct impact on the SDF (or investors' marginal utility in the IMRS): If the regime switching shocks generate movements in the bond return and the SDF in the same direction, the covariance is positive and the risk premium is negative as the bond offers investors a hedge against the risk of regime shifts. On the other hand, if regime shifts generate movements in the bond return and the SDF in opposite directions, the covariance is negative and the risk premium will be positive. In this case, MS makes bonds riskier because they decrease the asset's return when investors' marginal utility is high.

Dai and Singleton (2003) have generalized the seminal intuitions in Bansal and Zhou (2002) with reference to MSTSMs. Using model-implied yields simulated from several popular families of DTSMs, they assess their goodness-of-fit by their ability to replicate the historical regressions showing that holding-period returns on bonds are predictable using yield curve variables (as in Campbell & Shiller, 1991), in violation of the EH.[42] Dai and Singleton write the general pricing framework as a specification of the physical process for an $M \times 1$ vector of state variables $\mathbf{X}(t)$

$$d\mathbf{X}(t) = \mu^{\mathbb{P}}(\mathbf{X}, t)dt + \sigma_Y(\mathbf{z}, t)d\mathbf{Z}(t)$$

and an SDF

$$\frac{d\mathscr{M}(t)}{\mathscr{M}(t)} = -i(\mathbf{X}, t)dt - \lambda'(\mathbf{X}, t)d\mathbf{Z}(t)$$

where $i(\mathbf{X},t)$ is the instantaneous riskless rate, $d\mathbf{Z}(t)$ a vector of M independent Brownian motions, and $\lambda(\mathbf{X},t)$ the M-vector of market prices of risk. For simplicity, they take the risk factors driving $\mathscr{M}$ and $\mathbf{X}(t)$ to be the same, that is, they impose market completeness. For a fixed income

security with a continuous coupon rate $h(\mathbf{X},t)$ for $t<T$ and terminal payoff $g(\mathbf{X}(T))$ at date T, its price at date $t<T$ can be expressed in terms of the SDF as

$$P(\mathbf{X},t),t) = E_t\left[\int_t^T \frac{\mathscr{M}(s)}{\mathscr{M}(t)} h(\mathbf{X}(s),s)ds\right] + E_t\left[\frac{\mathscr{M}(T)}{\mathscr{M}(t)} g(\mathbf{X}(T))\right]$$

which can alternatively be written as a problem in which the conditional expectation is taken with respect to a risk neutral measure $\mathbb{Q}$ under which $\mathbf{X}(t)$ follows the process

$$d\mathbf{X}(t) = [\mu^{\mathbb{P}}(t) - \sigma_Y(\mathbf{X},t)\lambda(\mathbf{X},t)]dt + \sigma_Y(\mathbf{X},t)d\mathbf{Z}^{\mathbb{Q}}(t)$$

$$\text{and } P(\mathbf{X}(t),t) = E_t^{\mathbb{Q}}\left[\int_t^T \exp(-r(s)ds)h(\mathbf{X}(s),s)ds + \exp(-r(T)T)E_t^{\mathbb{Q}}[g(X(T))\right]$$

Within this general framework, Dai and Singleton develop a general family of MSTSMs by assuming that the evolution of regimes is described by a $(K+1)$-state continuous time Markov chain with a state-dependent $(K+1)\times(K+1)$ rate matrix $\Lambda^{\mathbb{P}}$ in which all rows sum to zero. For simplicity, they assume that regimes are observable. Intuitively, $\Lambda^{\mathbb{P}}[i,j]dt$ represents the probability of moving from regime i to regime j over the next short interval of time. Under MS, the SDF becomes

$$\frac{d\mathscr{M}(t)}{\mathscr{M}(t)} = -i(S_t;\mathbf{X},t)dt - \lambda'(S_t;\mathbf{X},t)d\mathbf{Z}(t) - \sum_{j=0}^{K} \varsigma_t^j \left(dI_{\{S_t=j\}} - \sum_{i=0}^{K} I_{\{S_t=i\}} R^{\mathbb{P}}[i,j]dt\right)$$

where $\lambda(S_t;\mathbf{X},t)$ is the vector of market prices of diffusion risk and $\varsigma^j(S_t;\mathbf{X},t)$ contains the $K+1$ market prices of a shift from the current regime S_t to regime j an instant later. MSTSMs in which regimes carry a specific risk price imply a very flexible specification of bond risk premia that are time-varying both because they are a function of the underlying state variables as in standard affine models and also because the coefficients of the diffusion risk premia vary across regimes and regimes are themselves priced. The risk-neutral distribution of the short rate is governed by $i(S_t=i;\mathbf{X},t) = \delta_0^i + (\delta_1^i)'\mathbf{X}(S_t;t)$, whereas the risk-neutral process for $\mathbf{X}(t)$ follows an affine diffusion with regime-dependent drifts and volatilities:[43]

$$d\mathbf{X}(t) = \left[\sum_{j=0}^{K} I_{\{S_t=j\}} \mathbf{K}_j^{\mathbb{Q}}(\theta_j^{\mathbb{Q}} - \mathbf{X}(t))\right]dt + \left[\sum_{j=0}^{K} I_{\{S_t=j\}} \text{diag}(\alpha_j^k + (\beta_j^k)'\mathbf{X}(t))\right]d\mathbf{Z}^{\mathbb{Q}}(t)$$

Dai and Singleton prove that the regime-dependent prices of a T-maturity zero coupon bond, $P_{S_t}(t,T)$, $S_t = 0, 1, \ldots, K$, can be determined by solving a system of $K+1$ partial differential equations under $K+1$ terminal conditions $P_{S_t}(T,T) = 1$. These PDEs involve a rate matrix for the conditional Markov chain under the risk-neutral measure,

$$\Lambda^{\mathbb{Q}}[i,j;\mathbf{X},t] = (1 - \varsigma^j(S_t = i;\mathbf{X},t))\Lambda^{\mathbb{P}}[i,j] \quad \text{for } i \neq j \tag{8}$$

and $\Lambda^{\mathbb{Q}}[i,j;\mathbf{X},t] = -\sum_{j\neq i}^{\mathbb{Q}}\Lambda[i,j;\mathbf{X},t]$, which is in general not diagonal. Moreover, Eq. (8) implies that even when $\Lambda^{\mathbb{P}}[i,j]$ is constant over time – as assumed by Dai and Singleton – provided the market prices of MS risk, $\varsigma^j(S_t;\mathbf{X},t)$ are time-varying, then $\Lambda^{\mathbb{Q}}[i,j;\mathbf{X},t]$ becomes time-varying. Additionally, $1 - \Lambda^{\mathbb{Q}}[i,j;\mathbf{X},t]/\Lambda^{\mathbb{P}}[i,j] = \varsigma^j(S_t = i;\mathbf{X},t)$ implying that the (complement to one) of the ratio between the risk-neutral and the physical transition probabilities has a peculiar interpretation: the market price of regime switching risk in correspondence to the initial state i. However the solution for bond prices will incorporate a regime-shifting nature even when $\varsigma^j(S_t;\mathbf{X},t) = 0$ for $j = 0, 1, \ldots, K$ (as in Bansal & Zhou, 2002) as excess bond returns may still be time varying through the coefficients $\{\alpha_j^k, \beta_j^k\}_{j=0}^K$.

Dai and Singleton prove in their paper that allowing both the diffusion risk prices $\lambda(S_t;\mathbf{X},t)$ to follow an MS process and $\varsigma^j(S_t;\mathbf{X},t) > 0$ is crucial not only for elegance reasons, but also in empirical terms. This is obvious from the empirical work presented to try and match Campbell and Shiller's regression-style violations of the EH: when MS risk is not priced and risk premia are restricted to be the same in different regimes, the fact that regime dependence of the bond risk premium is driven entirely by the regime-dependence of volatility prevents MSTSMs to capture violations of the EH. When (as in Bansal & Zhou, 2002) the market price of risk is allowed to vary across regimes – even when regime shifts are not specifically priced – the model fits the evidence of bond return predictability very accurately.

A related paper is Dai, Singleton, and Yang (2007, DSY) who have developed and empirically implemented an arbitrage-free MSTSM with priced regime-shift risks. Therefore they extend Bansal and Zhou's (2002) study of an (approximate) discrete-time "CIR-style" MSTSM to the case in which MS risk is priced in equilibrium and regimes are observable to investors.[44] DSY's strategy is straightforward: they proceed by parameterizing the risk-neutral distribution of the factors so as to ensure closed-form solutions for bond prices, and then overlay flexible specifications of the market prices of risk to describe the historical distribution of bond yields. The risk factors are assumed to follow a discrete-time Gaussian process, and regime shifts are governed by a discrete-time Markov process with TVTPs

(under the historical measure $\mathbb{P}$), according to a Gray-style logistic function that depends on the risk factors underlying changes in the shape of the yield curve. Within each regime, the short rate is assumed to be a linear affine function of a vector $\mathbf{X}_t$ of three risk factors, $r_{t,S_t} = \delta_{0,S_t} + \delta_1 \mathbf{X}_t$, where $\mathbf{X}_t$ follows a Gaussian VAR with constant conditional variances, and the market prices of factor risks depend on $\mathbf{X}_t$ as in Duffee (2002):

$$\mathbf{X}_{t+1} = \mu^{\mathbb{Q}}_{t,S_t} + \mathbf{\Sigma}_{S_t} \varepsilon_{t+1} \quad \varepsilon_{t+1} \sim \text{NID}(\mathbf{0}, \mathbf{I}_3)$$

where $\mu^{\mathbb{Q}}_{t,s} \equiv E[\mathbf{X}_{t+1}|S_t = s] = \mathbf{X}_t + \mathbf{K}^{\mathbb{Q}}(\boldsymbol{\theta}^{\mathbb{Q}}_s - \mathbf{X}_t)$ is the risk-neutralized drift function, $\boldsymbol{\theta}^{\mathbb{Q}}_s$ follows an MS process, and $S_t = L,H$. There are two regimes characterized by low (L) and high (H) volatility. The transition probabilities are restricted to be constant under the risk-neutral measure $\mathbb{Q}$.[45] This framework implies that, under $\mathbb{P}$, the conditional volatilities of both factors and bond yields may be state-dependent, that is, conditional volatilities are stochastic. For simplicity, agents are assumed to know the regime they are currently in. This leads to regime-dependent risk-neutral pricing of zero-coupon bonds,

$$P^s_{t,n} = \exp\{-A_{n,s} - \varphi'_n \mathbf{X}_t\}$$

$$A_{n+1,s} = \delta_{0,s} + (\mathbf{K}^{\mathbb{Q}}\boldsymbol{\theta}^{\mathbb{Q}}_s)'\varphi_n - \frac{1}{2}\varphi'_n \mathbf{\Sigma}_s \mathbf{\Sigma}'_s \varphi_n - \ln\left(\sum_{j=1}^{2} \Pr^{\mathbb{Q}}(S_{t+1} = j|S_t = s) e^{-A_{n,j}}\right) \quad A_{0,s} = 0$$

$$\varphi_{n+1} = \delta_1 + \varphi_n - (\mathbf{K}^{\mathbb{Q}})'\varphi_n \quad \varphi_0 = \mathbf{0}$$

to an analytic representation of the likelihood function of bond yields, and to a natural decomposition of bond risk premia into components corresponding to MS and linear affine, continuous factor risks. The conditional distributions of $(\mathbf{X}_t, S_{t+1})$ under $\mathbb{P}$ and $\mathbb{Q}$ are linked, under the assumption of no arbitrage opportunities, by the SDF $\mathscr{M}_{t,t+1}$ underlying the time t valuation of payoffs at date $t+1$:

$$\mathscr{M}_{t,t+1} = \exp\left\{-r_t - \varsigma(\mathbf{X}_t, S_t; S_{t+1}) - \frac{1}{2}\mathbf{\Lambda}'(\mathbf{X}_t, S_t)\mathbf{\Lambda}(\mathbf{X}_t, S_t) - \mathbf{\Lambda}'(\mathbf{X}_t, S_t)\mathbf{\Sigma}^{-1}(S_t)[\mathbf{X}_{t+1} - \mu^{\mathbb{Q}}(S_t)]\right\}$$

where $\varsigma(\mathbf{X}_t, S_t; S_{t+1})$ is the market price of regime shift risk from S_t to S_{t+1}, and $\mathbf{\Lambda}(\mathbf{X}_t, S_t)$ is the vector of market price(s) of factor risks. The SDF

depends implicitly on the regimes $(S_t; S_{t+1})$, because agents know both the regime S_{t+1} and the regime from which they have switched, S_t. One can show that the distribution of the factors $\mathbf{X}_{t+1}$ is also Gaussian with conditional mean $\mu_t^{\mathbb{P}} = \mu_t^{\mathbb{Q}} + \mathbf{\Sigma}_t^{-1}\mathbf{\Lambda}_t$ and variance $\mathbf{\Sigma}_t\mathbf{\Sigma}'_t$. At this point, DSY extend the essentially affine, Gaussian model of Duffee (2002) to the case of multiple regimes by assuming that

$$\mathbf{\Lambda}(\mathbf{X}_t, S_t) = \mathbf{\Sigma}_{S_t}^{-1}(\lambda_{0,S_t} + \lambda_{1,S_t}\mathbf{X}_t)$$

In standard affine models, the market price of (diffusion) risk is assumed to be proportional to the volatility of the state variables $\mathbf{X}_t$. Such a structure is intuitive: risk compensation goes to zero as risk goes to zero. However, since variances are nonnegative, this specification limits the variation of the compensations that investors anticipate to receive when bearing a risk: since the compensation is bounded below by zero, it cannot change sign over time. This restriction, however, is relaxed in the essentially affine models of Duffee (2002) by including λ_0, an extension that Dai and Singleton (2003) show to be crucial for empirical models to account for the failure of the EH.

Using monthly data on U.S. Treasury zero-coupon bond yields for a sample 1972–2003 modeled as a two-regime, three-factor MSTSM, DSY show that priced, state-dependent MS risk plays a critical role in capturing time variations in risk premia, and document notable differences in the behavior of the factor components of risk premia across high- and low-volatility regimes. The ML estimates of the parameters show that variances in the H regime are all larger than their counterparts, so that omitting MS risk may lead single-regime models to understate the fluctuations in excess returns during the periods of transitions between regimes, and to overstate the volatility of factor risk premiums and excess returns during less turbulent times. An LRT of the null hypothesis that the transition matrix is constant under the physical measure $\mathbb{P}$ – which is equivalent to a test of zero market prices for regime switching risks, because of DSY's result that the MS market price of regime switching risk may be measured as the log-ratio of time-varying transition probabilities under $\mathbb{P}$ and of constant transition probabilities under $\mathbb{Q}$ – suggests a strong rejection at conventional significance levels. The filtered regime probabilities implied by the MSTSM confirm the widely documented observation that regime H tends to be associated with recessions, but the signals given by the asset pricing model in DSY are stronger than in standard, "descriptive" two-state models such as those in Ang and Bekaert (2002b, 2002c), in the sense that the filtered probabilities get closer to 1 during recessions than those computed from

models estimated under the physical measure $\mathbb{P}$. The state dependence of the TVTPs under $\mathbb{P}$ is shown to capture an interesting asymmetry in the cyclical behavior of interest rates. If we view regime H as capturing periods of downturns and regime L as periods of expansions, this finding can be viewed as a manifestation of the well-documented asymmetry in U.S. business cycles: recoveries tend to take longer than contractions. The parameters governing the dynamics of the transition matrix under $\mathbb{P}$ imply that the probability of switching from regime L to regime H increases as the short-term yields or the slope of the yield curve increase. The shapes of the term structure of volatility of bond yield changes are also very different across regimes, with the well-known hump being largely a low-volatility regime phenomenon. The model does a very good job at matching the first and second unconditional moments of the data, as the sample curves fall well within the two standard deviation bands of the simulated curves. Finally, and in the same spirit as Bekaert et al. (2001), DSY stress that, following the result in Duffee (2002) and Dai and Singleton (2002), only sufficiently persistent and variable factor risk premia in affine models may shed light on the empirical failure of the EH. Their MSTSM resolves the EH puzzles summarized in Campbell and Shiller (1991).

It must be stressed that a potential weakness of DSY's regime switching DTSM is that the within-regime conditional variances of the factor process are constant. Even though DSY stress that their experience with estimating single-regime affine DTSMs is that the conditional volatility in bond yields induced by conditional volatility in the factors tends to be small relative to the volatility of excess returns, it may be important to further accommodate rich patterns of regime dependence of the vector of market prices of the risk factors, for instance of an MS ARCH type. Similarly, one problem with MSTSMs is a potential tension between the transition probabilities between regimes and the market price of MS risk: a considerable fraction of DSY's asset pricing insights rely on a one-to-one mapping from the log-ratio of the TVTPs under $\mathbb{P}$ and of the constant transition probabilities under $\mathbb{Q}$ to the market price of MS, that however rests on the maintained hypothesis that the transition probabilities are constant under $\mathbb{Q}$. Therefore one cannot rule that a rejection of the hypothesis of zero risk premia for regime shift risk may spuriously derive from the fact that transition probabilities are in fact not constant under $\mathbb{Q}$. It may be possible to loosen this restriction with more general models (i.e., with TVTPs and zero market price of regime-switching risk), but probably at the cost of not being able to obtain a closed form solution to bond prices.

Ang, Bekaert, and Wei (2008, ABW) is a recent paper that overcomes Dai et al.'s (2007) assumption that regimes are observable to investors. ABW develop a similar MSTSM framework but devote their efforts not only to the pricing of nominal bonds, but also in investigating the dynamics of the real term structure of interest rates, decomposing changes in nominal interest rates in movements in real interest rates, expected inflation, or the inflation risk premium. Relative to DSY, ABW use a more flexible MS structure with four regimes and link the regime-dependent parameters at time $t+1$ not to S_t but to S_{t+1} which is unobservable, as typical in the empirical literature. This implies that the conditional variances of ABW factors embed a jump term reflecting the difference in conditional means across regimes which is absent in DSW by construction.[46] In their paper, ABW infer the dynamics of real rates, expected inflation, and inflation risk premiums, using a model with three key features: (i) absence of arbitrage opportunities; (ii) MS behavior with an emphasis on disentangling the real and nominal sources of the regime switches; (iii) flexible time-varying risk premiums crucial for matching time-varying bond premia (see, e.g., Dai & Singleton, 2002). Identification is obtained by using a no-arbitrage term structure model that imposes restrictions on the nominal yields: the movements of long-term yields are linked to the dynamics of both short-term yields and inflation. ABW introduce two different regime variables, $S_t^f \in \{1,2\}$ affecting the drift and variance of the latent factor process, and $S_t^\pi \in \{1,2\}$ affecting the drift and variance of the inflation process. The former is therefore a nominal MS component – because S_t^f enters the conditional mean of inflation, this regime potentially affects expected inflation and can capture nonlinear expected inflation components not directly related to past inflation realizations – while the latter is a real MS component.[47] As usual, an MSM characterized by two Markov chains S_t^f and S_t^π can be re-written using an aggregate regime variable $S_t \in \{1,2,3,4\}$ to account for all possible combinations, according to a Cartesian product rule. To reduce the number of parameters in a 4×4 transition probability matrix, ABW consider two restricted models of the correlation between S_t^f and S_t^π: in a first case, independence is imposed; in an alternative case they specify a restricted form of the transition probability matrix so that S_{t+1}^π depends on S_t^f as well as the previous inflation environment, but future f_{t+1} regimes depend only on S_t^f. Intuitively, this specification captures the fact that aggressively high real rates, for example according to an activist Taylor rule in the implementation of monetary policies, captured by a specific S_t^f regime, could successfully stave off a regime of high inflation.[48]

Using 1-, 4-, 12- and 20-quarter maturity zero-coupon U.S. risk-free rates as yield data, and CPI inflation for the period 1952:Q2-2004:Q4, ABW find that of all the models, only their four-state MSVAR(1) DTSM fits the mean, variance, and autocorrelogram of inflation. Inflation features a relatively low first-order autocorrelation coefficient with very slowly decaying higher-order autocorrelations. Although the four-regime model passes all moment tests, serial correlation residual tests strongly reject all other models at a 1% level. However, only few models also fit the moments of long-term yields and yield spreads: of them, only the four-state models also match the inflation moments. The price of risk for the q_t factor is negative but imprecisely estimated. The prices of risk for the f_t factor are both significantly different from zero and significantly different across the two regimes. Moreover, they have a different sign in each regime, which may induce different term structure slopes across the regimes. The transition probability matrix shows that the S_t^f regimes are persistent with probabilities $\Pr(S_{t+1}^f = 1|S_t^f = 1) = 0.93$ and $\Pr(S_{t+1}^f = 2|S_t^f = 2) = 0.77$. The probability $\Pr(S_{t+1}^\pi = 1|S_{t+1}^f = 1, S_t^\pi = 1)$ is estimated to be one: Conditional on a period with a negative f_t and relatively high inflation (regime 1), the U.S. economy appears to be unable to transition to a period of lower expected inflation unless the f_t regime also shifts to the higher mean regime. The price of risk factor q_t is relatively highly correlated with both inflation and the nominal short rate, but shows little correlation with the nominal spread. In other words, q_t can be interpreted as a level factor. The MS term structure factor f_t is highly correlated in absolute value with the nominal spread, so f_t is a slope factor. In terms of ABW four-state characterization, the first regime is a low real rate-high inflation regime, where both real rates and inflation are not very volatile. The U.S. economy spends most of the time in this normal regime. The volatilities of real short rates, inflation compensation, and nominal short rates are all lowest in regime 1. The regime with the second-largest ergodic probability is regime 3, which is also a low real rate regime. In this regime, the mean of inflation compensation is the highest. Thus, the U.S. economy spends around 90% of the time in low real rate environments. Regimes 2 and 4 are characterized by relatively high and volatile real short rates. The inflation compensation in these regimes is relatively low. Regimes 2 and 4 are also associated with downward-sloping term structures of *real* yields. The transition probability estimates imply that passing through a downward-sloping real yield curve is necessary to reach the regime with relatively low inflation. Finally, regime 4 has the highest volatility of real rates, inflation compensation, and nominal rates. Unconditionally, the term structure of real rates assumes a fairly flat shape

around 1.3%, with a slight hump, peaking at a 1-year maturity. However, there are some regimes in which the real rate curve is downward sloping. Real rates are quite variable at short maturities but smooth and persistent at long maturities. There is no significant real term spread. Finally, ABW's model matches an unconditional upward-sloping nominal yield curve by generating an inflation risk premium that is increasing in maturity. The decompositions of nominal yields into real yields and inflation components at various horizons indicate that variation in inflation compensation (expected inflation and inflation risk premia) explains about 80% of the variation in nominal rates at both short and long maturities.

Bansal, Tauchen, and Zhou (2004, BTZ) have stressed that the success of MSTSMs opens the possibility that this class of DTSMs may be able to capture the dynamics of risk premia on bonds. Besides the common strategy of understanding bond risk premia in the form of deviations from the EH characterized as negative slope coefficient in regressions of yield changes on yield spreads à la Campbell and Shiller (1991) – as we have seen, addressed in an MS framework by Bekaert et al. (2001) – another form of violation of the EH is that the forward rate can predict excess bond returns (Fama & Bliss, 1987). In particular, Cochrane and Piazzesi (2005) have documented that using multiple forward rates to predict bond excess returns generates very high predictability scores, with adjusted R^2s of around 30%, and coefficients (δ_i^n) of multiple forward rate regressors forming a now famous "tent-shaped" pattern – that is, the δ_i^n at first increase and then decrease, as n grows – related to the maturity of the forward rate:

$$x_{t+12}^n \equiv \ln P_{t+12}^{n-1} - \ln P_t^n - y(t,1) = \delta_0^n + \delta_1^n y(t,1) + \sum_{i=2}^{5} \delta_i^n f(t,i) + \varepsilon_{t+12}^n \quad n = 2,\ldots,5$$

where x_{t+12}^n is the excess return of an n-year bond and $f(t,i) = \ln P_t^{n-1} - \ln P_t^n$ is the forward rate. BTZ set out to account for this predictability evidence from the perspective of latent factor term structure models, in the form of a simple, two-factor Bansal and Zhou (2002)-style MSTSM. Notice that the existence of predictability from forward rate to excess bond returns is easily explained by any DTSM with time-varying risk premia. However, the interesting challenge consists of explaining the tent-shaped pattern of the slope coefficients when multiple forward rates are used as regressors. Using, 6-month and 5-year U.S. Treasury yields for a 1964–2001 sample, BTZ find that MSTSMs can simultaneously justify the size and nature of bond return

predictability and the transition dynamics of yields. More specifically, MSMs can reproduce the high predictability and the tent-shaped regression coefficients documented by Cochrane and Piazzesi (2005). Their results suggest that the prediction capability of forward rates for excess returns may be explained by two or three linear factors, whereas the tent pattern of regression coefficients appears to be due to the MS nature of the yield curve. This insight is in no way trivial because while Duffee (2002) had shown that allowing more flexible (essentially affine) specification of the risk premium parameters for simple, affine conditional Gaussian factor model can dramatically improve its ability to match the predictability of excess returns. Yet, when estimated on Bansal and Zhou's data, the best performing model in Duffee – with three Gaussian factors and eight market price-of-risk parameters – is rejected by a χ^2 EMM test with a p-value of 0.0006. Additionally, Duffie's single-state essentially affine model cannot reproduce the tent shape of the predictability regression coefficients.

Monfort and Pegoraro (2007) have recently generalized these results on the importance of building MS risk prices within the SDF and developed a generalized DTSM which captures simultaneously a number of key features, such as MSVAR factor processes – possibly of a nonhomogeneous kind, that is, with TVTPs – an exponential-affine specification of the SDF with time-varying and regime-dependent risk premia, and closed (or quasi-closed) form formulas for zero-coupon bonds and interest rate derivative prices when yields are restricted to be positive. This ambitious objectives are reached by matching the historical distribution and the SDF in order to get a *cumulative autoregressive* (CAR) risk-neutral joint dynamics for the factors and the regimes, and by using the property of the CAR family of being able to incorporate lags and regimes. An N-dimensional process $\mathbf{x}_t$ follows a CAR(p) if the distribution of $\mathbf{x}_{t+1}$ given the past values $\mathbf{X}_t \equiv [\mathbf{x}_t \ \ \mathbf{x}_{t-1} \ \ \ldots]'$ admits a real Laplace transform,

$$E[\exp(\mathbf{u}'\mathbf{x}_{t+1})|\mathbf{X}_t] = \exp\left[\sum_{j=1}^{p} a_j(\mathbf{u})'\mathbf{x}_{t+1-j} + b(\mathbf{u})\right] \quad \forall \mathbf{u} \in \mathbb{R}^N$$

where $\{a_j(\mathbf{u})\}_{j=1}^{p}$ and $b(\mathbf{u})$ are nonlinear functions. The existence of this Laplace transform in the neighborhood of $\mathbf{u} = \mathbf{0}$, implies that all the conditional moments exist, and that the conditional expectations and covariance matrices (and all conditional cumulants) are affine functions of $\mathbf{x}_t$. In fact, it is straightforward to notice that a K-state homogeneous Markov Chain, $S_{t+1} = 1, \ldots, K$, can be represented as a vector state ξ_{t+1} that is also a CAR(1) process. The log-Laplace transform – that is inherently controlling

the conditional cumulants of $\mathbf{x}_{t+1}$ – will then be affine in $\mathbf{X}_t$ which implies that all the conditional cumulants and, in particular, the conditional mean and the conditional covariance matrix of $\mathbf{x}_{t+1}$ are affine in $\mathbf{X}_t$ (see Bertholon, Monfort, & Pegoraro, 2008). More interestingly, when $\mathbf{x}_t$ follows a CAR(p) in which $b(\mathbf{u}) = \tilde{\mathbf{b}}(\mathbf{u})'\lambda$, where $\tilde{\mathbf{b}}(\mathbf{u})$ is a $M \times 1$ vector of functions and the $M \times 1$ vector λ is a linear function of p lags of ξ_{t+1}, then the conditional distribution of $\mathbf{x}_{t+1}$ given $\mathbf{x}_t$ and ξ_{t+1} has a Laplace transform given by:

$$E[\exp(\mathbf{u}'\mathbf{x}_{t+1})|\mathbf{X}_t, \xi_{t+1}] = \exp\left[\sum_{j=1}^{p} a_j(\mathbf{u})'\mathbf{x}_{t+1-j} + \tilde{\mathbf{b}}(\mathbf{u})'\boldsymbol{\Lambda}\boldsymbol{\Xi}_t\right] \quad \forall \mathbf{u} \in \mathbb{R}^N$$

where Λ is a $M \times (p+1)K$ matrix and $\boldsymbol{\Xi}_t \equiv [\xi'_t \ \xi'_{t-1} \ \ldots \ \xi'_{t-p}]'$. For instance, when the physical ($\mathbb{P}$ measure) dynamics of a single, latent factor x_{t+1} is given by a Gaussian MSVAR(p)

$$x_{t+1} = \mu(\boldsymbol{\Xi}_t) + \sum_{j=1}^{p} a_j(\boldsymbol{\Xi}_t)x_{t+1-j} + \sigma(\boldsymbol{\Xi}_t)\varepsilon_{t+1} \quad \varepsilon_{t+1} \sim \text{NID}(0,1)$$

where $\boldsymbol{\Xi}_t$ represents p lags of a K-state nonhomogeneous Markov chain such that $\Pr(\xi_{t+1} = \mathbf{e}_j|\xi_t = \mathbf{e}_i; \mathbf{X}_t) = \pi(\mathbf{e}_j, \mathbf{e}_i; \mathbf{X}_t)$ and the SDF is an exponential affine function of the variables (x_{t+1}, $\mathbf{x}_{t+1}$) but with coefficients depending on the information at time t,

$$\mathcal{M}_{t,t+1} = \exp\left\{-\mathbf{c}'\mathbf{X}_t - \mathbf{d}'\boldsymbol{\Xi}_t + [\gamma_0(\boldsymbol{\Xi}_t) + \gamma_1(\boldsymbol{\Xi}_t)'\mathbf{X}_t]\varepsilon_{t+1} - \frac{1}{2}[\gamma_0(\boldsymbol{\Xi}_t) + \gamma_1(\boldsymbol{\Xi}_t)'\mathbf{X}_t]^2 - [\delta(\boldsymbol{\Xi}_t, \mathbf{X}_t)]'\xi_{t+1}\right\}$$

where $\delta'(\boldsymbol{\Xi}_t, \mathbf{X}_t)$ is a vector of regime-specific risk correction coefficients, this is a single-factor extension to the multilag case of the MS essentially affine specification proposed by Dai et al. (2007). As a result, the risk-neutral dynamics of the process $[x_{t+1} \ \xi_{t+1}]'$ under $\mathbb{Q}$ is

$$f_{t+1} = \mu(\boldsymbol{\Xi}_t) + \gamma_0(\boldsymbol{\Xi}_t)\sigma(\boldsymbol{\Xi}_t) + \sum_{j=1}^{p}[a_j(\boldsymbol{\Xi}_t) + \gamma_{1j}(\boldsymbol{\Xi}_t)\sigma(\boldsymbol{\Xi}_t)]x_{t+1-j} + \sigma(\boldsymbol{\Xi}_t)\varepsilon_{t+1} \quad \varepsilon_{t+1} \sim \text{NID}(0,1)$$

$$\Pr^{\mathbb{Q}}(\xi_{t+1} = \mathbf{e}_j|\xi_t = \mathbf{e}_i; \mathbf{X}_t) = \pi(\mathbf{e}_j, \mathbf{e}_i; \mathbf{X}_t)\exp[-(\delta(\boldsymbol{\Xi}_t, \mathbf{X}_t))'\mathbf{e}_j] = \pi^{\mathbb{Q}}(\mathbf{e}_j, \mathbf{e}_i)$$

Note that $\Pr^{\mathbb{Q}}(\xi_{t+1} = \mathbf{e}_j|\xi_t = \mathbf{e}_i; \mathbf{X}_t)$ may become time-varying even when $\pi(\mathbf{e}_j, \mathbf{e}_i; \mathbf{X}_t) = \pi(\mathbf{e}_j, \mathbf{e}_i)$ is not, similarly to DSY's results. To get closed-form

(or anyway manageable) pricing expressions, Monfort and Pegoraro proceed then by imposing that the risk-neutral dynamics is also MS Gaussian (extended) CAR(p) (see also Bertholon et al., 2008), which implies that the process for $[x_{t+1}\ \xi_{t+1}]'$ under $\mathbb{Q}$ has to satisfy the restrictions $\sigma(\mathbf{\Xi}_t) = \mathbf{v}'\mathbf{\Xi}_t$, a $Kp \times 1$ vector $\mathbf{m}$ exists such that

$$\gamma_0(\mathbf{\Xi}_t) = \frac{\mathbf{m}'\mathbf{\Xi}_t - \mu(\mathbf{\Xi}_t)}{\mathbf{v}'\mathbf{\Xi}_t}$$

(for a given historical stochastic drift $\mu(\mathbf{\Xi}_t)$ and stochastic volatility $\mathbf{v}'\mathbf{\Xi}_t$, the coefficient $\gamma_0(\mathbf{\Xi}_t)$ is indexed by the free parameter vector $\mathbf{m}$), a set of coefficients $\{a_j^{\mathbb{Q}}\}_{j=1}^{p}$ can be found such that

$$\gamma_{1j}(\mathbf{\Xi}_t) = \frac{a_j^{\mathbb{Q}} - a_j(\mathbf{\Xi}_t)}{\mathbf{v}'\mathbf{\Xi}_t} \quad j = 1, \ldots, p$$

and a homogeneous, time-invariant underlying Markov chain, $\Pr^{\mathbb{Q}}(\xi_{t+1} = \mathbf{e}_j | \xi_t = \mathbf{e}_i; \mathbf{X}_t) = \pi^{\mathbb{Q}}(\mathbf{e}_j, \mathbf{e}_i)$ exists that implies

$$\exp[(\delta(\mathbf{\Xi}_t, \mathbf{X}_t))'\mathbf{e}_j] = \frac{\pi(\mathbf{e}_j, \mathbf{e}_i; \mathbf{X}_t)}{\pi^{\mathbb{Q}}(\mathbf{e}_j, \mathbf{e}_i)} \Rightarrow \delta_j(\mathbf{\Xi}_t, \mathbf{X}_t) = \frac{\pi(\mathbf{e}_j, \xi_t; \mathbf{X}_t)}{\pi^{\mathbb{Q}}(\mathbf{e}_j, \xi_t)} \quad j = 1, 2, \ldots, K$$

This last implication means that the state-dependent MS risk premia are simply ratios of historical ($\mathbb{P}$) and risk-neutralized ($\mathbb{Q}$) transition probabilities. Pegoraro and Monfort's MSVAR DTSM model nests Ang et al.'s (2008) Gaussian MSVAR(1) term structure model driven by a homogeneous Markov chain where MS is not priced; Dai et al.'s (2007) trivariate Gaussian MSVAR(1) with time-varying Markov chain probabilities under the risk neutral measure for the general case of p vector autoregressive lags; TVTPs with a generic number of regimes K; and priced MS risk. In qualitative terms, Monfort and Pegoraro show that (Gaussian) MSVAR(1) price at date t of a zero-coupon bond with residual maturity $n \geq 1$ is $B_{t,n} = \exp\{\mathbf{C}'_n\mathbf{X}_t + \mathbf{D}'_n\mathbf{\Xi}_t\}$, where the vectors $\mathbf{C}_n$ and $\mathbf{D}_n$ satisfy recursive equations that can be easily computed starting from the initial conditions $\mathbf{C}_n = \mathbf{D}_n = \mathbf{0}$.

Monfort and Pegoraro (2007) propose to estimate their MSVAR(p) DTSM following a consistent two-step procedure. In the first step, using the observations on the endogenous factor yields denoted by $\mathbf{x}_t$, the vector of historical parameters under $\mathbb{P}$ may be estimated by the maximization of the likelihood function calculated by means of a standard Hamilton filter. In the second step, using observations on yields with maturities different from those used in the first step, and for a given estimate of the conditional historical (regime-dependent) covariance matrix, one can estimate by

nonlinear least squares the vector of risk-neutral parameters (under $\mathbb{Q}$) by minimizing the sum of squared fitted errors between the observed and theoretical yields, with constraints imposed to guarantee the absence of arbitrage opportunities on the yields determining the factors. The latent variable Ξ_t appearing in the yield-to-maturity formulas is extracted using smoothed probabilities, for each regime and each date. Empirical results using standard monthly U.S. Treasury yield data show that the introduction of multiple VAR lags and MS in the historical and risk-neutral dynamics of two observable factors (the short rate and the spread between the long and the short rate) leads to MSTSMs which are able to fit the yield curve and to explain violations of the EH over both the short and long horizon as well as or better than competing models such as two- and three-factor CIR models and the two-factor MS CIR proposed by Bansal and Zhou (2002).

CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

There are at least two lessons to be learnt from my survey of MS in the asset allocation and asset pricing literatures. First, I have collected robust evidence that modelling MS dynamics in asset returns, their underlying fundamentals, or in the fundamental pricing measure (SDF) linking asset prices to fundamentals and preferences, makes a first-order difference to our understanding of key financial phenomena. Importantly, this survey has specifically emphasized that this differential capabilities rely in no way on "overfitting" the data using highly nonlinear frameworks such as MSMs. On the opposite, MS in financial returns finds intuitive justification in the presence of regime shifts in the dynamics of key economic aggregates (the so-called fundamentals) that underlie asset prices. Importantly, once MS are modelled as part of the fundamental process, then realistic asset return properties, inclusive of regime shifts and pervasive nonlinearities may be generated. Additionally, a recent literature has taught us than even abstracting from an ability to pick what the exact fundamentals underlying asset prices may be, simple MS dynamics in the SDF will carry over to the risk-neutral measure used to price assets. This is in my view the novel implication that has blossomed in the literature: MS methods are no longer just interesting econometrics tools, but as we understand the deeper origins of MS as part of the primitives – preferences, technologies, beliefs – that

financial markets compound in equilibrium prices, also our econometric skills progress accordingly.

Second, a few specific and interesting points have recurred frequently enough to be noted here. 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 with-regime unit root. This makes short-term rates very persistent, even though the existence of a stationary regime is sufficient for them to be stationary in overall terms. 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. However, thanks to the hard work of dozens of researchers, we now understand that especially in an MS framework, this does not necessarily contradict the basic tenet that more risk should be compensated by higher risk premia. To the opposite, 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. Another important lesson goes to teach to financial econometricians an important dose of caution when they approach data that are likely to contain persistent regimes. In this case, peso problems may cast in considerable doubt any simplistic empirical research design that does not carefully take into account that asset prices at t may often reflect a rational expectation for breaks and regime shifts in fundamentals that – because of their very persistence – may not be adequately represented in commonly available data. Equivalently, as one would expect, the use of MSMs does encourage an increasing taste for long and adequately representative data sets.

What is most exciting at the end of such a trip is to take stock on the body of existing knowledge to be able to set our sights forward, to look at future extensions and unresolved questions and problems. Of course, it would also be presumptuous to seriously think that one individual researcher be able to summarize in a handful of closing pages everything that remains to be done when it comes to MS modelling of financial data. Instead of waving my hands around a long list of issues or soon-to-tackled research questions, I therefore prefer to focus on two additional papers that – even if probably not providing all the interesting answers and/or developing the necessary tools – have recently set the stage for new and

even more ambitious efforts concerning the scope of application of MSMs in finance.

A first unresolved issue concerns the development of a general, encompassing GEAPM with Markov regimes that could map primitive assumptions on preferences, technology, and possibly a small set of clearly understood frictions (e.g., transaction costs, indivisibility, and asymmetric information) into the existence of MS dynamics in a few fundamental series (e.g., consumption, dividends, and possibly monetary policy reaction functions) and from there into the equilibrium price process of a range (all) traded assets or portfolios. Although my own assessment is that we are currently far from this ambitious mark, a few recent papers have marked progress in this direction (see, e.g., Ang et al., 2008; Veronesi & Yared, 1999, among the papers reviewed above). One good example of this line of work – both in terms of its potential and of its limitations – is a recent paper by Elliott, Miao, and Yu (2007, EMY) who have returned on the issue of whether and how MSMs may solve standing asset pricing puzzles using a simple continuous time framework. They propose an MS extension of a CIR representative agent GEAPM of an economy with a single physical good in which there are finite number of N risky assets and one riskless asset. There are two possible regimes, a "good" state and a "bad" state (expansions and recessions). The dynamics of the two regimes is captured by a time homogeneous continuous Markov Chain S_t with a rate of transition matrix $\mathbf{\Xi}$ with diagonal elements $-\lambda_1$ and $-\lambda_2$, and S_t independent of $d\mathbf{Z}_t$. The drift and volatility matrices are stochastic and driven by the state of the economy, S_t. Using an extension of typical, intertemporal CAPM-style methods, EMY find closed-form solutions for the equilibrium instantaneous interest rate, for the SDF, and determine a partial differential equation satisfied by the price of all contingent claims written on the risky asset. EMY also characterize in quasi closed-form optimal consumption and portfolio weights (before imposing equilibrium conditions) for the case of power utility. Although their results are promising, their work makes a number of primitive assumptions on the MS dynamics of important objects (e.g., the value function) that I am hopeful could be soon grounded in deeper efforts at economic modelling.

A second, exciting set of applications of MS in financial economics concerns asset pricing, and in particular the potential that MSMs have for a deeper understanding of a number of cross-sectional asset pricing puzzles, that is, features of the empirical dispersion of how risk seems to be rewarded by different asset classes and portfolios that remain hard to reconcile with

existing GEAPMs. One first and interesting attempt at using the properties of MS returns to explain the cross-section of U.S. stock returns has been made by Ozoguz (2009), who argues that when investors do not have perfect knowledge of the processes associated with macrolevel variables or stock dividends but instead must learn on them using Bayes rule, the variance of their own beliefs (their *Bayesian uncertainty*) introduces a new source of time-variation in the investment opportunity set. Ozoguz uses this intuition to explore whether and how the level of uncertainty is related to expected returns, and how uncertainty risk is priced in the cross-section of stock returns. Because Bayesian uncertainty is an elusive concept, Ozoguz uses empirical proxies to measure investors' conditional beliefs and uncertainty about the aggregate state of the economy that are derived from simple two-state MSMs with TVTPs (for market returns or aggregate output). Motivated by these theoretical insights, Ozoguz studies a model with two new state variables, the conditional (filtered) probability $\pi_t \equiv \Pr(S_t = good|\mathscr{F}_t)$ and investors' uncertainty ($UC_t \equiv \pi_t(1-\pi_t)$). UC_t has a straightforward interpretation as an uncertainty index: when investors are uncertain about the future direction of the economy, suggesting that their beliefs linger around $\pi_t = 0.5$, UC_t approaches its maximum value of 0.25. This implies the following conditional multifactor representation of expected returns in the cross-section:

$$E_t[r^i_{t+1} - r^f_t] = \beta^m_{i,t}\lambda_{m,t} + \beta^\pi_{i,t}\lambda_{\pi,t} + \beta^{UC}_{i,t}\lambda_{UC,t}$$

where $\beta^m_{i,t}$ the loading on the excess market return, $\beta^\pi_{i,t}$ the asset's sensitivity to changes in π, $\beta^{UC}_{i,t}$ the asset's sensitivity to uncertainty risk, defined as the sensitivity of the return on asset i to an unanticipated change in uncertainty, $\lambda_{m,t}$ the price of market risk, and $\lambda_{\pi,t}$, $\lambda_{UC,t}$ denote the risk premia associated with changing investor beliefs and uncertainty risk. Ozoguz finds that investors' uncertainty about the state of the economy has a negative impact on asset valuations both at the aggregate market level and at the portfolio level, with substantial variation across portfolios sorted by size, book-to-market, and past returns. In the spirit of the intertemporal CAPM, Ozoguz also tests whether there is a premium associated with uncertainty risk, measured as the covariance between stock returns and unexpected changes in investors' own uncertainty. Ozoguz reports that there exists a positive, albeit nonlinear, relationship between the level of uncertainty and expected returns. This drive toward linking the dynamics in beliefs to the cross-section of asset returns may hold the key to new and significant advances in modern asset pricing theory.

NOTES

1. $\varkappa$ is the percentage increase in the certainty equivalent wealth from moving from strategy $\{\check{\mathbf{w}}_{t+j}\}$ to the optimal strategy $\{\hat{\omega}_{t+j}\}$. Here $\{\hat{\omega}_{t+j}\}_{j=0}^{T-t-1}$ is the sequence of portfolio weights under the optimal strategy while $\{\check{\omega}_{t+j}\}_{j=0}^{T-t-1}$ is the sequence under some other strategy that is suboptimal because it is constrained in some way. The first temporal sequence of weights induces a value function $\hat{Q}_{t,T}$, whereas the second sequence induces a value function $\check{Q}_{t,T}$.

2. In the special case of IID returns, $\mathbf{r}_{t+1}$, Samuelson (1969) shows that for CRRA utility the portfolio weights are constant ($\hat{\omega}_t = \hat{\omega}\ \forall t$), and the T-horizon problem becomes equivalent in solving the myopic $T=1$ one-period problem. However, when returns are not IID, the portfolio weights can be broken down into a myopic and a hedging component (see Merton, 1971). The myopic component is the solution from solving the one-period problem. The hedging component results from the investor's desire to hedge against unfavorable changes in the investment opportunity set.

3. A G-point quadrature rule for a function $h(\mathrm{u})$, $\mathrm{u} \in \mathbb{R}^{N-1}$, over the probability density $f(\mathbf{u})$ is a set of points $\{u_g\}_{g=1}^{G}$ and corresponding quadrature weights $\{w_g\}_{g=1}^{G}$ such that

$$\int\int \ldots \int h(\mathrm{u}) f(\mathrm{u}) d\mathrm{u} \simeq \sum_{g=1}^{G} h(u_g) w(u_g)$$

When regimes are observable, it follows that conditioning on the regime being $S_{t+1}=k$, asset returns at time $t+1$ will have a multivariate Gaussian distribution and in this case it is well-known that a G_k quadrature rule (in which the number of points is allowed to be state-dependent) yields accurate approximations.

4. When the return distributions of the assets depend on some instruments $\mathbf{z}_t$ (e.g., the U.S. short-term interest rate) at time t, the distribution of the returns is a function of both the regime and the realization of the instrument at time t. In this case, the probability density function of $\mathbf{r}_{t+1}$ conditional on S_{t+1} becomes $f(\mathbf{r}_{t+1}|S_{t+1},\mathbf{z}_t)$ and one needs to keep track of both the regime and the realizations of the predictors $\mathbf{z}_t$. However, Ang and Bekaert show that the quadrature numerical approach illustrated for the case of observable regimes extends to the mixed case in which there is both linear predictability and MS.

5. Despite the poor statistical significance levels for the difference in correlations across regimes, the MSM picks up the higher correlations during extreme downturn events. The (exceedance) correlations in the data exhibit a pronounced asymmetric pattern, with negative exceedance correlations higher than positive exceedance correlations. These patterns are easily reproduced by multivariate MSMs. On the contrary, simple IID Normal distributions and asymmetric multivariate GARCH models fail to match the empirical exceedance correlation asymmetry.

6. For instance, an investor with a risk aversion level of 5 and a 1-year horizon must be compensated with a $\varkappa = 1.04$ (1.16) cents in regime 1 (2) for holding an IID portfolio instead of the optimal MS portfolios.

7. Guidolin and Ria (2011) is a related paper in which closed-form solutions for means and variances to be fed to the minimum-variance frontier problem are explicitly characterized. Moreover, differently from Ang and Bekaert (2004), Guidolin and Ria assume that regimes are unobservable to investors and therefore characterize the efficient frontier when there is high uncertainty about the nature of the regime.

8. As pointed out by Detemple et al. (2003), numerical schemes based either on grid approximation of partial differential equations or on quadrature discretization of the state space suffer from a dimensionality curse that Monte Carlo simulation methods can help alleviate.

9. For instance, Perez-Quiros and Timmermann (2000) use a bivariate model to capture MS regimes in the distribution of small and large stocks' returns and find that a simple stylized trading rule generates superior Sharpe ratios during recessions.

10. GT (2005) report similar evidence on U.K. stock and bond excess returns data, although in this case a model specification search leads to the adoption of a slightly simpler three-state MSIH model. On British data, the annualized CER of ignoring MS is even larger than with U.S. data.

11. Ang and Bekaert (2004) also assume that transition probabilities between the two regimes are time-varying and affected by short-term rates. Consequently, the short rate predicts transitions in the regime, and hence, it implies time variation in expected returns, as in much of the literature in which when short-term interest rates are low, subsequent equity returns tend to be high. A constrained model in which the "stayer" transition probabilities are constant is strongly rejected.

12. David (1997) investigates optimal consumption and portfolio rules in a two-state continuous time model. Both Honda and David assume for technical reasons that the volatility coefficient is constant, which is a restriction rejected by the data.

13. However an interaction effect exists: ambiguity increases the relative importance of the intertemporal hedging demand, as a percentage of the total demand for the risky stock.

14. It is well known since Constantinides (1986) that when investment opportunities are constant over time, the presence of transaction costs significantly changes optimal consumption and investment strategies. For example, in the presence of transaction costs, continuous trade incurrs infinite transaction costs, and thus even a small transaction cost can dramatically decrease the frequency of trade. However, it has long seemed unlikely that transaction costs may play an important role in explaining the cross-sectional patterns of expected returns, the equity premium puzzle, or the small stock risk premium. In particular, Constantinides (1986) finds that the liquidity premium (i.e., the maximum expected return an investor is willing to exchange for zero transaction cost) is small relative to transaction costs and concludes that transaction costs only have a second-order effect on asset pricing.

15. For simplicity, they assume that death occurs at the first jump time of an independent Poisson process with intensity δ. The investor is restricted to have finite consumption paths and to remain solvent for all $t \geq 0$.

16. The liquidity premium to transaction cost ratio could be well above one. The consideration of a stochastic investment opportunity set makes this ratio typically more than 4 times and in many cases 10 times higher than what Constantinides finds. However, JKLL cautiously admit that even when the investment opportunity set is

time-varying, the magnitude of liquidity premia cannot be large enough to fully explain the equity premium puzzle of Mehra and Prescott (1985).

17. The large family of VARs (in principle, a total of 1,024 different models) is obtained by investigating the implied dynamic recursive portfolio choices and the resulting recursive OOS performances for all combinations one can form using 7 predictors besides lagged values of asset returns themselves, and experimenting with 4 alternative lag orders, $p = 1$, 2, 4, and 12. The seven predictors are typical in the finance literature and consist of widely employed macro-finance variables, that is, the dividend yield, the riskless term spread, the default spread between Baa and Aaa corporate bond yields, the CPI inflation rate, the nominal riskless 3-month *T*-bill rate, the rate of growth of industrial production, and the unemployment rate.

18. However, it must be stressed that papers such as Ang and Bekaert's (2002a) or Guidolin and Timmermann (2005, 2007) have tackled these very issues through smart applications of the delta methods or parametric bootstrap to tease out the distribution of the portfolio measures computed in their papers.

19. A related paper is Guidolin and Nicodano (2009), who document that preferences for positive skewness and negative excess kurtosis would account for the modest optimal (as well as, observed) weight of small capitalization stocks in internationally diversified portfolio.

20. This makes Mayfield's model different from Turner et al. (1989) and most subsequent literature, where the current state is not known and must, instead, be learned. Mayfield remarks that his model is in the spirit of Merton (1980), where agents have access to continuous returns data over a discrete interval of time such that they are able to estimate the variance of the underlying data generating process to any degree of precision required. However, such continuous time process for switching returns is left unspecified.

21. Interestingly, because in reality we observe returns and not the underlying risk premium, if the expected increase in wealth associated with a return to the low-volatility state is sufficiently large, then the expected intrastate returns in the high-volatility state can be negative even though the risk premium is positive, $\gamma\sigma_H^2 + \hat{\pi}_H \ln(1 + J_H)[1 - (1 + K_H)^{-\gamma}] > 0$.

22. In this case, the direction of causality runs opposite to that of volatility feedback, with the size of the change in volatility being somewhat proportional to the size of the price movement. Thus, if the leverage hypothesis were the driving force behind the negative relationship between return volatility and realized returns, we would expect to find lingering ARCH effects in the residuals from a model that only captures large discrete changes in market volatility, contrary to common empirical findings (see Bekaert & Wu, 2000).

23. $\mathcal{F}_t^* \supseteq \mathcal{F}_{t-1}$ means that investors observe past returns at the beginning of the trading period and obtain information through the process of trading about the volatility regime.

24. For instance, when high average past innovations predict higher persistence of the first regime, then negative (positive) averages lead to increases (decreases) in the probability of switching to the high-volatility regime from the low-volatility regime. Hence, shocks are negatively correlated with the subsequent volatility regime. This negative correlation can be interpreted as a further volatility feedback channel.

25. BKN stress that the two necessary conditions for volatility feedback are satisfied: persistent volatility and positive price of risk, μ_1, takes a value of 0.056 with a standard error of 0.020.

26. D_t is different from a standard Markov state S_t because the states for D_t are terminal in the sense that on exit they never recur and because the model parameters associated with the volatility process, the unconditional variance, and the unconditional mean change when D_t switches.

27. Alternative specifications of preferences may deliver these empirical features even when consumption growth follows an ARMA process. For instance, habit persistence as in Campbell and Cochrane (1999) and recursive utility as in Epstein and Zin (1989) match other features of stock return data, particularly the magnitude and volatility of the equity premium. Both approaches permit a separation between the IMRS and the inverse of the relative risk aversion coefficient, contrary to CRRA preferences. However, the principal effects of these generalizations are on the volatility of the SDF, not on the correlation between the SDF and equity returns. For example, under recursive utility, the SDF depends on both consumption growth and the market return; consequently, covariations with both these quantities determine the risk premium.

28. Although MS multifractals can accommodate any marginal distribution with positive support and unit mean, for simplicity, Calvet and Fisher restrict themselves to binomial processes and tightly parameterize the transition probabilities ρ_j as $\rho_j = 1 - (1 - \rho_k)^{b^{j-k}}$. Calvet and Fisher (2001) introduce this specification through the discretization of a Poisson arrival process. ρ_k controls the persistence of the highest-frequency component and the parameter b determines the spacing between components.

29. Because given these specifications, their baseline calibration implies a risk aversion of about 35, Calvet and Fisher later use the fact that small but persistent variations in the drift and volatility of consumption have been empirically documented by, for example, Bansal and Yaron (2004), to propose a generalization of their model to MS log-consumption growth rates. They find that by incorporating long-run risks in consumption, they can use a lower risk aversion of $\gamma = 10$ to match the equity premium and still generate a substantial contribution of dividend volatility feedback.

30. In practice, the two-state MSIH for real dividend growth rates reveals that the mean is not significantly different from zero in state 0. An LRT confirms that the variances of dividends are significantly different in the two regimes.

31. This is not inconsistent with stationarity of the process because stationarity within each regime is generally neither necessary nor sufficient for second-order stationarity of an MSVAR process (see Francq & Zakoan, 2001). Hence, despite the occasional nonstationary behaviour of the error correction variable, the equilibrium error can be globally stationary, provided that p_{11}, p_{22}, and the AR coefficient in regime 1 satisfy appropriate restrictions.

32. In section "Markov Regimes or Bubbles?" we have already mentioned one such set of asset pricing puzzles, the finding that stock (more generally, asset) returns and prices often tend to display excess volatility given the variability that is justfied by observed fundamentals (see, e.g., LeRoy & Porter, 1981; Shiller, 1981).

33. Differently from Mehra and Prescott (1985) (or earlier papers, such as Cecchetti et al., 1990) – where dividends on unlevered equity are identically equal to capital income, which is identically equal to total income, which is identically equal to consumption – Abel allows aggregate consumption and aggregate dividends to differ from each other. Obviously, aggregate consumption can deviate from aggregate dividends in a GEAPM if there is labor income in addition to capital income.

34. Bonomo, Garcia, Meddahi, and Tedongap (2011) have generalized results in Hung (1994) to the case of generalized disappointment aversion.

35. Von Neumann-Morgenstern time additive expected utility functions imply that the reciprocal of the CRRA coefficient equals the EIS. Kreps and Porteus establish a nonexpected utility framework which is capable of distinguishing the CRRA coefficient from EIS. However, Kocherlakota (1990) and Weil (1989) demonstrate that relaxing the restriction between the CRRA coefficient and EIS is by itself insufficient to resolve key asset pricing puzzles.

36. This phenomenon has been often described as the "Great Moderation." Stock and Watson (2003) conclude that the decline in volatility has occurred broadly across sectors, it has appeared in employment growth, consumption growth, inflation, and sectoral output growth, as well as in GDP growth, both in United States. and international data. Reductions in standard deviations are on the order of 60–70% relative to the 1970s and 1980s, and the marked change seems to be better described as a structural break, or regime shift, than a gradual, trending decline. However, casual evidence seems to show that such a break may have been over-turned by the surge in volatility that has accompanied the financial crisis and deep recession of 2008–2009.

37. A related point had already been made by Mayfield (2004) with reference to the Great Depression: as a result of a structural shift in the likelihood of future high-volatility periods, the simple historical average of excess market returns may have substantially overstated the magnitude of the market risk premium for the period since the Great Depression. When Mayfield augments his MSM to account for time-variation in transition probabilities, in the form of a single structural shift in the transition probabilities governing the evolution of the two volatility states, he finds evidence of a decline in the expected duration of the high-volatility state after 1940 (from 19 to less than 2 months).

38. For instance, an influential paper by Brown and Dybvig (1986) shows that empirically, the parameters of a CIR model change considerably across time, a feature that is consistent with MS.

39. The justification of this exponential-affine specification is now well documented in the asset pricing literature: this form naturally appears in equilibrium models like consumption-based asset pricing models either with habit formation or with Epstein–Zin preferences (see, e.g., Cochrane, 2005); in continuous time security market models, the discretized version of the SDF is always exponential-affine as shown in Gourieroux and Monfort (2007).

40. Naik and Lee (1997) had proposed a version of this family of MSTSMs in which λ does not depend on the current regime, while $\kappa_{S_{t+1}}$, $\theta_{S_{t+1}}$, and $\sigma_{S_{t+1}}$ do. Bansal and Zhou (2002) compare their model against this simpler benchmark and find that the latter is largely outperformed.

41. Campbell and Shiller (1991, p. 505) reported that "the slope of the term structure almost always gives a forecast in the wrong direction for the short-term change in the yield on the longer bond …", that is, the coefficient θ_1 is normally well below its EH-restricted value of one (in fact, it tends to be significantly negative) in the regression

$$r_{t+m,n-m} - r_{t,n} = \theta_0 + \theta_1 \frac{m}{n-m}(r_{t,n} - r_{t,m}) + u_{t+m},$$

where $r_{t+k,q}$ is the spot rate measured at time $t+k(k \geq 0)$ for a bond maturing at time $t+m$ $(m \geq 1)$, with $m > k$.

42. Dai and Singleton discuss four families of DTSMs: affine, quadratic-Gaussian (when the short-term rate is assumed to be a quadratic function of the factors), nonaffine stochastic volatility models with jumps, and multiple-regime MSTSMs.

43. If δ_1^i is regime-dependent, an analytical solution for the yield curve is unavailable. Bansal and Zhou (2002) and Wu and Zeng (2008) assume that δ_1^i depends on regimes and obtain the term structure of interest rates using log-linear approximations.

44. However, because of the CIR-type structure chosen by Bansal and Zhou (2002), DSY's model is not nested in or nesting Bansal and Zhou's.

45. As it turns out, the parameters measuring the price of regime switching risk may be measured as the log-ratio of TVTPs under $\mathbb{P}$ and as constant transition probabilities under $\mathbb{Q}$.

46. The three-factor structure imposed by ABW is more tightly parameterized than in DSY: ABW employ a three-factor representation of yields in which one factor, π_t – inflation – is observed; the other two factors are unobservable; the second factor, f_t, represents a latent MS term structure factor; the third factor is q_t, a latent time-varying but regime-invariant price of risk factor. Moreover, while the conditional means and volatility of f_t and π_t follow an MS process, the conditional mean and volatility of q_t do not. The feedback (vector autoregressive) parameters for all variables in the companion form also do not switch across regimes. Thanks to these restrictions, ABW's model produces closed-form solutions for bond prices.

47. Previously, regimes in real rates (Garcia & Perron, 1996) and regimes in inflation (Evans & Lewis, 1995b) had been only separately considered.

48. Bikbov and Chernov (2008) is an alternative example of an MSTSM in which the factors are assumed to be observable (i.e., they are identified ex ante) and bond yield data are simply used to infer the MS dynamics of the factors. Their model posits time-homogeneous MS in the volatility of exogenous output and inflation shocks, in the monetary policy rule, and in the volatility of monetary shocks. However, in Bibkov and Chernov MS is itself not priced.

REFERENCES

Abel, A. (1994). Exact solutions for expected rates of return under Markov regime switching: Implications for the equity premium. *Journal of Money, Credit, and Banking*, *26*, 345–361.

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., & Bekaert, G. (2002c). Short rate nonlinearities and regime switches. *Journal of Economic Dynamics and Control*, *26*, 1243–1274.

Ang, A., & Bekaert, G. (2004). How regimes affect asset allocation. *Financial Analysts Journal*, *60*(March/April), 86–99.

Ang, A., Bekaert, G., & Wei, M. (2008). Term structure of real rates and expected inflation. *Journal of Finance*, *63*, 797–849.

Ang, A., & Timmermann, A. (2011). *Regime changes and financial markets*. NBER Working paper No. 17182.

Backus, D., Gregory, A., & Zin, S. (1989). Risk premiums in the term structure: Evidence from artificial economies. *Journal of Monetary Economics*, *24*, 371–399.

Bae, J., Kim, C. J., & Nelson, C. (2007). Why are stock returns and volatility negatively correlated? *Journal of Empirical Finance*, *14*, 41–58.

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

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

Bansal, R., Tauchen, G., & Zhou, H. (2004). Regime-shifts, risk premiums in the term structure, and the business cycle. *Journal of Business and Economic Statistics*, *22*, 396–409.

Bansal, R., & Yaron, A. (2004). Risks for the long-run: A potential resolution of asset pricing puzzles. *Journal of Finance*, *59*, 1481–1509.

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

Barberis, N. (2000). Investing for the long run when returns are predictable. *Journal of Finance*, *55*, 225–264.

Barro, R. (2006). Rare disasters and asset markets in the twentieth century. *Quarterly Journal of Economics*, *121*, 823–866.

Bekaert, G., Hodrick, R., & Marshall, D. (2001). Peso problem explanations for term structure anomalies. *Journal of Monetary Economics*, *48*, 241–270.

Bekaert, G., & Wu, G. (2000). Asymmetric volatility and risk in equity markets. *Review of Financial Studies*, *13*, 1–42.

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

Bikbov, R., & Chernov, M. (2008). *Monetary policy regimes and the term structure of interest rates*. CEPR Discussion Paper no. DP7096.

Black, F. (1972). Capital market equilibrium with restricted borrowing. *Journal of Business*, *45*, 444–445.

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

Bonomo, M., & Garcia, R. (1994). Can a well-fitted equilibrium asset pricing model produce mean reversion? *Journal of Applied Econometrics*, *9*, 19–29.

Bonomo, M., & Garcia, R. (1996). Consumption and equilibrium asset pricing: An empirical assessment. *Journal of Empirical Finance*, *3*, 239–265.

Bonomo, M., Garcia, R., Meddahi, N., & Tedongap, R. (2011). Generalized disappointment aversion, long-run volatility risk, and asset prices. *Review of Financial Studies*, *24*, 82–122.
Brown, S., & Dybvig, P. (1986). The empirical implications of the cox, ingersoll, ross theory of the term structure of interest rates. *Journal of Finance*, *41*, 617–630.
Butler, K., & Joaquim, D. (2002). Are the gains from international equity portfolio diversification exaggerated? The influence of downside risk in bear markets. *Journal of International Money and Finance*, *21*, 981–1011.
Calvet, L., & Fisher, A. (2001). Forecasting multifractal volatility. *Journal of Econometrics*, *105*, 27–58.
Calvet, L., & Fisher, A. (2002). Multifractality in asset returns: Theory and evidence. *Review of Economics and Statistics*, *84*, 381–406.
Calvet, L., & Fisher, A. (2004). How to forecast long-run volatility: Regime-switching and the estimation of multifractal processes. *Journal of Financial Econometrics*, *2*, 49–83.
Calvet, L., & Fisher, A. (2007). Multifrequency news and stock returns. *Journal of Financial Economics*, *86*, 178–212.
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.
Campbell, J., & Hentschel, L. (1992). No news is good news. *Journal of Financial Economics*, *31*, 281–318.
Campbell, J., & Shiller, R. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies*, *1*, 195–227.
Campbell, J., & Shiller, R. (1991). Yield spreads and interest rate movements: A bird's eye view. *Review of Economic Studies*, *58*, 495–514.
Campbell, J., & Viceira, L. (2002). *Strategic asset allocation: Portfolio choice for long-term investors*. Oxford: Oxford University Press.
Cecchetti, S., Lam, P., & Mark, N. (1990). Mean reversion in equilibrium asset prices. *American Economic Review*, *80*, 398–418.
Cecchetti, S., Lam, P., & Mark, N. (2000). Asset pricing with distorted beliefs: Are equity returns too good to be true? *American Economic Review*, *90*, 787–805.
Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, *86*, 221–241.
Cochrane, J. (2005). *Asset pricing*. Princeton, NJ: Princeton University Press.
Cochrane, J., & Piazzesi, M. (2005). Bond risk premia. *American Economic Review*, *95*, 138–160.
Constantinides, G. (1986). Capital market equilibrium with transaction costs. *Journal of Political Economy*, *94*, 842–862.
Cox, J., Ingersoll, J., & Ross, S. (1985). An intertemporal general equilibrium model of asset prices. *Econometrica*, *53*, 363–384.
Dai, Q., & Singleton, K. (2002). Expectations puzzles, time-varying risk premia, and affine models of the term structure. *Journal of Financial Economics*, *63*, 415–441.
Dai, Q., & Singleton, K. (2003). Term structure dynamics in theory and reality. *Review of Financial Studies*, *16*, 631–678.
Dai, Q., Singleton, K., & Yang, W. (2007). Are regime shifts priced in U.S. treasury markets? *Review of Financial Studies*, *20*, 1669–1706.
David, A. (1997). Fluctuating confidence in stock markets: Implications for returns and volatility. *Journal of Financial and Quantitative Analysis*, *32*, 427–462.
David, A., & Veronesi, P. (2009). *What ties return volatilities to price valuations and fundamentals*. Working paper, University of Chicago, Chicago.

De Santis, G., & Imrohoroglu, S. (1997). Stock returns and volatility in emerging financial markets. *Journal of International Money and Finance*, *6*, 561–579.

Detemple, J., Garcia, R., & Rindisbacher, M. (2003). A Monte Carlo method for optimal portfolios. *Journal of Finance*, *58*, 401–446.

Detemple, J., Garcia, R., & Rindisbacher, M. (2005). Representation formulas for Malliavin derivatives of diffusion processes. *Finance and Stochastics*, *9*, 349–367.

Dittmar, R. (2002). Nonlinear pricing Kernels, Kurtosis preference, and evidence from the cross section of equity returns. *Journal of Finance*, *57*, 369–403.

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

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

Elliott, R., Miao, H., & Yu, J. (2007). *General equilibrium asset pricing under regime switching*. Working Paper, University of Calgary.

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.

Epstein, L., & Zin, S. (1989). Substitution risk aversion and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica*, *57*, 937–968.

Evans, M., & Lewis, K. (1995a). Do long-term swings in the dollar affect estimates of the risk premia? *Review of Financial Studies*, *8*, 709–742.

Evans, M., & Lewis, K. (1995b). Do expected shifts in inflation affect estimates of the long-run Fisher relation? *Journal of Finance*, *50*, 225–253.

Fama, E., & Bliss, R. (1987). The information in long-maturity forward rates. *American Economic Review*, *77*, 680–692.

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.

French, K., & Poterba, J. (1991). Investor diversification and international equity markets. *American Economic Review*, *81*, 222–226.

French, K., Schwert, W., & Stambaugh, R. (1987). Expected stock returns and volatility. *Journal of Financial Economics*, *19*, 3–29.

Froot, K., & Obstfeld, M. (1991). Intrinsic bubbles: The case of stock prices. *American Economic Review*, *81*, 1189–1214.

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

Gallant, A., & Tauchen, G. (1996). Which moment to match? *Econometric Theory*, *12*, 657–681.

Garcia, R., & Perron, P. (1996). An analysis of the real interest rates under regime shifts. *Review of Economics and Statistics*, *78*, 111–125.

Glosten, L., Jagannathan, R., & Runkle, D. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, *48*, 1779–1801.

Gourieroux, C., & Monfort, A. (2007). Econometric specifications of stochastic discount factor models. *Journal of Econometrics*, *136*, 509–530.

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

Grossman, S., & Shiller, R. (1981). The determinants of the variability of stock market prices. *American Economic Review*, *71*, 222–227.

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., & Hyde, S. (2010). *Can VAR models capture regime shifts in asset returns? A long-horizon strategic asset allocation perspective*. Working paper No. 608, Manchester Business School.

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). Bingley, UK: 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., & Ria, F. (2011). Regime shifts in mean-variance efficient frontiers: Some international evidence. *Journal of Asset Management*, forthcoming.

Guidolin, M., & Rinaldi, F. (2010). A Simple model of trading and pricing risky assets under ambiguity: Any lessons for policy-makers? *Applied Financial Economics*, *20*, 105–135.

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

Guidolin, M., & Timmermann, A. (2008). International asset allocation under regime switching, skew and Kurtosis preferences. *Review of Financial Studies*, *21*, 889–935.

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

Harrison, P., & Zhang, H. (1999). An investigation of the risk and return relation at long horizons. *Review of Economics and Statistics*, *81*, 1–10.

Harvey, C., & Siddique, A. (2000). Conditional skewness in asset pricing tests. *Journal of Finance*, *55*, 1263–1295.

Honda, T. (2003). Optimal portfolio choice for unobservable and regime-switching mean returns. *Journal of Economic Dynamics and Control*, *28*, 45–78.

Hung, M.-W. (1994). The interaction between nonexpected utility and asymmetric market fundamentals. *Journal of Finance*, *49*, 325–343.

Jang, B.-G., Koo, H. K., Liu, H., & Loewenstein, M. (2007). Liquidity premia and transaction costs. *Journal of Finance*, *62*, 2329–2366.

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

Kim, C.-J., Morley, J., & Nelson, C. (2004). Is there a positive relationship between stock market volatility and the equity premium. *Journal of Money, Credit, and Banking*, *36*, 336–360.

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.

Kocherlakota, N. (1990). Disentangling coefficient of relative risk aversion from elasticity of intertemporal substitution: An irrelevance result. *Journal of Finance*, *45*, 175–191.
Kreps, D., & Porteus, K. (1978). Temporal resolution of uncertainty and dynamic choice theory. *Econometrica*, *46*, 185–200.
Krolzig, H. M. (1997). *Markov-switching vector autoregressions: Modeling, statistical inference, and application to business cycle analysis*. Berlin: Springer.
Le Roy, S., & Porter, R. (1981). The present-value relation: Tests based on implied variance bounds. *Econometrica*, *49*, 555–574.
Lettau, M., & Ludvigson, S. (2001). Resurrecting the (C)CAPM: A cross-sectional test when risk premia are time-varying. *Journal of Political Economy*, *109*, 1238–1287.
Lettau, M., Ludvigson, S., & Wachter, J. (2008). The declining equity premium: What role does macroeconomic risk play? *Review of Financial Studies*, *21*, 1653–1687.
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.
Lucas, R. (1978). Asset prices in an exchange economy. *Econometrica*, *46*, 1426–1446.
Lynch, A., & Balduzzi, P. (2000). Predictability and transaction costs: The impact on rebalancing rules and behavior. *Journal of Finance*, *55*, 2285–2309.
Mayfield, S. (2004). Estimating the market risk premium. *Journal of Financial Economics*, *73*, 465–496.
Mehra, R., & Prescott, E. (1985). The equity premium: A puzzle. *Journal of Monetary Economics*, *15*, 145–162.
Merton, R. (1971). Optimal consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory*, *3*, 373–413.
Merton, R. (1973). An intertemporal asset pricing model. *Econometrica*, *41*, 867–888.
Merton, R. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics*, *8*, 323–361.
Monfort, A., & Pegoraro, F. (2007). Switching VARMA term structure models. *Journal of Financial Econometrics*, *5*, 105–153.
Naik, V., & Lee, M. (1997). Yield curve dynamics with discrete shifts in economic regimes: Theory and estimation. *Working paper*, University of British Columbia.
Ozoguz, A. (2009). Good times or bad times? Investors' uncertainty and stock returns. *Review of Financial Studies*, *22*, 4377–4422.
Perez-Quiros, G., & Timmermann, A. (2000). Firm size and cyclical variations in stock returns. *Journal of Finance*, *55*, 1229–1262.
Psaradakis, Z., Sola, M., & Spagnolo, N. (2004). On Markov error-correction models, with an application to stock prices and dividends. *Journal of Applied Econometrics*, *19*, 69–88.
Ramchand, L., & Susmel, R. (1998). Volatility and cross correlation across major stock markets. *Journal of Empirical Finance*, *5*, 397–416.
Rietz, T. (1988). The equity premium: A solution. *Journal of Monetary Economics*, *22*, 117–131.
Samuelson, P. (1969). Lifetime portfolio selection by dynamic stochastic programming. *Review of Economics and Statistics*, *51*, 239–246.
Schwert, G. (1989). Why does stock market volatility change over time? *Journal of Finance*, *44*, 1115–1153.
Shiller, R. (1981). Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review*, *71*, 421–436.

Shiller, R. (1989). *Market volatility*. Cambridge, MA: MIT Press.
Stock, J., & Watson, M. (2003). Has the business cycle changed? Evidence and explanations. In: M. Gertler & K. Rogoff (Eds.), *NBER macroeconomics annual: 2002*. Cambridge, MA: MIT Press.
Tesar, L., & Werner, I. (1995). Home bias and high turnover. *Journal of International Money and Finance*, *14*, 457–492.
Timmermann, A. (2000). Moments of Markov switching models. *Journal of Econometrics*, *96*, 75–111.
Tu, J. (2010). Is regime switching in stock returns important in portfolio decisions? *Management Science*, *56*, 1198–1215.
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.
van Dijk, D., Terasvirta, T., & Franses, P. H. (2000). Smooth transition autoregressive models: A survey of recent developments. Erasmus University Rotterdam, Econometric Institute Research Report EI2000-23/A.
Veronesi, P. (1999). Stock market overreaction to bad news in good times: A rational expectations equilibrium model. *Review of Financial Studies*, *12*, 975–1007.
Veronesi, P., & Yared, F. (1999). *Short and long horizon term and inflation risk premia in the U.S. term structure: Evidence from an integrated model for nominal and real bond prices under regime shifts*. CRSP Working Paper No. 508.
Weil, P. (1989). The equity premium puzzle and the risk-free rate puzzle. *Journal of Monetary Economics*, *24*, 401–421.
Whitelaw, R. (1994). Time variations and covariations in the expectation and volatility of stock market returns. *Journal of Finance*, *49*, 515–541.
Whitelaw, R. (2000). Stock market risk and return: An equilibrium approach. *Review of Financial Studies*, *13*, 521–547.
Wu, S., & Zeng, Y. (2008). *An econometric model of the term structure of interest rates under regime-switching risk*. Working paper, University of Kansas.