Modelling, Estimating and Forecasting Financial Data under Regime (Markov) Switching

Massimo Guidolin
Dept. of Finance, Bocconi University

1. **Introduction**

This chapter marks a strong discontinuity when compared to previous chapters: even though our goals remain the same, i.e.,

- model and forecast the conditional first and second moments (variances and correlations, besides conditional means) of asset returns, and
- adopt an active approach when necessary, which means that the modelling effort ought to be spent on the vector of asset returns when appropriate,

in this chapter we focus on models in which the dynamic process followed by such conditional moments may undergo sudden shifts of regimes. Ironically, the discontinuity mentioned above concerns the fact this chapter is about discontinuities in dynamic time series processes. Just to fix ideas, while in earlier chapters we have always focused on dynamic models in which parameters and therefore the nature of the underlying processes was constant over time, for instance

$$R_{t+1} = \mu + \phi R_t + \epsilon_{t+1} \qquad \epsilon_{t+1} \sim N(0, \sigma^2), \tag{1}$$

in this chapter we work with models of the type, say

$$R_{t+1} = \mu_{S_{t+1}} + \phi_{S_{t+1}} R_t + \epsilon_{t+1} \qquad \epsilon_{t+1} \sim N(0, \sigma^2_{S_{t+1}}), \tag{2}$$

where S_{t+1} is a stochastic variable ("S" here stands for *state*) that is allowed to change over time. Clearly, (1) represents a standard, Gaussian AR(1) model (see chapter 2); on the contrary, (2) will be defined as a regime switching (henceforth, RS) AR(1) model with (conditionally) Gaussian shocks. Although to exactly model the dynamics of S_{t+1} represents the deep point of this chapter, one example could be: $S_{t+1} = 0$ until time T_0 (i.e., between $t = 1$ and $t = T_0$); $S_{t+1} = 1$ from time $t = T_0 + 1$ until the end of the sample, $t = T$. Therefore while for $t = 1, 2, ..., T_0$ the AR(1) model

$$R_{t+1} = \mu_0 + \phi_0 R_t + \epsilon_{t+1} \qquad \epsilon_{t+1} \sim N(0, \sigma^2_0)$$

applies, for $t = T_0 + 1, T_0 + 2, ..., T$ it will be the model

$$R_{t+1} = \mu_1 + \phi_1 R_t + \epsilon_{t+1} \qquad \epsilon_{t+1} \sim N(0, \sigma^2_1)$$

that generates/fits the data, with $\mu_0 \neq \mu_1$, $\phi_0 \neq \phi_1$, and $\sigma_0^2 \neq \sigma_1^2$. Note that when you "switch" (indeed) from (1) to (2), not only the parameters entering the conditional mean function,

$$E_t[R_{t+1}] = \mu_{S_{t+1}} + \phi_{S_{t+1}} R_t$$

become RS, but the same occurs to the conditional variance function:

$$Var_t[R_{t+1}] = E_t\{(R_{t+1} - E_t[R_{t+1}])^2\} = E_t\{\epsilon_{t+1}^2\} = \sigma_{S_{t+1}}^2.$$

As we shall see, this family of time series models has the potential to render both conditional means and conditional second moments time-varying functions, depending on the state variable S_{t+1}. An obvious example of an event sufficiently important to trigger "a T_0" is Lehman's bankruptcy in the Fall of 2008.

More generally, the idea underlying RS models is simple enough: because in finance we care not just for modelling the relationships among variables/quantities of interest but also about forecasting the target quantities (as you now understand, not only conditional mean returns, but also variances or correlations), if and when such relationships are subject to *instability* over time, then such instability also needs to be modelled and predicted. RS models are a set of relatively recent and innovative statistical tools that are used to detect and predict instability (the discontinuities we have referred to above) in statistical relationships. It is easy to give a number of well-motivated, popular examples for how such instability manifests itself: in this chapter, we shall discuss in depth the idea of bull and bear regimes alternating in financial markets, and their economic meaning; the recent financial crisis has shown that most financial markets are characterized by variations in their liquidity state over time; it is well known that exchange rates tend to alternate protracted periods of depreciation and appreciation, which means that it is often possible to detect visible, persistent swings in currencies' trends; there is an ever expanding literature on the presence and the origins of regimes in monetary policy (e.g., as dictated by the personality of governors and chairmen, think of Ben Bernanke taking over Alan Greenspan's chair, or Mario Draghi taking the helm of the ECB from Jean-Claude Trichet).

The financial econometrics literature has captured the idea that statistical relationships may be unstable but that such instability may contain sufficient structure to allow us modelling (and often, prediction) in many different ways. There is a literature on regime switching models driven by observable economic variables, sometimes in the form of *thresholds*. Another literature has instead modelled regimes as driven by latent (unobservable) stochastic variables, usually with a Markov structure; in this case we shall speak of *Markov switching* (MS) models, which is one of the key technical aspects of this chapter. However, there is also a literature that has examined ways to test for the presence of *structural breaks* in parameters, which are unpredictable break-points; recently (but using Bayesian techniques that are beyond the reach of our technical treatment), techniques to also predict the effects of future structural breaks have been proposed.[1]

[1]See e.g., Guidolin, Ravazzolo, and Tortora (2013) on how to forecast stock, bond, and real estate returns when their process is subject to random breaks and a resarcher wants to account for this in an APT-style framework. Your former

Section 2 synthetically shows that in the presence of instability—in particular, when regimes/breaks may be predicted—standard, simple regressions are insufficient quantitative tools. Section 3 provides a short primer to RS models (threshold, smooth transitions) that are however not the more widespread MS models. In this Section we place particular emphasis on switching regressions. Section 4 introduces the basics concerning the structure and specification of MS models. Because these are special cases of RS models driven by a first-order Markov chain, this Section also contains a quick and heuristic review of what Markov chains are and of what properties we need them to possess for our econometric modelling effort to be effective. Section 5 explains how one should go about estimating MS models. This is a rather technical section: in practice, nobody really solves with paper and pencils the first-order conditions reported in Section 5 and yet a user of econometrics is as good as her understanding of what math coprocessors are crunching out inside your computer or server. Section 6 explains how one should forecast returns from MS models and one related complication that makes most of the predictions commonly computed simple (but sometimes useful) approximations to the process.

Section 7 deals with model selection—in particular, with the delicate task of picking the appropriate number of regimes—and diagnostic checks, i.e., one does one assess whether a MS model successfully fits the data at hand. This section also contains a multivariate example that emphasizes the MS may be used to forecast correlations, and discusses a few extensions that show that MS have been recently combined with the GARCH and DCC models of chapters 4-6. Section 8 shows that MS models naturally produce (both unconditional and conditional) distributions for asset returns that are not normal, and such present rich and interesting skewness and kurtosis, even when return shocks are simply assumed to be IID normal over time. Section 9 explains how it is possible to amend plain-vanilla MS models for them not to contradict any fundamental asset pricing principles, such as the one that higher risk ought to lead to higher risk premia. This section represents a sort of side-show and has the main goal of reassuring users of MS econometrics that their basic finance knowledge is not under threat. Section 10 presents three applications that were discussed during the lecture, i.e., using MS VAR models to study contagion, predictability, and in risk management applications. Appendices A-C present a few additional technical details that may be useful to understand the origin of claims that appear in the main text. Appendix D presents a fully worked out set of examples in Matlab®.

2. A Naive Approach: When Regressions Are No Longer Enough

A naïve approach is to model the instability in the process followed by financial return data simply using dummy variables in "regression-type" analysis: One regime applies before the break or regime switch, the other afterwards. For instance, one estimates (say, by OLS)

$$R_{t+1} = [\mu_0 I_{\{t \leq T_0\}} + \mu_1 I_{\{t > T_0\}}] + [\phi_0 I_{\{t \leq T_0\}} + \phi_1 I_{\{t > T_0\}}] R_t + \epsilon_{t+1} \quad \epsilon_{t+1} \sim N(0, [\sigma_0^2 I_{\{t \leq T_0\}} + \sigma_1^2 I_{\{t > T_0\}}])$$

colleague Carlo Magnani (2012) has written an MSc. thesis that extends GRT's framework to compare commercial with residential real estate, with reference to the subprime crisis in the United States.

where $I_{\{t\leq T_0\}}$ and $I_{\{t>T_0\}}$ are standard indicator variables:

$$I_{\{t\leq T_0\}} = \begin{cases} 1 & \text{if } t \leq T_0 \\ 0 & \text{if } t > T_0 \end{cases} \qquad I_{\{t>T_0\}} = \begin{cases} 1 & \text{if } t > T_0 \\ 0 & \text{if } t \leq T_0 \end{cases} .$$

However, this way of proceeding makes sense if and only if T_0 is known for sure. This has two problematic effects: first, dummy regressions can only be used to estimate RS parameters conditioning on some other method having been used early on to infer that T_0 triggers a regime shift or structural shift in parameters; such a method remains mysterious;[2] second, even assuming that such a method to isolate breaks or regime shifts may exist, using dummy variables will not allow us to predict future instability, for instance the average duration of the current regime (i.e., when it is likely to end).[3] When forecasting asset returns or their properties is your crucial objective, estimating simple regressions will be insufficient.

3. A Short Overview of Threshold and Smooth Transition Models

Econometricians have developed methods in which instability is stochastic, it has structure, and as such it can be predicted. This reflects the intuition that such regime shifts are caused by some imperfectly predictable forces that produce the change in the dynamic process of asset returns. Hence, rather than following the "dummy/break-the-regression approach" in Section 2, it is more elegant to assume that there is some larger model encompassing all "regressions" (better, dynamic time series models) across all possible states. For instance, using the same example already proposed in the Introduction, the new framework may be

$$R_{t+1} = \mu_{S_{t+1}} + \phi_{S_{t+1}} R_t + \epsilon_{t+1} \qquad \epsilon_{t+1} \sim N(0, \sigma^2_{S_{t+1}}),$$

where S_{t+1} is a stochastic variable that may take $K \geq 1$ values, $S_{t+1} = 1, ..., K \ \forall t \geq 0$. Of course, when $K = 1$, the model simplifies down to the classical (1); when $K \geq 2$, we speak of either a RS or of a MS model. However, specifying (2) provides at this point an incomplete description of the time series properties of the data: a complete description of the probability law governing the data requires a probabilistic model of what causes any time variation in S_{t+1}. The econometrics literature offers three key examples of frameworks with regimes:[4] (i) threshold models; (ii) smooth transition models; (iii) MS models. Sometimes, but this is a rather subjective labeling, we call RS models frameworks of the type (i) and (ii) in opposition to MS models that are admittedly more popular and successful, at least in quantitative financial applications. In this section, we devote some limited space to a brief description of what the models grouped under (i) and (ii) are, in case you ever encounter them.

[2] If you were to object that structural change must consist of something rather major and therefore visible, I will ask you to alert me (call collect, do not worry) next time you spot stock markets switching *for a fact* (i.e., you must be able to observe that with no uncertainty) from a bear to a bull market or viceversa. Nah...

[3] Either T_0 is obvious to predict, which is however an embarrassing claim because the change in parameter values would then have been discounted by all traders in the market a long time before, or T_0 becomes completely unpredictable which is equivalent to surrender to the randomness of market states.

[4] This is just a heuristic classification: once you leave the special case of linear models, there are infinite non-linear models—and among them, RS and MS specifications—you can choose from.

In a threshold model, S_{t+1} assumes K values in dependence of the value taken at time t by some threshold variable x_t (typically either one or a combination of a few of them) for instance:

$$S_{t+1} = \begin{cases} 1 & \text{if } x_t \leq x_1^* \quad (R_{t+1} = \mu_1 + \phi_1 R_t + \epsilon_{t+1} \quad \epsilon_{t+1} \sim N(0, \sigma_1^2)) \\ 2 & x_1^* < x_t \leq x_2^* \quad (R_{t+1} = \mu_2 + \phi_2 R_t + \epsilon_{t+1} \quad \epsilon_{t+1} \sim N(0, \sigma_2^2)) \\ \vdots & \vdots \qquad\qquad\qquad\qquad \vdots \\ K & x_K^* < x_t \quad (R_{t+1} = \mu_K + \phi_K R_t + \epsilon_{t+1} \quad \epsilon_{t+1} \sim N(0, \sigma_K^2)) \end{cases},$$

where x_1^*, x_2^*, ..., x_K^* are estimable threshold parameters that are simply required to exceed the minimum value in the sample for x_t and to be inferior to the maximum for x_t. What is x_t? An example could be represented by the Federal Funds (FF) rate, at least in the U.S.: when $K = 2$, then you may think that

$$R_{t+1} = \mu_{high} + \phi R_t + \epsilon_{t+1} \quad \epsilon_{t+1} \sim N(0, \sigma_{low}^2)$$

when $x_t \leq x^*$ (loose monetary policy, high expected returns and low risk), i.e., when the FF rate is below some threshold x^* that will need to be estimated; otherwise, when $x_t > x^*$ (tight policy),

$$R_{t+1} = \mu_{low} + \phi R_t + \epsilon_{t+1} \quad \epsilon_{t+1} \sim N(0, \sigma_{high}^2)$$

which implies low expected returns and high risk.

More generally, the threshold regression model of Tong (1983) allows for abrupt switching depending on whether the transition variable is above or below a threshold:

$$\begin{aligned} R_{t+1} &= [I_t\alpha_1 + (1 - I_t)\alpha_2] + [I_t\boldsymbol{\beta}_1 + (1 - I_t)\boldsymbol{\beta}_2]'\mathbf{X}_t + \epsilon_{t+1} \\ \epsilon_{t+1} &\sim N(0, [I_t\sigma_1^2 + (1 - I_t)\sigma_2^2]) \qquad I_t = \begin{cases} 1 & \text{if} \quad g(\mathbf{X}_t) > c \\ 0 & \text{if} \quad g(\mathbf{X}_t) \leq c \end{cases}, \end{aligned} \tag{3}$$

i.e., each of the two regimes applies in dependence on whether $g(\mathbf{X}_t)$ exceeds or not a threshold c (to be estimated), where $g : \mathcal{R}^m \to \mathcal{R}$ is a function that converts the current values of the predictors in $\mathbf{X}_t$ into a value to be compared with the threshold c_j.[5]

In a *smooth transition* model, S_{t+1} is drawn from some cumulative probability distribution (CDF) $F(\cdot)$ whose domain is suitably partitioned into K non-overlapping sub-intervals to allow us to simulate the regime S_{t+1}. One possibility is that in practice S_t is drawn from a discrete probability distribution $F(S_t; x_t)$ that can take K values and that also depends on some threshold variable x_t; $F(S_t; x_t)$ is then a CDF that gives you $\Pr(S_t = 1)$, $\Pr(S_t = 2)$,..., $\Pr(S_t = K)$, for instance:

$$\begin{aligned} \Pr(S_t &= 1; x_t) = F(1; x_t) \\ \Pr(S_t &= 2; x_t) = F(2; x_t) - F(1; x_t) \\ &\ldots \\ \Pr(S_t &= K; x_t) = F(K; x_t) - F(K - 1; x_t). \end{aligned}$$

[5]Of course, when the function $g(\cdot)$ reduces to a selector that "extracts" one variable from $\mathbf{X}_t$, then the regime is defined simply on the basis of the extracted variable.

In this case, it is the change of the shape of $F(\cdot; x_t)$ as x_t changes that affects the probability of regimes and hence their dynamics. One may ask why is this model called "smooth": the reason is that x_t no longer deterministically determines the state, but simply the CDF that affects the assessment of the probability of regimes.

Whilst threshold regressions impart an abrupt non-linear behavior depending on whether the threshold variable(s) is above or below the threshold value, the smooth-transition variant allows for possible gradual movements among regimes, and is able to capture two types of adjustment. First, the parameters of the model change depending upon whether the transition variables is above or below the transition value. Second, the parameters of the model change depending upon the distance between the transition variable and the transition value. For instance, a general smooth transition regression (STR) model is given by:

$$\begin{aligned} R_{t+1} &= \alpha_1 + (\boldsymbol{\beta}_1)'\mathbf{X}_t + [\alpha_2 - \alpha_1 + (\boldsymbol{\beta}_2)'\mathbf{X}_t - (\boldsymbol{\beta}_1)'\mathbf{X}_t]F(\mathbf{e}_i'\mathbf{X}_t) + \epsilon_{t+h}^j \\ \epsilon_{t+1} &\sim N(0, [\sigma_1^2 + (\sigma_2^2 - \sigma_1^2)F(\mathbf{e}_i'\mathbf{X}_t)]), \end{aligned} \tag{4}$$

where $0 \leq F(\mathbf{e}_i'\mathbf{X}_t) \leq 1$ is the transition function and the ith variable in $\mathbf{X}_t$ (selected by the product $\mathbf{e}_i'\mathbf{X}_t$) acts as the transition variable.[6] One may also think of generalizing $F(\mathbf{e}_i'\mathbf{X}_t)$ to $F(g(\mathbf{X}_t))$, where $g : \mathcal{R}^m \to \mathcal{R}$, a function that converts the current, time values of the predictors in $\mathbf{X}_t$ into a value to be fed into the transition function. The smooth transition is perhaps theoretically more appealing over the simple threshold models that impose an abrupt switch in parameter values because only if all traders act simultaneously will this be the observed outcome. For a market of many traders acting at slightly different times a smooth transition model is more appropriate. For instance, it may be true that high Fed funds rates (ff_t) have a negative effect on future stock returns only when monetary policy is strongly tightening, meaning that $\mathbf{e}_i'\mathbf{X}_t$ selects Δff_t and that $F(\mathbf{e}_i'\mathbf{X}_t) \simeq 1$ for very high values of Δff_t; at the same it may be sensible that high Fed funds rates forecast positive future stock returns only for extremely negative values of Δff_t, for which $F(\mathbf{e}_i'\mathbf{X}_t) \simeq 0$. In intermediate situations of $\Delta ff_t \simeq 0$, $F(\mathbf{e}_i'\mathbf{X}_t)$ could take intermediate values so that the effect of Δff_t on R_{t+1} will be captured by a weighted combination of elements in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

The STR model allows different types of market behavior depending on the nature of the transition function. Among the possible transition functions, the logistic has received considerable attention in the literature and is given by the following, where the full model is referred to as the Logistic STR (or LSTR) model:

$$F(\mathbf{e}_i'\mathbf{X}_t) = \frac{1}{1 + \exp(-\rho(\mathbf{e}_i'\mathbf{X}_t - c))} \quad \rho_j > 0, \tag{5}$$

where ρ_j is the smoothing parameter, and c_j the transition parameter, both to be estimated. This function allows the parameters to change monotonically with $\mathbf{e}_i'\mathbf{X}_t$. As $\rho \to \infty$, $F(\mathbf{e}_i'\mathbf{X}_t)$ becomes a

[6] We have already introduced in earlier chapters the fact that the vector $\mathbf{e}_i$ is a special array that contains a 1 in the ith position, and zeroes elsewhere.

standard dummy function:

$$F(\mathbf{e}_i'\mathbf{X}_t) = \begin{cases} 1 & \text{if} \quad \mathbf{e}_i'\mathbf{X}_t > c \\ 0 & \text{if} \quad \mathbf{e}_i'\mathbf{X}_t \leq c \end{cases}$$

and (5) reduces to a threshold regression model. As $\rho \to 0$, (4)-(5) becomes linear because switching is impossible.[7]

A peculiar issue in estimating STR models concerns the smoothing parameter, ρ, the estimation of which may be problematic. In the LSTR model, a large ρ results in a steep slope of the transition function at c, thus a large number of observations in the neighborhood of c are required to estimate ρ accurately. As a result convergence of ρ may be slow, with relatively large changes in ρ having only a minor effect upon the shape of the transition function. When applying these non-linear models, another key issue is the choice of the transition variable.

4. **Markov Switching: Definition and Basic Properties**

MS represents the most widely applied and best known case of RS model in both finance and macroeconomics. Moreover, it is certain that RS and MS models represent by themselves the most important example of *non-linear time series* models of current application.[8] Because in our lectures we have dealt with both the case of univariate as well as multivariate MS models, in this chapter we present the general case of a multivariate model. In particular, we consider the case of a $N \times 1$ random vector of returns, $\mathbf{R}_t$. However, especially to convey the necessary intuition, we also present a few examples based on the limit case of $N = 1$, in case these make your understanding easier; yet, it is should be clear that there are very few or no differences between the cases of $N = 1$ and $N > 1$, apart from the need to use vector and matrices in the latter case. Suppose that the random vector collecting returns on N different assets follows a K-state Markov switching (MS) $VAR(p)$ process with heteroskedastic component, compactly $MS(I)VARH(K,p)$,

$$\mathbf{R}_{t+1} = \boldsymbol{\mu}_{S_{t+1}} + \sum_{j=1}^{p} \mathbf{A}_{j,S_{t+1}} \mathbf{R}_{t+1-j} + \boldsymbol{\Omega}_{S_{t+1}}^{1/2} \boldsymbol{\epsilon}_{t+1} \quad S_{t+1} = 1, 2, ..., K. \tag{6}$$

with $\boldsymbol{\epsilon}_{t+1} \sim \text{IID } N(\mathbf{0}, \mathbf{I}_N)$.[9] The acronym MS(I)VARH means "Markov switching", "Vector autoregressive", "heteroskedastic" model;[10] K is the number of regimes that you are free to specify (or test for,

[7] An alternative functional form governing the switches is represented by the exponential, with the resulting model referred to as the Exponential STR (or ESTR) model:

$$F(\mathbf{e}_i'\mathbf{X}_t) = 1 - \exp(-\rho(\mathbf{e}_i'\mathbf{X}_t - c)^2) \qquad \rho > 0$$

where the parameters change symmetrically about c with $\mathbf{e}_i'\mathbf{X}_t$. If $\rho \to \infty$ or $\rho \to 0$ the ESTR model becomes linear, while non-linearities require intermediate values for ρ. This model implies that the dynamics obtained for values of the transition variable close to c differ from those obtained for values that largely differ from c.

[8] In spite of this possible difference among MS and RS models, in this chapter the concepts of regime and state are now used interchangeably. This derives from the fact that after all MS represents one special case of RS.

[9] Assume the absence of roots outside the unit circle, thus making the process stationary. See chapter 3 for related definitions and explanations.

[10] The "I" in paranthesis is probably superfluous, but it stands there to emphasize that in (6) also the intercept is regime-dependent. In what follows, we shall often simplify the acronym omitting the "I" when this causes no ambiguity.

when needed, see Section 7) and p is the number of autoregressive lags that you can select (or again, test for). $\boldsymbol{\mu}_{S_{t+1}}$ collects the N regime-dependent intercepts, while the p alternative $N \times N$ $\{\mathbf{A}_{j,S_{t+1}}\}_{j=1}^{p}$ vector autoregressive matrices capture regime-dependent VAR-type effects at lags $j = 1, 2, ..., p$.[11] This means that with p VAR lags and K regimes, there are a total of pK matrices to deal with, each potentially containing—unless restrictions are imposed—N^2 parameters to estimate. The (lower triangular) matrix $\boldsymbol{\Omega}_{S_{t+1}}^{1/2}$ represents the factor applicable to state S_{t+1} in a state-dependent Choleski factorization of the variance covariance matrix of asset returns $\boldsymbol{\Omega}_{S_{t+1}}$:[12]

$$\boldsymbol{\Omega}_{S_{t+1}}^{1/2}(\boldsymbol{\Omega}_{S_{t+1}}^{1/2})' = \boldsymbol{\Omega}_{S_{t+1}} \equiv Var[\mathbf{R}_{t+1}|\Im_t, S_{t+1}]$$

$\Im_t$ denotes time t information of all past observations and states (filtered states, see below). Note that $\boldsymbol{\Omega}_{S_{t+1}}^{1/2}$ is in no way the matrix of square roots of the elements of the full covariance matrix $\boldsymbol{\Omega}_{S_{t+1}}$ (if so, how would you deal with potentially negative covariances?).[13] Obviously, a non-diagonal $\boldsymbol{\Omega}_{S_{t+1}}^{1/2}$ makes the N asset returns simultaneously cross-correlated, thus capturing simultaneous comovements between returns on different assets. Conditionally on the unobservable state S_{t+1}, (6) defines a standard Gaussian reduced form VAR(p) model, which is the meaning of $\boldsymbol{\epsilon}_{t+1} \sim$ IID $N(\mathbf{0}, \mathbf{I}_N)$. This means that if one were to take S_{t+1} as given and observable (we shall not of course, in practice), then between time t and $t+1$, (6) would become a VAR(p) similar to those already encountered in chapter 3.

We also assume that $K > 1$ alternative hidden states are possible and that they influence both the conditional mean, the conditional variance, and the conditional correlation structures characterizing the multivariate process in (6), $S_{t+1} = 1, 2, ..., K \ \forall t$. This regime variable is latent (also said hidden or unobservable) in the sense that even at time t both the agent/investor of our models and the econometrician fail to observe S_{t+1}: at most they can both use the methods that will be described below to produce data-driven inferences on the nature of S_{t+1} over time. Basically, the same sample data concerning the N variables in $\mathbf{R}_{t+1}$ are used to also produce inferences on the sample path followed by $\{S_t\}_{t=1}^{T}$, besides producing standard inferences on the parameters, see Section 5.1. Finally, given past regimes, S_{t+1} is assumed to be independent of any other random variable indexed at time $t+1$, and in particular S_{t+1} is independent of z_{t+1}.

Several special cases of (6) are often used in finance, for instance the simple MSVARH(K, 1) case:

$$\mathbf{R}_{t+1} = \boldsymbol{\mu}_{S_{t+1}} + \mathbf{A}_{S_{t+1}}\mathbf{R}_{t+1-j} + \boldsymbol{\Omega}_{S_{t+1}}^{1/2}\boldsymbol{\epsilon}_{t+1} \quad S_{t+1} = 1, 2, ..., K,$$

which is a simple VAR(1) with K regimes. Of course, in the literature, the case of $K = 2$ tends to be the most common, even though Guidolin (2012) explains why there is nothing special or magical

[11] Here VAR is the acronym for "vector autoregressive" and this has little to do with the acronym VaR (notice the lowercase "a"), which means value-at-risk.

[12] $Var[\mathbf{R}_{t+1}|\Im_t, S_{t+1}]$ in the expression that follows is a covariance matrix that conditions on time t information, but the structure of which depends on the regime at time $t+1$, S_{t+1}. We should indeed emphasize that all matrices of parameters in (6) are made to depend on the regime at time $t+1$, S_{t+1}. This regime-dependent covariance matrix has been called on purpose $\boldsymbol{\Omega}_{S_{t+1}}$ and not $\boldsymbol{\Sigma}_{S_{t+1}}$ to distinguish it from the GARCH-type covariance matrix, $\boldsymbol{\Sigma}_{t+1}$.

[13] In fact, $\boldsymbol{\Omega}_{S_{t+1}}^{1/2}$ is a lower triangular matrix appropriately defined according to an algorithm that is implemented in most software packages (sure enough, in Matlab). Section 10.1 shows one example for the $N = 2$ case.

about setting $K = 2$, especially when N is relatively large.[14] Interestingly, especially when daily and weekly returns data are used, it is not uncommon to find that the data actually support a choice of $p = 0$, which reduces the model to a MSIH(K) (or MSIH($K, 0$)):

$$\mathbf{R}_{t+1} = \boldsymbol{\mu}_{S_{t+1}} + \boldsymbol{\Omega}_{S_{t+1}}^{1/2} \boldsymbol{\epsilon}_{t+1} \quad \boldsymbol{\epsilon}_{t+1} \sim \text{IID } N(\mathbf{0}, \mathbf{I}_N).$$

However, in the literature you also find many cases in which $p = 0$ works at all frequencies. The reason is that when $K >> 2$, it is possible that our common perception of $p > 1$ being needed in standard single-state VAR(p) models may be caused by our omitting the presence of regimes in the dynamics of asset returns.[15] Clearly, in the univariate case (6) becomes a simpler MSAR(K, p), where the "V" indicating a vector process has been dropped because $N = 1$:

$$R_{t+1} = \mu_{S_{t+1}} + \sum_{j=1}^{p} a_{j,S_{t+1}} R_{t+1-j} + \sigma_{S_{t+1}} \epsilon_{t+1} \quad S_{t+1} = 1, 2, ..., K,$$

where $\epsilon_{t+1} \sim$ IID $N(0, 1)$ and $\sigma_{S_{t+1}}$ has now become a regime-specific volatility. One example of a two-state bivariate heteroskedastic VAR(1) model is:

$$\begin{aligned} \begin{bmatrix} R_{t+1}^1 \\ R_{t+1}^2 \end{bmatrix} &= \begin{bmatrix} \mu_{S_{t+1}}^1 \\ \mu_{S_{t+1}}^2 \end{bmatrix} + \begin{bmatrix} a_{S_{t+1}}^{11} & a_{S_{t+1}}^{12} \\ a_{S_{t+1}}^{21} & a_{S_{t+1}}^{22} \end{bmatrix} \begin{bmatrix} R_t^1 \\ R_t^2 \end{bmatrix} + \\ &+ \begin{bmatrix} \sigma_{1,S_{t+1}} & 0 \\ \rho_{12,S_{t+1}} \sigma_{2,S_{t+1}} & \sqrt{1 - (\rho_{12,S_{t+1}})^2} \sigma_{2,S_{t+1}} \end{bmatrix} \begin{bmatrix} \epsilon_{t+1}^1 \\ \epsilon_{t+1}^2 \end{bmatrix}, \end{aligned}$$

where $a_{S_{t+1}}^{11}$ and $a_{S_{t+1}}^{22}$ are MS AR(1) coefficients, while $a_{S_{t+1}}^{12}$ and $a_{S_{t+1}}^{21}$ capture the regime-specific cross-serial correlation effects of R_t^2 on R_{t+1}^1 and of R_t^1 on R_{t+1}^2, respectively. The matrix

$$\begin{bmatrix} \sigma_{1,S_{t+1}} & 0 \\ \rho_{12,S_{t+1}} \sigma_{2,S_{t+1}} & \sqrt{1 - (\rho_{12,S_{t+1}})^2} \sigma_{2,S_{t+1}} \end{bmatrix}$$

is a bivariate Choleski factor. Moreover

$$\begin{aligned} &\begin{bmatrix} \sigma_{1,S_{t+1}} & 0 \\ \rho_{12,S_{t+1}} \sigma_{2,S_{t+1}} & \sqrt{1 - (\rho_{12,S_{t+1}})^2} \sigma_{2,S_{t+1}} \end{bmatrix} \cdot \begin{bmatrix} \sigma_{1,S_{t+1}} & \rho_{12,S_{t+1}} \sigma_{2,S_{t+1}} \\ 0 & \sqrt{1 - (\rho_{12,S_{t+1}})^2} \sigma_{2,S_{t+1}} \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{1,S_{t+1}}^2 & \underbrace{\rho_{12,S_{t+1}} \sigma_{1,S_{t+1}} \sigma_{2,S_{t+1}}}_{\sigma_{12,S_{t+1}}} \\ \rho_{12,S_{t+1}} \sigma_{1,S_{t+1}} \sigma_{2,S_{t+1}} & (\rho_{12,S_{t+1}})^2 \sigma_{2,S_{t+1}}^2 + [1 - (\rho_{12,S_{t+1}})^2] \sigma_{2,S_{t+1}}^2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{1,S_{t+1}}^2 & \sigma_{12,S_{t+1}} \\ \sigma_{12,S_{t+1}} & \sigma_{2,S_{t+1}}^2 \end{bmatrix} = \boldsymbol{\Omega}_{S_{t+1}}. \end{aligned}$$

[14] Think about collecting in $\mathbf{R}_{t+1}$ three different assets or portfolios, each characterized by two specific, not perfectly synchroneous regimes. Then one ought to expect to find $2^3 = 8$ regimes if the assets or portfolios are truly different. For instance, Guidolin and Timmermann (2006) use two U.S. stock portfolios and one U.S. Treasury portfolio to find that because the stocks are largely in synch, four regimes—i.e., two for stocks and two for bonds—are needed to fully characterize their data.

[15] Just to complete the list of possibilities, Ang and Bekaert (2002) have used weekly data to make famous a MSH(K) model, $\mathbf{R}_{t+1} = \boldsymbol{\mu} + \boldsymbol{\Omega}_{S_{t+1}}^{1/2} \boldsymbol{\epsilon}_{t+1}$. In principle it is also possible to envision the use of homoskedastic MSI(K) models, $\mathbf{R}_{t+1} = \boldsymbol{\mu}_{S_{t+1}} + \boldsymbol{\Omega}^{1/2} \boldsymbol{\epsilon}_{t+1}$ with constant covariance matrix. However, the empirical evidence of conditional heteroskedasticity is empirically so overwhelming that the instances in which MSI(K) has been found to appropriately fit the data are limited.

Finally, a typical and famous univariate, two-state MSIH(2) model is:

$$R_{t+1} = \mu_{S_{t+1}} + \sigma_{S_{t+1}} \epsilon_{t+1} \quad S_{t+1} = 1, 2,$$

where $\mu_1 < \mu_2$ and $\sigma_1 > \sigma_2$, which supports the interpretation of regime 1 as a "bear state" of high variance and of regime 2 as a "bull state" also characterized by lower volatility. For instance, Figure 1 shows such an estimation result for three alternative U.S. *excess* return (stock and bond) NYSE (New York Stock Exchange) portfolios, as obtained by Guidolin and Timmermann (2006a):

Parameter	Large caps	Small caps	Bonds
	Panel A: two-state AR(0) models		
μ_1	−0.0083	0.0045	0.0015
μ_2	0.0097	0.0109	−0.0012
σ_1	0.0641	0.0852	0.0246
σ_2	0.0335	0.0360	0.0070
p_{11}	0.7298	0.8910	0.9721
p_{22}	0.9424	0.9218	0.9196
Log-likelihood	996.3292	804.2038	1394.8273

Figure 1: MSIH(2,0) parameter estimates for U.S. stock and bond portfolios, monthly 1954-1999 U.S. data

In case you are wondering how it may be possible that the highest (lowest) mean return regime may imply the lowest (highest) risk, at least as measured by portfolio variance here, this will be discussed in Section 9.[16]

Consider instead a few univariate MSIAH(2,1) models for the same portfolios as in Figure 1.

Parameter	Large caps	Small caps	Bonds
	Panel B: two-state AR(1) models		
μ_1	−0.0239	0.0042	0.0012
μ_2	0.0154	0.0070	−0.0007
a_1	0.4400	0.1555	0.0645
a_2	−0.1639	0.2553	0.2989
σ_1	0.0444	0.0873	0.0247
σ_2	0.0347	0.0366	0.0071
p_{11}	0.3819	0.8768	0.9757
p_{22}	0.8521	0.9285	0.9315
Log-likelihood	993.5284	816.2982	1399.0809

Figure 2: MSIAH(2,1) parameter estimates for U.S. stock and bond portfolios, monthly 1954-1999 U.S. data

In Figure 2, expected excess returns, risk (as measured by state-specific volatility), as well as the

[16] You are possibly already objecting that, unless one is dealing with the market portfolio (here, the large capitalization stocks portfolio), it is mistaken to measure (systematic) risk using variance only. Stay tuned on this point too. Your former colleague Maria Luisa Magli (2013) has shown that when MS is taken into account, residual non-systematic risk stops indeed to be priced in stock returns.

persistence of returns all depend on the unobservable Markov state variable that may take two values.[17] For instance, large capitalization stock returns are more persistent in the second state than they are in the first state; for Treasury bonds, the opposite applies. Note that in these applications from Guidolin and Timmermann (2006a), as depicted in Figures 1 and 2, the two regimes are always very persistent, in the sense that estimates of $\hat{p}_{kk}$ (sometimes called the "stayer" probabilities) always largely exceed 0.5, meaning that you are more likely to remain in the initial regime than to switch out of it. For instance, in the case of 10-year Treasury returns, we have $\hat{p}_{11} = 0.97$ and $\hat{p}_{22} = 0.93$: this means that from the good (positive risk premium) state, one has only a 0.03 probability of switching to the bad state and 0.97 to remain, between t and $t+1$; from the bad (negative risk premium) state, such probabilities are 0.07 and 0.93, respectively. These considerations on the estimates of the main diagonal of the transition matrix $\mathbf{P}$ imply that when one estimates MS (vector) autoregressive models, two notions of persistence emerge and these may even be conflicting. On the one hand, persistence is captured by the usual autoregressive parameters, such as the estimates of the a coefficients in Figure 2. On the other hand, non-linear persistence in (6) is always captured by the implied persistence level of the Markov chain that intuitively stems from the size of the estimated transition probabilities on the main diagonal of $\hat{\mathbf{P}}$.[18] An interesting finding of applications of MSIAH(K,p) models to financial time series, also at relatively high frequencies, such as weekly, is that it is *not* true that asset returns are generally not serially correlated; they are except for a few particular states. For instance, in Figure 2, we see that large caps excess returns are highly and positively serially correlated in regime 1 ($\hat{a}_1 = 0.44$) but rather negatively serially correlated in regime 2 ($\hat{a}_2 = -0.16$). It is then not surprising that when one ignores the existence of regimes (i.e., when $K = 1$ is imposed without additional thoughts or tests), she tends to find one single $\hat{a} \simeq 0$ and not statistically significant: if you take $\hat{a}_1 = 0.44$ and $\hat{a}_2 = -0.16$ and you average them (maybe using their ergodic, long-run state probabilities defined below), you are bound to find a small positive number that is often unlikely to be statistically significant.[19]

When N is large, (6) implies the estimation of a large number of parameters:

$$K[N + pN^2 + N(N+1)/2 + (K-1)].$$

In this formula, KN is the number of regime-specific intercepts that need to be estimated; KpN^2 is the total number of regime-specific VAR matrix parameters; $KN(N+1)/2$ is the total number of

[17] For instance, conditioning on being and remaining (foreover, which is counter-intuitive of course) in a regime $k = 1, 2$, you could compute the regime-specific risk premium as

$$E[R_{t+1} - R^f | S_{t+1} = k] = \frac{\mu_k}{1 - a_k}.$$

[18] Formally, such a non-linear persistence derives from the size of the eigenvalues of $\hat{\mathbf{P}}'$ in the VAR representation used below in (9).

[19] It is like asking what is on average the weather like in Milan: averaging the 200 sunny days with the 150 days of rain and overcast conditions, you get an answer—cloudy with chances of sporadic, timid rain—that is not really an accurate one (that is a good forecast for London, not Milan). The source of the problem is clear: in Milan one tends to notice the prevalence of at least two clearly defined regimes, and avaraging across them to just report one simple answer discards most of the useful information.

regime-specific lower triangular Cholesky factor parameters that are needed; finally, $K(K-1)$ is the number of elements that can be estimated in the transition matrix, when the by-row summing up constraints are taken into account. Because the saturation ratio is simply the ratio between the total number of observations available for estimation (NT) and the total number of parameters, (6) implies a saturation ratio of

$$\frac{NT}{K[N + pN^2 + N(N+1)/2 + (K-1)]}.$$

For instance, for $K = 2$, $N = 8$, and $p = 1$ (the parameters characterizing some of the applications in Guidolin and Ono, 2006), this implies the estimation of 218 parameters and—with 35 years of monthly data—a saturation ratio of $(35 \times 12 \times 8)/218 = 15.4$ that, as we know, is much less than reassuring. Of course, not all MS models imply such low saturation ratios. For instance, for the same example a simpler MSIH(2) model (i.e., when $p = 0$) leads to a saturation ratio of $(35 \times 12 \times 8)/90 = 37.3$ which is quite an acceptable one, even though the burden to proceed to the estimation of 90 parameters remains considerable. However, Section 5 will introduce an iterative estimation scheme (called E-M algorithm) that makes this task possible.

MS models are known to capture central statistical features of asset returns. For instance, differences in conditional means across regimes enter the higher moments such as variance, skewness, and kurtosis. In particular, the variance is not simply the average of the variances across the two regimes: the difference in means also imparts an effect because the switch to a new regime contributes to volatility; this difference in regime means also generates non-zero conditional skewness. Section 8 performs these calculations in detail. Finally, differences in means in addition to differences in variances can generate persistence in levels as well as squared values akin to volatility persistence observed in many return series. Again differences in means play an important role in generating autocorrelation in first moments: without such differences, the autocorrelation will be zero. In contrast, volatility persistence can be induced either by differences in means or by differences in variances across regimes. In both cases, the persistence tends to be greater, the stronger the combined persistence, as measured by the diagonal transition probabilities collected in $\mathbf{P}$.[20] For instance, consider the simple case in which $K = 2$ and $\Pr(S_{t+1} = 1|S_t) = \Pr(S_{t+1} = 1) = \pi_1$ and $\Pr(S_{t+1} = 2|S_t) = \Pr(S_{t+1} = 2) = 1 - \pi_1$. This model does not represent a Markov chain switching process: it is a special, simpler case in which the probabilities of each of the two regimes are independent of the past regimes and information. In this case, we talk about IID *mixture distributions*. Yet, even in this case combining two normal densities delivers arbitrary skewness and excess kurtosis, as shown by the simulated density (once you simulate,

[20] This is the sense in which Marron and Wand (1992) emphasize that *mixtures* of normal distributions provide a flexible family that can be used to approximate many distributions. A mixtures of normals refers to a weighted sum of normal densities, in which the weights are themselves random. In the case of MS, such weights are given by the random state probabilities inferred over time, see Section 8. Mixtures of normals can also be viewed as a nonparametric approach to modeling the return distribution if the number of states, K, is allowed to grow with the sample size.

you can fit it using your favorite kernel density estimator) in Figure 3.

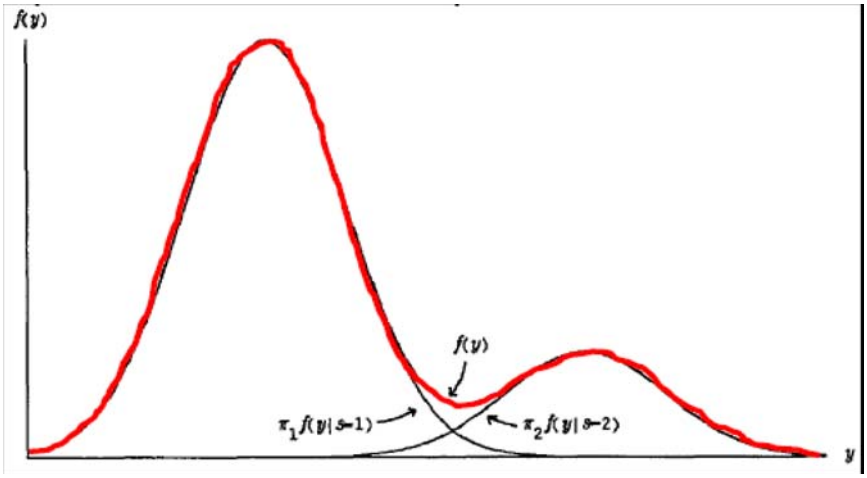


Figure 3: Mixture density with $y_t|S_t{=}\,1 \sim N(0,1)$, $y_t|S_t{=}\,2 \sim N(4,1)$ and $\Pr(S_t{=}\,1) = 0.8$

In fact, in Figure 3, the departure from normality caused by the mixture is so strong to take the form of an obvious bimodality. However, a mixture of two Gaussian random variables need not have a bimodal appearance: Gaussian mixtures can also produce a uni-modal density, and still allow skewness and kurtosis to differ from those of a single-regime Gaussian benchmark, see for example Figure 4. Therefore Markov models can clearly capture non-normalities in the data and can be useful in many risk management applications. Section 10.3 tackles this issue again in greater depth.

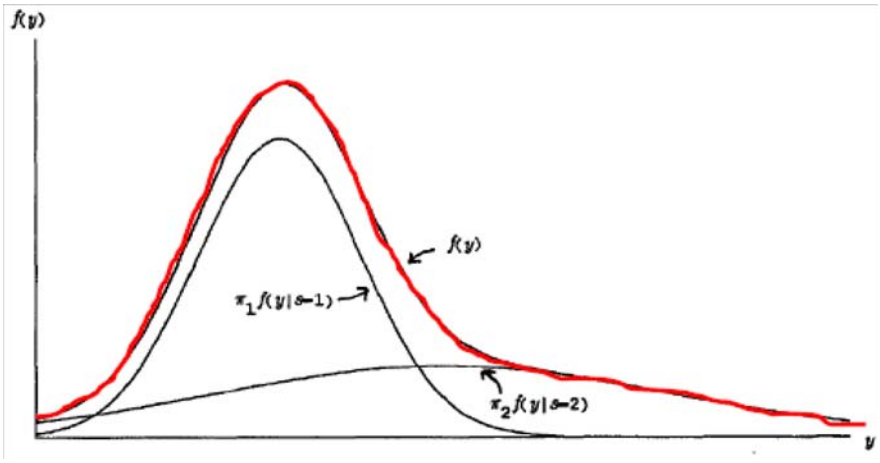


Figure 4: Mixture density with $y_t|S_t{=}\,1 \sim N(0,1)$, $y_t|S_t{=}\,2 \sim N(2,8)$ and $\Pr(S_t{=}\,1) = 0.6$

You will have already reckoned that the concept of MS model invokes the familiar notion of a Markov random variable (chain). What is the link between MS models and the well-known Markov chains analyzed in earlier courses, such as Derivatives I? MS models are defined in this way because of the crucial assumption that the unobservable state S_t is generated by a discrete-state, homogeneous, irreducible and ergodic first-order Markov chain such that:[21]

$$\Pr\left(S_t = j|\{S_j\}_{j=1}^{t-1}, \{\mathbf{R}_\tau\}_{\tau=1}^{t-1}\right) = \Pr\left(S_t = j|S_{t-1} = i\right) = p_{ij} \in (0,1), \tag{7}$$

[21]The assumption of a first-order Markov process is not especially restrictive, since a higher order Markov chain can always be reparameterized as a higher dimensional (i.e., when the number of regimes K is expanded to $K^* > K$) first-order Markov chain.

where p_{ij} is the generic $[i,j]$ element of the $K \times K$ *transition matrix* $\mathbf{P}$. Therefore in p_{ij}, the first index denotes that starting regime, the one that applies at time $t-1$, while the second index refers to the "landing" regime, that the system reaches at time t. $\mathbf{P}$ is a transition matrix because it collects the probabilities that the Markov chain follows when "transitioning" among alternative regimes. Clearly, the Markov nature of S_t derives from the fact that given all past information from both the regimes (albeit unobservable) and the return data—i.e., $\{S_j\}_{j=1}^{t-1}, \{\mathbf{R}_\tau\}_{\tau=1}^{t-1}$—the probability of the subsequent states simply depends on the most recent set of states recorded in the system, $\Pr\left(S_t = j | \{S_j\}_{j=1}^{t-1}, \{\mathbf{R}_\tau\}_{\tau=1}^{t-1}\right) = \Pr\left(S_t = j | S_{t-1} = i\right)$. In this sense, it is as if all the "memory" in (6) is simply retained through the most recent state, $S_{t-1} = i$.

For instance, in the special case of $K = 3$, if you have obtained an estimate of $\mathbf{P}$ equal to

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.88 & 0.09 & 0.03 \\ 0.01 & 0.96 & 0.03 \\ 0.23 & 0 & 0.77 \end{bmatrix},$$

and regime 1 is a bear state, regime 2 is a normal state, and regime 3 is a bull state, all regimes are rather persistent with estimated durations of 8, 25 and 4 months, respectively. However, $\hat{\mathbf{P}}$ also displays a peculiar structure by which while from the bear state it is possible to switch both to the normal and to the bull state (and the same applies to the normal regime, even though this state is by itself very persistent), from the bull state the market can only crash back into the bear state without transitioning through the normal state. This type of structure tends to mimic the occurrence of boom/bust dynamics, in which strongly bullish—even "bubbly" periods, at least in an empirical sense—later collapse into bear regimes of declining prices and negative returns.

We now explain in detail the five characterizations/properties that we have attributed above to the Markov chain process followed by S_t:

1. S_t follows a *discrete* Markov chain because it can take only a finite number K of regimes.

2. The Markov chain is a *first-order* one because $\Pr\left(S_t = j | \{S_j\}_{j=1}^{t-1}, \{\mathbf{R}_\tau\}_{\tau=1}^{t-1}\right) = \Pr\left(S_t = j | S_{t-1} = i\right)$; as already discussed, the current state is only affected by the state one period ago. However, this assumption is not critical because even though one would have $\Pr\left(S_t = j | \{S_j\}_{j=1}^{t-1}, \{\mathbf{R}_\tau\}_{\tau=1}^{t-1}\right) = \Pr(S_t = j | S_{t-1} = i, S_{t-2} = q)$, if you re-define $\ddot{S}_t = [S_t \; S_{t-1}]$ then it is clear that

$$\Pr\left(\ddot{S}_t = j | \{\ddot{S}_j\}_{j=1}^{t-1}, \{\mathbf{R}_\tau\}_{\tau=1}^{t-1}\right) = \Pr(\ddot{S}_t = j | \ddot{S}_{t-1} = i),$$

   i.e., any hth order Markov chain can be re-written as a first-order chain after re-defining the chain to include $h \geq 2$ "copies" of the original states, for a total of K^h total regimes.

3. *Ergodicity* implies the existence of a stationary $K \times 1$ vector of probabilities $\bar{\boldsymbol{\xi}}$ satisfying

$$\bar{\boldsymbol{\xi}} = \mathbf{P}' \bar{\boldsymbol{\xi}}. \tag{8}$$

This equation states that if the system in (6) were to be started from a vector configuration for probabilities $\bar{\boldsymbol{\xi}}$, this would be simply copied by the multiplication $\mathbf{P}'\bar{\boldsymbol{\xi}}$ in finding $\bar{\boldsymbol{\xi}}$ again. The meaning of such multiplication is easily seen when $\boldsymbol{\pi}$ is a unit vector $\mathbf{e}_j$, $j = 1, 2, ..., K$:[22]

$$\mathbf{P}'\boldsymbol{\pi} = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{K1} \\ p_{12} & p_{22} & \cdots & p_{K2} \\ \vdots & \cdots & \ddots & \vdots \\ p_{1K} & p_{2K} & \cdots & p_{KK} \end{bmatrix} \mathbf{e}_j = \begin{bmatrix} p_{j1} \\ p_{j2} \\ \vdots \\ p_{jK} \end{bmatrix},$$

i.e., the product gives the vector of (predicted) probabilities of switching from a fixed, initial regime j to each of the other possible regimes, besides the (predicted) probability of (6) remaining in regime j, p_{jj}. This example illustrates the sense in which (8) defines a $K \times 1$ vector of ergodic, also called *long-run or unconditional state probabilities*: if you start the system from a configuration of current state probabilities equal to $\bar{\boldsymbol{\xi}}$, then your prediction for the probabilities of the different regimes one-period forward is identical to $\bar{\boldsymbol{\xi}}$ itself, i.e., it is as if the system (6) has indeed reached a steady-state. Appendix A shows that $\bar{\boldsymbol{\xi}}$ can also be interpreted as the average, long-run time of occupation of the different regimes by the Markov chain, i.e. (at least heuristically), as

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} I_{\{S_t=j\}} = \bar{\boldsymbol{\xi}}'\mathbf{e}_j \qquad j = 1, 2, \ ..., \ K,$$

where $\bar{\boldsymbol{\xi}}'\mathbf{e}_j$ simply selects the jth element of the ergodic probability vector $\bar{\boldsymbol{\xi}}$.

An alternative way to think about ergodicity can be developed by first defining $\boldsymbol{\delta}_t$ as a $K \times 1$ vector made of zeros except for the jth element that equals 1 to signal $S_t = j$ and 0 otherwise. Then, when S_t follows a first-order Markov chain, it is possible to show (see Hamilton, 1994) that

$$\boldsymbol{\delta}_{t+1} = \mathbf{P}'\boldsymbol{\delta}_t + \mathbf{v}_{t+1} \tag{9}$$

where $\mathbf{v}_{t+1}$ is some error term with $E_t[\mathbf{v}_{t+1}] = E_t[\mathbf{v}_{t+H}] = 0$, $\forall H \geq 1$.[23] (9) represents a sort of VAR(1)-like representation of (a function of) the Markov state variable, $\boldsymbol{\delta}_t$. (9) is useful because it gives you a easy way to forecast the state in a MS model:

$$E_t[\boldsymbol{\delta}_{t+1}] = E_t[\mathbf{P}'\boldsymbol{\delta}_t + \mathbf{v}_{t+1}] = \mathbf{P}'\boldsymbol{\delta}_t,$$

which is exactly the $\mathbf{P}'\boldsymbol{\pi}$ predictive multiplication used above. Moreover

$$E_t[\boldsymbol{\delta}_{t+2}] = E_t[\mathbf{P}'\boldsymbol{\delta}_{t+1} + \mathbf{v}_{t+2}] = \mathbf{P}'E_t[\boldsymbol{\delta}_{t+1}] = (\mathbf{P}'\mathbf{P}')\boldsymbol{\delta}_t,$$

so that one can establish by induction that

$$E_t[\boldsymbol{\delta}_{t+H}] = (\mathbf{P}')^H\boldsymbol{\delta}_t,$$

[22] Note that the following expression uses the transpose of $\mathbf{P}$ and not $\mathbf{P}$ itself. Therefore because the rows of $\mathbf{P}$ need to sum to 1 by construction, obviously the same applies to sums across columns of $\mathbf{P}'$, which is used in what follows.

[23] Technically, we say that $\mathbf{v}_{t+1}$ is a martingale difference sequence.

where $(\mathbf{P}')^H \equiv \prod_{j=1}^{H} \mathbf{P}'$. At this point, a Markov chain (hence, the associated MS model) is ergodic if and only if[24]

$$p\lim(\mathbf{P}')^H \boldsymbol{\delta}_t = \bar{\boldsymbol{\xi}},$$

i.e., if a constant limit for the prediction as the forecast horizon diverges can be found that does not depend on what time t is. By construction, $\bar{\boldsymbol{\xi}}\boldsymbol{\iota}_K = 1$ (Appendix A provides details on this calculation). Note that $(\mathbf{P}')^H$ as defined above does not yield the same result as taking powers of each individual element of $\mathbf{P}'$. For instance, while the matrix of squares of a transposed transition matrix gives

$$\begin{bmatrix} 0.95^2 & 0.19^2 \\ 0.05^2 & 0.81^2 \end{bmatrix} = \begin{bmatrix} 0.9025 & 0.0361 \\ 0.0025 & 0.6561 \end{bmatrix},$$

the product of matrices yields

$$\begin{bmatrix} 0.95 & 0.19 \\ 0.05 & 0.81 \end{bmatrix} \cdot \begin{bmatrix} 0.95 & 0.19 \\ 0.05 & 0.81 \end{bmatrix} = \begin{bmatrix} 0.9120 & 0.3344 \\ 0.0088 & 0.6656 \end{bmatrix} \neq \begin{bmatrix} 0.9025 & 0.0361 \\ 0.0025 & 0.6561 \end{bmatrix}.$$

4. The Markov chain process followed by S_t is time-*homogeneous* because $\mathbf{P}$ is a constant matrix over time, i.e., p_{ij} does not change for all pairs i and j. This is not a superfluous point because more complex time-varying transition probability models with a dynamic transition matrix $\mathbf{P}$ have been studied by econometricians and appear to be particularly loved by financial economists (see Guidolin, 2012, for references and a discussion). Equivalently, in these models S_t follows a time-heterogeneous Markov chain, so that $p_{jj,t}$ becomes itself a function of time.[25]

5. Finally, *irreducibility* of the Markov chain implies that $\bar{\boldsymbol{\xi}} > \mathbf{0}$, meaning that all unobservable states are possible over time and no absorbing states or cycles among states exist. Consider for instance the case $K = 3$, then

$$\breve{\mathbf{P}} = \begin{bmatrix} p_{11} & p_{12} & 0 \\ p_{21} & p_{22} & 0 \\ 0 & p_{32} & p_{33} \end{bmatrix} = \begin{bmatrix} p_{11} & 1-p_{11} & 0 \\ 1-p_{22} & p_{22} & 0 \\ 0 & 1-p_{33} & p_{33} \end{bmatrix}$$

implies that it is impossible to reach state 3 from the other two states: as soon as one leaves regime 3, because $p_{33} \in (0,1)$ but $p_{i3} = 0$ for $i = 1, 2$, it becomes impossible to ever return again to state 3. Therefore, the third element of $\bar{\boldsymbol{\xi}}$ will have to be zero because $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} I_{\{S_t=3\}} = 0$.

In practice, the matrix $\mathbf{P}$ is unknown and hence $\bar{\boldsymbol{\xi}}$ can be at most estimated given knowledge/estimates of $\mathbf{P}$ extracted from the (full-sample) information set $\Im_T = \{\mathbf{R}_\tau\}_{\tau=1}^{T}$. For simplicity, we will denote as $\bar{\boldsymbol{\xi}}$ such an "estimated" vector of ergodic (unconditional) state probabilities**.** Appendix A shows that besides representing the vector of ergodic probabilities, $\bar{\boldsymbol{\xi}}$ also represents the vector of unconditional

[24]Here plim means "limit in probability" as $T \to \infty$.

[25]Note the twisted roller-coaster effect: in this case, the model that is being used to capture instability in statistical relationships becomes itself unstable, i.e., we would therefore model and forecast the instability of instability, i.e., instability^2.

probabilities, i.e., the average frequencies of the K different regimes as the sample size $T \to \infty$, sometimes also denoted as $\boldsymbol{\pi}$. The Appendix also offers one example of actual calculation of $\bar{\boldsymbol{\xi}}$ that you should be familiar with. In the special case of $K = 2$, one obtains that

$$\bar{\xi}_1 = \frac{1 - p_{22}}{2 - p_{11} - p_{22}} \qquad \bar{\xi}_2 = \frac{1 - p_{11}}{2 - p_{11} - p_{22}}.$$

Appendix A shows that the formula $\mathbf{P}'\bar{\boldsymbol{\xi}} = \bar{\boldsymbol{\xi}}$ implies these expressions for $\bar{\xi}_1$ and $\bar{\xi}_2$, where obviously $\bar{\xi}_1 + \bar{\xi}_2 = 1$.

4.1. *One three-state example*

Before moving to the more technical Sections 5 and 6 of this chapter, it is useful to take a look at one more example to get additional "feeling" for what the estimation outputs from a MS model may look like. The spirit is also to educate you to the fact that—although this remains less common—there is nothing magical or unavoidable in setting $K = 2$ in a MS exercise. Sure enough, $K = 2$ is easier and implies the estimation of less parameters than $K = 3$, but nothing in the universe is going to scream for vengeance in case you happen to find that your data require $K \geq 3$. Figure 3 shows the estimation outputs of 6 alternative univariate three-state MS models applied to 1954-1999 monthly US data on large stocks, small stocks, and long-term (10-year) Treasury bond excess returns, once more from Guidolin and Timmermann (2006a).

Parameter	Large caps	Small caps	Bonds	Large caps	Small caps	Bonds
	Panel C: three-state AR(0) models			Panel D: three-state AR(1) models		
μ_1	−0.0169	−0.0245	0.0029	−0.0289	−0.0155	0.0000
μ_2	0.0061	0.0121	−0.0014	0.0057	0.0070	−0.0003
μ_3	0.0371	0.0867	0.0006	0.0306	0.1106	0.0026
a_1	NA	NA	NA	0.3804	0.1215	0.0948
a_2	NA	NA	NA	−0.0290	0.2612	0.5497
a_3	NA	NA	NA	−0.2615	−0.3356	0.0486
σ_1	0.0722	0.0744	0.0337	0.0452	0.0753	0.0170
σ_2	0.0354	0.0365	0.0056	0.0300	0.0359	0.0029
σ_3	0.0181	0.0762	0.0181	0.0371	0.0726	0.0334
p_{11}	0.7356	0.8578	0.9799	0.4578	0.8776	0.9809
p_{22}	0.9663	0.9232	0.9206	0.9562	0.9347	0.8932
p_{33}	0.6716	0.4533	0.9726	0.7155	0.3433	0.9800
p_{12}	0.0017	0.0011	0.0069	0.0079	0.0014	0.0118
p_{21}	0.0313	0.0645	0.0001	0.0418	0.0592	0.1067
p_{31}	0.0052	0.0029	0.0077	0.2129	0.0082	0.0116
Log-likelihood	1004.7285	814.9706	1420.7636	1005.6759	826.5749	1429.0516

Figure 3: MSIAH(3,p) parameter estimates for U.S. stock and bond portfolios, monthly 1954-1999 U.S. data

Columns 2-4 concern estimates of MSIH(3,0) models in which there are no autoregressive components; columns 5-7 concerns estimates of MSIAH(3,1) models. To save space (and also because we do not known yet how to compute p-values for a MS model, or to perform estimation), we have omitted standard errors of the individual parameter estimates, similarly to Figures 1 and 2. In the case of equities, and independently of the model estimated, there are three states with a natural economic interpretation: a bad, bear regime of negative mean excess returns; a normal regime of positive but moderate mean excess returns; and a (strong) bull regime of exceptionally high mean excess returns.

Once more, and this remains puzzling at least in the case of large cap excess returns which are quite similar (i.e., highly correlated) to excess returns on the market portfolio, volatility is higher in the bear regime than in the two remaining regimes. In fact, in the case of large caps, as the estimate of μ increases across regimes, the estimate of σ declines: less risk maps into higher risk premia. In the case of excess bond returns, the match with the properties described above for stock portfolios is only partial:[26] in the case of bonds, the highest risk premium state also carries the highest variance and the ordering of the $\hat{\mu}_k$ estimates is the same as the ordering for the $\hat{\sigma}_k$, $k = 1, 2, 3$, which is sensible. The Markov switching estimates of the AR(1) coefficients in columns 4-7 confirm what we had observed before: regimes exist in which linear persistence is strong and statistically significant; on the contrary, at least in the case of equities, the bad, bear regimes imply negative and (you may check) statistically significant negative AR(1) coefficients, which means that lower excess returns today forecast a rebound, higher excess returns, in the subsequent period. Finally, all regimes, especially in the MSIH(3,0) case, are persistent, similarly to what was reported in Figures 1 and 2. Figure 4 reports the ex-post smoothed probabilities of the three states for large and small cap stock portfolios from the MSIH(3,0) models.[27]

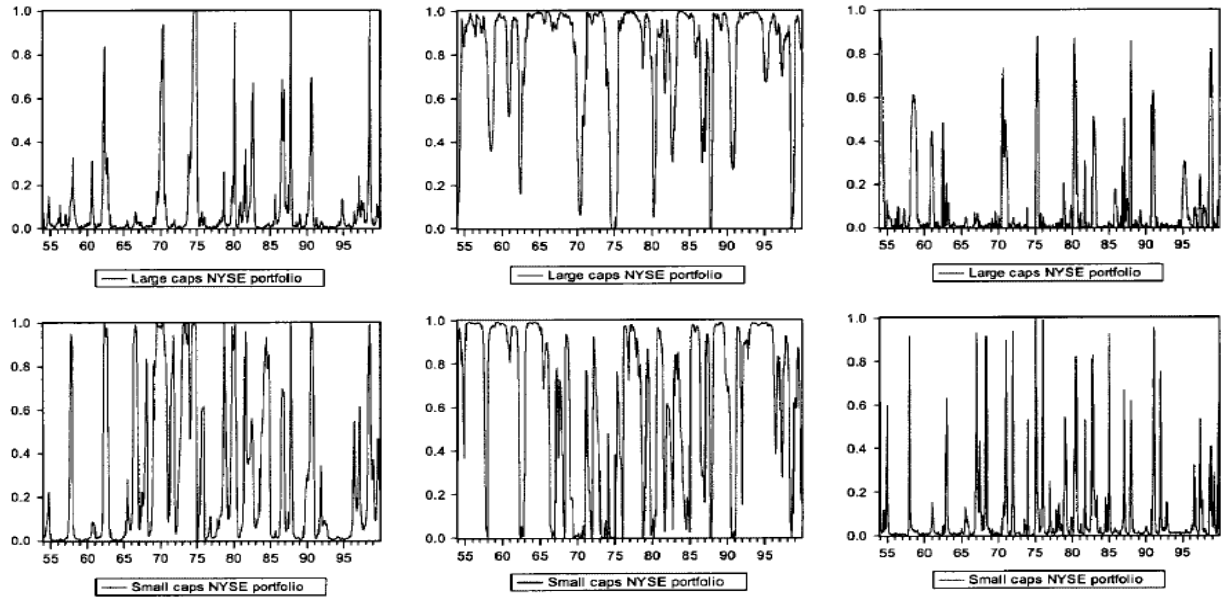


Figure 4: MSIH(3,0) smoothed probability estimates for large and small capitalization U.S. stock portfolios

In Figure 4, the two rows of plots refer to large and small cap excess returns, respectively. The three columns refer each to the three regimes. All regimes are also visibly persistent, especially the second, normal state: when you enter such a regime, you tend to stay there on average for a few years. Moreover, the smoothed probabilities of regime 2 is rather similar across small and large caps. Regimes 1 and especially 3 tend instead to be "spikier", although if you look closely some persistence (average duration is anyway several months) appears also in this case. In the case of bear state 1, this is most

[26] Make no mistake: the ordering and labeling of regimes is completely arbitrary, i.e., also in the case of excess bond returns, the estimates of the regime-specific means can be sorted in the same way we did in the case of stocks.

[27] As Section 5.1 will clarify, the smoothed probabilities represent full-sample, complete information inferences on the probability of each of the three regimes at each point in time t.

interesting: indeed when you enter such a bad regime of declining stock prices, you tend to remain there with probability $\hat{p}_{11} = 0.74$ in the case of large caps and $\hat{p}_{11} = 0.86$ in the case of small caps. Applying standard results from Poisson distributions, you have that the average durations of a bear regime are:[28]

$$Avg.\text{ duration large(1)} = \frac{1}{1-0.74} = 3.8 \text{ months} \quad Avg.\text{ duration small(1)} = \frac{1}{1-0.86} = 7 \text{ months},$$

respectively. Of course, to a risk managers, to know that markets will be likely to remain bearish for the next 4 or even 7 months may be incredibly useful. Average duration calculations confirm the high persistent of regime 2 for both large and small cap stocks:

$$Avg.\text{ duration large(2)} = \frac{1}{1-0.97} = 29.7 \text{ months} \quad Avg.\text{ duration small(2)} = \frac{1}{1-0.92} = 13 \text{ months}.$$

Analogous calculations find that the average durations for regimes 3 are 3 and 2 months, for large and small stocks, respectively.

Figure 5 concludes showing the smoothed probabilities estimated from a MSIH(3,0) model in the case of excess bond returns.

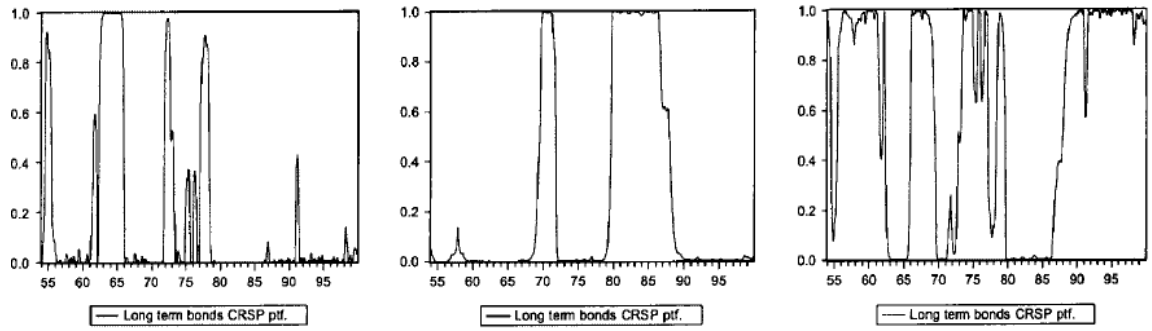


Figure 5: MSIH(3,0) smoothed probability estimates for U.S. 10-year Treasuries

Clearly, also in the case of bonds all the regimes are highly persistent, as confirmed by both the pictures and the large estimates of $\hat{p}_{kk}$, $k = 1, 2, 3$. Indeed the corresponding average durations in the case of bonds are 49.8, 12.6, and 36.5 months, respectively. Regime 2 tends to appear more episodically but when it does, it is highly persistent. Of course a comparison of Figures 4 and 5 shows that the regime dynamics in U.S. stock and bond excess returns appears to be rather different. Guidolin and Timmermann (2006, 2007) elaborate on the implications of such differences and their asset management implications.

5. **Estimation of MS models**

MS(I)VARH models are estimated by maximum likelihood. However some assumptions have to be imposed to guarantee at least the local identifiability of the parameters (collected in a vector $\boldsymbol{\theta}$) under estimation.[29] The vector $\boldsymbol{\theta}$ collects all regime-dependent parameters in $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$, $\{\mathbf{A}_{j,k}\}_{j=1}^{p}{}_{k=1}^{K}$, and

[28] Given a regime $j = 1, 2, 3$ with estimated "stayer" probability $\hat{p}_{jj} < 1$, the average estimated duration, i.e., the expected time spent in each regime, is defined as $duration(j) = 1/(1 - \hat{p}_{jj})$.

[29] Later on $\boldsymbol{\theta}$ will be expanded to also include the elements of the transition matrix $\mathbf{P}$, to form a new vector γ. However, the conditions discussed here simply concern $\boldsymbol{\theta}$ in the definition given below.

$\{\mathbf{\Omega}_k^{1/2}\}_{k=1}^K$. Roughly speaking, local identifiability means that at least in a neighborhood of the true but unknown vector of parameters $\boldsymbol{\theta}_0$ that generates the data in (6), it must be true that $\boldsymbol{\theta}_0$ is also the vector of parameters that maximizes the log-likelihood function. Krolzig (1997, pp. 93-95) generalizes results in Leroux (1992) to show that under the assumption of multivariate Gaussian shocks (to the measurement equation, see Appendix B), MSIVARH models are identifiable up to any arbitrary relabeling of unobservable states.

Estimation is performed through the EM (Expectation-Maximization) algorithm proposed by Dempster et al. (1977) and Hamilton (1990), a filter that allows the iterative calculation of the one-step ahead forecast of the state vector $\boldsymbol{\xi}_{t+1|t}$ given the information set $\Im_t$ and the consequent construction of the log-likelihood function of the data. The algorithm is dividend in two logical steps, the Expectation and the Maximization steps. Start from the model written in state-space form (see Appendix B for an explanation of what this means, but even a very superficial understanding of this aspect will not prevent you from following the argument below),

$$\begin{aligned} \mathbf{R}_t &= \mathbf{X}_t \mathbf{A} \boldsymbol{\xi}_t + \mathbf{\Sigma}_K \left((\boldsymbol{\xi}_t) \otimes \mathbf{I}_N \right) \boldsymbol{\epsilon}_t \\ \boldsymbol{\xi}_{t+1} &= \mathbf{P}' \boldsymbol{\xi}_t + \mathbf{v}_{t+1}. \end{aligned}$$

Here $\mathbf{X}_t$ is a $N \times (Np+1)$ matrix of predetermined variables with structure $[1 \ \mathbf{R}'_{t-1} ... \mathbf{R}'_{t-p}] \otimes \boldsymbol{\iota}_N$, $\mathbf{A}$ is a $(Np+1) \times NK$ matrix collecting the VAR parameters, both means or intercepts and autoregressive coefficients, in all regimes

$$\mathbf{A} = \begin{bmatrix} \boldsymbol{\mu}_1^{'} & \boldsymbol{\mu}_2^{'} & \cdots & \boldsymbol{\mu}_K^{'} \\ \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{p1} & \mathbf{A}_{p2} & \cdots & \mathbf{A}_{pK} \end{bmatrix},$$

$\mathbf{\Sigma}_K$ is a $N \times NK$ matrix collecting all the possible K "square root" (Choleski decomposition) covariance matrix factors $[\Omega_1^{1/2} \ \Omega_2^{1/2} \ ... \ \Omega_K^{1/2}]$ such that $\forall t$, $\mathbf{\Sigma}_K \left(\boldsymbol{\xi}_t \otimes \mathbf{I}_N \right) \left(\boldsymbol{\xi}_t \otimes \mathbf{I}_N \right)' \mathbf{\Sigma}'_K = \Omega_{S_t}$, the covariance matrix of the asset return innovations $\boldsymbol{\epsilon}_t$. For the sake of argument, assume that all the parameters of the model in $\mathbf{A}$ and $\mathbf{\Sigma}_M$ are known. Because our goal is to perform estimation of $\{\boldsymbol{\mu}_k\}_{k=1}^K$, $\{\mathbf{A}_{j,k}\}_{j=1 \ k=1}^{p \quad K}$, $\{\mathbf{\Omega}_k^{1/2}\}_{k=1}^K$ among the other parameters, and these enter $\mathbf{A}$ and $\mathbf{\Sigma}_M$, we shall see below how this assumption is later removed. We separately describe the expectation and maximization steps in Sections 5.1 and 5.2, and then bring them together in Section 5.3.

5.1. *The expectation step: filtered and smoothed probabilities*

The expectation step consists of taking parameter estimates from the previous maximization step as given (call it $\boldsymbol{\theta}$) and in computing both the time series sequence of filtered probability vectors, $\{\hat{\boldsymbol{\xi}}_{t|t}\}_{t=1}^T$, and the time series sequence of smoothed probability vectors, $\{\hat{\boldsymbol{\xi}}_{t|T}\}_{t=1}^T$, with the latter depending on the former. The fact that one needs to use $\{\hat{\boldsymbol{\xi}}_{t|t}\}_{t=1}^T$ and $\{\hat{\boldsymbol{\xi}}_{t|T}\}_{t=1}^T$ to extract inferences concerning the dynamics of regimes over time (technically, concerning $\{\boldsymbol{\delta}_t\}_{t=1}^T$) derives from the latent nature of $\{S_t\}_{t=1}^T$ and therefore $\{\boldsymbol{\delta}_t\}_{t=1}^T$ in a MS model.

Algorithmically, the expectation step is the outcome of a few smart applications of Bayes' law that allow us to recursively derive a sequence of *filtered* probability distributions and then (going backwards) a sequence of *smoothed* probability distributions. What filtered and smoothed probabilities are and how these are inferred from the data is explained below. Starting from a *prior* on the $K \times 1$ vector of probabilities $\boldsymbol{\xi}_t$, $\forall t \geq 1$, defined as[30]

$$\Pr\left(\boldsymbol{\xi}_t|\Im_{t-1}\right) = \sum_{\boldsymbol{\xi}_{t-1}} \Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t-1}\right) \Pr\left(\boldsymbol{\xi}_{t-1}|\Im_{t-1}\right).$$

This prior simply takes the time $t-1$ posterior $\Pr\left(\boldsymbol{\xi}_{t-1}|\Im_{t-1}\right)$ defined below and turns it into a new prior, $\Pr\left(\boldsymbol{\xi}_t|\Im_{t-1}\right)$. Note that the elements of $\Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t-1}\right)$ are simply the elements of the transition matrix $\mathbf{P}$. The *posterior* distribution of $\boldsymbol{\xi}_t$ given $\Im_t = \{\Im_{t-1}, \mathbf{R}_t\}$, $\Pr\left(\boldsymbol{\xi}_t|\Im_t\right)$, is then given by[31]

$$\Pr\left(\boldsymbol{\xi}_t|\Im_t\right) = \frac{\Pr\left(\mathbf{R}_t|\boldsymbol{\xi}_t, \Im_{t-1}\right) \Pr\left(\boldsymbol{\xi}_t|\Im_{t-1}\right)}{\Pr\left(\mathbf{R}_t|\Im_{t-1}\right)}, \tag{10}$$

where $\Pr\left(\mathbf{R}_t|\Im_{t-1}\right) = \sum_{\boldsymbol{\xi}_t} \Pr\left(\mathbf{R}_t, \boldsymbol{\xi}_t|\Im_{t-1}\right) = \sum_{\boldsymbol{\xi}_t} \Pr\left(\mathbf{R}_t|\boldsymbol{\xi}_t, \Im_{t-1}\right) \Pr\left(\boldsymbol{\xi}_t|\Im_{t-1}\right)$ is the unconditional likelihood of the current observation given its past. For compactness it can also be expressed as

$$\boldsymbol{\eta}_t' \hat{\boldsymbol{\xi}}_{t|t-1} = \boldsymbol{\iota}_N' \left(\boldsymbol{\eta}_t \odot \hat{\boldsymbol{\xi}}_{t|t-1}\right)$$

where $\odot$ denotes the element by element (Hadamard) product and the $K \times 1$ vector $\boldsymbol{\eta}_t$ collects the possible log-likelihood values as a function of the realized state:[32]

$$\boldsymbol{\eta}_t \equiv \begin{bmatrix} p(\mathbf{R}_t|\boldsymbol{\xi}_t = \mathbf{e}_1, \Im_{t-1}) \\ p(\mathbf{R}_t|\boldsymbol{\xi}_t = \mathbf{e}_2, \Im_{t-1}) \\ \vdots \\ p(\mathbf{R}_t|\boldsymbol{\xi}_t = \mathbf{e}_K, \Im_{t-1}) \end{bmatrix} = \begin{bmatrix} (2\pi)^{-1/2}|\boldsymbol{\Omega}_1|^{-1/2} \exp\left[(\mathbf{R}_t - \mathbf{X}_t\mathbf{A}\mathbf{e}_1)\boldsymbol{\Omega}_1^{-1}(\mathbf{R}_t - \mathbf{X}_t\mathbf{A}\mathbf{e}_1)\right] \\ (2\pi)^{-1/2}|\boldsymbol{\Omega}_2|^{-1/2} \exp\left[(\mathbf{R}_t - \mathbf{X}_t\mathbf{A}\mathbf{e}_2)\boldsymbol{\Omega}_2^{-1}(\mathbf{R}_t - \mathbf{X}_t\mathbf{A}\mathbf{e}_2)\right] \\ \vdots \\ (2\pi)^{-1/2}|\boldsymbol{\Omega}_K|^{-1/2} \exp\left[(\mathbf{R}_t - \mathbf{X}_t\mathbf{A}\mathbf{e}_K)\boldsymbol{\Omega}_K^{-1}(\mathbf{R}_t - \mathbf{X}_t\mathbf{A}\mathbf{e}_K)\right] \end{bmatrix}.$$

Of course (10) is nothing but Bayes' rule applied to our problem. At this point, the vector of *filtered* probabilities, $\hat{\boldsymbol{\xi}}_{t|t}$, corresponds to the discrete probability distribution of the possible states perceived on the basis of the information set $\Im_t$:

$$\hat{\boldsymbol{\xi}}_{t|t} = \frac{\boldsymbol{\eta}_t \odot \hat{\boldsymbol{\xi}}_{t|t-1}}{\boldsymbol{\iota}_K' \left(\boldsymbol{\eta}_t \odot \hat{\boldsymbol{\xi}}_{t|t-1}\right)}. \tag{11}$$

[30] In the expression below, $\sum_{\boldsymbol{\xi}_{t-1}}(\cdot)$ denotes the summation over all the elements of $\boldsymbol{\xi}_{t-1}$. For instance, when $K = 2$:

$$\Pr\left(\xi_t^1|\Im_{t-1}\right) = \sum_{k=1}^{2} \Pr\left(\xi_t^1|\xi_{t-1}^k\right) \Pr\left(\xi_{t-1}^k|\Im_{t-1}\right).$$

A prior distribution on some random vector $\boldsymbol{\xi}_t$ simply collects your initial views on what sensible values for the elements of $\boldsymbol{\xi}_t$ are. This is of course a good point to stop and review what Bayes' law is from your undergraduate textbooks and notes.

[31] In a Bayesian problem, the posterior distribution of the random vector $\boldsymbol{\xi}_t$ collects your views after you have observed the data up to time t (here it is called $\Im_t$), and therefore reflects a mixture between your initial priors and the data, as summarized by their likelihood function, in this case $\Pr\left(\mathbf{R}_t|\boldsymbol{\xi}_t, \Im_t\right)$.

[32] The Hadamard product is a bit different from the Kronecker product. Carefully observe the following example, that echoes a similar example in chapter 6 for the Kronecker product:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \odot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} \\ a_{21}b_{21} & a_{22}b_{22} \end{bmatrix}.$$

Programmers also refer to $\odot$ as the "dot product".

A filtered probability is the best assessment of (inference on) the current state, based on real time information. Of course, $\hat{\boldsymbol{\xi}}'_{t|t}\boldsymbol{\iota}_K = 1$, the filtered probability at time t they all sum up to 1. The expressions in (10) and (11) emphasize that the filtered probability of being in regime $k = 1, 2, ..., K$ at time t is the ratio between: the sum of the probabilities of reaching regime k from each of the K possible regimes, including k itself, scaled (divided by) the total probability of $\mathbf{R}_t$, given all past information.

This algorithm is completed by the transition equation that implies that

$$E_t[\boldsymbol{\xi}_{t+1}|\Im_t] \equiv E_t[\boldsymbol{\xi}_{t+1}] = \hat{\boldsymbol{\xi}}_{t+1|t} = \mathbf{P}'\hat{\boldsymbol{\xi}}_{t|t}, \tag{12}$$

i.e., the predicted probability vector $\hat{\boldsymbol{\xi}}_{t+1|t}$—note, what you expect the state probabilities will be at time $t+1$ on the basis of current information $\Im_t$—is simply $\mathbf{P}'$ times the vector of filtered probabilities at time t. Of course, this mimics the $E_t[\boldsymbol{\delta}_{t+1}] = \mathbf{P}'\boldsymbol{\delta}_t$ recursion illustrated in Section 4, when $\boldsymbol{\delta}_t$ is replaced by $\hat{\boldsymbol{\xi}}_{t|t}$, which means that at time t—just because the states are unobservable—you are not sure of the nature of the starting regime and as such you use the inferred $\hat{\boldsymbol{\xi}}_{t|t}$ from the previous step of the algorithm. Assuming that the initial state probability vector $\hat{\boldsymbol{\xi}}_{1|0}$ is unknown and must be estimated, (11)-(12) define an iterative algorithm that allows one to generate a sequence of filtered state probability vectors $\{\hat{\boldsymbol{\xi}}_{t|t}\}_{t=1}^{T}$.[33]

The filtered probabilities are the product of a limited information technique, since despite the availability of a sample of size T, each $\hat{\boldsymbol{\xi}}_{t|t}$ is filtered out of the information set $\Im_t$ only, ignoring $\{\mathbf{R}_\tau\}_{\tau=t+1}^{T}$. However, once the full time series of filtered probabilities $\{\hat{\boldsymbol{\xi}}_{t|t}\}_{t=1}^{T}$ has been calculated, Kim's (1994) algorithm is easily implemented to recover the sequence of *smoothed* probability distributions $\{\hat{\boldsymbol{\xi}}_{t|T}\}_{t=1}^{T}$ by iterating the following algorithm backwards, starting from the filtered (and smoothed) probability distribution $\hat{\boldsymbol{\xi}}_{T|T}$ produced by (11)-(12). Observe that

$$\begin{aligned}
\hat{\boldsymbol{\xi}}_{t|T} &= \Pr\left(\boldsymbol{\xi}_t|\Im_T\right) = \sum_{\boldsymbol{\xi}_{t+1}} \Pr\left(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1}|\Im_T\right) \quad \text{(by the definition of probability)} \\
&= \sum_{\boldsymbol{\xi}_{t+1}} \Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1}, \Im_T\right) \Pr\left(\boldsymbol{\xi}_{t+1}|\Im_T\right) \quad \text{(by the definition of joint probability)} \\
&= \sum_{\boldsymbol{\xi}_{t+1}} \Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1}, \Im_t, \{\mathbf{R}_\tau\}_{\tau=t+1}^{T}\right) \Pr\left(\boldsymbol{\xi}_{t+1}|\Im_T\right) \\
&= \sum_{\boldsymbol{\xi}_{t+1}} \frac{\Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1}, \Im_t\right) \Pr\left(\{\mathbf{R}_\tau\}_{\tau=t+1}^{T}|\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1}, \Im_t\right)}{\Pr\left(\{\mathbf{R}_\tau\}_{\tau=t+1}^{T}|\boldsymbol{\xi}_{t+1}, \Im_t\right)} \Pr\left(\boldsymbol{\xi}_{t+1}|\Im_T\right) \\
&= \sum_{\boldsymbol{\xi}_{t+1}} \Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t+1}, \Im_t\right) \Pr\left(\boldsymbol{\xi}_{t+1}|\Im_T\right) \\
&= \sum_{\boldsymbol{\xi}_{t+1}} \frac{\Pr\left(\boldsymbol{\xi}_t|\Im_t\right) \Pr\left(\boldsymbol{\xi}_{t+1}|\boldsymbol{\xi}_t, \Im_t\right)}{\Pr\left(\boldsymbol{\xi}_{t+1}|\Im_t\right)} \Pr\left(\boldsymbol{\xi}_{t+1}|\Im_T\right)
\end{aligned}$$

[33]This assumption implies that $\hat{\boldsymbol{\xi}}_{1|0}$ is a $K \times 1$ vector that must be estimated. A simpler alternative is postulate that the stochastic process had start from a deterministic but unknown state S_0 that must be estimated along with the remaining parameters (in practice it is $\boldsymbol{\delta}_0$ that is estimated). Alternatively, $\hat{\boldsymbol{\xi}}_{1|0}$ might be assumed to correspond to the stationary unconditional probability distribution such that $\bar{\boldsymbol{\xi}} = \mathbf{P}'\bar{\boldsymbol{\xi}}$.

because the first-order Markov structure implies that $\Pr\left(\{\mathbf{R}_\tau\}_{\tau=t+1}^{T}|\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+1}, \Im_t\right) = \Pr(\{\mathbf{R}_\tau\}_{\tau=t+1}^{T}|\boldsymbol{\xi}_{t+1}, \Im_t)$. Hence $\hat{\boldsymbol{\xi}}_{t|T}$ can be re-written as

$$\hat{\boldsymbol{\xi}}_{t|T} = \left(\mathbf{P}'\left(\hat{\boldsymbol{\xi}}_{t+1|T} \oslash \hat{\boldsymbol{\xi}}_{t+1|t}\right)\right) \odot \hat{\boldsymbol{\xi}}_{t|t}, \tag{13}$$

where $\oslash$ denotes element-by-element division and $\Pr\left(\boldsymbol{\xi}_{t+1}|\boldsymbol{\xi}_t, \Im_t\right)$ equals by construction the transition matrix driving the first order Markov chain.[34] (13) is initialized by setting $t = T - 1$ thus obtaining

$$\hat{\boldsymbol{\xi}}_{T-1|T} = \left(\mathbf{P}'\left(\hat{\boldsymbol{\xi}}_{T|T} \oslash \hat{\boldsymbol{\xi}}_{T|T-1}\right)\right) \odot \hat{\boldsymbol{\xi}}_{T-1|T-1}$$

and so forth, proceeding backwards until $t = 1$.[35]

What is the deep difference between filtered and smoothed probability (vectors)? Clearly, while the filtered $\{\hat{\boldsymbol{\xi}}_{t|t}\}_{t=1}^{T}$ condition on information up to time t, smoothed probabilities $\{\hat{\boldsymbol{\xi}}_{t|T}\}_{t=1}^{T}$ condition on the entire sample and hence reflect more information. Therefore a smoothed probability represents an ex-post measure of the state of the model at time t, where $t << T$ is possible. A filtered probability provides instead a recursive, real time assessment (filter) on the current state. One example that may ease you into an understanding of the difference comes from comparing the two questions:

- Given what I know about what the weather has been like during the past few weeks, what is chance of recording a high atmospheric pressure today (also given observed conditions today)? This requires a real-time, recursive assessment akin to the calculation underlying a filtered probability.

- Given the information on the weather in the past 12 months and up to today, what was the chance of a high atmospheric pressure today 4 months ago? This requires a full-information, but backward-looking assessment that employs data that were not yet available 4 months ago.

Obviously, finance people tend to operate in real time, to focus on forecasting future market conditions, and as such they tend to care more for filtered probabilities than for smoothed ones, even though it is clear that the two concepts always coincide at the end of all available data.[36] In fact, using (12), the focus frequently goes to the vector of predicted H-step ahead probabilities, with $H \geq 1$:

$$E_t[\boldsymbol{\xi}_{t+H}] = \hat{\boldsymbol{\xi}}_{t+H|t} = (\mathbf{P}')^H \hat{\boldsymbol{\xi}}_{t|t}.$$

On the contrary, the smoothed probabilities correspond to the logical approach of historians to assessing events: using all the available information at time T, the researcher wants to understand what the

[34]The element-wise division operator $\oslash$ is defined similarly to the Hadamard "dot product", for instance:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \oslash \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}/b_{11} & a_{12}/b_{12} \\ a_{21}/b_{21} & a_{22}/b_{22} \end{bmatrix}.$$

[35]Note that $\hat{\boldsymbol{\xi}}_{T|T}$ and $\hat{\boldsymbol{\xi}}_{T-1|T-1}$ will be known from the application of Hamilton's smoothing algorithm, and that $\hat{\boldsymbol{\xi}}_{T|T-1} = \mathbf{P}'\hat{\boldsymbol{\xi}}_{T-1|T-1}$.

[36]You will easily recognize that when $t = T$, the end of the available sample, $\hat{\boldsymbol{\xi}}_{t|T} = \hat{\boldsymbol{\xi}}_{T|T} = \hat{\boldsymbol{\xi}}_{t|t}$, i.e., filtered and smoothed coincide by definition at the very end of the sample.

probability of the K different regimes had been at time $t < T$. Clearly, using information posterior to time T may easily make our understanding of events more accurate and interesting. Yet, the fact remains that such a probabilistic assessment would not have been available to investors at time $t < T$, i.e., in real time.[37] However, we will see in Section 5.2 that smoothed probabilities also play a crucial role in ML estimation of MS models.

5.2. *The maximization step*

What follows is not for people with a fragile health. Please read and meditate under medical care. The point of taking a look at the conditions and results that follow is that it is important to have some idea for what happens behind the curtains of Matlab's routines. Call $\boldsymbol{\theta}$ the vector collecting all the parameters appearing in the measurement equation and $\boldsymbol{\rho}$ the vector collecting the transition probabilities in $\mathbf{P}$, i.e. $\boldsymbol{\theta} \equiv [vec(\mathbf{A}) | vec(\boldsymbol{\Sigma}_K)]$ and $\boldsymbol{\rho} \equiv vec(\mathbf{P})$. The matrices of regime-dependent parameters are all "vectorized" to make $\boldsymbol{\theta}$ into a simpler object.[38] Write the likelihood function of our sample of N asset returns as

$$L\left(\{\mathbf{R}_t\}_{t=1}^T | \{\boldsymbol{\xi}_t\}_{t=1}^T, \boldsymbol{\theta}\right) = \sum_{\{\boldsymbol{\xi}_t\}_{t=1}^T} \prod_{t=1}^T p(\mathbf{R}_t | \boldsymbol{\xi}_t, \Im_{t-1}; \boldsymbol{\theta}) \Pr\left(\boldsymbol{\xi}_t | \boldsymbol{\xi}_0; \boldsymbol{\rho}\right) \tag{14}$$

where $\Pr\left(\boldsymbol{\xi}_t | \boldsymbol{\xi}_0; \boldsymbol{\rho}\right) = \sum_{S_0=1}^K \xi_{S_0} \prod_{t=1}^T p_{S_{t-1},S_t}$ and the first summation spans the space defined by

$$\boldsymbol{\xi}_1 \otimes \boldsymbol{\xi}_2 \otimes ... \otimes \boldsymbol{\xi}_T$$

for a total of K^T possible combinations. In words, this means that in principle the log-likelihood function forces you to sum over all possible paths/evolutions of regime probabilities between $t = 1$ and $t = T$. As we know, when the shocks to (6) are assumed to be multivariate normal (as they are most of the time), then the density function is

$$p(\mathbf{R}_t | \boldsymbol{\xi}_t, \Im_{t-1}; \boldsymbol{\theta}) = \boldsymbol{\eta}_t \odot \boldsymbol{\xi}_t$$

where the kth element of $\boldsymbol{\eta}_t$ is defined as $(2\pi)^{-1/2} |\boldsymbol{\Omega}_k|^{-1/2} \exp\left[(\mathbf{R}_t - \mathbf{X}_t \mathbf{A} \mathbf{e}_k) \boldsymbol{\Omega}_k^{-1} (\mathbf{R}_t - \mathbf{X}_t \mathbf{A} \mathbf{e}_k)\right]$, i.e., the multivariate normal density. At this point, the parameters $[\boldsymbol{\theta}' \ \boldsymbol{\rho}']'$ can be derived by maximization of (14) subject to the natural constraints:

$$\mathbf{P} \boldsymbol{\iota}_K = \boldsymbol{\iota}_K \text{ (rows sum to 1)} \quad \boldsymbol{\xi}_0' \boldsymbol{\iota}_K = 1 \text{ (probabilities sum to one)} \tag{15}$$

$$\boldsymbol{\rho} \geq \mathbf{0}, \ \boldsymbol{\xi}_0 \geq \mathbf{0}, \text{ and } \boldsymbol{\Sigma}_K \mathbf{e}_k \text{ is (semi-)positive definite } \forall k = 1, 2, ..., K. \tag{16}$$

[37] Suppose one of you advisors tries and market some product/strategy that exploits a MS model and he/she relies on a backtesting exercise based on smoothed and not filtered probabilities. The product yields amazing alpha over the backtesting sample: would you buy it?

[38] In principle one ought to take $vec(\cdot)$ of parameters and also remove duplicate parametes that appear in all symmetric objects. However, because we are reasoning here in terms of lower triangular Choleski factors of regime-dependent covariance matrices, this caution seems largely superfluous. However, in the case of $\boldsymbol{\rho} \equiv vec(\mathbf{P})$, the summing-up constraints that apply to the matrix $\mathbf{P}$ usually reduce the size of $\boldsymbol{\rho}$ to be less than $K^2 \times 1$ (to how many?).

At this point, it is common to assume that the "non-negativity" constraints in (16) are satisfied and to take the first-order conditions (FOCs) of a Lagrangian function that explicitly enforces the adding-up constraints:

$$L^*\left(\{\mathbf{R}_t\}_{t=1}^T|\{\boldsymbol{\xi}_t\}_{t=1}^T,\boldsymbol{\theta},\boldsymbol{\rho}\right)=\sum_{\{\boldsymbol{\xi}_t\}_{t=1}^T}\prod_{t=1}^T p(\mathbf{R}_t|\boldsymbol{\xi}_t,\Im_{t-1};\boldsymbol{\theta})\Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_0;\boldsymbol{\rho}\right)-\boldsymbol{\lambda}_1'\left(\mathbf{P}\boldsymbol{\iota}_K-\boldsymbol{\iota}_K\right)-\lambda_2\left(\boldsymbol{\xi}_0'\boldsymbol{\iota}_K-1\right) \tag{17}$$

However, some additional work on the FOCs derived from (17) show a few interesting aspects of the ML estimator.

If you differentiate the logarithm of (17) with respect to $\boldsymbol{\theta}$, this gives the so-called *score function*,

$$\begin{aligned}\frac{\partial \ln L^*(\boldsymbol{\theta},\boldsymbol{\rho})}{\partial\boldsymbol{\theta}'} &= \frac{1}{L(\boldsymbol{\theta},\boldsymbol{\rho})}\sum_{\{\boldsymbol{\xi}_t\}_{t=1}^T}\frac{\partial\prod_{t=1}^T p(\mathbf{R}_t|\boldsymbol{\xi}_t,\Im_{t-1};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_0;\boldsymbol{\rho}\right)\\ &= \frac{1}{L(\boldsymbol{\theta},\boldsymbol{\rho})}\sum_{\{\boldsymbol{\xi}_t\}_{t=1}^T}\frac{\partial\ln\left[\prod_{t=1}^T p(\mathbf{R}_t|\boldsymbol{\xi}_t,\Im_{t-1};\boldsymbol{\theta})\right]}{\partial\boldsymbol{\theta}'}\prod_{t=1}^T p(\mathbf{R}_t|\boldsymbol{\xi}_t,\Im_{t-1};\boldsymbol{\theta})\Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_0;\boldsymbol{\rho}\right)\\ &= \sum_{\{\boldsymbol{\xi}_t\}_{t=1}^T}\sum_{t=1}^T\Pr\left(\boldsymbol{\xi}_t|\Im_T;\boldsymbol{\theta},\boldsymbol{\rho}\right)\frac{\partial\ln p(\mathbf{R}_t|\boldsymbol{\xi}_t,\Im_{t-1};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'},\end{aligned}$$

because from the definition of conditional probability

$$\frac{\prod_{t=1}^T p(\mathbf{R}_t|\boldsymbol{\xi}_t,\Im_{t-1};\boldsymbol{\theta})\Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_0;\boldsymbol{\rho}\right)}{\sum_{\{\boldsymbol{\xi}_t\}_{t=1}^T}\prod_{t=1}^T p(\mathbf{R}_t|\boldsymbol{\xi}_t,\Im_{t-1};\boldsymbol{\theta})\Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_0;\boldsymbol{\rho}\right)}=$$

$$=\frac{\prod_{t=1}^T p(\mathbf{R}_t|\boldsymbol{\xi}_t,\Im_{t-1};\boldsymbol{\theta})\Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_0;\boldsymbol{\rho}\right)}{L(\boldsymbol{\theta},\boldsymbol{\rho})}=\Pr\left(\boldsymbol{\xi}_t|\Im_T;\boldsymbol{\theta},\boldsymbol{\rho}\right).$$

Therefore

$$\sum_{t=1}^T\boldsymbol{\xi}_{t|T}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\rho}})\frac{\partial\ln\boldsymbol{\eta}_t(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}'}=\mathbf{0}' \tag{18}$$

provides the first set of FOCs with respect to (w.r.t.) $\boldsymbol{\theta}$. Notice that these conditions involve the smoothed probabilities of the state vector, $\{\hat{\boldsymbol{\xi}}_{t|T}\}_{t=1}^T$ and not the filtered probabilities as one may naively come to expect. The reason lies in the math shown above. At this point, (18) simply represents a smoothed probability-weighted standard ML vector FOC, $\partial\ln\boldsymbol{\eta}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'=\mathbf{0}'$. This means that in practice, the special structure of the MS model allows us to perform standard, multivariate normal PDF-based estimation, with the only caution that because each observation $\mathbf{R}_t$ carries a different vector of probabilities of coming from any of the K regimes, when the log-likelihood is computed, it must be weighted by the time series of the smoothed probabilities, $\{\hat{\boldsymbol{\xi}}_{t|T}\}_{t=1}^T$.

The FOCs w.r.t. the transition probabilities are determined as follows. Because

$$\begin{aligned}\frac{\partial\ln L(\boldsymbol{\theta},\boldsymbol{\rho})}{\partial\boldsymbol{\rho}'} &= \frac{1}{L(\boldsymbol{\theta},\boldsymbol{\rho})}\sum_{\{\boldsymbol{\xi}_t\}_{t=1}^T}\frac{\partial\Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_0;\boldsymbol{\rho}\right)}{\partial\boldsymbol{\rho}'}\prod_{t=1}^T p(\mathbf{R}_t|\boldsymbol{\xi}_t,\Im_{t-1};\boldsymbol{\theta})\\ &= \frac{1}{L(\boldsymbol{\theta},\boldsymbol{\rho})}\sum_{\{\boldsymbol{\xi}_t\}_{t=1}^T}\frac{\partial\ln\Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_0;\boldsymbol{\rho}\right)}{\partial\boldsymbol{\rho}'}\prod_{t=1}^T p(\mathbf{R}_t|\boldsymbol{\xi}_t,\Im_{t-1};\boldsymbol{\theta})\Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_0;\boldsymbol{\rho}\right)\end{aligned}$$

$$= \sum_{\{\boldsymbol{\xi}_t\}_{t=1}^T} \sum_{t=1}^{T} \frac{\partial \ln \Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_0;\boldsymbol{\rho}\right)}{\partial \boldsymbol{\rho}'} \Pr\left(\boldsymbol{\xi}_t|\Im_T;\boldsymbol{\theta},\boldsymbol{\rho}\right),$$

for each component p_{ij} of $\boldsymbol{\rho}$ this implies:

$$\begin{aligned}
\frac{\partial \ln L(\boldsymbol{\theta},\boldsymbol{\rho})}{\partial p_{ij}} &= \sum_{t=1}^{T} \sum_{\boldsymbol{\xi}_{t-1}=\mathbf{e}_i} \sum_{\boldsymbol{\xi}_t=\mathbf{e}_j} \frac{\partial \ln \Pr\left(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t-1};\boldsymbol{\rho}\right)}{\partial p_{ij}} \Pr\left(\boldsymbol{\xi}_t,\boldsymbol{\xi}_{t-1}|\Im_T;\boldsymbol{\theta},\boldsymbol{\rho}\right) \\
&= \sum_{t=1}^{T} \sum_{\boldsymbol{\xi}_{t-1}=\mathbf{e}_i} \sum_{\boldsymbol{\xi}_t=\mathbf{e}_j} \frac{1}{p_{ij}} I_{\{\boldsymbol{\xi}_{t-1}=\mathbf{e}_i,\boldsymbol{\xi}_t=\mathbf{e}_j\}} \Pr\left(\boldsymbol{\xi}_t,\boldsymbol{\xi}_{t-1}|\Im_T;\boldsymbol{\theta},\boldsymbol{\rho}\right) \\
&= \sum_{t=1}^{T} \sum_{\boldsymbol{\xi}_{t-1}=\mathbf{e}_i} \sum_{\boldsymbol{\xi}_t=\mathbf{e}_j} \frac{\Pr(\boldsymbol{\xi}_{t-1}=\mathbf{e}_i,\boldsymbol{\xi}_t=\mathbf{e}_j|\Im_T;\boldsymbol{\theta},\boldsymbol{\rho})}{p_{ij}},
\end{aligned}$$

which originates the vector expression

$$\frac{\partial \ln L(\boldsymbol{\theta},\boldsymbol{\rho})}{\partial \boldsymbol{\rho}'} = \left(\sum_{t=1}^{T} \left(\hat{\boldsymbol{\xi}}_{t|T}^{(2)}\right)'\right) \oslash \boldsymbol{\rho}'$$

where $\hat{\boldsymbol{\xi}}_{t|T}^{(2)}$ is a $K^2 \times 1$ vector of (smoothed) probabilities concerning the matrix of state perceptions $\boldsymbol{\xi}_{t-1|T}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\rho}}) \otimes \boldsymbol{\xi}_{t|T}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\rho}})$, capturing how these regime beliefs move between $t-1$ and t. Because the K adding-up restrictions in $\mathbf{P}\boldsymbol{\iota}_K = \boldsymbol{\iota}_K$ can equivalently be written as $(\boldsymbol{\iota}_K' \otimes \mathbf{I}_K)\boldsymbol{\rho} = \boldsymbol{\iota}_K$, it follows that the FOCs can be written as

$$\frac{\partial L^*(\boldsymbol{\theta},\boldsymbol{\rho})}{\partial \boldsymbol{\rho}'} = \left(\sum_{t=1}^{T} \left(\hat{\boldsymbol{\xi}}_{t|T}^{(2)}\right)'\right) \oslash \hat{\boldsymbol{\rho}}' - \hat{\boldsymbol{\lambda}}_1'(\boldsymbol{\iota}_K' \otimes \mathbf{I}_K) = \mathbf{0}'.$$

In other words,

$$\hat{\boldsymbol{\rho}} = \left(\sum_{t=1}^{T} \left(\hat{\boldsymbol{\xi}}_{t|T}^{(2)}\right)\right) \oslash (\boldsymbol{\iota}_K \otimes \hat{\boldsymbol{\lambda}}_1)$$

implying

$$(\boldsymbol{\iota}_K' \otimes \mathbf{I}_K) \left(\sum_{t=1}^{T} \left(\hat{\boldsymbol{\xi}}_{t|T}^{(2)}\right)\right) \oslash (\boldsymbol{\iota}_K \otimes \hat{\boldsymbol{\lambda}}_1) = \left(\sum_{t=1}^{T} \left(\hat{\boldsymbol{\xi}}_{t|T}\right)\right) \oslash \hat{\boldsymbol{\lambda}}_1 = \boldsymbol{\iota}_K$$

so that $\hat{\boldsymbol{\lambda}}_1 = \sum_{t=1}^{T} \hat{\boldsymbol{\xi}}_{t|T}$ obtains.[39] Finally, we have

$$\hat{\boldsymbol{\rho}} = \left(\sum_{t=1}^{T} \left(\hat{\boldsymbol{\xi}}_{t|T}^{(2)}\right)\right) \oslash \left(\boldsymbol{\iota}_K \otimes \left(\sum_{t=1}^{T} \hat{\boldsymbol{\xi}}_{t|T}\right)\right), \tag{19}$$

which is a highly nonlinear function of estimated smoothed probabilities.

Appendix C explains how you should go about derive the MLE for the initial state probability vector $\boldsymbol{\xi}_{1|0}$, which happens to be given by a boundary condition (i.e., the MLE exactly satisfies one of the constraints):

$$\hat{\boldsymbol{\xi}}_{1|0} = \underset{1 \le i \le K}{\arg\max} \boldsymbol{\iota}_K' \prod_{t=1}^{T} K_t(\boldsymbol{\theta}) \mathbf{e}_i \tag{20}$$

Note that the FOCs (18)-(20) all depend on smoothed probabilities $\hat{\boldsymbol{\xi}}_{t|T} \equiv \Pr\left(\boldsymbol{\xi}_t|\Im_T;\boldsymbol{\theta},\boldsymbol{\rho}\right)$ and therefore they all present a high degree of non-linearity in the parameters $[\boldsymbol{\theta}\ \rho]'$. Therefore the FOCs have to be solved numerically.

[39] $(\boldsymbol{\iota}_K' \otimes \mathbf{I}_K)\left(\sum_{t=1}^{T} \left(\hat{\boldsymbol{\xi}}_t^{(2)}\right)\right)$ produces a $K \times 1$ vector with i-th element $\sum_{t=1}^{T} \hat{\boldsymbol{\xi}}_t$. $(\boldsymbol{\iota}_K' \otimes \mathbf{I}_K)$ is the a communication (conversion) matrix that converts probability distributions over $\boldsymbol{\xi}_{t-1|T}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\rho}}) \otimes \boldsymbol{\xi}_{t|T}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\rho}})$ into a distribution over $\boldsymbol{\xi}_t$ only.

5.3. *The EM algorithm*

The expectation and maximization steps can be combined and used in an iterative fashion with the objective of solving numerically (18)-(20). Starting with arbitrary initial values $\tilde{\boldsymbol{\theta}}^0$, $\tilde{\boldsymbol{\rho}}^0$, and $\tilde{\boldsymbol{\xi}}^0_{1|0}$, the expectation step is applied first, thus obtaining a time series sequence of smoothed probability distributions $\{\hat{\boldsymbol{\xi}}^1_{t|T}\}_{t=1}^T$. Given these smoothed probabilities, (19) is then used to calculate $\tilde{\boldsymbol{\rho}}^1$, (18) to derive $\tilde{\boldsymbol{\theta}}^1$, and (20) to find the maximizing $\tilde{\boldsymbol{\xi}}^1_{1|0}$.[40] Based on $\tilde{\boldsymbol{\theta}}^1$, $\tilde{\boldsymbol{\rho}}^1$, and $\tilde{\boldsymbol{\xi}}^1_{1|0}$, the expectation step can be applied again to find a new sequence of smoothed probability distributions $\{\hat{\boldsymbol{\xi}}^2_{t|T}\}_{t=1}^T$.[41] This starts the second iteration of the algorithm. The algorithm keeps being iterated until convergence, i.e. until $[\tilde{\boldsymbol{\theta}}^l \ \tilde{\boldsymbol{\rho}}^l]' \simeq [\tilde{\boldsymbol{\theta}}^{l-1} \ \tilde{\boldsymbol{\rho}}^{l-1}]'$, or

$$[\tilde{\boldsymbol{\theta}}^l \tilde{\boldsymbol{\rho}}^l]' - [\tilde{\boldsymbol{\theta}}^{l-1} \tilde{\boldsymbol{\rho}}^{l-1}]' \simeq \mathbf{0},$$

which means that simple tests will have to be applied to check whether two subsequent iterations have essentially left the corresponding estimates unaltered, so that (say)

$$\sqrt{\sum_j (\tilde{\theta}^l_j - \tilde{\theta}^{l-1}_j)^2 + \sum_i (\tilde{\rho}^l_j - \tilde{\rho}^{l-1}_j)^2} < \bar{\varepsilon},$$

where $\bar{\varepsilon}$ is a number chosen to be small (e.g., typically, $1e^{-04}$). Note that these conditions normally exclude the estimate for $\boldsymbol{\xi}_{1|0}$. At that point one simply sets $\hat{\boldsymbol{\theta}}_T^{MLE} = \tilde{\boldsymbol{\theta}}^l$, $\hat{\boldsymbol{\rho}}^{MLE} = \tilde{\boldsymbol{\rho}}^l$, and $\hat{\boldsymbol{\xi}}^{MLE}_{1|0} = \tilde{\boldsymbol{\xi}}^1_{1|0}$. Importantly, by construction, when the EM iterations are applied, the likelihood function increases at each step and reaches an approximate maximum in correspondence to convergence (see Baum et al., 1970).

[40] Krolzig (1997, pp.103-107) shows that in MS models with autoregressive coefficients that are not switching, estimates of $vec(\mathbf{A})$ (here recall that the matrix $\mathbf{A}$ contains all intercepts and vector autoregressive coefficients that depend on the K regimes) can be easily obtained in closed-form by appropriately setting up the log-likelihood function as in a GLS problem. As for the estimates of $vec(\boldsymbol{\Sigma}_K)$, since the (expected) log-likelihood function can be expressed as

$$\begin{aligned}\ell(\boldsymbol{\theta}|\{\mathbf{R}_t\}_{t=1}^T) \quad \propto \quad & \frac{1}{2}\sum_{k=1}^{K^T}\sum_{t=1}^{T} \ln|\boldsymbol{\Omega}_k^{-1}|\hat{\boldsymbol{\xi}}_{kt|T} + \\ & -\frac{1}{2}\sum_{k=1}^{K^T}\sum_{t=1}^{T}(\mathbf{R}_t - \mathbf{X}_t\mathbf{A}_k)\boldsymbol{\Omega}_k^{-1}(\mathbf{R}_t - \mathbf{X}_t\mathbf{A}_k)\hat{\boldsymbol{\xi}}_{kt|T}\end{aligned}$$

this leads to

$$\frac{\partial \ell(\boldsymbol{\theta}|\{\mathbf{R}_t\}_{t=1}^T)}{\partial \boldsymbol{\Omega}_k^{-1}} = \frac{1}{2}\boldsymbol{\Omega}_k \sum_{t=1}^{T}\hat{\boldsymbol{\xi}}_{kt|T} +$$
$$-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{R}_t - \mathbf{X}_t\mathbf{A}_k)(\mathbf{R}_t - \mathbf{X}_t\mathbf{A}_k)' = \mathbf{O}$$

and

$$\hat{\boldsymbol{\Omega}}_k(\mathbf{A}_k) = \left(\sum_{t=1}^{T}\hat{\boldsymbol{\xi}}_{kt|T}\right)^{-1}\sum_{t=1}^{T}(\mathbf{R}_t - \mathbf{X}_t\mathbf{A}_k)(\mathbf{R}_t - \mathbf{X}_t\mathbf{A}_k)'.$$

[41] Please note this has nothing to do with $\hat{\boldsymbol{\xi}}^{(2)}_{t|T}$.

5.4. *Properties and inference concerning ML estimators*

As for the properties of the resulting ML estimators, for ergodic, univariate MS models with autoregressive components, it has been proven by Karlsen (1990) that R_t in

$$R_{t+1} = \mu_{S_{t+1}} + \sum_{j=1}^{p} a_{j,S_{t+1}} R_{t+1-j} + \sigma_{S_{t+1}} \epsilon_{t+1} \quad S_{t+1} = 1, 2, ..., K,$$

where $\epsilon_{t+1} \sim$ IID $N(0,1)$, follows a strong mixing process (i.e., some sufficient technical property of the corresponding stochastic process) whose stationarity is implied by the stationarity of the homogenous Markov chain $\boldsymbol{\xi}_t$, so that the functional central limit theorem may be used to derive the asymptotic distribution of $[\hat{\boldsymbol{\theta}}_T^{MLE} \hat{\boldsymbol{\rho}}^{MLE}]'$. Leroux (1992) has formally proved the consistency of MLE for MSIVAR(K,0) processes, what we have also called MSI and MSIH processes. More generally, under standard regularity conditions (such as identifiability, ergodicity and the fact that the true parameter vector does not fall on the boundaries established by (15) and (16)) we can at least speculate—however because in finance MSIH models are very popular, it is good to know that for them the result is mathematically exact—the consistency and asymptotic normality of the ML estimator $\hat{\boldsymbol{\gamma}} = [\hat{\boldsymbol{\theta}}_T^{MLE} \ \hat{\boldsymbol{\rho}}^{MLE}]'$:[42]

$$\sqrt{T}\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right) \xrightarrow{D} N\left(\mathbf{0}, \mathcal{I}_a(\boldsymbol{\gamma})^{-1}\right) \tag{21}$$

where $\mathcal{I}_a(\boldsymbol{\gamma})$ is the asymptotic information matrix,

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

Three alternative sample estimators of $\mathcal{I}_a(\boldsymbol{\gamma})$ providing estimates $\widetilde{Var}(\hat{\boldsymbol{\gamma}})$ are available and commonly employed:

1. An estimator based on the conditional scores:

$$\mathcal{I}_1(\hat{\boldsymbol{\gamma}}) = T^{-1} \sum_{t=1}^{T} \left[\mathbf{h}_t(\hat{\boldsymbol{\gamma}})\right] \left[\mathbf{h}_t(\hat{\boldsymbol{\gamma}})\right]' \qquad \mathbf{h}_t(\hat{\boldsymbol{\gamma}}) = \frac{\partial \ln p(\mathbf{R}_t|\Im_{t-1};\hat{\boldsymbol{\gamma}})}{\partial\boldsymbol{\gamma}}. \tag{22}$$

2. Alternatively, it is possible to numerically calculate the second partial derivative of the log-likelihood function w.r.t. to the estimated parameters, simply

$$\mathcal{I}_2(\hat{\boldsymbol{\gamma}}) = -T^{-1} \sum_{t=1}^{T} \left[\frac{\partial^2 \ln p(\mathbf{R}_t|\Im_{t-1};\hat{\boldsymbol{\gamma}})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}\right]. \tag{23}$$

3. Finally, it may happen that (22) and (23) widely differ in finite samples. Although this might simply reveal a poor numerical approximation of the second partial derivative of the log-likelihood function, it might also be a sign of model misspecification. In this case, the "sandwich" quasi-maximum likelihood estimator of the information matrix proposed by White (

[42]Notice though that the estimator for $\hat{\boldsymbol{\xi}}_{1|0}$ is inconsistent due to the binary nature of its components. Later we call Γ the space in which the vector of parameters γ is defined.

4. 1982) may be preferable:

$$\widetilde{Var}(\hat{\boldsymbol{\gamma}}) = T^{-1}\left[\mathcal{I}_2(\hat{\boldsymbol{\gamma}})\left(\mathcal{I}_1(\hat{\boldsymbol{\gamma}})\right)^{-1}\mathcal{I}_2(\hat{\boldsymbol{\gamma}})\right].$$

As a consequence of these results on consistency and asymptotic normality, and with one important exception, standard inferential procedures are available to test statistical hypotheses with relevant economic content. Starting with the usual aspects of testing procedures, assuming asymptotic normality for $\hat{\boldsymbol{\gamma}}$, as implied by (21), the three classical tests are available. Call $\phi : \mathcal{R}^q \to \mathcal{R}^r$ a function that imposes $q - r$ restrictions on the q-dimensional parameter vector $\boldsymbol{\theta}$. Note that $\boldsymbol{\theta}$ is a sub-vector of $\boldsymbol{\gamma}$. We want to test $H_0 : \phi(\boldsymbol{\gamma}) = \mathbf{0}$ vs. $H_1 : \phi(\boldsymbol{\gamma}) \neq \mathbf{0}$ *under the assumption that under both hypotheses the number of regimes K is identical.*[43] Such a null hypothesis–in fact this may be a vector of hypotheses, as signalled by the fact that $\phi(\boldsymbol{\gamma}) = \mathbf{0}$ is $\mathcal{R}^q \to \mathcal{R}^r$—may be tested using three alternative procedures that you have encountered in your undergraduate statistics. First, Lagrange Multiplier (LM) tests are undoubtedly the preferred tests as they only require the estimation of the restricted model. While the cumulative scores,

$$\mathbf{s}_T(\hat{\boldsymbol{\theta}}) \equiv \sum_{t=1}^{T} \mathbf{h}_t(\hat{\boldsymbol{\theta}}) = \sum_{t=1}^{T} \frac{\partial \ln p(\mathbf{R}_t|\Im_{t-1};\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}}$$

of an *unrestricted* model have zero mean vector by construction, as these correspond to the FOCs for the vector $\boldsymbol{\theta}$**,** the scores of the *restricted* model obtained by maximum likelihood and imposing $\phi(\boldsymbol{\theta}) = \mathbf{0}$ can be used to obtain the standard test statistic:

$$LM \equiv \mathbf{s}_T(\tilde{\boldsymbol{\theta}}_r)' \left[\widetilde{Var}(\tilde{\boldsymbol{\theta}}_r)\right]^{-1} \mathbf{s}_T(\tilde{\boldsymbol{\theta}}_r) \xrightarrow{d} \chi^2_r$$

where $r \equiv \text{rank}\big(\partial\phi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'\big)$ and $\tilde{\boldsymbol{\theta}}_r$ denotes the restricted estimator.[44] The idea is that if the restriction is rejected by the data, while $\mathbf{s}_T(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ by construction, $\mathbf{s}_T(\tilde{\boldsymbol{\theta}}_r)$ will be large. Therefore a suitable weighted sum of squares of such restricted, non-zero scores over your sample ought to be large. Here the weighting is performed using the estimated covariance matrix of the restricted estimates, $\widetilde{Var}(\tilde{\boldsymbol{\theta}}_r)$, which can be computed in one of the three ways listed above. If such a weighted sum of squares deviations of the restricted scores from zero is large, then given some pre-specified size of the test, $\mathbf{s}_T(\tilde{\boldsymbol{\theta}}_r)'[\widetilde{Var}(\tilde{\boldsymbol{\theta}}_r)]^{-1}\mathbf{s}_T(\tilde{\boldsymbol{\theta}}_r)$ will exceed the critical value under a χ^2_r and cause—as it should—a rejection of the null hypothesis.

As an alternative, the Likelihood Ratio (LR) test may be employed,

$$LR \equiv 2\left[\ln L(\hat{\boldsymbol{\theta}}) - \ln L(\tilde{\boldsymbol{\theta}}_r)\right] \xrightarrow{d} \chi^2_r,$$

[43] Hypotheses involving elements of $\boldsymbol{\rho}$ set equal to zero cannot be entertained as simply as the ones in the main text as they fall on the boundaries of the parameter space and imply a change in the number of the regimes. However other hypotheses involving $\boldsymbol{\rho}$ can be tested without special caution, for instance the important statistical hypothesis of independent regime switching (i.e., $\mathbf{P}$ has rank one), when $p_{ij} = p_j$ independently of the initial state i. In this case, all the columns of $\mathbf{P}$ contain identical numbers. As you may recall, a $K \times K$ square matrix in which all columns are identical, trivially has a rank of 1.

[44] For instance, a test of the hypothesis of homoskedasticity ($H_0 : vec(\boldsymbol{\Omega}_i) = vec(\boldsymbol{\Omega}_K)$ $i = 1, 2, ..., K$) implies $r = (K-1)N(N+1)/2$ restrictions (because of the symmetry of a covariance matrix) and can be formulated as a set of linear restrictions on the matrix $\boldsymbol{\Sigma}_K$.

where $\ln L(\hat{\boldsymbol{\theta}})$ is the maximized log-likelihood under the unrestricted model and $\ln L(\tilde{\boldsymbol{\theta}}_r)$ is the maximized log-likelihood under the restricted one. Although very simple to compute and understand, this test requires the estimation of both the restricted and the unrestricted models, which for N large enough, can be quite cumbersome and require a host of diagnostic checks on the performance of the EM algorithm in locating a truly global maximum of the likelihood function. However, it remains the case that a LR test is logically very simple: under the null hypothesis $H_0 : \phi(\boldsymbol{\gamma}) = \mathbf{0}$ imposes a restriction involving $\boldsymbol{\theta}$. If this restriction is rejected by the data (they are false), then maximizing the log-likelihood subject to a false constraint will prevent us from reaching the true maximum of the log-likelihood. It is like running carrying a heavy weight—you will end up being much slower than you otherwise would. Therefore $\ln L(\tilde{\boldsymbol{\theta}}_r)$ will be considerable inferior to $\ln L(\hat{\boldsymbol{\theta}})$ and $2[\ln L(\hat{\boldsymbol{\theta}}) - \ln L(\tilde{\boldsymbol{\theta}}_r)]$ will be large. If this is the case, given some pre-specified size of the test (what we sometimes call the "significance of the test" causing the dead statisticians to roll in their graves), $2[\ln L(\hat{\boldsymbol{\theta}}) - \ln L(\tilde{\boldsymbol{\theta}}_r)]$ will exceed the critical value under a χ^2_r and cause—as it should—a rejection of the null hypothesis.

Finally standard t and F statistics can be calculated using a Wald test (really, to call them t or F is inappropriate and equivalent to name an object based on some of its properties, like a car "the polluter" or an econometrics professor "the confuser"). Under asymptotic normality of the unrestricted ML estimator $\hat{\boldsymbol{\theta}}$, and assuming the function $\phi(\boldsymbol{\theta})$ is smooth and one-to-one, one can prove that[45]

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

$$Wald \equiv T(\phi(\hat{\boldsymbol{\theta}}))' \left[\left.\frac{\partial \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \widetilde{Var}(\hat{\boldsymbol{\theta}}) \left.\frac{\partial \phi'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]^{-1} \phi(\hat{\boldsymbol{\theta}}) \xrightarrow{D} \chi^2_r.$$

Interestingly, also a Wald test has an asymptotic chi-square distribution with a number of degrees of freedom equal to the number r of restrictions that you want to test. However, this is not surprising, as you know that asymptotically, as the number of degrees of freedom goes to ∞, a t statistic converges in distribution to a normal; moreover, the definition of $Wald$ given above employs a quadratic form that is a weighted square of normals and it is well known that a weighted sum of r squared normals has a χ^2_r distribution. The idea is that if the restrictions captured by $\phi(\boldsymbol{\theta}) = \mathbf{0}$ are satisfied by $\hat{\boldsymbol{\theta}}$, then in correspondence to $\phi(\hat{\boldsymbol{\theta}}) \simeq \mathbf{0}$ and as such the quadratic form

$$(\phi(\hat{\boldsymbol{\theta}}))' \left[\left.\frac{\partial \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \widetilde{Var}(\hat{\boldsymbol{\theta}}) \left.\frac{\partial \phi'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]^{-1} \phi(\hat{\boldsymbol{\theta}}) \simeq \mathbf{0}$$

so that the null will not be rejected for most/all choices of size of the test under a χ^2_r. If, one the contrary, the quadratic form defined above (and built around the inverse of the covariance matrix of $\phi(\hat{\boldsymbol{\theta}})$ gives large values, then given some pre-specified size of the test, it will exceed the critical value under a χ^2_r and cause—as it should—a rejection of the null hypothesis.

[45] This follows from the fact that if $\hat{\boldsymbol{\theta}}$ is an ML estimator, then under suitable technical conditions, also $\phi(\hat{\boldsymbol{\theta}})$ is an ML estimator of $\phi(\boldsymbol{\theta})$, and as such consistent and asymptotically normal.

For instance, the hypothesis that in (6) the matrices of autoregressive coefficients are regime independent can be written as:

$$\begin{bmatrix} \mathbf{O}_N & \mathbf{O}_N & \cdots & \mathbf{I}_N & -\mathbf{I}_N & \mathbf{O}_N & \cdots & \mathbf{O}_N \\ \mathbf{O}_N & \mathbf{O}_N & \cdots & \mathbf{O}_N & \mathbf{I}_N & -\mathbf{I}_N & \cdots & \mathbf{O}_N \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{O}_N & \mathbf{O}_N & \cdots & \mathbf{O}_N & \mathbf{O}_n & \mathbf{O}_N & \cdots & \mathbf{O}_N \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \mathbf{A}'_{11}\mathbf{e}_1 \\ \vdots \\ \mathbf{A}'_{1K}\mathbf{e}_N \\ \vdots \\ \mathbf{A}'_{pK}\mathbf{e}_N \end{bmatrix} = \mathbf{R}vec(\mathbf{A}) = \mathbf{0}$$

and implies the (F) test statistic:

$$T\hat{\boldsymbol{\theta}}'\mathbf{R}' \left[\mathbf{R}\widehat{Var}(\hat{\boldsymbol{\theta}})\mathbf{R}'\right]^{-1} \mathbf{R}\hat{\boldsymbol{\theta}}.$$

This is a simple F statistic because in this case $\phi(\boldsymbol{\theta})$ defines a linear function, as shown by the use of linear algebra to express the constraints.

What is the difference between LM, LR, and Wald tests? Which one should you be using, given a set of null hypotheses $\phi(\boldsymbol{\theta}) = \mathbf{0}$ that you would like to test? First, note that all the inferential results concerning the distribution of the test statistics listed above simply hold asymptotically, when $T \to \infty$: that was the meaning of the claim that LM, LR, $Wald \xrightarrow{D} \chi^2_r$. Second, but then this automatically answers your question, i.e., because we know how the test statistics behave only as the sample size grows without bounds, and therefore the idea is to assume that this the case, then using any of the three specific test procedures becomes a matter of indifference. However, as already mentioned above, there was a time in which—because a LM test imply a need to estimate (6) only under the restrictions implied by $\phi(\boldsymbol{\theta}) = \mathbf{0}$ (which often means to estimate less parameters than one would find in $\boldsymbol{\theta}$)—CPU-deprived researchers developed a strong preference for LM tests.[46] Yet, it turns out that in general LM tests have rather poor small sample properties, which means that among the three tests, these are the ones converging in distribution to χ^2_r more slowly than the other test statistics do. Finally, when it comes to a choice between LM and Wald tests, it must be added that we still lack of sufficient knowledge for which of these two tests may perform best in small samples for MS models. However, because of their clear intuitive meaning and their direct reliance on the maximized log-likelihood function, many quant researchers tend to have a preference for LR tests, even though these imply estimating two different MS models, one unrestricted and the other one restricted.

The only exception to these methods to test hypotheses concerns the *number of non-zero rows* of the transition matrix $\mathbf{P}$, i.e. the number of regimes K**.** In this case, even under the assumption of asymptotic normality of the ML estimator $\hat{\boldsymbol{\gamma}}$, standard testing procedures suffer from non-standard asymptotic distributions of the test statistics due to the existence of *nuisance parameters* under the

[46] You may object that also under Wald tests, you shall need to estimate only the unrestricted model. This is correct, but the complication here arises from the need to estimate the quantity $\partial\phi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$.

null hypothesis. We defer the discussion of this important and challenging inferential procedures until the end of Section 7.

6. **Forecasting with MS Models**

Under a mean squared prediction error (MSPE) criterion, the required algorithms are relatively simple in spite of the nonlinearity of this class of processes. A MSPE criterion has a simple meaning: you care for minimizing the *square* of forecast errors,

$$\boldsymbol{\eta}_{t+H} \equiv \mathbf{R}_{t+H} - (\mathbf{R}^f_{t+H}|\Im_t),$$

where $\mathbf{R}^f_{t+H}|\Im_t$ is a $N \times 1$ vector of forecasts that simply condition on the information available at time t. Such a criterion is so deeply ingrained in our way of thinking, that stating it may even seem superfluous (it is not, but that is a different story): for instance, minimizing $\sum_{t=1}^{T} \boldsymbol{\eta}_t' \boldsymbol{\eta}_t$ is in some way the standard objective of ordinary least squares, OLS.

Under such a MSPE criterion and appealing once more to the state-space representation in Appendix B, yields rather intuitive results. Ignoring for the time being the issue of parameter uncertainty, i.e. the fact that the parameters of the MS process are unknown and must therefore be estimated (see Section 5), the function minimizing the MSPE is the standard conditional expectation function: $\mathbf{R}^f_{t+H}|\Im_t = E[\mathbf{R}_{t+H}|\Im_t]$.[47] For instance, in the case of one-step ahead forecasts, we have:

$$E[\mathbf{R}_{t+1}|\Im_t] = \mathbf{X}_{t+1}\hat{\mathbf{A}}\left(\hat{\boldsymbol{\xi}}_{t+1|t} \otimes \boldsymbol{\iota}_N\right)$$

where $\mathbf{X}_{t+1} \equiv [1\ \mathbf{R}_t' ... \mathbf{R}_{t-p+1}'] \otimes \boldsymbol{\iota}_N$, $\hat{\mathbf{A}}$ collects the estimated conditional mean parameters of the system, and $\hat{\boldsymbol{\xi}}_{t+1|t}$ is the one-step ahead, predicted latent state vector to be filtered out of the available information set $\Im_t$ according to the known transition equation

$$\hat{\boldsymbol{\xi}}_{t+1|t} = \hat{\mathbf{P}}'\hat{\boldsymbol{\xi}}_{t|t},$$

where also the transition matrix $\mathbf{P}$ will have to be estimated. Here, although this has been already discussed, it may be useful a reminder of what the expression $\mathbf{X}_{t+1}\hat{\mathbf{A}}$ really means. Because $\mathbf{X}_t$ is a $N \times (Np+1)$ matrix of predetermined variables with structure $[1\ \mathbf{R}_{t-1}' ... \mathbf{R}_{t-p}'] \otimes \boldsymbol{\iota}_N$ and $\mathbf{A}$ is a $(Np+1) \times NK$ matrix collecting the VAR parameters, both means or intercepts and autoregressive coefficients, in all regimes

$$\hat{\mathbf{A}} = \begin{bmatrix} \hat{\boldsymbol{\mu}}_1' & \hat{\boldsymbol{\mu}}_2' & \cdots & \hat{\boldsymbol{\mu}}_K' \\ \hat{\mathbf{A}}_{11} & \hat{\mathbf{A}}_{12} & \cdots & \hat{\mathbf{A}}_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{A}}_{p1} & \hat{\mathbf{A}}_{p2} & \cdots & \hat{\mathbf{A}}_{pK} \end{bmatrix},$$

[47] Because we have obtained predictions from conditional expectations as a matter or routine in this course, this means that the MSPE criterion has been maintained all along.

then

$$\mathbf{X}_{t+1}\hat{\mathbf{A}} = \underset{N\times(Np+1)}{([1\ \mathbf{R}_t'...\mathbf{R}_{t-p+1}']} \otimes \boldsymbol{\iota}_N) \underset{(Np+1)\times NK}{\begin{bmatrix} \hat{\boldsymbol{\mu}}_1' & \hat{\boldsymbol{\mu}}_2' & \cdots & \hat{\boldsymbol{\mu}}_K' \\ \hat{\mathbf{A}}_{11} & \hat{\mathbf{A}}_{12} & \cdots & \hat{\mathbf{A}}_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{A}}_{p1} & \hat{\mathbf{A}}_{p2} & \cdots & \hat{\mathbf{A}}_{pK} \end{bmatrix}} =$$

$$= \begin{bmatrix} \hat{\boldsymbol{\mu}}_1' + \sum_{j=1}^p \mathbf{R}_{t+1-j}\hat{\mathbf{A}}_{1j} & \hat{\boldsymbol{\mu}}_2 + \sum_{j=1}^p \mathbf{R}_{t+1-j}'\hat{\mathbf{A}}_{2j} & ... & \hat{\boldsymbol{\mu}}_k + \sum_{j=1}^p \mathbf{R}_{t+1-j}'\hat{\mathbf{A}}_{kj} \\ \hat{\boldsymbol{\mu}}_1' + \sum_{j=1}^p \mathbf{R}_{t+1-j}\hat{\mathbf{A}}_{1j} & \hat{\boldsymbol{\mu}}_2 + \sum_{j=1}^p \mathbf{R}_{t+1-j}'\hat{\mathbf{A}}_{2j} & ... & \hat{\boldsymbol{\mu}}_k + \sum_{j=1}^p \mathbf{R}_{t+1-j}'\hat{\mathbf{A}}_{kj} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\boldsymbol{\mu}}_1' + \sum_{j=1}^p \mathbf{R}_{t+1-j}\hat{\mathbf{A}}_{1j} & \hat{\boldsymbol{\mu}}_2 + \sum_{j=1}^p \mathbf{R}_{t+1-j}'\hat{\mathbf{A}}_{2j} & ... & \hat{\boldsymbol{\mu}}_k + \sum_{j=1}^p \mathbf{R}_{t+1-j}'\hat{\mathbf{A}}_{kj} \end{bmatrix}.$$

It follows that

$$E[\mathbf{R}_{t+1}|\Im_t] = \mathbf{X}_{t+1}\hat{\mathbf{A}}\left(\hat{\mathbf{P}}'\hat{\boldsymbol{\xi}}_{t|t} \otimes \boldsymbol{\iota}_N\right). \tag{24}$$

For instance in the univariate case of $N = 1$ and $p = 1$, i.e., of a MS(I)ARH$(K, 1)$ model, we have:

$$E[R_{t+1}|\Im_t] = \sum_{k=1}^{K}(\mu_k + a_k R_t)\hat{\xi}_{t+1|t}^k,$$

where $\hat{\xi}_{t+1|t}^k = \hat{\boldsymbol{\xi}}_{t+1|t}'\mathbf{e}_k$, i.e., the kth element of the vector of predicted probabilities $\hat{\boldsymbol{\xi}}_{t+1|t}'$. Clearly, this expression simply means that one forecasts returns conditioning on each of the K regimes, and then each of this state-specific predictions is weighted by the appropriate predicted probabilities. As usual, when it comes to forecast conditional means, the regime-specific second moments are irrelevant because under a standard MS set up, the errors are IID $N(\mathbf{0}, \boldsymbol{\Omega}_{S_{t+1}})$ and hence have zero mean (vector).

However, for $h > 1$-step ahead forecasts the task is much more challenging as: (1) $\mathbf{X}_{t+H}$ is unknown and must be predicted itself; (2) $E[\mathbf{X}_{t+H}|\Im_t]$ involves sequences of predictions $\{E[\mathbf{R}_{t+1}|\Im_t]$,..., $E[\mathbf{R}_{t+H-1}\ |\Im_{t+T-2}]\}$ and as such $\{\hat{\boldsymbol{\xi}}_{t+1|t}, ..., \hat{\boldsymbol{\xi}}_{t+H-1|t}\}$ which are likely to impress patterns of cross-correlation to the unconditional values of the parameters to be used, because of the presence of regime switching. For instance, for $H = 2$, $p = 1$, and ignoring the presence of an intercept term, we have

$$\begin{aligned} E[\mathbf{R}_{t+2}|\Im_t] &= E\left[\left(\mathbf{R}_{t+1}' \otimes \boldsymbol{\iota}_N\right)\hat{\mathbf{A}}\left(\boldsymbol{\xi}_{t+2} \otimes \boldsymbol{\iota}_N\right)|\Im_t\right] \\ &= E\left[\left(\left(\left(\mathbf{R}_t' \otimes \boldsymbol{\iota}_N\right)\hat{\mathbf{A}}\left(\boldsymbol{\xi}_{t+1} \otimes \boldsymbol{\iota}_N\right) + \boldsymbol{\Sigma}_K\left(\boldsymbol{\xi}_{t+1} \otimes \mathbf{I}_N\right)\boldsymbol{\varepsilon}_t\right) \otimes \boldsymbol{\iota}_N'\right)\hat{\mathbf{A}}\left(\boldsymbol{\xi}_{t+2} \otimes \boldsymbol{\iota}_N\right)|\Im_t\right] \\ &= E\left[\left(\left(\mathbf{R}_t' \otimes \boldsymbol{\iota}_N\right)\hat{\mathbf{A}}\left(\boldsymbol{\xi}_{t+1} \otimes \boldsymbol{\iota}_N\right) \otimes \boldsymbol{\iota}_n' + \boldsymbol{\Sigma}_K\left(\boldsymbol{\xi}_{t+1} \otimes \mathbf{I}_N\right)\boldsymbol{\varepsilon}_t \otimes \boldsymbol{\iota}_N'\right)\hat{\mathbf{A}}\left(\boldsymbol{\xi}_{t+2} \otimes \boldsymbol{\iota}_N\right)|\Im_t\right] \\ &= E\left[\left(\left(\mathbf{R}_t' \otimes \boldsymbol{\iota}_N\right)\hat{\mathbf{A}}\left(\boldsymbol{\xi}_{t+1} \otimes \boldsymbol{\iota}_N\right) \otimes \boldsymbol{\iota}_N'\right)\hat{\mathbf{A}}\left(\boldsymbol{\xi}_{t+2} \otimes \boldsymbol{\iota}_N\right)|\Im_t\right] \end{aligned}$$

which is not the product of the conditional expectations $[(\mathbf{R}_t' \otimes \boldsymbol{\iota}_N)\hat{\mathbf{A}}(\hat{\boldsymbol{\xi}}_{t+1|t} \otimes \boldsymbol{\iota}_N) \otimes \boldsymbol{\iota}_N']\hat{\mathbf{A}}(\hat{\boldsymbol{\xi}}_{t+2|t} \otimes \boldsymbol{\iota}_N)$ as the future state vectors $\boldsymbol{\xi}_{t+1}$ and $\boldsymbol{\xi}_{t+2}$ are correlated, $\boldsymbol{\xi}_{t+2} = \mathbf{P}'\boldsymbol{\xi}_{t+1} + \mathbf{v}_{t+2}$. However, in applied work it is customary to follow the suggestion of Doan et al. (1984) consisting in the substitution of the sequence of predicted values $\{E[\mathbf{R}_{t+1}|\Im_t], ..., E[\mathbf{R}_{t+H-1}|\Im_t]\}$ in place of $\{E[\mathbf{R}_{t+1}|\Im_t], ..., E[\mathbf{R}_{t+H-1}|\Im_{t+T-2}]\}$. In this case (24) generalizes to generic $H > 2$-step ahead predictions:

$$\breve{E}[\mathbf{R}_{t+H}|\Im_t] = E[\mathbf{X}_{t+H}|\Im_t]\hat{\mathbf{A}}\left[(\hat{\mathbf{P}}')^H\hat{\boldsymbol{\xi}}_{t|t} \otimes \boldsymbol{\iota}_N\right],$$

which in practice gives a recursive formula since $E[\mathbf{X}_{t+H}|\Im_t]$ forces one to forecast a sequence of future $\mathbf{R}_{t+i}$ values, $i = 1, ..., H-1$. For instance, in the univariate case of $N = 1$ and $p = 1$, i.e., of a MS(I)ARH$(K, 1)$ model, we have that

$$E[R_{t+2}|\Im_t] = \sum_{k=1}^{K}(\mu_k + a_k E[R_{t+1}|\Im_t, \hat{\xi}^k_{t+1|t}])\hat{\xi}^k_{t+2|t+1}$$

and this not the same as

$$\begin{aligned}
\breve{E}[R_{t+2}|\Im_t] &= \sum_{k=1}^{K}(\mu_k + a_k E[R_{t+1}|\Im_t])\hat{\xi}^k_{t+2|t} \\
&= \sum_{k=1}^{K}\mu_k\hat{\xi}^k_{t+2|t} + \sum_{k=1}^{K}a_k E[R_{t+1}|\Im_t]\hat{\xi}^k_{t+2|t} \\
&= \sum_{k=1}^{K}\mu_k\hat{\xi}^k_{t+2|t} + \sum_{k=1}^{K}a_k\left(\sum_{k=1}^{K}(\mu_k + a_k R_t)\hat{\xi}^k_{t+1|t}\right)\hat{\xi}^k_{t+2|t} \\
&= \sum_{k=1}^{K}\mu_k\hat{\xi}^k_{t+2|t} + \sum_{k=1}^{K}a_k\left(\sum_{j=1}^{K}\mu_j\hat{\xi}^j_{t+1|t}\right)\hat{\xi}^k_{t+2|t} + \sum_{k=1}^{K}a_k\left(\sum_{j=1}^{K}a_j R_t\hat{\xi}^j_{t+1|t}\right)\hat{\xi}^k_{t+2|t}.
\end{aligned}$$

However, $\breve{E}[R_{t+2}|\Im_t]$ is what is reported in most applied work.

7. **Model Selection and Diagnostic Checks**

Compared to the standard econometric methods you are familiar with, MS models pose one obvious, additional problem: selecting the appropriate number of regimes, $K \geq 1$. When $K = 1$, a MS model boils down to a standard, homoskedastic VAR(p):

$$\begin{aligned}
\mathbf{R}_{t+1} &= \boldsymbol{\mu}_1 + \sum_{j=1}^{p}\mathbf{A}_{j,1}\mathbf{R}_{t+1-j} + \boldsymbol{\Omega}_1^{1/2}\boldsymbol{\epsilon}_{t+1} \quad \boldsymbol{\epsilon}_{t+1} \sim IIDN(\mathbf{0}, \mathbf{I}_N) \text{ or} \\
&= \boldsymbol{\mu} + \sum_{j=1}^{p}\mathbf{A}_j\mathbf{R}_{t+1-j} + \boldsymbol{\Omega}^{1/2}\boldsymbol{\epsilon}_{t+1},
\end{aligned}$$

where the index always equal to 1 can be dropped. Therefore, a first important divide occurs at the choice of whether $K = 1$ or $K \geq 2$. Once $K \geq 2$ has been established, then one may even worry about whether more than two regimes may be needed. The problem is then: how do we test for the appropriate number of regimes, or in any event proceed to select them? Of course, selecting the number of regimes should not be perceived as a problem, something else to worry about, but instead as an enormous opportunity to make the model as flexible as the data ask for.

The problem with the choice of the number of states is that under any number of regimes smaller than the starting value K^*, there are a few structural parameters of the unrestricted model—the elements of the transition probability matrix associated to the rows that correspond to "disappearing states"—that can take any values without influencing the resulting likelihood function. We say that these parameters become a nuisance to the estimation. The result is that the presence of these nuisance

parameters gives the likelihood surface so many degrees of freedom that computationally one can never reject the null that the nonnegative (better, positive) values of those parameters were purely due to sampling variation.[48] For instance, suppose you start with a MSIH(3) model,

$$\mathbf{R}_{t+1} = \boldsymbol{\mu}_{S_{t+1}} + \boldsymbol{\Omega}_{S_{t+1}}^{1/2} \boldsymbol{\epsilon}_{t+1} \quad \boldsymbol{\epsilon}_{t+1} \sim \text{IID } N(\mathbf{0}, \mathbf{I}_N).$$

$S_{t+1} = 1, 2, 3$ and you want to test whether $K = 2 < K^* = 3$ may be optimal. Suppose you are to compare the maximized log-likelihood obtained from the three-state model, $\ln L(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\rho}}; K = 3)$, to the log-likelihood of the restricted model in which $K = 2$ so that $\boldsymbol{\mu}_3$ and $\boldsymbol{\Omega}_3^{1/2}$ and especially the (1,3), (2,3), (3,1), (3,2), and (3,3) elements of the transition matrix can be set to any value without affecting $\ln L(\hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\rho}}_r; K = 2)$. Unfortunately, this makes the standard LR test invalid, in the sense that even in larger and larger samples, the distribution of the LR statistic fails to converge to a known χ_r^2. There are however a number of ideas in the literature on how to deal with this nuisance parameters issues:

1. Hansen (1992) proposes to see the likelihood as a function of the unknown and non-estimable nuisance parameters so that the asymptotic distribution is generated in each case numerically (i.e., by simulation) from a grid of transition and regime-dependent nuisance parameters. The test statistic becomes then

   $$LW_T \leq \sup_{\boldsymbol{\rho}' \in \mathcal{P}} LW_T(\boldsymbol{\rho})$$

   where the right-hand side converges in distribution to a function of a Brownian bridge (which is a Brownian motion in which the coefficients are themselves functions of other Brownian motions). In most of the cases, a closed form expression cannot be found and the bound must be calculated by simulation and becomes data-dependent. Hansen's way to empirically compute the p-values of LR tests is logically straightforward but computationally intensive and related to a set of statistical techniques generally called *bootstrap*.

2. Also Davies (1977) bounds the LR test but avoids the problem of estimating the nuisance parameters and of resorting to simulations, deriving instead an upper bound for the significance level of the LR test under nuisance parameters:[49]

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

3. Davidson and MacKinnon's (1981) J test for non-nested models can be also applied, since MS models with K and $K-1$ regimes are logically nested but cannot be treated as such on a

[48] Mathematically, the presence of unidentified nuisance parameters implies that the scores become identically zero and that the covariance matrix is singular.

[49] The bound holds if the likelihood function has a single peak (i.e., only one stationary point). A related test is proposed by Wolfe (1971) and applied in finance by Turner et al. (1989). The modified LR test is:

$$LR^{Wolf} = -\frac{2}{T}(T-3)\left[\ln L(\hat{\gamma}) - \ln L(\tilde{\gamma}_r)\right] \xrightarrow{d} \chi_r^2$$

where $\tilde{\gamma}_r$ is obtained under the null of simple multivariate normality and $r = K(K-1)$ since in the absence of regime switching there are $K(K-1)$ which cannot be estimated.

mathematical basis.[50] The test is implemented by estimating the model with K and $K-1$ states and calculating their full information "fitted" values, $\tilde{\mathbf{R}}_t^{(j)} = \mathbf{X}_t \tilde{\mathbf{A}}^{(j)} \tilde{\boldsymbol{\xi}}_{t|T}^{(j)}$. Then one can estimate the regression

$$\mathbf{R}_t = (\mathbf{I}_N - \boldsymbol{\Upsilon})\mathbf{X}_t \hat{\mathbf{A}} \hat{\boldsymbol{\xi}}_t^{(K-1)} + \boldsymbol{\Upsilon} \tilde{\mathbf{R}}_t^{(K)} + \boldsymbol{\varepsilon}_t.$$

The p-value of an F-test for the matrix of coefficients $\boldsymbol{\Upsilon}$ gives the p-value for the null of $K-1$ regimes against the alternative of K regimes. The intuition is that if tests cannot reject the null that the matrix $\boldsymbol{\Upsilon} \simeq \mathbf{O}$, then this means that once the fitted return values produced by a $K-1$-state model have been computed, no significant explanatory power may be further derived from the fitted values of a larger, K-state MS model. On the opposite if the null hypotheses that $\boldsymbol{\Upsilon} = \mathbf{O}$ can be rejected at a given size of the (Wald) test, then it means that there is evidence of the fact that one also needs the Kth regime in order to explain returns, so that K regimes may be preferred to $K-1$.

A practical alternative to these tests to diagnose the number of regimes appropriate in a MS model consists of the use of *information criteria*. As already discussed in chapter 4, these are *penalized measures of fit* which trade-off in-sample fit with parsimony, i.e., whose value increases as the fit to the data improves but also decreases as the number of estimated parameters increase.[51] As you will recall, in a ML set up, in the same way in which the $\bar{R}^2$ is based on the application of penalties to the classical coefficient of determination (R^2), information criteria are based on the concept of applying additional penalty terms to the maximized log-likelihood. Their general structure is:

$$-(\text{Maximized Log-Lik}) + l(dim(\hat{\boldsymbol{\gamma}})),$$

where $l(\cdot)$ is a penalty function, and $dim(\hat{\boldsymbol{\gamma}})$ is the notation for a counter of the number of different parameters in to be estimated in $\boldsymbol{\gamma} \in \Gamma$. The negative sign attached to the maximized log-likelihood is due to the fact that, as we have seen, most numerical optimization software actually minimize the negative of the log-likelihood function. Because the maximized log-likelihood is multiplied by -1 while the penalty has been added, it is clear that empirically we shall select models that actually *minimize* information criteria, not maximize them. Three information criteria are widely employed:

- The Bayes-Schwartz information criterion (BIC): $-2\mathcal{L}(\hat{\boldsymbol{\theta}}) + (dim(\hat{\boldsymbol{\theta}}) ln(T)/T)$; this criterion is known to select rather parsimonious models and it appears to be very popular in the applied literature.

[50] Two models are nested if one can go from model A to model B just by "turning off" (i.e., setting to zero) a few of the parameters. For instance a ARCH(1) model is nested in a GARCH(1,1) because the former obtains from the latter just by setting $\beta = 0$. However, mathematically, a MS model with K^* states cannot be obtained from one with K states simply by setting a sub-set of the parameters to zero; on the opposite, the latter become unidentified. As such, the former MS model is not nested with the latter.

[51] Since your early age you have been familiar with one such measure, the adjusted R^2 (often denoted as $\bar{R}^2$) which, indeed, penalizes the standard R^2 with a measure of the parameter vector dimension to prevent that big models have an unfair advantage over smaller, tightly parameterized ones. Why do we value parsimony? Because in general terms the forecasting performance of a model improves as the number of parameters used to fit the data in sample declines—i.e., smaller models tend to perform better than bigger ones do.

- The Akaike information criterion (AIC): $-2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2(dim(\hat{\boldsymbol{\theta}})/T)$; this criterion is also popular because it has optimal asymptotic properties (it is consistent), although it is also known to select too large non-linear models in small samples.

- The Hannan-Quinn information criterion (H-Q):$-2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2[dim(\hat{\boldsymbol{\theta}}) \log(\log(T))/T]$; this criterion has been shown to perform very strongly in small samples and for non-linear models; numerically, it can be shown that it represents a compromise between BIC and AIC.

Interestingly, few papers have addressed the issue of the small-sample and asymptotic performance of these information criteria specifically for the case of MS models. Because these measures rely on the same conditions employed in the asymptotic theory of the LR test, their small and large sample properties are, likewise, largely unknown.

Once a restricted set of (or more simply, one) MS models has been estimated, either the need of further improvements could arise as the result of a few diagnostic checks or the best model will be chosen based on the success of such checks. Although the EM algorithm naturally delivers estimates of the parameters $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\xi}}^1_{1|0}$, besides the smoothed sequence of probability distributions $\{\hat{\boldsymbol{\xi}}_{t|T}\}_{t=1}^T$ and would therefore lead to define the (smoothed) residuals as

$$\hat{\boldsymbol{\epsilon}}_t = \mathbf{R}_t - \mathbf{X}_t \hat{\mathbf{A}} \hat{\boldsymbol{\xi}}_{t|T},$$

these are not well suited to the use in diagnostic checks as they are full-sample random statistics and hence they structurally overestimate the explanatory power of MS. On the contrary the one-step predictions errors

$$\hat{\boldsymbol{\eta}}_{t|t-1} = \mathbf{R}_t - \mathbf{X}_t \hat{\mathbf{A}} \widehat{\mathbf{P}}' \hat{\boldsymbol{\xi}}_{t-1|t-1}$$

are limited information statistics (being based on filtered probabilities) and uncorrelated with the information set $\Im_{t-1}$ since $E[\mathbf{R}_t|\Im_{t-1}] = \mathbf{X}_t \hat{\mathbf{A}} \widehat{\mathbf{P}}' \hat{\boldsymbol{\xi}}_{t-1|t-1}$ and therefore form a martingale difference sequence $E[\hat{\boldsymbol{\eta}}_{t|t-1}|\Im_{t-1}] = \mathbf{0}$. Therefore standard tests of this hypothesis (such as Portmanteau tests of no serial correlation) could be used in order to detect any deviation from the martingale structure.[52] Here, recall that $E[\hat{\boldsymbol{\eta}}_{t|t-1}|\Im_{t-1}] = \mathbf{0}$ really means that none of the information contained in $\Im_{t-1}$ can help forecast subsequent prediction errors, so that $E[\hat{\boldsymbol{\eta}}_{t|t-1}|\Im_{t-1}] = \mathbf{0}$ implies the possibility of testing restrictions such as $E[\hat{\boldsymbol{\eta}}_{t|t-1}\hat{\boldsymbol{\eta}}'_{t-h|t-h-1}|\Im_{t-1}] = \mathbf{O}\ \forall h \geq 1$ or $E[\hat{\boldsymbol{\eta}}_{t|t-1} g(\hat{\boldsymbol{\eta}}'_{t-h|t-h-1})|\Im_{t-1}] = \mathbf{O}\ \forall h \geq 1$ where $g(\cdot)$ is any function that extracts information from $\Im_{t-1}$.

In the presence of MS heteroskedastic components, researchers in finance have also suggested to check whether the smoothed, standardized residuals contain any residual ARCH effects. Standard LM-type as well as Ljiung-Box tests can be applied. This is a way to check whether MS variance is

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

sufficient to capture most of the dynamics in volatility, else explicit ARCH-type modeling (even of a MS nature, see Section 7.2) would be required.[53]

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

$$RCM_1 = 100\frac{K^2}{T}\sum_{t=1}^{T}\prod_{k=1}^{K}\hat{\xi}_{t|T}^{k},$$

i.e., the sample average of the products of the smoothed state probabilities. Clearly, when a MS model offers precise indications on the nature of the regime at each time t, the implication is that for at least one value of $k = 1, ..., K$, $\hat{\xi}_{t|T}^{k} \simeq 1$ so that $\sum_{k=1}^{K}\hat{\xi}_{t|T}^{k} \simeq 0$ because most other smoothed probabilities are zero. Therefore a good MS model will imply $RCM_1 \simeq 0$.[54] However, when applied to models with $K > 2$, RCM_1 has one obvious disadvantage: a model can imply an enormous degree of uncertainty on the current regime, but still have $\sum_{k=1}^{K}\hat{\xi}_{t|T}^{k} \simeq 0$ for most values of t. For instance, when $K = 3$, it is easy to see that if $\hat{\xi}_{t|T}^{1} = 1/2$, $\hat{\xi}_{t|T}^{2} = 1/2$, and $\hat{\xi}_{t|T}^{3} = 0\ \forall t$, then $RCM_1 = 0$ even though this remains a rather uninformative switching model to use in practice. As a result, it is rather common to witness that as K exceeds 2, almost all switching models (good and bad) will automatically imply values of RCM_1 that decline towards 0. Guidolin (2009) proposes a number of alternative measures that may shield against this type of problems, for instance

$$RCM_2 = 100\left[1 - \frac{K^{2K}}{(K-1)^2}\frac{1}{T}\sum_{t=1}^{T}\prod_{k=1}^{K}\left(\hat{\xi}_{t|T}^{k} - \frac{1}{K}\right)^2\right].$$

7.1. *One multivariate MS example*

Before resuming the process of introducing new notions and better dealing with a few loose ends that have been left behind, we shall pause again one example, this time of a multivariate nature. In essence, we now want to *simultaneously* capture the time series dynamics in U.S. large cap, small cap, and 10-year Treasury monthly excess returns, over a 1954-1999 sample. These are the same data underlying Figures 1-5 above. The difference is that now we want to develop not three different univariate MS concerning one series at the time, but instead one unique tri-variate model for all excess returns jointly considered. Figure 6 presents a table in which we perform a number of the model specification tests that have been discussed above. For instance, the p-values of the LR tests concerning the number of

[53] Under the null of regime switching, the resulting asset returns have non-linear stochastic structures that could show up in significant ARCH-type tests even in the absence of truly ARCH effects in the data generating process.

[54] On the opposite, the worst possible MSM has $\hat{\xi}_{t|T}^{1} = ... = \hat{\xi}_{t|T}^{K} = 1/K$ so that $\sum_{k=1}^{K}\hat{\xi}_{t|T}^{k} = 1/K^2$ and $RCM_1 = 100$. Therefore $RCM_1 \in [0, 100]$ and lower values are to be preferred to higher ones.

regimes are corrected for the presence of nuisance parameters in the way indicated by Davies (1977).[55]

Model (k,p)	Number of parameters	Log-likelihood	LR test for linearity	Hannan−Quinn
		Base model: MSIA(1,0)		
MMSIA(1,0)	9	3290.82	NA	−11.8632
MMSIA(1,1)	18	3314.34	NA	−11.9099
MMSIA(1,2)	27	3314.72	NA	−11.8618
		Base model: MSIA(2,0)		
MMSIA(2,0)	14	3316.24	50.8244 (0.000)	−11.8552
MMSIAH(2,0)	20	3392.79	203.9312 (0.000)	−12.1592
MMSIAH(2,1)	38	3436.99	245.2865 (0.000)	−12.2213
MMSIAH(2,2)	56	3438.51	253.5739 (0.000)	−12.1285
		Base model: MSIA(3,0)		
MMSIA(3,0)	21	3340.86	100.0658 (0.000)	−11.9643
MMSIAH(3,0)	33	3418.03	254.4206 (0.000)	−12.1639
MMSIAH(3,1)	60	3468.10	307.5043 (0.000)	−12.1871
MMSIAH(3,2)	87	3480.08	336.7194 (0.000)	−12.0721
		Base model: MSIA(4,0)		
MMSIA(4,0)	30	3380.29	178.9327 (0.000)	−12.0471
MMSIAH(4,0)	48	3462.91	344.1803 (0.000)	−12.2263
MMSIAH(4,1)	84	3517.36	406.0404 (0.000)	−12.2054
MMSIAH(4,2)	120	3554.56	485.6775 (0.000)	−12.1218
MMSIAH(4,3)	156	3589.30	550.8718 (0.000)	−12.0291

Figure 6: Model specification search for U.S. stock and bond portfolios, monthly 1954-1999 U.S. data

A few remarks are in order. First, the highest maximized log-likelihood is simply given by the biggest available model that is taken to the data, in this case a very rich MSIAH(4,3) model that implies the estimation of 156 parameters with 1,656 observations (saturation ratio is then 10.6 only). This is why one should penalize the maximized log-likelihood: ruling out numerical problems in the maximization performed by your Math package, it is easy to inflate the log-likelihood by expanding the number of parameters, which does not mean this a good idea either in a statistical or in a financial economics perspective. When you penalize the log-likelihood, for instance in this case by computing the Hannan-Quinn criterion, the minimum is reached at -12.2263 by a much more parsimonious MSIH(4,0) model characterized by 48 parameters to estimate (which gives a saturation ratio of 34.5, which appears to be rather comfortable). Interestingly, this model is characterized by four different states. The closest model seems to appear a slightly smaller MSIAH(2,1) model that achieves a H-Q score of -12.2213. Finally, in this case it is clear that formally testing for the number of regimes using an adjusted LR test always rejects the null of $K = 1$ against $K \geq 2$ and that this happens for $K = 2$, 3 and 4.

Figure 7 reports instead parameter estimates for the MSIH(4,0) that was selected by the H-Q criterion. You can read a summary interpretation of what the four different regimes may be taken to represent next to the regime labels in the Figure. Interestingly, asset volatilities are strongly time-varying but tend also in this trivariate model to be higher in the two extreme regimes (especially in the bear regime) than they are in the slow growth and bull states. Moreover, also dynamic correlations become now time-varying because these depend on the Markov state variable. Correlations between large and small stocks are very high and highly statistically significant in bear markets, but at least bonds provide considerable hedging. Equity correlations are instead smaller in regimes 3 and 4, when space to exploit diversification benefits seem to exist.

Figure 8 shows the smoothed probabilities computed from the MSIH(4,0) estimated in Figure 7. Note that there is now only one set of smoothed probability plots (differently from the 9 figures that

[55] In the table, you read about MMSIAH(K,p) models becase the first M stands for "multivariate". You can simply disregard the first M for practical purposes.

have appeared in Figures 4 and 5), one for each of the $K = 4$ regimes. While regimes 2 and 3 are rather persistent (with average durations of 6.8 and 8.5 months, respectively), regime 4 is less persistent (3.2 months), and the first bear state is not persistent at all, capturing situations of unpredictable market crashes. A casual inspection confirms that the period captured by high smoothed probabilities of regime 1 approximately correspond to those that have already appeared in Figure 4 for large and small stocks. The same is true of all other regimes. The implication is that bond-specific regimes seem to be not reflected by the behavior of the unique, multivariate Markov state S_t that has been estimated here or, equivalently that such a S_t is more affected by S_t^{large} and S_t^{small} than it is by S_t^{bond}. Why this may be the case and what its effects may be for portfolio and risk management is a topic currently under investigation.

	Panel B: four-state model			
	Large caps	Small caps		Long-term bonds
1. Mean excess return				
Regime 1 (crash)	−0.0510 (0.0146)	−0.0810 (0.0219)		−0.0131 (0.0047)
Regime 2 (slow growth)	0.0069 (0.0027)	0.0008 (0.0033)		0.0009 (0.0016)
Regime 3 (bull)	0.0116 (0.0032)	0.0167 (0.0048)		−0.0023 (0.0007)
Regime 4 (recovery)	0.0226 (0.0055)	0.0458 (0.0098)		0.0098 (0.0033)
2. Correlations/volatilities				
Regime 1 (crash):				
Large caps	0.1625***			
Small caps	0.8233***	0.2479***		
Long-term bonds	−0.4060*	−0.2590		0.0902***
Regime 2 (slow growth):				
Large caps	0.1118***			
Small caps	0.7655***	0.1099***		
Long-term bonds	0.2043***	0.1223		0.0688***
Regime 3 (bull):				
Large caps	0.1133***			
Small caps	0.6707***	0.1730***		
Long-term bonds	0.1521	−0.0976		0.0261***
Regime 4 (recovery):				
Large caps	0.1479***			
Small caps	0.5013***	0.2429***		
Long-term bonds	0.3692***	−0.0011		0.1000***
3. Transition probabilities	Regime 1	Regime 2	Regime 3	Regime 4
Regime 1 (crash)	0.4940 (0.1078)	0.0001 (0.0001)	0.0241 (0.0417)	0.4818
Regime 2 (slow growth)	0.0483 (0.0233)	0.8529 (0.0403)	0.0307 (0.0110)	0.0682
Regime 3 (bull)	0.0439 (0.0252)	0.0701 (0.0296)	0.8822 (0.0403)	0.0038
Regime 4 (recovery)	0.0616 (0.0501)	0.1722 (0.0718)	0.0827 (0.0498)	0.6836

Figure 7: MLE estimates of MSIH(4,0) model for U.S. stock and bond portfolios, monthly 1954-1999 U.S. data

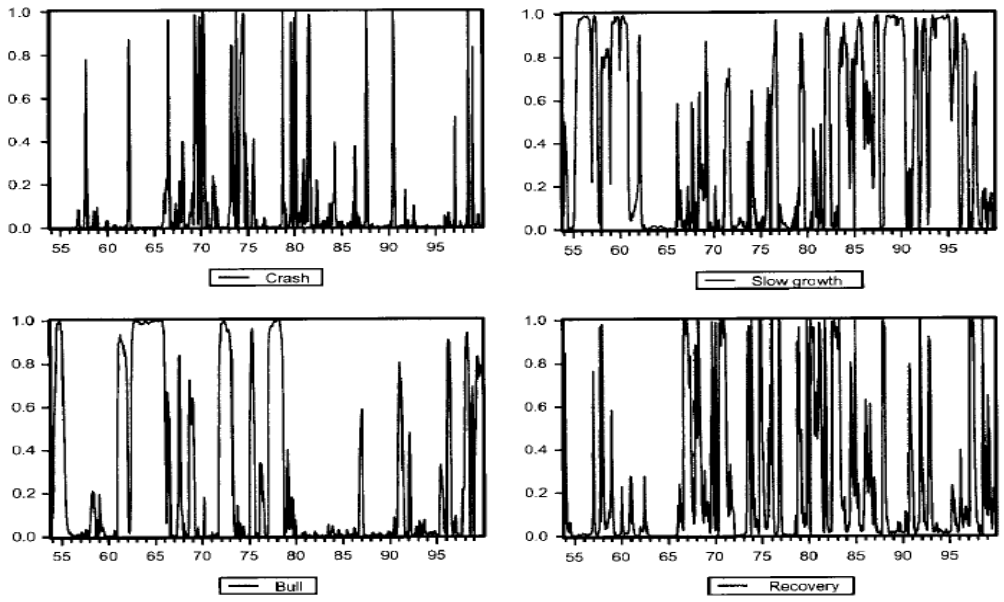


Figure 8: Smoothed probabilities from MSIH(4,0) model for U.S. stocks and bonds, monthly 1954-1999 U.S. data

7.2. *Extensions: Markov switching ARCH*

One may wonder about the positioning of MS models within the logical path covered so far in chapters 5-6. The path was marked by three incremental steps: (i) even after fitting relatively sophisticated GARCH models, standardized residuals from financial returns often remain non-Gaussian with evidence of thick tails and asymmetries, so that in chapter 5 we have developed methods to model non-Gaussian returns; (ii) to deploy active risk management methods, you need to model correlations, besides variances; (iii) DCC models are the most promising multivariate models of heteroskedastic dynamics in second moments. Instead of representing a separate approach, MS models perfectly fit the logical sequence marked by (i)-(iii) above, in at least two ways:

- MS models represent distinct, practical and powerful solutions to problems (i)-(ii) above;
- MS methods can be easily combined with everything else we have seen in chapters 5-6 and therefore are not in contradiction with the development in (iii) above.

As for the first point, we have already seen in Section 4 that in general MS models generate strong non-normalities in asset returns, such as non-zero skewness and excess kurtosis. Moreover, our example in Section 7.1 has shown that multivariate MS models may lead to the estimation of regime switching correlations, which is clearly relevant to (ii) above.

As for the second bullet point, although at some frequencies—mostly monthly, when the residuals of well-specified MS models often reject the need of also introducing ARCH effects—MS directly competes with GARCH, at high (daily, weekly) frequencies MS, ARCH, DCC, and t-Student variants are compatible with Markov switching. For instance, efforts have been made to produce MS models with switching ARCH and GARCH effects; the same applies to DCC models. Although GARCH models driven by normally distributed innovations and their numerous extensions can account for a substantial portion of both the volatility clustering and excess kurtosis found in financial return series, a GARCH-type model has yet to be constructed for which the filtered residuals consistently fail to exhibit clear-cut signs of non-normality. On the contrary, it appears that the vast majority of GARCH models, when fitted to returns over weekly and shorter horizons, imply quite heavy-tailed conditional innovation distributions. A natural solution has consisted of developing GARCH frameworks that incorporate the original assumption of normal innovations but in which the conditional distribution is mixture of normals, as under MS. As for the frequency, the empirical result that seems to rule in most of applied econometrics holds: the higher the frequency, the higher the chances that MS GARCH may genuinely be needed, with little peril of over-fitting the data. As a rule of thumb, most papers that analyze daily or weekly data, normally specify some form of MS GARCH process; at a monthly frequency, there is much more uncertainty as to what the right choice may be;[56] at quarterly or annual frequencies, strong evidence of both regimes and ARCH seems unlikely.

[56]For instance, using U.K. equity and bond data, Guidolin and Timmermann (2005) formally test for (bivariate) ARCH effects in a three-state MSIH-type model and found that the null of no ARCH cannot be rejected. On U.S. monthly equity data, Guidolin and Timmermann (2007), have reported similar evidence in a four-state model.

What does a MS GARCH look like? Cai (1994) develops a MS ARCH model to examine the issue of volatility persistence in monthly excess returns of 3-month T-bills: Cai was concerned that the high volatility persistence commonly reported from ARCH models might be spuriously inflated by the presence of a small number of regime shifts. Cai proposed to model occasional shifts in the long-run, ergodic variance of a MS ARCH process. In this case, the conditional variance is no longer determined by an exact linear combination of past squared shocks, as in a standard ARCH: the intercept in the conditional variance is allowed to change in response to occasional discrete shifts. Thus the model is able to retain the volatility-clustering feature of ARCH and, in addition, to capture the discrete shifts in the intercept in the conditional variance that may cause spurious persistence in the process. In the simplest of the two-regime cases explored by Cai (1994), his MSIAR(2,1) ARCH process is:

$$\begin{aligned} R_{t+1} &= \mu_{S_{t+1}} + \phi(R_t - \mu_{S_{t+1}}) + \sqrt{\sigma^2_{t+1}}\epsilon_{t+1} \qquad \epsilon_{t+1} \sim \text{IID } N(0,1) \\ \sigma^2_{t+1} &= \omega_{S_{t+1}} + \sum_{i=1}^{q} \alpha_i \epsilon^2_{t+1-i}, \quad \omega_{S_{t+1}},\ \alpha_i \geq 0, \end{aligned}$$

where $S_t = 1, 2$ follows a first-order, homogeneous and irreducible two-state Markov chain. A related, but slightly different approach is Hamilton and Susmel's (1994) who have proposed a (SWARCH) model in which changes in regime are captured as changes in the scale of the ARCH process,

$$\begin{aligned} R_{t+1} &= \mu + \sqrt{\delta_{S_{t+1}}}\sqrt{\sigma^2_{t+1}}\epsilon_{t+1} \qquad \epsilon_{t+1} \sim \text{IID } N(0,1) \\ \sigma^2_{t+1} &= \omega + \sum_{i=1}^{q} \alpha_i \epsilon^2_{t+1-i}, \quad \alpha_i \geq 0,\ S_t = 0, 1, 2, \end{aligned}$$

so that ϵ_t follows a standard ARCH(p) process and the MS component concerns the scaling factor $\delta_{S_{t+1}}$. This is obviously different (and in some sense more powerful) than Cai's MS ARCH in which a shift to the volatile regime only affects the unconditional (long-run) variance, while in Hamilton and Susmel's SWARCH also the dynamic process of conditional variance is affected. This model is flexible enough to attribute most of the persistence in the volatility of stock returns to the persistence of the low-, moderate-, and high-volatility regimes, which typically last for several years.

Both of these models simply focus on augmenting ARCH with regimes. In a way, this is natural because the point of the literature has been to show that the high persistence of asset return volatilities often reported in the GARCH literature may have been spuriously inflated by the presence of regime shifts and/or breaks. As we have seen in chapter 4, the reason why Bollerslev (1986) had proposed the GARCH generalization of ARCH was to increase the persistence of the ARCH conditional heteroskedastic family within a parsimonious parameterization. Therefore, the early prominence of MS ARCH models over MS GARCH models should not come a surprise. However, one may still wonder how we should go about specifying and estimating MS GARCH models. Unfortunately, combining the MS model with GARCH induces tremendous complications in estimation. As a result of the particular lag structure of a GARCH model—by which all past lags of squared shocks affect conditional variance—the standard equations characterizing the EM algorithm for MS parameter estimation would

depend on the entire history of the Markov states through the smoothed probabilities. Because each of the Markov states may take K values, this implies a total of K^T probabilities that need to be computed and stored, which would make most MS GARCH models extremely difficult to estimate for sample sizes of more than 100 observations. Direct maximum likelihood estimation (i.e., not based on the EM algorithm) via a nonlinear filter also turned out to be practically infeasible. Gray (1996) has developed a two-state generalized MS ARCH model for the U.S. short-term riskless nominal interest rate (1-month T-bill, i_t):

$$\begin{aligned} \Delta i_{t+1} &= \mu_{S_{t+1}} + \phi_{S_{t+1}} i_t + \sqrt{\sigma^2_{t+1}}\epsilon_{t+1} \qquad \epsilon_t \sim \text{IID } N(0,1) \\ \sigma^2_{t+1} &= \omega_{S_{t+1}} + \alpha_{S_{t+1}}\epsilon^2_t + \beta_{S_{t+1}}\sigma^2_t, \end{aligned} \tag{25}$$

$(S_t = 1, 2)$ which implies an infinite memory because $Var_t[\Delta i_{t+1}|S_{t+1}] = \omega_{S_{t+1}} + \alpha_{S_{t+1}}\epsilon^2_t + \beta_{S_{t+1}} Var_{t-1}[\Delta i_t|S_t]$, which can be solved backwards to show that conditional variance depends on the entire history of shocks to the short-term rate, ϵ_0, $\epsilon_1, ..., \epsilon_t$. Gray tackles the problem of path dependence in MS GARCH adopting an approach that preserves the essential nature of GARCH and yet allows tractable estimation. Under conditional normality, the variance of changes in the short rate at time t is given by

$$\bar{\sigma}^2_t = E_{t-1}[(\Delta i_t)^2] - \{E_{t-1}[\Delta i_t]\}^2 = \Pr(S_t{=}2|\mathcal{F}_{t-1})[\mu^2_2 + \sigma^2_t(S_{t-1}{=}2|\mathcal{F}_{t-1})] +$$
$$+[1 - \Pr(S_t{=}2|\mathcal{F}_{t-1})][\mu^2_1 + \sigma^2_t(S_t{=}1|\mathcal{F}_{t-1})] - \{\Pr(S_t{=}2|\mathcal{F}_{t-1})\mu_2 - [1 - \Pr(S_t{=}2|\mathcal{F}_{t-1})]\mu_1\}^2,$$

which is not path-dependent and corresponds to a difference of averages across regimes (with probabilities given by filtered probabilities) of the the first and second moments. This value of $\bar{\sigma}^2_t$ can now be used in the MS GARCH (1,1) specification (25) to replace $\sigma^2_t(S_t)$.

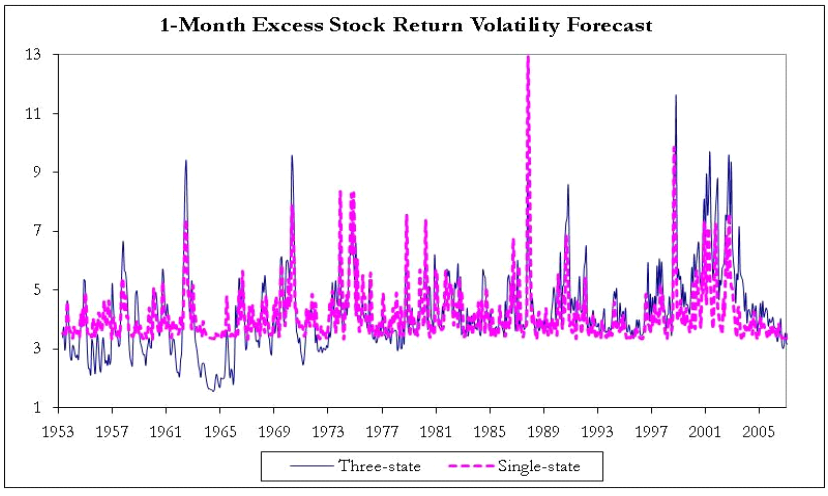


Figure 9: 1-month predicted volatility from single- vs. three-state GARCH models of U.S. excess stock returns

For instance, Figure 9 below is copied from Guidolin (2009) and shows the differences in filtered volatility for U.S. excess stock monthly returns between a standard GARCH(1,1) model and a three-regime MS GARCH model with structure similar to (25). Visibly, their dynamics appears to be similar,

but the three-state GARCH allows one-month predicted volatility to sometimes go below the level allowed by the single-state model in the first portion of the sample, while during the bear market period of 1999-2002, the opposite occurs—some volatility spikes are estimated under MS that fail to be visible under a plain-vanilla GARCH. Guidolin (2009) shows that volatility forecasts obtained incorporate regime switching are considerably more accurate than those obtained from standard methods.

The extensions discussed above only concern univariate ARCH and GARCH models. What about their multivariate counterparts? Pelletier (2006) has recently proposed an extension of Bollerslev's (1990) constant conditional correlation (CCC) multivariate framework to incorporate MS dynamics in the conditional variance and covariance functions. As it is well known, applied econometricians face considerable identification and numerical problems when they try to write truly multivariate models of volatility and correlations. This is because not only must the variances be positive, the variance matrix must also be positive semi-definite at every point in time. Another important problem is the curse of dimensionality; because applied researchers need models that can be applied to more than a few time series, it is problematic to directly generalize the plain vanilla univariate GARCH models, and even relatively parsimonious and carefully constructed multivariate variations such as the BEKK model do suffer from a tendency to over-parameterize the estimation problem. Chapter 6 has discussed these issues at length. Similarly to a standard DCC model, Pelletier's (2006) regime switching dynamic correlation (RSDC) model decomposes the covariances into standard deviations and correlations, but these correlations are allowed to change over time as they follow a MS model:

$$\mathbf{R}_{t+1} = \mathbf{\Sigma}_{t+1}^{1/2}\boldsymbol{\epsilon}_{t+1} \qquad \boldsymbol{\epsilon}_t \text{ IID } (\mathbf{0}, \mathbf{I}_N) \qquad \mathbf{\Sigma}_{t+1} = \mathbf{D}_{t+1}\mathbf{\Gamma}_{S_{t+1}}\mathbf{D}_{t+1},$$

where $\mathbf{D}_{t+1}$ is a diagonal matrix composed of the standard deviations (Pelletier simply suggests that each of them may follow a standard univariate GARCH(1,1) process) of N return series and the regime-dependent matrix $\mathbf{\Gamma}_{S_{t+1}}$ contains the correlations that are assumed to be constant within a regime but different across regimes. This feature implies that in the evaluation of the likelihood, the correlation matrix can only take K possible values so we only have to invert K times a $N \times N$ matrix, which—especially when the number of time series is large—can be a computational advantage over models such as a DCC, where a different correlation matrix has to be inverted for every observation.

Pelletier shows that the RSDC model has many interesting properties. First, it is easy to impose that the variance matrices are PSD. Second, it does not suffer from a curse of dimensionality because it can be estimated with a two-step procedure. Third, when combined with a ARMACH model for standard deviations (here the ACH stands for absolute conditional heteroskedasticity),

$$\sigma_{t+1} = \nu + \phi|R_t| + \psi\sigma_t$$

(in its simplest (1,1) version), this correlation model allows analytic computation of multi-step ahead conditional expectations of the whole variance matrix.[57] Finally, by modelling time variation in corre-

[57] Although it may seem natural (as in chapter 6) to adopt a GARCH(1,1) to model the process of univariate conditional variances, because the covariance is the product of a correlation and the square-root of the product of two variances, the square-root introduces non-linearities that will prohibit analytic computation of conditional expectations of correlations.

lations as a MS model, the variances and covariances are not bounded which is the case when they are the ones following a regime switching. Estimation is made simpler by adopting a two-step quasi-ML estimation procedure as in chapter 6: in a first step, we can estimate the univariate volatility models and in a second step, we can estimate the parameters in the correlation matrix conditional on the first step estimates. In an empirical application to exchange rate data, Pelletier also shows that a simple two-state RSDCM can produce a better fit than the celebrated DCC model.

8. **Non-Normal Returns under MS Mixture: Implied Conditional Moments**

In Section 4 we have generically argued that *mixtures* of normals driven by MS models may lead to strong non-normalities in returns. This makes MS an interesting alternative to other routes explored in chapter 5 to generate realistic skewness and excess kurtosis. Although these issues are rather algebra-intensive and can actually best seen also from rich sets of simulations, some insight may be gained from considering a simple univariate MSIH(K) model with $K = 2$, written as

$$R_{t+1} = S_{t+1}\mu_1 + (1 - S_{t+1})\mu_0 + [S_{t+1}\sigma_1 + (1 - S_{t+1})\sigma_0]z_{t+1} \qquad z_{t+1} \sim N(0,1),$$

in which $S_{t+1} = 0, 1$ is *unobservable* at all points in time. You can easily see that in this special $K = 2$ case, $S_{t+1}\mu_1 + (1-S_{t+1})\mu_0$ reproduces the $\mu_{S_{t+1}}$ dependence; the same applies to $S_{t+1}\sigma_1 + (1-S_{t+1})\sigma_0$.

Let's start by checking moments for the benchmark, single-regime case in which $K = 1$. Because these will be important below, we compute both unconditional and conditional moments. When $K = 1$, it is as if $S_t = 0$ always, which means there is only one regime and μ and σ may loose the pedix that refers to the regime. Therefore, when we perform calculations for time t conditional moments and for unconditional moments, respectively, we have:

$$\begin{aligned}
E_t[R_{t+1}] &= E_t[\mu + \sigma z_{t+1}] = \mu + \sigma E_t[z_{t+1}] = \mu \\
E[R_{t+1}] &= E[\mu + \sigma z_{t+1}] = \mu + \sigma E[z_{t+1}] = \mu \\
Var_t[R_{t+1}] &= Var_t[\mu + \sigma z_{t+1}] = \sigma^2 Var_t[z_{t+1}] = \sigma^2 \\
Var[R_{t+1}] &= Var[\mu + \sigma z_{t+1}] = \sigma^2 Var[z_{t+1}] = \sigma^2 \\
Skewness_t[R_{t+1}] &= \frac{E_t[(R_{t+1} - E_t[R_{t+1}])^3]}{(Var_t[R_{t+1}])^{1.5}} = \frac{E_t[(\mu + \sigma z_{t+1} - \mu)^3]}{\sigma^3} \\
&= \frac{\sigma^3 E_t[z_{t+1}^3]}{\sigma^3} = 0 \text{ (as } z_{t+1} \sim N(0,1)) \\
Skewness[R_{t+1}] &= \frac{E[(\mu + \sigma z_{t+1} - \mu)^3]}{\sigma^3} = \frac{\sigma^3 E[z_{t+1}^3]}{\sigma^3} = 0 \\
Exkurt_t[R_{t+1}] &= \frac{E_t[(R_{t+1} - E_t[R_{t+1}])^4]}{(Var_t[R_{t+1}])^2} - 3 = \frac{E_t[(\mu + \sigma z_{t+1} - \mu)^4]}{\sigma^4} - 3 \\
&= \frac{\sigma^4 E_t[z_{t+1}^4]}{\sigma^4} - 3 = 0 \text{ (as } z_{t+1} \sim N(0,1)) \\
Exkurt[R_{t+1}] &= \frac{E_t[(\mu + \sigma z_{t+1} - \mu)^4]}{\sigma^4} - 3 = \frac{\sigma^4 E[z_{t+1}^4]}{\sigma^4} - 3 = 0
\end{aligned}$$

Because $z_{t+1} \sim N(0,1)$ and $R_{t+1} = \mu + \sigma z_{t+1}$, and σ is constant, we have that R_{t+1} has a normal

conditional and unconditional distribution.

Things are a tad more involved when $K = 2$. In this case, when you apply the conditioning, you will also need to condition with respect to the current state, S_t:

$$\begin{aligned}
E[R_{t+1}|S_t] &= E[S_{t+1}\mu_1 + (1 - S_{t+1})\mu_0 + (S_{t+1}\sigma_1 + (1 - S_{t+1})\sigma_0)z_{t+1}|S_t] \\
&= E[S_{t+1}|S_t]\mu_1 + E[(1 - S_{t+1})|S_t]\mu_0 + E[S_{t+1}|S_t]E[\sigma_1 z_{t+1}|S_t] + \\
&\quad + E[(1 - S_{t+1})|S_t]E[\sigma_0 z_{t+1}|S_t] \\
&= \Pr(S_{t+1} = 1|S_t)\mu_1 + (1 - \Pr(S_{t+1} = 1|S_t))\mu_0 \\
E[R_{t+1}] &= E[S_{t+1}\mu_1 + (1 - S_{t+1})\mu_0 + (S_{t+1}\sigma_1 + (1 - S_{t+1})\sigma_0)z_{t+1}] \\
&= E[S_{t+1}]\mu_1 + E[(1 - S_{t+1})]\mu_0 + E[S_{t+1}]E[\sigma_1 z_{t+1}] + E[(1 - S_{t+1})]E[\sigma_0 z_{t+1}] \\
&= \pi_1\mu_1 + (1 - \pi_1)\mu_0
\end{aligned}$$

where π_1 is the unconditional probability of regime 1, and $(1 - \pi_1)$ is the unconditional probability of regime 2.[58] Insofar as $\pi_1 \neq \Pr(S_{t+1}|S_t)$ and $(1 - \pi_1) \neq 1 - \Pr(S_{t+1}|S_t)$, clearly $E[R_{t+1}|S_t] \neq E[R_{t+1}]$ as the first moment will be a time-varying one. As for variances:

$$\begin{aligned}
Var_t[R_{t+1}] &= \Pr(S_{t+1} = 1|S_t)E[(\mu_1 + \sigma_1 z_{t+1} - \Pr(S_{t+1} = 1|S_t)\mu_1 - (1 - \Pr(S_{t+1} = 1|S_t))\mu_0)^2|S_t] + \\
&\quad + \Pr(S_{t+1} = 0|S_t)E[(\mu_0 + \sigma_0 z_{t+1} - \Pr(S_{t+1} = 1|S_t)\mu_1 - (1 - \Pr(S_{t+1} = 1|S_t))\mu_0)^2|S_t] \\
&= \Pr(S_{t+1} = 1|S_t)E[((1 - \Pr(S_{t+1} = 1|S_t))(\mu_1 - \mu_0) + \sigma_1 z_{t+1})^2|S_t] + \\
&\quad + \Pr(S_{t+1} = 0|S_t)E[(\Pr(S_{t+1} = 1|S_t)(\mu_0 - \mu_1) + \sigma_0 z_{t+1})^2|S_t] \\
&= \Pr(S_{t+1} = 1|S_t)(1 - \Pr(S_{t+1} = 1|S_t))^2(\mu_1 - \mu_0)^2 + (1 - \Pr(S_{t+1} = 1|S_t)) \times \\
&\quad \times (\Pr(S_{t+1} = 1|S_t))^2(\mu_1 - \mu_0)^2 + \Pr(S_{t+1} = 1|S_t)\sigma_1^2 + (1 - \Pr(S_{t+1} = 1|S_t))\sigma_0^2 \\
&= \Pr(S_{t+1} = 1|S_t)(1 - \Pr(S_{t+1} = 1|S_t))(\mu_1 - \mu_0)^2[(1 - \Pr(S_{t+1} = 1|S_t)) + \\
&\quad + (\Pr(S_{t+1} = 1|S_t))] + \Pr(S_{t+1} = 1|S_t)\sigma_1^2 + (1 - \Pr(S_{t+1} = 1|S_t))\sigma_0^2 \\
&= \Pr(S_{t+1}{=}1|S_t)\sigma_1^2 + (1 - \Pr(S_{t+1}{=}1|S_t))\sigma_0^2 + \Pr(S_{t+1}{=}1|S_t)(1 - \Pr(S_{t+1}{=}1|S_t))(\mu_1 - \mu_0)^2
\end{aligned}$$

because $[(1 - \Pr(S_{t+1} = 1|S_t)) + (\Pr(S_{t+1} = 1|S_t))] = 1$. Instead

$$\begin{aligned}
Var[R_{t+1}] &= \pi_1 E[(\mu_1 + \sigma_1 z_{t+1} - \pi_1\mu_1 - (1 - \pi_1)\mu_0)^2] + \\
&\quad + (1 - \pi_1)E[(\mu_0 + \sigma_0 z_{t+1} - \pi_1\mu_1 - (1 - \pi_1)\mu_0)^2] \\
&= \pi_1 E[((1 - \pi_1)(\mu_1 - \mu_0) + \sigma_1 z_{t+1})^2] - (1 - \pi_1)E[\pi_1(\mu_1 - \mu_0) + \sigma_0 z_{t+1})^2]
\end{aligned}$$

[58] The reason why

$$E[S_{t+1}\sigma_j z_{t+1}|S_t] = E[S_{t+1}|S_t]E[\sigma_j z_{t+1}|S_t] = 0 \qquad j = 0, 1$$

is that given S_t, S_{t+1} is independent of any other random variable indexed at time $t + 1$, and in particular S_{t+1} is independent of z_{t+1} (just think of the way we have "manually simulated" returns from Markov switching at the very beginning of lecture 5). Moreover, $E[\sigma_1 z_{t+1}|S_t] = \sigma_1 E[z_{t+1}|S_t] = \sigma_1 E[z_{t+1}] = 0$ because $E[z_{t+1}] = 0$ by construction in a Markov switching model. The same applies to

$$E[S_{t+1}\sigma_j z_{t+1}] = E[S_{t+1}]E[\sigma_j z_{t+1}] = 0 \qquad j = 0, 1.$$

$$
\begin{aligned}
&= \pi_1(1-\pi_1)^2(\mu_1-\mu_0)^2 + (1-\pi_1)\pi_1^2(\mu_1-\mu_0)^2 + \pi_1\sigma_1^2 + (1-\pi_1)\sigma_0^2 \\
&= \pi_1(1-\pi_1)(\mu_1-\mu_0)^2[(1-\pi_1)+\pi_1)] + \pi_1\sigma_1^2 + (1-\pi_1)\sigma_0^2 \\
&= \pi_1\sigma_1^2 + (1-\pi_1)\sigma_0^2 + \pi_1(1-\pi_1)(\mu_1-\mu_0)^2.
\end{aligned}
$$

In both cases, notice that

$$
\begin{aligned}
Var_t[R_{t+1}] &\neq \Pr(S_{t+1}=1|S_t)\sigma_1^2 + (1-\Pr(S_{t+1}=1|S_t))\sigma_0^2 \\
Var_t[R_{t+1}] &\neq \pi_1\sigma_1^2 + (1-\pi_1)\sigma_0^2,
\end{aligned}
$$

with the difference represented by the terms $\Pr(S_{t+1}=1|S_t)(1-\Pr(S_{t+1}=1|S_t))(\mu_1-\mu_0)^2$ in the case of the conditional variance and $\pi_1(1-\pi_1)(\mu_1-\mu_0)^2$ in the case of the unconditional variance. This means that in a MSIH$(2,0)$ model, not only the regime-specific variances will be weighted in the overall variances across regimes, but also the (squared) size of the between-regime "jumps" in regime-specific means, $\mu_1-\mu_0$, will contribute to the variability of the process.

We now move to compute conditional and unconditional skewness:

$$
\begin{aligned}
E_t[(R_{t+1}-E_t[R_{t+1}])^3] &= \Pr(S_{t+1}=1|S_t)E[(\mu_1+\sigma_1 z_{t+1}-\Pr(S_{t+1}=1|S_t)\mu_1-(1-\Pr(S_{t+1}=1|S_t))\mu_0)^3|S_t]+ \\
&+ \Pr(S_{t+1}=0|S_t)E[(\mu_0+\sigma_0 z_{t+1}-\Pr(S_{t+1}=1|S_t)\mu_1-(1-\Pr(S_{t+1}=1|S_t))\mu_0)^3|S_t] \\
&= \Pr(S_{t+1}=1|S_t)(1-\Pr(S_{t+1}=1|S_t))^3(\mu_1-\mu_0)^3 + (1-\Pr(S_{t+1}=1|S_t))\times \\
&\quad \times(\Pr(S_{t+1}=1|S_t))^3(\mu_0-\mu_1)^3 + \sigma_1^3 E[z_{t+1}^3|S_t] + \sigma_0^3 E[z_{t+1}^3|S_t]+ \\
&+ 3\Pr(S_{t+1}=1|S_t)(1-\Pr(S_{t+1}=1|S_t))(\mu_1-\mu_0)\sigma_1^2 E[z_{t+1}^2|S_t]+ \\
&\quad - 3\Pr(S_{t+1}=1|S_t)(1-\Pr(S_{t+1}=1|S_t))(\mu_1-\mu_0)\sigma_0^2 E[z_{t+1}^2|S_t] \\
&= \Pr(S_{t+1}=1|S_t)(1-\Pr(S_{t+1}=1|S_t))(\mu_1-\mu_0)^3[(1-\Pr(S_{t+1}=1|S_t))^2 - (\Pr(S_{t+1}=1|S_t))^2] \\
&\quad + 3\Pr(S_{t+1}=1|S_t)(1-\Pr(S_{t+1}=1|S_t))(\mu_1-\mu_0)(\sigma_1^2-\sigma_0^2)
\end{aligned}
$$

where $E[z_{t+1}^3|S_t]=0$, $E[z_{t+1}^2|S_t]=1$, $E[(1-\Pr(S_{t+1}=1|S_t))^2(\mu_1-\mu_0)^2\sigma_1 z_{t+1}|S_t] = E[(\Pr(S_{t+1}=1|S_t))^2\ (\mu_0-\mu_1)^2\ \sigma_0 z_{t+1}|S_t]=0$, so that

$$
Skewness_t[R_{t+1}] = (\mu_1-\mu_0)\frac{\xi_{1,t+1}(1-\xi_{1,t+1})\left\{(\mu_1-\mu_0)^2[(1-\xi_{1,t+1})^2-\xi_{1,t+1}^2]+3(\sigma_1^2-\sigma_0^2)\right\}}{[\xi_{1,t+1}\sigma_1^2+(1-\xi_{1,t+1})\sigma_0^2+\xi_{1,t+1}(1-\xi_{1,t+1})(\mu_1-\mu_0)^2]^{3/2}},
$$

where we have shortened the notation by defining $\xi_{1,t+1}\equiv\Pr(S_{t+1}=1|S_t)$, as in Section 5. Similarly, straightforward but tedious algebra reveals that

$$
Skewness[R_{t+1}] = (\mu_1-\mu_0)\frac{\pi_1(1-\pi_1)[(\mu_1-\mu_0)^2[(1-\pi_1)^2-\pi_1^2]+3(\sigma_1^2-\sigma_0^2)]}{[\pi_1\sigma_1^2+(1-\pi_1)\sigma_0^2+\pi_1(1-\pi_1)(\mu_1-\mu_0)^2]^{3/2}}.
$$

This finding is very interesting:

$$
\begin{aligned}
Skewness_t[R_{t+1}] &\neq 0 \text{ if and only if } \mu_1\neq\mu_0 \\
Skewness[R_{t+1}] &\neq 0 \text{ if and only if } \mu_1\neq\mu_0,
\end{aligned}
$$

i.e., you need switching in conditional means in order for non-zero skewness to obtain. However, it is also clear that even when $\mu_1 \neq \mu_0$ it is possible for both conditional and unconditional skewness coefficient to be zero when (this is just a sufficient condition): (i) $\xi_{1,t+1} = 0$ or $\pi_1 = 0$; (ii) $\xi_{1,t+1}$=1$or\pi_1 = 1$. The two sets of restrictions do not carry the same meaning though, as $\pi_1 = 0$ or 1 really means you are not facing a MS model, in the sense that the underlying MC may be clearly reduced to a single state, while $\xi_{1,t+1} = 0$ or 1 just means that as of time t you are certain that in the following period you are either in the first regime or in the second.[59]

Finally, we deal with conditional and unconditional excess kurtosis:

$$
\begin{aligned}
E_t[(R_{t+1}-E_t[R_{t+1}])^4] &= \Pr(S_{t+1}=1|S_t)E[(\mu_1+\sigma_1 z_{t+1}-\Pr(S_{t+1}=1|S_t)\mu_1-(1-\Pr(S_{t+1}=1|S_t))\mu_0)^4|S_t]+ \\
&+ \Pr(S_{t+1}=0|S_t)E[(\mu_0+\sigma_0 z_{t+1}-\Pr(S_{t+1}=1|S_t)\mu_1-(1-\Pr(S_{t+1}=1|S_t))\mu_0)^4|S_t] \\
&= \Pr(S_{t+1}=1|S_t)(1-\Pr(S_{t+1}=1|S_t))^4(\mu_1-\mu_0)^4+(1-\Pr(S_{t+1}=1|S_t))\times \\
&\times(\Pr(S_{t+1}=1|S_t))^4(\mu_1-\mu_0)^4+6\Pr(S_{t+1}=1|S_t)(1-\Pr(S_{t+1}=1|S_t))^2(\mu_1-\mu_0)^2\sigma_1^2E[z_{t+1}^2|S_t]+ \\
&+6(1-\Pr(S_{t+1}=1|S_t))(\Pr(S_{t+1}=1|S_t))^2(\mu_1-\mu_0)^2\sigma_0^2E[z_{t+1}^2|S_t]+\sigma_1^4E[z_{t+1}^4|S_t]+\sigma_0^4E[z_{t+1}^4|S_t]+ \\
&= \Pr(S_{t+1}=1|S_t)(1-\Pr(S_{t+1}=1|S_t))(\mu_1-\mu_0)^4[(1-\Pr(S_{t+1}=1|S_t))^3+(\Pr(S_{t+1}=1|S_t))^3]+ \\
&+6\Pr(S_{t+1}=1|S_t)(1-\Pr(S_{t+1}=1|S_t))(\mu_1-\mu_0)^2[(1-\Pr(S_{t+1}=1|S_t))\sigma_1^2+\Pr(S_{t+1}=1|S_t)\sigma_0^2]+ \\
&+3\Pr(S_{t+1}=1|S_t)\sigma_1^4+3(1-\Pr(S_{t+1}=1|S_t))\sigma_0^4
\end{aligned}
$$

where $E[z_{t+1}|S_t] = E[z_{t+1}^3|S_t] = 0$, $E[z_{t+1}^2|S_t] = 1$, $E[z_{t+1}^4|S_t] = 3$, so that

$$
\begin{aligned}
Excess\ Kurt_t[R_{t+1}] &= \frac{\xi_{1,t+1}(1-\xi_{1,t+1})\left\{(\mu_1-\mu_0)^4[(1-\xi_{1,t+1})^3+\xi_{1,t+1}^3]+6(\mu_1-\mu_0)^2\right.}{[\xi_{1,t+1}\sigma_1^2+(1-\xi_{1,t+1})\sigma_0^2+\xi_{1,t+1}(1-\xi_{1,t+1})(\mu_1-\mu_0)^2]^2}+ \\
&+\frac{\left.[(1-\xi_{1,t+1})\sigma_1^2+\xi_{1,t+1}\sigma_0^2]\right\}+3\xi_{1,t+1}\sigma_1^4+3(1-\xi_{1,t+1})\sigma_0^4}{[\xi_{1,t+1}\sigma_1^2+(1-\xi_{1,t+1})\sigma_0^2+\xi_{1,t+1}(1-\xi_{1,t+1})(\mu_1-\mu_0)^2]^2}-3,
\end{aligned}
$$

where we have shortened the notation by defining $\xi_{1,t+1} \equiv \Pr(S_{t+1} = 1|S_t)$, as in the lectures. Similarly, straightforward but tedious algebra reveals that

$$
\begin{aligned}
Excess\ Kurt[R_{t+1}] &= \frac{\pi_1(1-\pi_1)\left\{(\mu_1-\mu_0)^4[(1\text{-}\pi_1)^3+\pi_1^3]+6(\mu_1-\mu_0)^2[(1-\pi_1)\sigma_1^2+\pi_1\sigma_0^2]\right\}}{[\xi_{1,t+1}\sigma_1^2+(1-\xi_{1,t+1})\sigma_0^2+\xi_{1,t+1}(1-\xi_{1,t+1})(\mu_1-\mu_0)^2]^2} \\
&+\frac{3\pi_1\sigma_1^4+3(1-\pi_1)\sigma_0^4}{[\xi_{1,t+1}\sigma_1^2+(1-\xi_{1,t+1})\sigma_0^2+\xi_{1,t+1}(1-\xi_{1,t+1})(\mu_1-\mu_0)^2]^2}-3.
\end{aligned}
$$

[59] If that seems more plausible, consider that based on the results in the lectures, we have that

$$\Pr(S_{t+1}=1|S_t) = (1-p_{00})\Pr(S_t=0|\Im_t)+p_{11}\Pr(S_t=1|\Im_t)$$

which can be 1 if and only if either $(1-p_{00})\Pr(S_t=0|\Im_t)=1$ (but that means that $p_{00}=0$), $p_{11}\Pr(S_t=1|\Im_t)=1$ (but that means that $p_{11}=1$), or the sum happens to be one. The first two cases do indicate problems with the irreducibility of the MC. The third case is more interestig. If

$$1 = (1-p_{00})\Pr(S_t=0|\Im_t)+p_{11}\Pr(S_t=1|\Im_t) = \mathbf{e}_2'\mathbf{P}'\boldsymbol{\xi}_t = \mathbf{e}_2'E_t[\boldsymbol{\xi}_{t+1}]$$

where $\boldsymbol{\xi}_t$ is the 2×1 vector that collects time t filtered probabilities and $\mathbf{e}_2 \equiv [0\ 1]'$, this means that as of time t you are forecasting with certainty that time $t+1$ will be dominated by regime 1. That is rather odd, if you think of it, and it may pose of periodicity of the underlying MC.

This finding is once more very interesting. First of all, notice that also in this case, when $\mu_0 = \mu_1$,

$$Ex\ Kurt_t[R_{t+1}] = \frac{3\xi_{1,t+1}\sigma_1^4 + 3(1-\xi_{1,t+1})\sigma_0^4}{[\xi_{1,t+1}\sigma_1^2 + (1\text{-}\xi_{1,t+1})\sigma_0^2]^2} - 3 = 3\left[\frac{\xi_{1,t+1}\sigma_1^4 + (1-\xi_{1,t+1})\sigma_0^4}{\xi_{1,t+1}^2\sigma_1^4 + (1\text{-}\xi_{1,t+1})^2\sigma_0^4 + 2\xi_{1,t+1}(1\text{-}\xi_{1,t+1})\sigma_0^2\sigma_1^2} - 1\right]$$

which is less than the expression found above: regime switching means simply adds to the excess kurtosis of a series. Moreover, in this case MS will generate positive excess kurtosis if and only if

$$\xi_{1,t+1}\sigma_1^4 + (1-\xi_{1,t+1})\sigma_0^4 > \xi_{1,t+1}^2\sigma_1^4 + (1-\xi_{1,t+1})^2\sigma_0^4 + 2\xi_{1,t+1}(1-\xi_{1,t+1})\sigma_0^2\sigma_1^2.$$

Moreover, notice that if one also has $\sigma_0^2 = \sigma_1^2 = \sigma^2$, then

$$Excess\ Kurt_t[R_{t+1}] = \frac{\sigma^4[3\xi_{1,t+1} + 3(1-\xi_{1,t+1})]}{\sigma^4[\xi_{1,t+1} + (1-\xi_{1,t+1})]^2} - 3 = 0,$$

as it should be because when $\mu_0 = \mu_1$ and $\sigma_0^2 = \sigma_1^2$, there is no MS left in the process.

Because in the single-regime case, the normality of the shocks z_{t+1} carries over to returns, it is sensible to ask what are the conditional and unconditional distributions of returns under the two-state MS process. Here the point is that even a simple two-state MSIH model such as the one in this question may generate substantial departures from normality. Given

$$R_{t+1} = S_{t+1}\mu_1 + (1-S_{t+1})\mu_0 + [S_{t+1}\sigma_1 + (1-S_{t+1})\sigma_0]z_{t+1} \qquad z_{t+1} \sim N(0,1),$$

in which $S_{t+1} = 0,\ 1$, it is clear that conditioning on S_{t+1}—which is equivalent to say that either the regime is observable (but that violates our assumptions) or that, again oddly, S_{t+1} may be perfectly predicted—$R_{t+1} \sim N(\mu_{S_{t+1}}, \sigma_{S_{t+1}}^2)$, which a simple Gaussian distribution. However, as we have stressed in the lectures, in a MS model, S_{t+1} is unobservable, while the case in which S_{t+1} may be perfectly predicted given time t information appears to be rather bizarre.[60] In fact, even if you were to somehow know what the current, time t regime S_t is, notice that in general $\Pr(S_{t+1} = j|S_t = i)$ represents the generic $[i,j]$ element of the transition matrix $\mathbf{P}$. If the Markov chain is ergodic and irreducible, you then know that $\Pr(S_{t+1} = j|S_t = i) < 1$, $i,j = 1,2$. Because of this fact the conditional distribution of R_{t+1} returns is:

$$f(R_{t+1}|\Im_t) = f(R_{t+1}|S_t) = \Pr(S_{t+1} = 1|S_t)\phi(\mu_1, \sigma_1^2) + (1 - \Pr(S_{t+1} = 1|S_t))\phi(\mu_0, \sigma_0^2),$$

where $\phi(\mu_1, \sigma_1^2)$ is a normal density function with mean μ_1 and variance σ_1^2. Such a density is called a *mixture*, with probabilistic and time-varying weights $\Pr(S_{t+1} = 1|S_t)$ and $(1 - \Pr(S_{t+1} = 1|S_t))$, of two normal densities and it is NOT itself a normal density. Therefore, even conditioning on time t information and on knowledge (still difficult to obtain) of the current state S_t, returns in a two-state MS will not have a normal distribution, unless $\mu_0 = \mu_1$ and $\sigma_0^2 = \sigma_1^2$, when (trivially)

$$\begin{aligned} f(R_{t+1}|\Im_t) &= f(R_{t+1}|S_t) = \Pr(S_{t+1} = 1|S_t)\phi(\mu, \sigma^2) + (1 - \Pr(S_{t+1} = 1|S_t))\phi(\mu, \sigma^2) \\ &= [\Pr(S_{t+1} = 1|S_t) + (1 - \Pr(S_{t+1} = 1|S_t))]\phi(\mu, \sigma^2) = \phi(\mu, \sigma^2). \end{aligned}$$

[60] Please read the previous footnote in case you have skipped it.

In fact, also notice that when $\mu_0 = \mu_1$ and $\sigma_0^2 = \sigma_1^2$, from results obtained above we have

$$Skewness_t[R_{t+1}] = (\mu - \mu)\frac{\xi_{1,t+1}(1\text{-}\xi_{1,t+1})\left\{(\mu-\mu)^2[(1\text{-}\xi_{1,t+1})^2 + \xi_{1,t+1}^2] + 3(\sigma^2-\sigma^2)\right\}}{[\xi_{1,t+1}\sigma^2 + (1-\xi_{1,t+1})\sigma^2 + \xi_{1,t+1}(1-\xi_{1,t+1})(\mu-\mu)^2]^{3/2}} = 0$$

$$\begin{aligned} Ex\ Kurt_t[R_{t+1}] = {} & \frac{\xi_{1,t+1}(1\text{-}\xi_{1,t+1})\left\{(\mu\text{-}\mu)^4[(1\text{-}\xi_{1,t+1})^3 + \xi_{1,t+1}^3] + 6(\mu\text{-}\mu)^2[(1\text{-}\xi_{1,t+1})\sigma^2 + \xi_{1,t+1}\sigma^2]\right\}}{[\xi_{1,t+1}\sigma^2 + (1-\xi_{1,t+1})\sigma^2 + \xi_{1,t+1}(1-\xi_{1,t+1})(\mu-\mu)^2]^2} \\ & + \frac{3\xi_{1,t+1}\sigma^4 + 3(1\text{-}\xi_{1,t+1})\sigma^4}{[\xi_{1,t+1}\sigma^2 + (1-\xi_{1,t+1})\sigma^2 + \xi_{1,t+1}(1-\xi_{1,t+1})(\mu-\mu)^2]^2} - 3 = 0, \end{aligned}$$

which is consistent with the conclusion that R_{t+1} follows a normal distribution.

As for the unconditional density of returns, i.e., the density of R_{t+1} not conditioning on any precise prior information, it is logical to state that absent any information on either S_t or at least $\Pr(S_t|\Im_t)$, the best assessment we can make of each of the regimes is simply that $\Pr(S_t = 1) = \pi_1$ and $\Pr(S_t = 0) = 1 - \pi_1$. Therefore, on average, returns will come π_1 percent of the time from $\phi(\mu_1, \sigma_1^2)$ and $(1 - \pi_1)$ percent of the time from $\phi(\mu_0, \sigma_0^2)$. The result is that the unconditional distribution of R_{t+1} is:

$$f(R_{t+1}) = \pi_1\phi(\mu_1, \sigma_1^2) + (1-\pi_1)\phi(\mu_0, \sigma_0^2),$$

which is another mixture (in this case, not time-varying, being unconditional) of two normal distributions and that, as we know, this will imply (assuming $\pi_1 \in (0, 1)$)

$$Skewness[R_{t+1}] = (\mu_1 - \mu_0)\frac{\pi_1(1\text{-}\pi_1)[(\mu_1-\mu_0)^2[(1-\pi_1)^2 + \pi_1^2] + 3(\sigma_1^2 - \sigma_0^2)]}{[\pi_1\sigma_1^2 + (1-\pi_1)\sigma_0^2 + \pi_1(1-\pi_1)(\mu_1-\mu_0)^2]^{3/2}} \neq 0$$

$$\begin{aligned} Ex\ Kurt[R_{t+1}] = {} & \frac{\pi_1(1\text{-}\pi_1)\left\{(\mu_1-\mu_0)^4[(1\text{-}\pi_1)^3 + \pi_1^3] + 6(\mu_1-\mu_0)^2[(1\text{-}\pi_1)\sigma_1^2 + \pi_1\sigma_0^2]\right\}}{[\xi_{1,t+1}\sigma_1^2 + (1-\xi_{1,t+1})\sigma_0^2 + \xi_{1,t+1}(1-\xi_{1,t+1})(\mu_1-\mu_0)^2]^2} + \\ & + \frac{3\pi_1\sigma_1^4 + 3(1-\pi_1)\sigma_0^4}{[\xi_{1,t+1}\sigma_1^2 + (1-\xi_{1,t+1})\sigma_0^2 + \xi_{1,t+1}(1-\xi_{1,t+1})(\mu_1-\mu_0)^2]^2} - 3 > 0 \end{aligned}$$

Additionally, when $\mu_0 \neq \mu_1$, notice that even the variance of $f(R_{t+1})$ fails to simply be the probability-weighted average of σ_1^2 and σ_0^2 because, as we know, $Var[R_{t+1}] = \pi_1\sigma_1^2 + (1-\pi_1)\sigma_0^2 + \pi_1(1-\pi_1)(\mu_1-\mu_0)^2$.

9. **Markov Switching and the Risk-Return Trade-Off**

Despite its key role in many applications, estimating and understanding the dynamics over time of the market risk premium has proven difficult. As you will recall from your theory of finance sequence, the market risk premium can be defined as the mean of market returns in excess of some risk-free rate, say $E[R_{t+1} - R^f]$.[61] For instance, even though classical finance theory suggests estimating the risk premium based on the theoretical relationship between mean returns and the contemporaneous variance of returns, for a long time empirical research has failed to document a significantly positive relationship between average returns and the filtered/predicted levels of market volatility (see e.g., Glosten, Jagannathan, and Runkle, 1993). In fact, a number of researchers have instead unveiled a

[61] Here, one may also stress a distinction between ex-ante and ex-post risk premia, with the ex-ante quantity being an expectation (i.e., population mean), and the ex-post one being an estimator of such a mean. One could also define a conditional risk premium, in obvious ways: $E[R_{t+1} - R^f|\Im_t]$.

negative relationship between volatility and market prices, the so-called *volatility feedback* effect. As already discussed in chapter 4 (where it was called leverage effect), this feedback effect refers to 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.

Because the aggregate stock market portfolio remains one of the most natural starting points to an understanding of asset pricing phenomena, it is surprising that there is still a good deal of controversy around the issue of how to measure risk at the market level. Recent empirical studies have documented two puzzling results. First, there is evidence of a weak, or even negative, relation between conditional mean returns and the conditional volatility of returns. Second, they document significant time variation in this relation. For instance, in a modified GARCH-in mean framework using post-World War II monthly data, Glosten et al. (1993) find that the estimated coefficient on volatility in a return/volatility regression is negative: a higher conditional volatility would depress the conditional risk premium, not the opposite. Or, equivalently, negative news that depress the risk premium, would increase conditional variance, which was already discussed in chapter 4.

More recently, Lettau and Ludvigson (2001) have provided evidence suggesting the failure to find a positive relationship between excess returns and market volatility may result from not controlling for shifts in investment opportunities, i.e., regimes. However, within applications of MS models in financial economics, this idea dates back at least to a seminal paper that had traced a connection between MS as a time series technique and asset pricing theory, Turner, Startz and Nelson (1989, henceforth TSN). TSN introduce a model of the aggregate market portfolio (the Standard and Poor's index) in which the *excess* return ($r_t \equiv R_t - R_t^f$) 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 itself is assumed to be generated by a first-order Markov process,

$$r_t = \mu_t + \epsilon_t \qquad \epsilon_t \; NID(0, \sigma^2_{S_t}),$$

where $\sigma_1^2 \geq \sigma_0^2$ and the conditional mean $\mu_t \equiv E[r_t|\Im_{t-1}]$ is discussed below. Of course this is an odd MSIH(2) model, in the sense that variance is MS in the usual way and the intercept varies according to some function that will also involve the Markov chain S_t. TSN develop two models based on the heteroskedastic structure discussed above. Each incorporates a different assumption about agents' information sets. In the first model, economic agents know (because they observe it) the realization of the Markov state process, even though the econometrician does not observe it. There are two risk premia 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 over the riskless rate 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:

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

The parameter estimates from this model suggest that whereas the first risk premium is positive, 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 $\sigma_1^2 \geq \sigma_0^2$ 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: parameter estimates indicate that agents require an increase in annual return over T-bills of approximately 10% to hold the risky asset in *low*-variance periods. The estimates also suggest, however, that the premium declines as the level of risk increases, that is, $\hat{\mu}_1 < \hat{\mu}_0$. Further, 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.

As we have seen in Section 4, Figure 1, this occurs also with reference to more recent, different data, such as those in Guidolin and Timmermann (2006a).

Parameter	Large caps	Small caps	Bonds
	Panel A: two-state AR(0) models		
μ_1	−0.0083	0.0045	0.0015
μ_2	0.0097	0.0109	−0.0012
σ_1	0.0641	0.0852	0.0246
σ_2	0.0335	0.0360	0.0070
p_{11}	0.7298	0.8910	0.9721
p_{22}	0.9424	0.9218	0.9196
Log-likelihood	996.3292	804.2038	1394.8273

Figure 1: MSIH(2,0) parameter estimates for U.S. stock and bond portfolios, monthly 1954-1999 U.S. data

Here the column that is relevant to this discussion is especially the first one, where the portfolio of large stocks (the top two annual deciles of the market capitalization distribution over time) is almost the same as TSN's S&P 500 index.

As already hinted at, misspecification is a likely explanation for this result. If agents are uncertain about the state, so that they are basing their decisions on forecasts of the regime 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 states. In each period, agents form probabilities of each possible state in the following period conditional on current and past excess returns, and use these probabilities in making their portfolio choices. Each period, investors update their prior beliefs about that period's state with current information using Bayes' rule, as in Section 5.1. 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. Agents' portfolio choice may be specified as a simple function of this probability:

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

where the constant, α, represents agents' required excess return for holding an asset in the low-variance state. Note that this is an intuitive and yet ad-hoc model: there is no reason for μ_t to depend linearly on the filtered probability of a high-variance state, $\Pr(S_t = 1|\mathcal{F}_{t-1})$. Yet, this simple model means that agents require an increase in the excess return in period t when faced with an increase in their prior probability that the high-variance state will prevail in that period, and this intuition is sufficiently sound for the model to represent a starting point. In fact, TSN generalize slightly this model to

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

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 for the added risk. 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 $\Pr(S_t = 1|\mathcal{F}_{t-1})$ represents the effect when agents anticipate as of time $t-1$ that the return of 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 predicted variance in this model is simply

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

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

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

is positive when $\Pr(S_t = 1|\mathcal{F}_{t-1}) < 0.5$, the expected excess return should be positive so that the parameter θ is positive. On the other hand, it could be that today's high-variance state, $S_t = 1$, was not anticipated in the previous period. In this case $\Pr(S_t = 1|\mathcal{F}_{t-1})$ 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 the event $S_t = 1$, 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 should be positive.

TSN also manage to establish the sign of a linear combination of the parameters. The risk premium in t is given by the expected value of r_t conditional on the current information set. Thus, the risk premium is

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

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

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

i.e., 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 next period's return will be drawn from the low-variance density, agents anticipate a monthly return of 5% percent. Likewise, if agents are certain next period's return will be drawn from the high-variance density, then agents will require a monthly return of 180% annually. 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 however only 0.0352. This means that in spite of that 180% spike in expectation during high-variance regimes, 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%.[62]

10. **Some Applications**

10.1. *Using MS models to study contagion*

MSVAR models are particularly suitable to model and study contagion dynamics. Contagion represents an important topic in empirical finance because studies concerning this phenomenon answer the question of whether it is possible to use performance in any market or country to forecast what will happen in other markets or countries. Typical questions are whether and how today's performance in the U.S. equity markets drive the performance in European markets in the subsequent period; or whether the current return in some ABS (asset-backed securities) market drives returns in other credit markets, or the corporate bond market. Typically, the literature has used simple, single-state ($K = 1$) VAR(p) models to model contagion. For instance, in the easiest case of a homoskedastic VAR(1) model

[62]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.

for $N = 2$ (say, the ABS market and the corporate bond market), one may try and estimate:[63]

$$\begin{aligned}
\begin{bmatrix} R_{t+1}^{ABS} \\ R_{t+1}^{CBond} \end{bmatrix} &= \boldsymbol{\mu} + \mathbf{A}\mathbf{R}_t + \boldsymbol{\Omega}^{1/2}\boldsymbol{\epsilon}_{t+1} = \begin{bmatrix} \mu^{ABS} \\ \mu^{CBond} \end{bmatrix} + \begin{bmatrix} a^{ABS} & a^{ABS,CBond} \\ a^{CBond,ABS} & a^{CBond} \end{bmatrix} \begin{bmatrix} R_t^{ABS} \\ R_t^{CBond} \end{bmatrix} + \\
&+ \begin{bmatrix} \sigma^{ABS} & 0 \\ \rho^{CBond,ABS}\sigma^{CBond} & \sqrt{1-(\rho^{CBond,ABS})^2}\sigma^{CBond} \end{bmatrix} \begin{bmatrix} \epsilon_{t+1}^{ABS} \\ \epsilon_{t+1}^{CBond} \end{bmatrix} \\
&= \begin{bmatrix} R_{t+1}^{ABS} = \mu^{ABS} + a^{ABS}R_t^{ABS} + a^{ABS,CBond}R_t^{CBond} + \sigma^{ABS}\epsilon_{t+1}^{ABS} \\ R_{t+1}^{CBond} = \mu^{CBond} + a^{CBond,ABS}R_t^{ABS} + a^{CBond}R_t^{CBond} + \rho^{CBond,ABS}\sigma^{CBond}\epsilon_{t+1}^{ABS} \\ + \sqrt{1-(\rho^{CBond,ABS})^2}\sigma^{CBond}\epsilon_{t+1}^{CBond} \end{bmatrix},
\end{aligned}$$

where $\boldsymbol{\Omega}^{1/2}$ is the Choleski decomposition of the return shock covariance. In such a VAR(1) model, contagion—say, from the ABS market to the corporate bond one—may occur in two ways: (i) because $a^{CBond,ABS} > 0$ and large, so that a large ABS return today affects subsequent, realized corporate bond returns; (ii) because $\rho^{CBond,ABS} > 0$ and large, so that a large shocks to ABS returns today contemporaneously affects realized corporate bond returns. Typically, the literature has found that such two channels are insufficient to capture the overall comovements of markets during periods of financial crisis.

One simple and yet intuitive extension of this simple VAR(1) framework, consists of linking some (or all) of its parameter vectors and matrices to the dynamics of a Markov chain state variable S_{t+1}:

$$\begin{aligned}
\begin{bmatrix} R_{t+1}^{ABS} \\ R_{t+1}^{CBond} \end{bmatrix} &= \boldsymbol{\mu}_{S_{t+1}} + \mathbf{A}_{S_{t+1}}\mathbf{R}_t + \boldsymbol{\Omega}_{S_{t+1}}^{1/2}\boldsymbol{\epsilon}_{t+1} = \begin{bmatrix} \mu_{S_{t+1}}^{ABS} \\ \mu_{S_{t+1}}^{CBond} \end{bmatrix} + \begin{bmatrix} a_{S_{t+1}}^{ABS} & a_{S_{t+1}}^{ABS,CBond} \\ a_{S_{t+1}}^{CBond,ABS} & a_{S_{t+1}}^{CBond} \end{bmatrix} \begin{bmatrix} R_t^{ABS} \\ R_t^{CBond} \end{bmatrix} \\
&+ \begin{bmatrix} \sigma_{S_{t+1}}^{ABS} & 0 \\ \rho_{S_{t+1}}^{CBond,ABS}\sigma_{S_{t+1}}^{CBond} & \sqrt{1-(\rho_{S_{t+1}}^{CBond,ABS})^2}\sigma_{S_{t+1}}^{CBond} \end{bmatrix} \begin{bmatrix} \epsilon_{t+1}^{ABS} \\ \epsilon_{t+1}^{CBond} \end{bmatrix} \\
&= \begin{bmatrix} R_{t+1}^{ABS} = \mu_{S_{t+1}}^{ABS} + a_{S_{t+1}}^{ABS}R_t^{ABS} + a_{S_{t+1}}^{ABS,CBond}R_t^{CBond} + \sigma_{S_{t+1}}^{ABS}\epsilon_{t+1}^{ABS} \\ R_{t+1}^{CBond} = \mu_{S_{t+1}}^{CBond} + a_{S_{t+1}}^{CBond,ABS}R_t^{ABS} + a_{S_{t+1}}^{CBond}R_t^{CBond} + \rho_{S_{t+1}}^{CBond,ABS}\sigma_{S_{t+1}}^{CBond}\epsilon_{t+1}^{ABS} \\ + \sqrt{1-(\rho_{S_{t+1}}^{CBond,ABS})^2}\sigma_{S_{t+1}}^{CBond}\epsilon_{t+1}^{CBond} \end{bmatrix},
\end{aligned}$$

where $S_{t+1} = 1, 2, ..., K$. This is of course a MSVARH(2,1) models for the bivariate vector that includes both ABS and corporate bond returns. Such a model implies three types of contagion effects:

1. *Simultaneous*, through the off-diagonal elements of $\boldsymbol{\Omega}_{S_{t+1}}^{1/2}$ that capture the dynamics across regimes of non-zero correlations.

2. Dynamic and *linear*, through the VAR components, which now become themselves potentially regime-switching.

3. Dynamic and *non-linear*, through the fact that the regime variable that drives the process of all variables in $\mathbf{R}_{t+1}$ is common to all variables, in this case both asset markets under consideration.

[63] You can easily think of conditional heteroskedastic extensions along the lines of chapter 6. However, this is not essential for the point we want to make here. Also, your lecture slides contain a similar but not identical example concerning the contagion from the U.S. to the Italian stock market.

The intuition is that if the vectors and matrices of parameters switch with similar effects on realized returns as S_{t+1} evolves over time, patterns of contagion that are neither linear nor exclusively related to the structure of the covariance matrix may be captured.

Interestingly, the third, MS-related contagion pattern may occur independently of all other patterns listed above. This means that one may have contagion even in the simple MSI(K) model:

$$\begin{aligned}\begin{bmatrix} R_{t+1}^{ABS} \\ R_{t+1}^{CBond} \end{bmatrix} &= \boldsymbol{\mu}_{S_{t+1}} + diag\{\sigma_{S_{t+1}}^{ABS}, \sigma_{S_{t+1}}^{CBond}\}\boldsymbol{\epsilon}_{t+1} \\ &= \begin{bmatrix} \mu_{S_{t+1}}^{ABS} \\ \mu_{S_{t+1}}^{CBond} \end{bmatrix} + \begin{bmatrix} \sigma^{ABS} & 0 \\ 0 & \sigma^{CBond} \end{bmatrix} \begin{bmatrix} \epsilon_{t+1}^{ABS} \\ \epsilon_{t+1}^{CBond} \end{bmatrix},\end{aligned}$$

provided that in the regimes in which $\mu_{S_{t+1}}^{ABS}$ is low, also $\mu_{S_{t+1}}^{CBond}$ and viceversa. Note that the previous example also stresses an implied capability of MS models: to capture and forecast time-varying variances and correlations, similarly to ARCH and DCC models.

There is one last form of contagion that has been explored in the finance literature: in our previous example, suppose that R_{t+1}^{CBond} is driven by a specific Markov state S_{t+1}^{CBond} and R_{t+1}^{ABS} by S_{t+1}^{ABS}. A form of interesting and testable contagion pattern is then whether:

$$S_{t+1}^{CBond} = S_t^{ABS},$$

i.e., whether the state in the corporate bond market at time $t+1$ is deterministically driven by the market state in the ABS market as of last period. The workout example in Appendix D shows one way in which this hypothesis can be formally tested.

10.2. *MS predictability*

MS models have had wide applications to the debate on the predictability of financial returns. Because MS models come in a variety of ways, in the following we consider a simple example to examine a few of the interesting issues that arise when you approach the quantitative modelling of predictability in a MS framework. Consider the (restricted) two state MSVARH(2,1) for US and Canadian stock returns in which lagged values of the US dividend yield predict stock returns in both markets, formally:

$$\mathbf{y}_{t+1} = \boldsymbol{\mu}_{S_{t+1}} + \mathbf{A}_{S_{t+1}}\mathbf{y}_t + \boldsymbol{\epsilon}_{t+1} \qquad \boldsymbol{\epsilon}_{t+1} \sim N(\mathbf{0}, \boldsymbol{\Omega}_{S_{t+1}}),$$

where S_t follows an ergodic, irreducible, first-order Markov chain with constant transition matrix $\mathbf{P}$ and $\mathbf{y}_t \equiv [R_t^{US}\ R_t^{Can}\ dy_t^{US}]'$. The MS VAR model is restricted because we assume that the US dividend yield is *not* directly (i.e., linearly, in a regression sense) affected by lagged values of stock returns in either the US or Canada. In explicit form, the model can be written as:

$$\begin{aligned} R_{t+1}^{US} &= \mu_{S_{t+1}}^{US} + a_{S_{t+1}}^{US,US} R_t^{US} + a_{S_{t+1}}^{US,Can} R_t^{Can} + a_{S_{t+1}}^{US,dy} dy_t^{US} + \epsilon_{t+1}^{US} \\ R_{t+1}^{Can} &= \mu_{S_{t+1}}^{Can} + a_{S_{t+1}}^{Can,US} R_t^{US} + a_{S_{t+1}}^{Can,Can} R_t^{Can} + a_{S_{t+1}}^{Can,dy} dy_t^{US} + \epsilon_{t+1}^{Can} \\ dy_{t+1}^{US} &= \mu_{S_{t+1}}^{dy} + a_{S_{t+1}}^{dy,dy} dy_t^{US} + \epsilon_{t+1}^{dy}. \end{aligned}$$

$$\mathbf{\Omega}_{S_{t+1}} = \begin{bmatrix} (\sigma^{US}_{S_{t+1}})^2 & \rho^{US,Can}_{S_{t+1}}\sigma^{US}_{S_{t+1}}\sigma^{Can}_{S_{t+1}} & \rho^{US,dy}_{S_{t+1}}\sigma^{US}_{S_{t+1}}\sigma^{dy}_{S_{t+1}} \\ \rho^{US,Can}_{S_{t+1}}\sigma^{US}_{S_{t+1}}\sigma^{Can}_{S_{t+1}} & (\sigma^{Can}_{S_{t+1}})^2 & \rho^{Can,dy}_{S_{t+1}}\sigma^{Can}_{S_{t+1}}\sigma^{dy}_{S_{t+1}} \\ \rho^{US,dy}_{S_{t+1}}\sigma^{US}_{S_{t+1}}\sigma^{dy}_{S_{t+1}} & \rho^{Can,dy}_{S_{t+1}}\sigma^{Can}_{S_{t+1}}\sigma^{dy}_{S_{t+1}} & (\sigma^{dy}_{S_{t+1}})^2 \end{bmatrix}.$$

The model implies a long list of estimable parameters: μ_1^{US}, μ_2^{US}, $a_1^{US,US}$, $a_2^{US,US}$, $a_1^{US,Can}$, $a_2^{US,Can}$, $a_1^{US,dy}$, $a_2^{US,dy}$, μ_1^{Can}, μ_2^{Can}, $a_1^{Can,US}$, $a_2^{Can,US}$, $a_1^{Can,Can}$, $a_2^{Can,Can}$, $a_1^{Can,dy}$, $a_2^{Can,dy}$, μ_1^{dy}, μ_2^{dy}, $a_1^{dy,dy}$, $a_2^{dy,dy}$, σ_1^{US}, σ_2^{US}, σ_1^{Can}, σ_2^{Can}, σ_1^{dy}, σ_2^{dy}, $\rho_1^{US,Can}$, $\rho_2^{US,Can}$, $\rho_1^{US,dy}$, $\rho_2^{US,dy}$, $\rho_1^{Can,dy}$, $\rho_2^{Can,dy}$, plus the two elements from the transition matrix:

$$\mathbf{P} = \begin{bmatrix} p_{11} & 1-p_{11} \\ 1-p_{11} & p_{22} \end{bmatrix}.$$

Notice that the elements are only two, because the rows of $\mathbf{P}$ need to sum to one. If you count them, this gives you a total of $2 \times 3 = 6$ parameters in the vector of intercepts, $\boldsymbol{\mu}_{S_{t+1}}$, $2 \times 7 = 14$ elements from the restricted VAR(1) matrix defined in the question,

$$\mathbf{A}_{S_{t+1}} = \begin{bmatrix} a^{US,US}_{S_{t+1}} & a^{US,Can}_{S_{t+1}} & a^{US,dy}_{S_{t+1}} \\ a^{Can,US}_{S_{t+1}} & a^{Can,Can}_{S_{t+1}} & a^{Can,dy}_{S_{t+1}} \\ 0 & 0 & a^{dy,dy}_{S_{t+1}} \end{bmatrix}$$

and $2 \times (3 \times 4)/2 = 12$ elements from the Markov switching covariance matrix. The total is $6 + 14 + 12 + 2 = 34$ parameters to be estimated, which may seem a lot but it is actually not, at least in a trivariate VAR-type model.

Suppose you somehow know—as normally this is not observable information, as we have seen in Sections 4 and 5—what regime will prevail at time $t+1$, call it s_{t+1} (this can be either 1 or 2). If we knew that next period the regime will be $S_{t+1} = s_{t+1}$, then forecasting returns (say, U.S. ones) one-period ahead is simple:

$$E[R^{US}_{t+1}|S_{t+1} = s_{t+1}] = \mu^{US}_{s_{t+1}} + a^{US,US}_{s_{t+1}} R^{US}_t + a^{US,Can}_{s_{t+1}} R^{Can}_t + a^{US,dy}_{s_{t+1}} dy^{US}_t,$$

as $E[\epsilon^{US}_{t+1}|S_{t+1} = s_{t+1}] = 0$. On the opposite, if the nature of the regime in $t+1$ were not known and unobservable, then:

$$\begin{aligned} E[R^{US}_{t+1}|S_t &= s_t] = \Pr(S_{t+1} = 1|S_t = s_t)\left[\mu^{US}_1 + a^{US,US}_1 R^{US}_t + a^{US,Can}_1 R^{Can}_t + a^{US,dy}_1 dy^{US}_t\right] + \\ &+ \Pr(S_{t+1} = 2|S_t = s_t)\left[\mu^{US}_2 + a^{US,US}_2 R^{US}_t + a^{US,Can}_2 R^{Can}_t + a^{US,dy}_2 dy^{US}_t\right]. \end{aligned}$$

It is easy to see that in general $E[R^{US}_{t+1}|S_{t+1} = s_{t+1}] \neq E[R^{US}_{t+1}|S_t = s_t]$. The exceptions may be summarized in two sets of conditions: (i) when there are no differences across regimes in conditional mean parameters, i.e., $\mu^{US}_1 = \mu^{US}_2$, $a^{US,US}_1 = a^{US,US}_2$, $a^{US,Can}_1 = a^{US,Can}_2$, $a^{US,dy}_1 = a^{US,dy}_2$; (ii) or when it happens that $\Pr(S_{t+1} = s_{t+1}|S_t = s_t) = 1$, i.e., from state the current $S_t = s_t$ one can only switch the state $S_{t+1} = s_{t+1}$, assuming this will occur. This shows that the unobservable nature of the regime becomes an essential and realistic feature of the practical use of MS models.

Another curiosity may concern the conditions under which you can state that Canadian stock returns do not depend in any *linear* fashion from US economic conditions, including both US stock

returns and dividend yields. Given

$$\begin{aligned} R_{t+1}^{Can} &= \mu_{S_{t+1}}^{Can} + a_{S_{t+1}}^{Can,US} R_t^{US} + a_{S_{t+1}}^{Can,Can} R_t^{Can} + a_{S_{t+1}}^{Can,dy} dy_t^{US} + \epsilon_{t+1}^{Can} \\ \begin{bmatrix} \epsilon_{t+1}^{US} \\ \epsilon_{t+1}^{Can} \\ \epsilon_{t+1}^{dy} \end{bmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} (\sigma_{S_{t+1}}^{US})^2 & \rho_{S_{t+1}}^{US,Can}\sigma_{S_{t+1}}^{US}\sigma_{S_{t+1}}^{Can} & \rho_{S_{t+1}}^{US,dy}\sigma_{S_{t+1}}^{US}\sigma_{S_{t+1}}^{dy} \\ \rho_{S_{t+1}}^{US,Can}\sigma_{S_{t+1}}^{US}\sigma_{S_{t+1}}^{Can} & (\sigma_{S_{t+1}}^{Can})^2 & \rho_{S_{t+1}}^{Can,dy}\sigma_{S_{t+1}}^{Can}\sigma_{S_{t+1}}^{dy} \\ \rho_{S_{t+1}}^{US,dy}\sigma_{S_{t+1}}^{US}\sigma_{S_{t+1}}^{dy} & \rho_{S_{t+1}}^{Can,dy}\sigma_{S_{t+1}}^{Can}\sigma_{S_{t+1}}^{dy} & (\sigma_{S_{t+1}}^{dy})^2 \end{bmatrix} \right) \end{aligned}$$

it is clear that you will need $a_1^{Can,US} = a_2^{Can,US} = 0$, $a_1^{Can,dy} = a_2^{Can,dy} = 0$, $\rho_1^{US,Can} = \rho_2^{US,Can} = 0$, and $\rho_1^{Can,dy} = \rho_2^{Can,dy} = 0$ from Canadian markets not to depend in any way on U.S. economic conditions. The last two sets of restrictions imply that shocks to either US stock markets or to the US dividend yield will fail to correlate with shocks to Canadian stock returns. Under these restrictions, the model clearly simplifies to:

$$\begin{aligned} R_{t+1}^{Can} &= \mu_{S_{t+1}}^{Can} + a_{S_{t+1}}^{Can,Can} R_t^{Can} + \epsilon_{t+1}^{Can} \\ \begin{bmatrix} \epsilon_{t+1}^{US} \\ \epsilon_{t+1}^{Can} \\ \epsilon_{t+1}^{dy} \end{bmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} (\sigma_{S_{t+1}}^{US})^2 & 0 & \rho_{S_{t+1}}^{US,dy}\sigma_{S_{t+1}}^{US}\sigma_{S_{t+1}}^{dy} \\ 0 & (\sigma_{S_{t+1}}^{Can})^2 & 0 \\ \rho_{S_{t+1}}^{US,dy}\sigma_{S_{t+1}}^{US}\sigma_{S_{t+1}}^{dy} & 0 & (\sigma_{S_{t+1}}^{dy})^2 \end{bmatrix} \right), \end{aligned}$$

and the Canadian stock market would appear to be completely isolated from US markets. However, note that because one single, common Markov state S_t drives both US and Canadian markets, some dependence would be left in this model. For instance, suppose that $\mu_1^{Can} < \mu_2^{Can}$ and $\mu_1^{US} < \mu_2^{US}$. Then it is natural to expect that low Canadian returns will tend to appear simultaneously with low US returns, impressing a positive correlation pattern that may derive from the existence of effects from US economic conditions onto Canadian equity valuations.[64]

10.3. *Value-at-Risk in (simple) MS models*

This subsection summarizes work in Guidolin and Timmermann (2006b) on the effectiveness and benefits of MS modelling in risk management applications. Consider the simple univariate MSIH(2, 0) model,

$$R_{t+1} = S_{t+1}\mu_1 + (1 - S_{t+1})\mu_0 + [S_{t+1}\sigma_1 + (1 - S_{t+1})\sigma_0]z_{t+1} \qquad z_{t+1} \sim N(0, 1),$$

in which $S_{t+1} = 0, 1$ is *unobservable* at all points in time. As we know from Section 8, this way of writing a MSIH(2,0) model is equivalent to (6) when $K = 2$. As usual, in the limit case in which $K = 1$, which is a benchmark single-state linear model, to compute (say) 1% VaR is identical to what has been done in chapter 5 and that has appeared a few times already in your lecture slides. Because when $K = 1$ the model collapses to a simple $R_{t+1} = \mu + \sigma z_{t+1}$ with $z_{t+1} \sim N(0, 1)$, it is straightforward

[64] However, no causality may be established. Although this would be hard to persuasively argue, such a positive correlation might even derive from effects from Canadian economic conditions on US stock valuations.

to obtain that:

$$\begin{aligned} 0.01 &= \Pr(R_{t+1} < -VaR_{t+1}^{0.01}(k=1)) = \Pr\left(\frac{R_{t+1}-\mu}{\sigma} < -\frac{VaR_{t+1}^{0.01}(k=1)+\mu}{\sigma}\right) \\ &= \Pr\left(z_{t+1} < -\frac{VaR_{t+1}^{0.01}(k=1)+\mu}{\sigma}\right) = \Phi\left(-\frac{VaR_{t+1}^{0.01}(k=1)+\mu}{\sigma}\right) \end{aligned}$$

so that, after defining $\Phi^{-1}(\cdot)$ as the inverse CDF of a standard normal distribution,

$$\begin{aligned} \Phi^{-1}(0.01) &= \Phi^{-1}\left(\Phi\left(-\frac{VaR_{t+1}^{0.01}(k=1)+\mu}{\sigma}\right)\right) = -\frac{VaR_{t+1}^{0.01}(k=1)+\mu}{\sigma} \\ &\Longrightarrow \quad VaR_{t+1}^{0.01}(k=1) = -\sigma\Phi^{-1}(0.01) - \mu. \end{aligned}$$

Now, moving to the $K = 2$ case, let's start from an approximate way to look at the problem of computing 1% VaR: one colleague in your risk management department is proposing to use the following *conditional* 1% VaR measure:

$$\begin{aligned} VaR_{t+1}^{0.01}(k &= 2) = -\left[\Pr(S_{t+1}=1|S_t)\sigma_1^2 + (1-\Pr(S_{t+1}=1|S_t))\sigma_0^2\right]\Phi^{-1}(0.01) + \\ &\quad -\left[\Pr(S_{t+1}=1|S_t)\mu_1 + (1-\Pr(S_{t+1}=1|S_t))\mu_0\right], \end{aligned}$$

in which the colleague is obviously conditioning with respect to the current state, S_t, but still applying a normal distribution result. Unfortunately, you should not agree with his/her proposal, or at least should clarify to the team that this is simply an approximation. The reason is that in Section 8 we have found that

$$\begin{aligned} f(R_{t+1}|\Im_t) &= f(R_{t+1}|S_t) = \Pr(S_{t+1}=1|S_t)\phi(\mu_1,\sigma_1^2) + (1-\Pr(S_{t+1}=1|S_t))\phi(\mu_0,\sigma_0^2) \\ &\neq \phi\left(\left[\Pr(S_{t+1}{=}1|S_t)\mu_1 + (1\text{-}\Pr(S_{t+1}{=}1|S_t))\mu_0\right], \Pr(S_{t+1}{=}1|S_t)\sigma_1^2 + (1\text{-}\Pr(S_{t+1}{=}1|S_t))\sigma_0^2\right), \end{aligned}$$

and that $f(R_{t+1}|\Im_t)$ does not actually follow a Normal distribution, but a probability-weighted mixture of two normal distributions which is itself not a Normal distribution. As a result, the way of proceeding to VaR calculations proposed by the colleague may turn out to be grossly incorrect as it employs $\Phi^{-1}(0.01)$, where the use of the standard normal CDF was previously coming from the fact that $R_{t+1} \sim N(\mu, \sigma^2)$. When this assumption breaks down, the procedure is clearly invalid. Moreover, you know from Section 8 that

$$\begin{aligned} Var_t[R_{t+1}] &= \Pr(S_{t+1}{=}1|S_t)\sigma_1^2 + (1\text{-}\Pr(S_{t+1}{=}1|S_t))\sigma_0^2 + \Pr(S_{t+1}{=}1|S_t)(1\text{-}\Pr(S_{t+1}{=}1|S_t))(\mu_1-\mu_0)^2 \\ &\neq \Pr(S_{t+1}=1|S_t)\sigma_1^2 + (1-\Pr(S_{t+1}=1|S_t))\sigma_0^2 \end{aligned}$$

unless $\mu_0 = \mu_1$, which is generally not the case in a MSIH(2,0) model.

After you have made your objection during his presentation, this colleague of yours revises his/her proposal to use the following *conditional* 1% VaR measure:

$$\begin{aligned} VaR_{t+1}^{0.01}(k &= 2) = -[\Pr(S_{t+1}{=}1|S_t)\sigma_1^2 + (1\text{-}\Pr(S_{t+1}=1|S_t))\sigma_0^2 + \Pr(S_{t+1}{=}1|S_t)(1\text{-}\Pr(S_{t+1}{=}1|S_t)) \times \\ &\quad \times(\mu_1-\mu_0)^2]\Phi^{-1}(0.01) - \left[\Pr(S_{t+1}=1|S_t)\mu_1 + (1-\Pr(S_{t+1}=1|S_t))\mu_0\right]. \end{aligned}$$

Your reaction should remain negative: unfortunately, making one claim "less wrong" does not make it correct. Even though it is now correct that

$$\begin{aligned} Var_t[R_{t+1}] &= \Pr(S_{t+1}{=}1|S_t)\sigma_1^2 + (1-\Pr(S_{t+1}{=}1|S_t))\sigma_0^2 + \Pr(S_{t+1}{=}1|S_t)(1-\Pr(S_{t+1}{=}1|S_t)) \times \\ &\quad \times(\mu_1-\mu_0)^2, \end{aligned}$$

the fact remains that

$$\begin{aligned} f(R_{t+1}|\Im_t) = f(R_{t+1}|S_t) &= \Pr(S_{t+1}=1|S_t)\phi(\mu_1,\sigma_1^2) + (1-\Pr(S_{t+1}=1|S_t))\phi(\mu_0,\sigma_0^2) \\ &\neq \phi([\Pr(S_{t+1}{=}1|S_t)\mu_1 + (1\text{-}\Pr(S_{t+1}{=}1|S_t))\mu_0], \\ &\left[\Pr(S_{t+1}{=}1|S_t)\sigma_1^2 + (1\text{-}\Pr(S_{t+1}{=}1|S_t))\sigma_0^2\right] + \Pr(S_{t+1}{=}1|S_t)(1\text{-}\Pr(S_{t+1}{=}1|S_t))(\mu_1-\mu_0)^2) \end{aligned}$$

so that VaR cannot be computed in that simply way.

Finally, it seems time for you to suggest how this should be done correctly. Here you may be in trouble, though: unfortunately there is no closed-form solution which means that you will have to resort to simulation-based (Monte Carlo) methods. The problem is that

$$f(R_{t+1}|S_t) = \Pr(S_{t+1}=1|S_t)\phi(\mu_1,\sigma_1^2) + (1-\Pr(S_{t+1}=1|S_t))\phi(\mu_0,\sigma_0^2)$$

fails to have a closed-form representation and as such it impossible to simply draw from some well-specified PDF or CDF. This means that your proof of the functional form of 1% VaR in

$$\begin{aligned} 0.01 &= \Pr(R_{t+1} < -VaR_{t+1}^{0.01}(k=1)) = \Pr\left(\frac{R_{t+1}-\mu}{\sigma} < -\frac{VaR_{t+1}^{0.01}(k=1)+\mu}{\sigma}\right) \\ &= \Pr\left(z_{t+1} < -\frac{VaR_{t+1}^{0.01}(k=1)+\mu}{\sigma}\right) = \Phi\left(-\frac{VaR_{t+1}^{0.01}(k=1)+\mu}{\sigma}\right) \end{aligned}$$

simply fails because it is not true that $\Pr\left(z_{t+1} < -\frac{VaR_{t+1}^{0.01}(k=1)+\mu}{\sigma}\right)$ can be measured using $\Phi(\cdot)$. What you can do is (very simply, indeed) the following. First, simulate a large number M of one-month returns assuming $S_t = 0$ from

$$R_{t+1} = S_{t+1}\mu_1 + (1-S_{t+1})\mu_0 + [S_{t+1}\sigma_1 + (1-S_{t+1})\sigma_0]z_{t+1} \qquad z_{t+1} \sim N(0,1),$$

when $S_{t+1} = 1$ with probability $p_{01} = (1-p_{00})$ and $S_{t+1} = 0$ with probability p_{00}. Call these M one-month ahead returns $\left\{R_{t+1}^m(S_t=0)\right\}_{m=1}^M$.[65] Second, simulate a large number M of one-month returns assuming $S_t = 1$ from

$$R_{t+1} = S_{t+1}\mu_1 + (1-S_{t+1})\mu_0 + [S_{t+1}\sigma_1 + (1-S_{t+1})\sigma_0]z_{t+1} \qquad z_{t+1} \sim N(0,1),$$

when $S_{t+1} = 1$ with probability p_{11} and $S_{t+1} = 0$ with probability $1-p_{11}$. Call these M one-month ahead returns $\left\{R_{t+1}^m(S_t=1)\right\}_{m=1}^M$. Finally, you need to "aggregate" this $2M$ simulations in a unique

[65] This means that when $S_{t+1} = 1$ you will simulate from $R_{t+1} = \mu_1 + \sigma_1 z_{t+1}$; when $S_{t+1} = 0$ you will simulate from $R_{t+1} = \mu_0 + \sigma_0 z_{t+1}$. How do you simulate a two-point (also called Bernoulli) random variable that takes value 1 with probability $1-p_{00}$ and 0 with probability p_{00}? Simple, you draw a uniform defined on [0,1] and you set $S_{t+1} = 1$ if the uniform draw is less than (or equal to) $1-p_{00}$, and you set $S_{t+1} = 0$ otherwise.

set, using:

$$R_{t+1}^m = \Pr(S_t = 1|\Im_t)R_{t+1}^m(S_t = 1) + (1 - \Pr(S_t = 1|\Im_t))R_{t+1}^m(S_t = 0) \quad m = 1, 2, ..., M.$$

At this point, your 1% VaR will be simply defined as: the simulated returns in the set $\left\{R_{t+1}^m\right\}_{m=1}^M$ that leaves exactly 1% of your total M simulations (after your aggregation step, i.e., $M/100$ simulations, which better be an integer) *below* the 1% VaR value.

Appendix A — More on Ergodic Markov Chains

Consider a K-state, first-order Markov chain (MC) with transition matrix with generic element $p_{ij} \equiv \Pr(S_{t+1} = j|S_t = i)$:

$$\mathbf{P} \equiv \begin{bmatrix} p_{11} & p_{12} & ... & p_{1K} \\ p_{21} & p_{22} & ... & p_{2K} \\ \vdots & ... & \ddots & \vdots \\ p_{K1} & p_{K2} & ... & p_{KK} \end{bmatrix}.$$

Notice that in general $\mathbf{P} \neq \mathbf{P'}$ although many of the claims that follow refer to $\mathbf{P'}$.[66] Suppose that one of the eigenvalues of $\mathbf{P'}$ is unity and that all other eigenvalues of $\mathbf{P'}$ are inside the unit circle (i.e., they are less than 1).[67] Then the MC is said to be ergodic and the $K \times 1$ vector of ergodic probabilities for the chain is denoted as $\bar{\boldsymbol{\xi}}$. This vector $\bar{\boldsymbol{\xi}}$ is defined as the eigenvector of $\mathbf{P'}$ associated with the unit eigenvalue, that is, the vector of ergodic probabilities $\bar{\boldsymbol{\xi}}$ satisfies $\mathbf{P'}\bar{\boldsymbol{\xi}} = \bar{\boldsymbol{\xi}}$ and it is normalized to sum to unity, (i.e., $\bar{\boldsymbol{\xi}}'\boldsymbol{\iota}_K = 1$, where $\boldsymbol{\iota}_K \equiv [1\ 1\ ...\ 1]'$ is a $K \times 1$ vector of ones).

First, noting that the eigenvalues of $\mathbf{P}$ and $\mathbf{P'}$ are identical by construction and using the standard properties of a (discrete) probability law (measure), it is easy to prove that $\mathbf{P}\boldsymbol{\iota}_K = \boldsymbol{\iota}_K$ so that at least one eigenvalue of $\mathbf{P'}$ is equal to one:

$$\mathbf{P}\boldsymbol{\iota}_k = \begin{bmatrix} p_{11} & p_{12} & ... & p_{1K} \\ p_{21} & p_{22} & ... & p_{2K} \\ \vdots & ... & \ddots & \vdots \\ p_{K1} & p_{K2} & ... & p_{KK} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} p_{11} + p_{12} + ... + p_{1K} \\ p_{21} + p_{22} + ... + p_{2K} \\ \vdots \\ p_{K1} + p_{K2} + ... + p_{KK} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

where the last equality derives from the law of total probability, i.e., the fact that starting from any state $S_t = i$, the sum of the probabilities of either staying in regime i or of switching to any other regime must always be 1:

$$\Pr(S_{t+1} = 1|S_t = i) + \Pr(S_{t+1} = 2|S_t = i) + ... + \Pr(S_{t+1} = K|S_t = i) = 1.$$

Recall now from your math courses that the expression $\mathbf{P}\boldsymbol{\iota}_K = \boldsymbol{\iota}_K$ is equivalent to the definition of one specific set of eigenvector/eigenvalue of a matrix $\mathbf{P}$, in the sense that $\mathbf{P}\boldsymbol{\iota}_K = \boldsymbol{\iota}_K$ identifies 1 as one of

[66] $\mathbf{P'}$ is in fact playing the role of matrix of vector autoregressive coefficients in the $\boldsymbol{\delta}_{t+1} = \mathbf{P'}\boldsymbol{\delta}_t + \mathbf{v}_{t+1}$ process seen in Section 4.

[67] The eigenvalues of the transition matrix $\mathbf{P'}$ for any K-state Markov chain are found from the solutions to $|\mathbf{P'} - \lambda\mathbf{I}_K|$ $= \det(\mathbf{P'} - \lambda\mathbf{I}_K) = \mathbf{0}$.

the eigenvalues of $\mathbf{P}$. Also notice that if $\mathbf{P}\boldsymbol{\iota}_K = \boldsymbol{\iota}_K$ holds, then also $\zeta\mathbf{P}\boldsymbol{\iota}_K = \zeta\boldsymbol{\iota}_K$, with $\zeta \in \mathcal{R}$ some scalar, which means that ζ will be an eigenvalue of $\mathbf{P}$ as well.

At this point, if $\mathbf{P}$ is the transition matrix for an ergodic Markov chain with K distinct eigenvalues, then

$$\lim_{T\to\infty}(\mathbf{P}')^T = \bar{\boldsymbol{\xi}}\boldsymbol{\iota}_K' = \begin{bmatrix} \xi_1 & \xi_1 & ... & \xi_1 \\ \xi_2 & \xi_2 & ... & \xi_2 \\ \vdots & ... & \ddots & \vdots \\ \xi_K & \xi_K & ... & \xi_K \end{bmatrix},$$

where $(\mathbf{P}')^T$ is the matrix $\mathbf{P}'$ multiplied by itself T times, i.e., $(\mathbf{P}')^T \equiv \prod_{\tau=1}^{T}\mathbf{P}'$. Recall that when the K eigenvalues are distinct, $\mathbf{P}'$ can always be written in the form $\mathbf{P}' = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1}$ where $\mathbf{Q}$ is a $K \times K$ matrix whose columns are the eigenvectors of $\mathbf{P}'$ and $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal contains the corresponding eigenvalues of $\mathbf{P}'$, sorted in descending order (so 1 will occupy the (1,1) position). It is elementary (try it with a $K = 2$ example) to show that

$$(\mathbf{P}')^T = \mathbf{Q}\boldsymbol{\Lambda}^T\mathbf{Q}^{-1}.$$

Since the (1,1) element of $\boldsymbol{\Lambda}$ is unity and all other elements of $\boldsymbol{\Lambda}$ are inside the unit circle, $\boldsymbol{\Lambda}^T$ converges to a matrix with unity in the (1, 1) position and zeros elsewhere. For instance

$$\lim_{T\to\infty}\begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & \lambda_2^T < 1 & ... & 0 \\ \vdots & ... & \ddots & \vdots \\ 0 & 0 & ... & \lambda_k^T < 1 \end{bmatrix} = \begin{bmatrix} \lim_{T\to\infty} 1 & 0 & ... & 0 \\ 0 & \lim_{T\to\infty}\lambda_2^T & ... & 0 \\ \vdots & ... & \ddots & \vdots \\ 0 & 0 & ... & \lim_{T\to\infty}\lambda_k^T \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & 0 & ... & 0 \\ \vdots & ... & \ddots & \vdots \\ 0 & 0 & ... & 0 \end{bmatrix}.$$

Hence,

$$\lim_{T\to\infty}(\mathbf{P}')^T = \mathbf{x}\mathbf{y}'$$

where $\mathbf{x}$ is the first column of $\mathbf{Q}$ and $\mathbf{y}'$ is the first row of $\mathbf{Q}^{-1}$. The first column of $\mathbf{Q}$ is the eigenvector of $\mathbf{P}'$ corresponding to the unit eigenvalue, which eigenvector was defined as $\bar{\boldsymbol{\xi}}$ in $\mathbf{P}'\bar{\boldsymbol{\xi}} = \bar{\boldsymbol{\xi}}$, so $\mathbf{x} = \bar{\boldsymbol{\xi}}$.[68] Moreover, the first row of $\mathbf{Q}^{-1}$, when expressed as a column vector, corresponds to the eigenvector of $\mathbf{P}$ associated with the unit eigenvalue, which eigenvector was seen to be *proportional* to the vector 1 in 1a, $\zeta\mathbf{P}\boldsymbol{\iota}_k = \zeta\boldsymbol{\iota}_k$, with ζ some scalar. Therefore $\mathbf{y} = \zeta\boldsymbol{\iota}_k$. At this point, substituting $\mathbf{x} = \bar{\boldsymbol{\xi}}$ and $\mathbf{y} = \zeta\boldsymbol{\iota}_k$ into the limit expression for $(\mathbf{P}')^T$ as $T \to \infty$, we have:

$$\lim_{T\to\infty}(\mathbf{P}')^T = \zeta\bar{\boldsymbol{\xi}}\boldsymbol{\iota}_k'.$$

[68] Here we have used without proof the fact that the first row of $\mathbf{Q}^{-1}$, when expressed as a column vector, corresponds to the eigenvector of $\mathbf{P}$ associated with the unit eigenvalue.

Because $(\mathbf{P}')^T$ can be interpreted as a matrix of (predicted) transition probabilities, each column must sum to unity. Thus, since the vector of ergodic probabilities $\bar{\boldsymbol{\xi}}$ was normalized by the condition that $\bar{\boldsymbol{\xi}}'\boldsymbol{\iota}_k = 1$, it follows that the normalizing -constant ζ must be unity, establishing that

$$\lim_{T\to\infty}(\mathbf{P}')^T = \bar{\boldsymbol{\xi}}\boldsymbol{\iota}_k'.$$

This means that as the forecast horizon for predicted transition probabilities T diverges, all the elements of the resulting T-step ahead transition matrix $\mathbf{P}^T$ will simply collapse to be identical to the ergodic, unconditional probabilities.

For instance, in the special case of $K = 3$, if you have obtained an estimate of $\mathbf{P}$ equal to

$$\hat{\mathbf{P}} = \begin{bmatrix} 0.88 & 0.09 & 0.03 \\ 0.01 & 0.96 & 0.03 \\ 0.23 & 0 & 0.77 \end{bmatrix},$$

the ergodic probabilities $\bar{\boldsymbol{\xi}}$ characterizing this three-state model can be derived resorting to a computer (just type "eigenvalues" in the Matlab on-line guide). It turns out that both $\hat{\mathbf{P}}$ and $\hat{\mathbf{P}}'$ share the same eigenvalues, i.e., 1, 0.87 and 0.74. Here we care only for the unit eigenvalue. Your math software will also inform you that the eigenvector of $\hat{\mathbf{P}}'$ associated to the unit eigenvalue is:

$$[0.3926\ 0.8834\ 0.1664]'$$

This eigenvector is not yet $\bar{\boldsymbol{\xi}}$ because it fails to have unit length. In fact the eigenvector ends up summing to 1.4424 while $\bar{\boldsymbol{\xi}}'\boldsymbol{\iota}_K = 1$ by definition. However, it is now sufficient to scale the eigenvector so to have unit length, which is done by simply dividing its entries by their sum, 1.4424. The resulting estimated (because implied by $\hat{\mathbf{P}}$) $\bar{\boldsymbol{\xi}}$ is:

$$\bar{\boldsymbol{\xi}} = [0.272\ 0.613\ 0.125]'.$$

Finally, the vector of ergodic probabilities can also be viewed as indicating the unconditional probability of each of the K different states, $\boldsymbol{\pi} = \bar{\boldsymbol{\xi}}$. We have seen in Section 4 that if we define $\boldsymbol{\delta}_{t+1}$ to be a $K \times 1$ vector that lists a 1 in its jth position if the MC is in state j at time t and 0 otherwise, then $E_t[\boldsymbol{\delta}_{t+1}] = E[\boldsymbol{\delta}_{t+1}|S_t] = \Pr(S_{t+1}|S_t)$ and will equal the ith column of the matrix $\mathbf{P}'$ if $S_t = i$. This is the vector of conditional probabilities of all possible K states, given $S_t = i$. Correspondingly, the unconditional probabilities of each of the K regimes may be defined as a vector $\Pr(S_{t+1})$:

$$\begin{aligned} E[\boldsymbol{\delta}_{t+1}] &= \Pr(S_{t+1}) = E[\mathbf{P}'\boldsymbol{\delta}_t + \mathbf{v}_{t+1}] \\ &= \mathbf{P}'\Pr(S_{t+1}) + E[\mathbf{v}_{t+1}] = \mathbf{P}'\Pr(S_{t+1}). \end{aligned}$$

Then it is clear that the vector $\Pr(S_{t+1})$ satisfies $\Pr(S_{t+1}) = \mathbf{P}'\Pr(S_{t+1})$. At this point, please compare with the definition of ergodic probabilities $\bar{\boldsymbol{\xi}} = \mathbf{P}'\bar{\boldsymbol{\xi}}$: clearly $\bar{\boldsymbol{\xi}} = \Pr(S_{t+1})$ so that $\bar{\boldsymbol{\xi}}$ can also be interpreted as the vector of long-run, unconditional probabilities for each of the K regimes. Alternatively, as seen in Section 4, because

$$\lim_{T\to\infty}(\mathbf{P}')^T = \boldsymbol{\pi}\boldsymbol{\iota}_K'.$$

and $\Pr(S_t) \equiv E[\boldsymbol{\delta}_t] = \lim_{T\to\infty} E[\boldsymbol{\delta}_{t+T}|S_t] = \lim_{T\to\infty}(\mathbf{P}')^T\boldsymbol{\delta}_t$, then

$$\Pr(S_t) = \boldsymbol{\pi}\boldsymbol{\iota}_K'\boldsymbol{\delta}_t = \boldsymbol{\pi}$$

as by construction $\boldsymbol{\iota}_K'\boldsymbol{\delta}_t = 1$. As a result, $\bar{\boldsymbol{\xi}} = [0.272\ 0.613\ 0.125]$ will also give the long-run, unconditional frequencies of the bear, normal, and bull phases of the market. As one would expect, the normal regime occurs on the majority of time, in excess of 60% of any long sample. The finding above that $\Pr(S) = \mathbf{P}'\Pr(S)$ extends more generally to show that

$$\begin{aligned} E_t[\boldsymbol{\delta}_{t+1}] &= \boldsymbol{\xi}_{t+1|t} = \Pr(S_{t+1}|\Im_t) = E[\mathbf{P}'\boldsymbol{\delta}_t + \mathbf{v}_{t+1}|\Im_t] \\ &= \mathbf{P}'\Pr(S_t|\Im_t) + E[\mathbf{v}_{t+1}|\Im_t] = \mathbf{P}'\boldsymbol{\delta}_t. \end{aligned}$$

Moreover, but to show it is a bit tedious, also the recursion $\boldsymbol{\xi}_{t+1|t} = \mathbf{P}\boldsymbol{\xi}_t$ holds.

Appendix B — State-Space Representation of a MS Model

The first step towards estimation and prediction of a MSVARH model is to put the model in state-space form. This Appendix offers a heuristic idea of what that means but it is in no way binding for the purposes of your exam preparation. Let's collect the information on the time t realization of the Markov chain in a random vector

$$\boldsymbol{\xi}_t \equiv \begin{bmatrix} I(S_t = 1) \\ I(S_t = 2) \\ \vdots \\ I(S_t = K) \end{bmatrix},$$

where $I(S_t = i)$ is a standard indicator variable. In practice the sample realizations of $\boldsymbol{\xi}_t$ will always consist of unit versors $\mathbf{e}_i$ characterized by a 1 in the ith position and by zero everywhere else. As we have seen in Section 4, another important property is that $E[\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t-1}] = \mathbf{P}'\boldsymbol{\xi}_{t-1}$. The state-space form is composed of two equations:

$$\begin{aligned} \mathbf{R}_t &= \mathbf{X}_t\mathbf{A}\left(\boldsymbol{\xi}_t \otimes \boldsymbol{\iota}_N\right) + \boldsymbol{\Sigma}_K\left(\boldsymbol{\xi}_t \otimes \mathbf{I}_N\right)\boldsymbol{\epsilon}_t \qquad \text{(measurement equation)} \\ \boldsymbol{\xi}_{t+1} &= \mathbf{F}\boldsymbol{\xi}_t + \mathbf{v}_{t+1} \qquad \text{(transition equation)} \end{aligned} \tag{26}$$

where $\mathbf{X}_t$ is a $N \times (Np+1)$ matrix of predetermined variables with structure $[1\ \mathbf{R}_{t-1}'...\mathbf{R}_{t-p}'] \otimes \boldsymbol{\iota}_n$, $\mathbf{A}$ is a $(Np+1) \times NK$ matrix collecting the VAR parameters, both means or intercepts and autoregressive coefficients, in all regimes

$$\mathbf{A} = \begin{bmatrix} \boldsymbol{\mu}_1' & \boldsymbol{\mu}_2' & \cdots & \boldsymbol{\mu}_K' \\ \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{p1} & \mathbf{A}_{p2} & \cdots & \mathbf{A}_{pK} \end{bmatrix},$$

$\boldsymbol{\Sigma}_K$ is a $N \times NK$ matrix collecting all the possible K "square root" (Choleski decomposition) covariance matrix factors $[\Omega_1^{1/2}\ \Omega_2^{1/2}\ ...\ \Omega_K^{1/2}]$ such that $\forall t$, $\boldsymbol{\Sigma}_K\left(\boldsymbol{\xi}_t \otimes \mathbf{I}_N\right)\left(\boldsymbol{\xi}_t \otimes \mathbf{I}_N\right)'\boldsymbol{\Sigma}_K' = \Omega_{S_t}$, the covariance

matrix of the asset return innovations $\boldsymbol{\epsilon}_t$. Moreover, $\boldsymbol{\epsilon}_t \sim$IID $N(\mathbf{0}, \mathbf{I}_N)$, and in the transition equation $\mathbf{v}_{t+1}$ is a zero-mean discrete random vector that can be shown to be a martingale difference sequence. Also, the elements of $\mathbf{v}_{t+1}$ are uncorrelated with $\boldsymbol{\epsilon}_{t+1}$ as well as $\boldsymbol{\xi}_{t-j}$, $\boldsymbol{\epsilon}_{t-j}$, $\mathbf{R}_{t-j}$, and $\mathbf{X}_{t-j}$ $\forall j \geq 0$. To operazionalize the dynamics state-space system (26), assume that the multivariate process (6) started with a random draw from the unconditional probability distribution $\bar{\boldsymbol{\xi}}$. Finally, from the definition of transition probability matrix in Section 3, it follows that since $E[\mathbf{v}_{t+1}|\boldsymbol{\xi}_t] = \mathbf{0}$ by assumption, then

$$E[\boldsymbol{\xi}_{t+1}|\boldsymbol{\xi}_t] = \mathbf{F}\boldsymbol{\xi}_t$$

implies that $\mathbf{F}$ corresponds to the transpose of the transition probability matrix, $\mathbf{P}'$.[69]

Appendix C — The Expression for the ML Estimator of the Initial State Probability Vector

Consistently with our assumption that $\boldsymbol{\xi}_{1|0}$ is an unknown $K \times 1$ vector of parameters that must be estimated, we now note that the likelihood function can be alternatively be written as:

$$\begin{aligned} L\left(\{\mathbf{R}_t\}_{t=1}^T|\{\boldsymbol{\xi}_t\}_{t=1}^T, \boldsymbol{\theta}\right) &= \prod_{t=1}^T \sum_{\{\boldsymbol{\xi}_t\}_{t=1}^T} p(\mathbf{R}_t|\boldsymbol{\xi}_t, \Im_{t-1}; \boldsymbol{\theta}) \Pr\left(\boldsymbol{\xi}_t|\Im_{t-1}; \boldsymbol{\theta}, \boldsymbol{\rho}\right) \\ &= \prod_{t=1}^T \boldsymbol{\eta}_t' \boldsymbol{\xi}_{t|t-1} = \prod_{t=1}^T \boldsymbol{\eta}_t' \mathbf{P}' \boldsymbol{\xi}_{t-1|t-1} \\ &= \boldsymbol{\iota}_K' \prod_{t=1}^T \text{diag}(\boldsymbol{\eta}_t) \mathbf{P}' \hat{\boldsymbol{\xi}}_{t-1|t-1} = \boldsymbol{\iota}_K' \prod_{t=1}^T \mathbf{K}_t \boldsymbol{\xi}_{1|0} \end{aligned}$$

where $K_t(\boldsymbol{\theta}) \equiv$diag$(\boldsymbol{\eta}_t(\boldsymbol{\theta}))\mathbf{P}'$ (see Krolzig, 1997, p. 81, for a proof of the last line). Since the likelihood function is linear in $\boldsymbol{\xi}_{1|0}$ the solution is a boundary one:

$$\hat{\boldsymbol{\xi}}_{1|0} = \underset{1 \leq i \leq K}{\arg\max} \boldsymbol{\iota}_K' \prod_{t=1}^T K_t(\boldsymbol{\theta}) \mathbf{e}_i.$$

Appendix D — A Matlab® Workout

Construction of your portfolio: You are a European investor and your reference currency is the Euro. Using monthly data in STOCKINT2013.XLS and/or derived .txt files that collect the data, construct monthly excess returns (*in Euros*) using the *two* price indices **DS Market-PRICE INDEXes** for *two* economies, Germany and the US.

[69] Observe that in general, this dynamic state-space model is neither linear (as the state vector $\boldsymbol{\xi}_t$ also influences the covariance matrix of the process) nor Gaussian, as the innovations driving the transition equation are non-Gaussian random variables.

1. For the sample period January 1988 - December 2008, plot the values of each of the two individual indices (in logarithmic terms) and the excess returns for each of the two indices when denominated in Euros. Make sure to include the dividends paid by each of the two indices in each of the monthly return series. Notice that the monthly data made available on the course web site also include data on the dividend yield on index i (i = GER, US, UK), $I_{i,t}$, defined as

$$dy_{i,t} \equiv \frac{D_{i,t}}{I_{i,t-1}}.$$

2. Over the same sample, estimate a two-state Markov switching model with no VAR component but regime-dependent covariance matrix (i.e., a MSVARH(2,0) also called MSIH(2,0) in our lectures) on *excess* returns data,

$$\begin{aligned} r_{us,t+1} &= \mu^{us}_{S_{t+1}} + \epsilon_{us,t+1} \\ r_{ger,t+1} &= \mu^{ger}_{S_{t+1}} + \epsilon_{ger,t+1} \\ \begin{bmatrix} \epsilon_{us,t+1} \\ \epsilon_{ger,t+1} \end{bmatrix} &\sim N\left(\mathbf{0}, \mathbf{\Omega}_{S_{t+1}}\right), \end{aligned}$$

   in which expected excess returns (i.e., the means), the variances and covariances all depend on the same two-state Markov chain S_{t+1}, with constant transition matrix across regimes. Use the function "MS_VAR_Fit" from the Markov Switching toolbox that has been provided through the class web site to print on your screen the parameter estimates obtained in the two regimes. How can you interpret—on the basis of the parameter estimates—the economic nature of the first regime? How about the second regime? Plot the dynamics of (i) expected excess returns, (ii) standard deviations, and (iii) the full-sample, ex-post smoothed probabilities implied by the two-state Markov chain. Finally, compute and plot the dynamics of the conditional correlations implied by the two-state model using only real time information (i.e., using filtered and not smoothed probabilities, analogously with what "MS_VAR_Fit" does automatically). In computing dynamic correlations, make sure to adjust for the effects on both variances and covariance of the joint presence of switches in expected excess returns, as explained in the lectures.

3. Use the dynamic variance-covariance matrix and the dynamic conditional means filtered from question 2 to build an in-sample, recursive dynamic Markowitz portfolio based on the simple expression

$$\mathbf{w}_t^{Markow} = \frac{1}{\lambda}[\hat{\mathbf{\Omega}}_t]^{-1}\hat{\boldsymbol{\mu}}_t,$$

   where $\hat{\boldsymbol{\mu}}_t \equiv [\hat{\mu}_{us} \;\; \hat{\mu}_{ger}]'$ and $\lambda = 0.2$ (this is of course a measure of aversion to risk). Plot the corresponding recursive, real-time portfolio weights (notice that because you are solving the problem using excess returns, what is not allocated to stocks must be allocated to the riskless asset, here a short-term euro-denominated bond).

4. Repeat question 2 for the case of a two-state (restricted) Markov switching VAR(1) model with

regime-dependent covariance matrix

$$\begin{aligned} r_{us,t+1} &= \mu^{us}_{S_{t+1}} + \phi^{us,us}_{S_{t+1}} r_{us,t} + \phi^{us,ger}_{S_{t+1}} r_{ger,t} + \epsilon_{us,t+1} \\ r_{ger,t+1} &= \mu^{ger}_{S_{t+1}} + \phi^{ger,us}_{S_{t+1}} r_{us,t} + \phi^{ger,ger}_{S_{t+1}} r_{ger,t} + \epsilon_{ger,t+1} \\ \begin{bmatrix} \epsilon_{us,t+1} \\ \epsilon_{ger,t+1} \end{bmatrix} &\sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{us,S_{t+1}} & 0 \\ 0 & \sigma^2_{ger,S_{t+1}} \end{bmatrix}\right), \end{aligned}$$

where the restriction consists of the fact that the covariances in both regimes are restricted to be zero, i.e., the only source of correlation in the system is the fact that the same Markov state variables drives the first two moments for both countries.[70] Plot the dynamics of (i) expected excess returns, (ii) standard deviations, and (iii) the full-sample, ex-post smoothed probabilities implied by the two-state Markov chain. Finally, compute and plot the dynamics of the conditional correlations implied by the two-state model using only real time information (i.e., using filtered and not smoothed probabilities). In computing dynamic correlations, make sure to adjust for the effects on both variances and covariance of the joint presence of switches in expected excess returns, as explained in the lectures. [*Hint*: Although this question can be done applying simple modifications to your use of "MS_VAR_Fit" in question 2, it is now a good exercise to try and use a different function, "MS_Regress_Fit"]

5. Use the dynamic variance-covariance matrix and the dynamic conditional means filtered from question 4 to build an in-sample, recursive dynamic Markowitz portfolio based on the simple expression

$$\mathbf{w}^{Markow}_t = \frac{1}{\lambda}[\hat{\mathbf{\Omega}}_t]^{-1}\hat{\boldsymbol{\mu}}_t,$$

where $\hat{\boldsymbol{\mu}}_t \equiv [\hat{\mu}_{us} \ \hat{\mu}_{ger}]'$ and $\lambda = 0.2$. Plot these recursive, real-time portfolio weights. Why are these more stable than what found in question 3?

6. With reference to the out-of-sample period January 2009 - December 2012, proceed to compute optimal weights for the two-state Markov switching model in questions 4-5. Perform the calculation in the following way: use the same estimated conditional mean parameters (the regime-switching intercepts and VAR parameters) and the regime-dependent covariance matrix parameters estimated in question 4, that you should have saved. Compute the dynamic means and covariance matrix on the basis of those parameter performing the updating on the basis of the out-of-sample forecast errors over the out-of-sample period. Importantly, you need to derive predicted regime probabilities from the end-of-sample smoothed probabilities using the formulas derived in the lectures. The weights will then come from the classical Markowitz formula. After obtaining the weights, compute the realized Sharpe ratios (for the pure equity, risky portfolio) over the out of sample period. Compare these realized Sharpe ratios with those that you would

[70] Notice that also in this case, the expected returns (i.e., both the intercepts and the AR(1) coefficients), the variances and covariances all depend on the same two-state Markov chain S_{t+1}, with constant transition matrix across regimes.

have achieved by simply investing all of your wealth in each of the three stock indices under consideration.

7. Going back to the sample period January 1988 - December 2008, estimate now two distinct, univariate Markov switching first-order autoregressive (MSARH(2,1)) models with regime-dependent variance for excess stock returns on the US and the German index:

$$r_{i,t+1} = \mu^i_{S^i_{t+1}} + \phi^i_{S^i_{t+1}} r_{i,t} + \phi^{-i}_{S^i_{t+1}} r_{-i,t} + \epsilon_{us,t+1} \qquad \epsilon_{i,t+1} \sim N\left(0, \sigma^2_{i,S^i_{t+1}}\right),$$

where i = US, Germany, and $-i$ means Germany if i = US, and $-i$ means US if i = Germany. Notice that the Markov chain S^i_{t+1} driving the switching dynamics in the two models is now country-specific, i.e., S^{us}_{t+1} follows a chain that is potentially different (possibly, independent) of S^{ger}_{t+1}. For each of the two countries, proceed to plot the dynamics of (i) expected excess returns, (ii) standard deviations, and (iii) the full-sample, ex-post smoothed probabilities implied by the two-state Markov chain.

8. Estimate now the same bivariate two-state MSVARH(2,1)) model as in question 4 (but just to make it interesting, please now use "MS_VAR_Fit" as in question 2):

$$\begin{aligned} r_{us,t+1} &= \mu^{us}_{S_{t+1}} + \phi^{us,us}_{S_{t+1}} r_{us,t} + \phi^{us,ger}_{S_{t+1}} r_{ger,t} + \epsilon_{us,t+1} \\ r_{ger,t+1} &= \mu^{ger}_{S_{t+1}} + \phi^{ger,us}_{S_{t+1}} r_{us,t} + \phi^{ger,ger}_{S_{t+1}} r_{ger,t} + \epsilon_{ger,t+1} \\ \begin{bmatrix} \epsilon_{us,t+1} \\ \epsilon_{ger,t+1} \end{bmatrix} &\sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{us,S_{t+1}} & 0 \\ 0 & \sigma^2_{ger,S_{t+1}} \end{bmatrix}\right). \end{aligned}$$

This is a restricted version of the pair of univariate models for US and Germany estimated in question 7 in which $S_{t+1} = S^{us}_{t+1} = S^{ger}_{t+1}$, i.e., a unique Markov chain is assumed to drive switches in both US and German data. In particular, the model of this question may be obtained from the model in question 7 when (i) the mean and variance parameters are set to be identical; (ii) $p_{11} = p^{us}_{11} = p^{ger}_{11}$ and $p_{22} = p^{us}_{22} = p^{ger}_{22}$ which amounts to imposing 4 equality restrictions.[71] Use a likelihood ratio test (LRT) to assess the null hypothesis that this restriction (formulated as a null hypothesis) cannot be rejected based on the available data. How do you interpret a rejection of this restriction? [*Hint*: Notice that because of the diagonal structure of the covariance matrix, the total log-likelihood for the pair of univariate models for US and Germany estimated in question 7 can be simply computed as the sum of the individually maximized log-likelihood functions. In order to work on this question, make sure to read the toolbox manual related to Markov switching, on how to constrain coefficients in estimation.]

9. Repeat point 7 above when the marginal distribution of the errors is assumed to follow a t-student distribution, i.e., the model is

$$r_{i,t+1} = \mu^i_{S^i_{t+1}} + \phi^i_{S^i_{t+1}} r_{i,t} + \phi^{-i}_{S^i_{t+1}} r_{-i,t} + \epsilon_{us,t+1} \qquad \epsilon_{i,t+1} \sim t\left(0, \sigma^2_{i,S^i_{t+1}}; d\right),$$

[71] Technically, one also needs the two country-specific regimes to be initialized to be identical at the beginning of the sample.

where i = US, Germany, and -i means Germany if i = US, and -i means US if i = Germany. For each of the two countries, proceed to plot the dynamics of (i) expected excess returns, (ii) standard deviations, and (iii) the full-sample, ex-post smoothed probabilities implied by the two-state Markov chain.

Solution

This solution is a commented version of the MATLAB code Markov_switching_2013.m posted on the course web site. Also in this case, all the Matlab functions needed for the correct functioning of the code have been included. The loading of the *monthly* data is performed by the usual lines of code:

filename=uigetfile('*.txt');
data=dlmread(filename);

The above two lines import only the numbers, not the strings, from a .txt file. The usual lines of code take care of the strings and are not repeated here. The same applies to the exchange rate transformations that have now become customary in the first part of our Matlab workouts.

1. Figure A1 plots the values of each of the two individual indices (in logarithmic terms) and the excess returns denominated in Euros. Although it is not the same because the indices are two and the sample period is different, this plot resembles the one in workout 3, chapter 6.

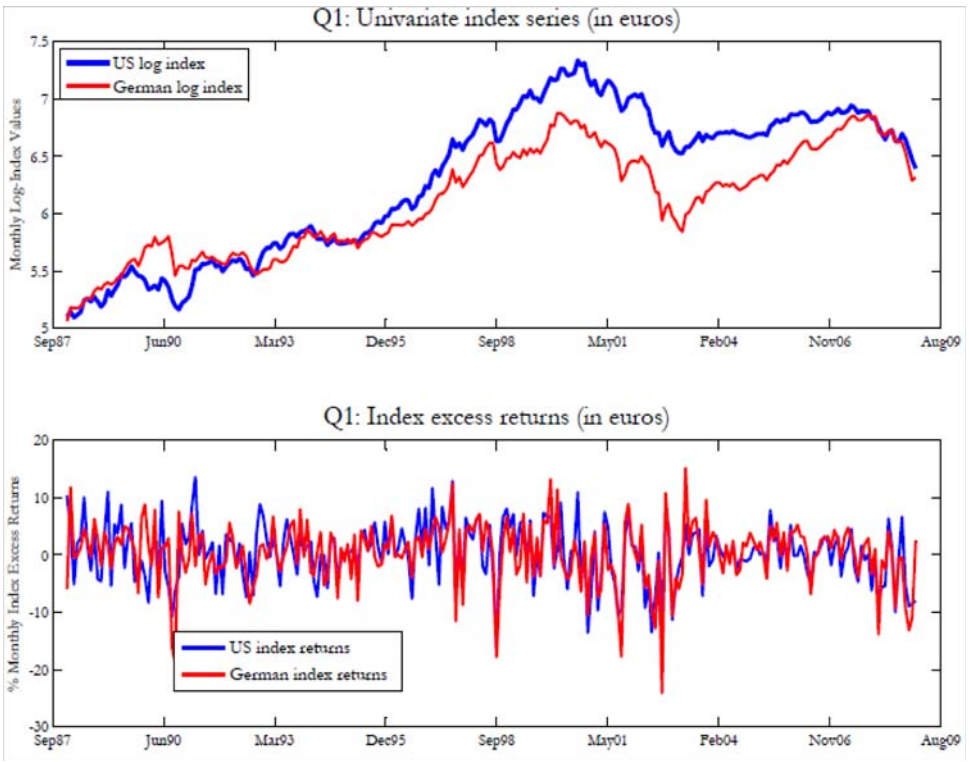


Figure A1:Monthly portfolio indices and returns expressed in euros

2. We now use the Perlin's *m_files* toolbox to estimate a two-state Markov switching model with no VAR component but regime-dependent covariance matrix (i.e., a MSIVARH(2,0) also called MSIH(2,0) in our lectures) on *excess* returns data,

$$\begin{aligned} r_{us,t+1} &= \mu^{us}_{S_{t+1}} + \epsilon_{us,t+1} \\ r_{ger,t+1} &= \mu^{ger}_{S_{t+1}} + \epsilon_{ger,t+1} \\ \begin{bmatrix} \epsilon_{us,t+1} \\ \epsilon_{ger,t+1} \end{bmatrix} &\sim N\left(\mathbf{0}, \mathbf{\Omega}_{S_{t+1}}\right). \end{aligned}$$

We do that by using the function "MS_VAR_Fit" from the toolbox. In fact, in code the following lines pass to toolbox the specification of the model and a few estimation options:

dep=R_eq; % Defines the dependent variables in system
nLag=0; % Number of vector autoregressive lags in 2x1 system (p)
k=2; % Number of states/regimes (K)
doIntercept=1; % Add intercept to equations (1= Yes; 0= No) (whether μ is MS)
advOpt.distrib='Normal'; % The Distribution assumption (only 'Normal' is allowed for MSVAR)
advOpt.std_method = 1; % Defining the method for calculation of standard errors.
advOpt.diagCovMat = 0; % Whether we will estimate by MLE also MS covariances (H feature)
advOpt.doPlots = 0; % Does not produce automatic plots (you are in charge of that!)
advOpt.printIter = 1; % When set to 0, does not print iterations to the screeen

[Spec_Out_1]=MS_VAR_Fit(dep,nLag,k,doIntercept,advOpt);

Some numerical optimization alterations and parameter estimates are therefore printed at the screen, as shown in Figure A2. This estimation output gives a wealth of information on the MS model. First, as you notice convergence is rather slow: unless you have are working on a server, an average 2-year old laptop may indeed take up to 4 minutes to complete estimation. This is becase the iterative EM algorithm that implements MLE in the case of MS models tends to be slower to converge because the need to iterate on both steps before the convergence criterion is eventually satisfied. The final maximized log-likelihood is then -1443.1062. The model implies the estimation 12 parameters—i.e., 4 different means, 6 elements of the two regime-specific covariance matrices (a total of 4 variances and 2 covariances), and 2 transition probabilities, p_{11} and p_{22}.

```
Sum log likelihood for MS Regression -->-1465.6386
Sum log likelihood for MS Regression -->-1465.6386
Sum log likelihood for MS Regression -->-1465.6386
```

...

```
Sum log likelihood for MS Regression -->-1443.1062
Sum log likelihood for MS Regression -->-1443.1062
Sum log likelihood for MS Regression -->-1443.1062

Calculating Standard Error Vector...


***** Numerical Optimization Converged *****

Final log Likelihood: -1443.1062
Number of estimated parameters: 14
Number of Equations in System: 2
Distribution Assumption -> Normal
Standard error calculation -> 1

***** Final Parameters for Equation #1 *****

     Intercept - Parameter Value (Standard Error, p value)
          State 1, Intercept = 1.13 (0.37,0.00)
          State 2, Intercept = -1.07 (0.69,0.12)
     Dependent Variable #1 - Parameter Value (Standard Error, p value)
     Dependent Variable #2 - Parameter Value (Standard Error, p value)

***** Final Parameters for Equation #2 *****

     Intercept - Parameter Value (Standard Error, p value)
          State 1, Intercept = 1.36 (0.32,0.00)
          State 2, Intercept = -1.50 (0.75,0.05)
     Dependent Variable #1 - Parameter Value (Standard Error, p value)
     Dependent Variable #2 - Parameter Value (Standard Error, p value)

---> Transition Probabilities Matrix (std. error, p-value) <---

      0.94 (0.06,0.00)   0.10 (0.05,0.03)
      0.06 (0.03,0.03)   0.90 (0.03,0.00)
```

Figure A2 : EM parameter estimates for MSIH(2,0) model for US and German excess stock returns

I know, the package states 14 but this derives from a mis-counting of the number of the free parameters appearing in $\hat{\mathbf{P}}$, which as we know is equal to 2, not 4 (because of the summing up constraint on the rows). The two state-dependent vectors of conditional mean excess returns (that we may call conditional risk premia) are (with p-values under the ML estimates; these p-values are obtained from Wald tests applied to invidual coefficients obtained in the way explain in Section 5):

$$\hat{\boldsymbol{\mu}}_{bull} = \begin{bmatrix} \hat{\mu}^{us}_{bull} \\ \hat{\mu}^{ger}_{bull} \end{bmatrix} = \begin{bmatrix} 1.13 \\ (0.00) \\ 1.36 \\ (0.00) \end{bmatrix} \qquad \hat{\boldsymbol{\mu}}_{bear} = \begin{bmatrix} \hat{\mu}^{us}_{bear} \\ \hat{\mu}^{ger}_{bear} \end{bmatrix} = \begin{bmatrix} -1.07 \\ (0.12) \\ -1.50 \\ (0.05) \end{bmatrix}.$$

In the bear regime, US risk premia are negative but not precisely estimated (hence one may consider to set them to zero, although we have already discussed that this is a bit rushed), but the German risk premium is negative and significant at 5%. In bull regime, both countries are characterized by positive and highly statistically significant risk premia. The estimated transition probability matrix characterizes both states as highly persistent with $\hat{p}_{bear,bear} = 0.90$ and $\hat{p}_{bull,bull} = 0.94$, and these are

both highly significant. Figure A3 shows that these estimated transition probabilities imply considerable average durations of 15.5 and 9.7 months, respectively. In fact, that bear states tend to last on average more than bull states do is a common finding in the literature. Figure A3 completes the picture by reporting the two regime-specific covariance matrices.[72] As noticed in Sections 4 and 9, also for these recent international equity data, the bear regime features variances that 3-4 times what is found in the bull state. Moreover, the implied state-specific correlations are:

$$\hat{\rho}_{bull} = \frac{\hat{\sigma}^{us,ger}_{bull}}{\hat{\sigma}^{us}_{bull}\hat{\sigma}^{ger}_{bull}} = \frac{7.86677}{\sqrt{14.63377}\sqrt{11.38909}} = 0.61 \quad \hat{\rho}_{bear} = \frac{\hat{\sigma}^{us,ger}_{bear}}{\hat{\sigma}^{us}_{bear}\hat{\sigma}^{ger}_{bear}} = \frac{33.81489}{\sqrt{43.10577}\sqrt{55.55218}} = 0.69.$$

The fact that international correlations grow during bear markets is also a well-known phenomenon. Of course, such linear correlations are just the tip of the iceberg, in the sense that another source of comovements between these two markets in this case comes from the fact that $\hat{\mu}^{us}_{bull} > \hat{\mu}^{us}_{bear}$ and $\hat{\mu}^{ger}_{bull} > \hat{\mu}^{ger}_{bear}$, i.e., the Markov state moves both intercepts in the same direction and at the same time, which makes the standard correlation a useful and yet imperfect measure of comovements across different markets.[73]

```
---> Expected Duration of Regimes <---

     Expected duration of Regime #1: 15.48 time periods
     Expected duration of Regime #2: 9.73 time periods

---> Covariance Matrix <---

State 1
      14.63377 (1.75778,0.00)    7.86677 ( NaN, NaN)
      7.86677 (1.35514,0.00)    11.38909 (1.48202,0.00)
State 2
      43.10577 (6.65956,0.00)    33.81489 ( NaN, NaN)
      33.81489 (6.16528,0.00)    55.55218 (7.68278,0.00)
```

Figure A3:EM parameter estimates for MSIH(2,0) model for US and German excess stock returns

As requested by the question, we also plot the dynamics of (i) expected excess returns, (ii) standard deviations, and (iii) the full-sample, ex-post smoothed probabilities implied by the two-state Markov chain in Figure A4. The smoothed state probabilities show a rather clear state definition with regimes going from values close to 0 to values close 1 and few periods of lingering uncertainty on the nature of the underlying regime. The main bear periods are characterized as late October 1987, 1989-1990, the Summer of 1998, several bouts during 2000-2003, and of course the great financial crisis of 2008-2009. The first two plots in Figure A4 show instead because both mean risk premia and volatilities are "in synch" across countries as far as the two regimes are concerned, both means and volatilities largely move together, reflecting the shapes of the evolution of smoothed probabilities in the third plot. Finally, in Figure A5 we have computed and plotted the dynamics of the conditional correlations implied by the two-state model using only real time information (i.e., using filtered and not smoothed

[72]Note that because the symmetry of covariance matrices, standard errors and p-values are not computed for the terms of the matrices that are simply copied across the main diagonal.

[73]However, how such correlations may be computed to take synchroneous regimes into account is an advanced topic.

probabilities, analogously with what "MS_VAR_Fit" does automatically).

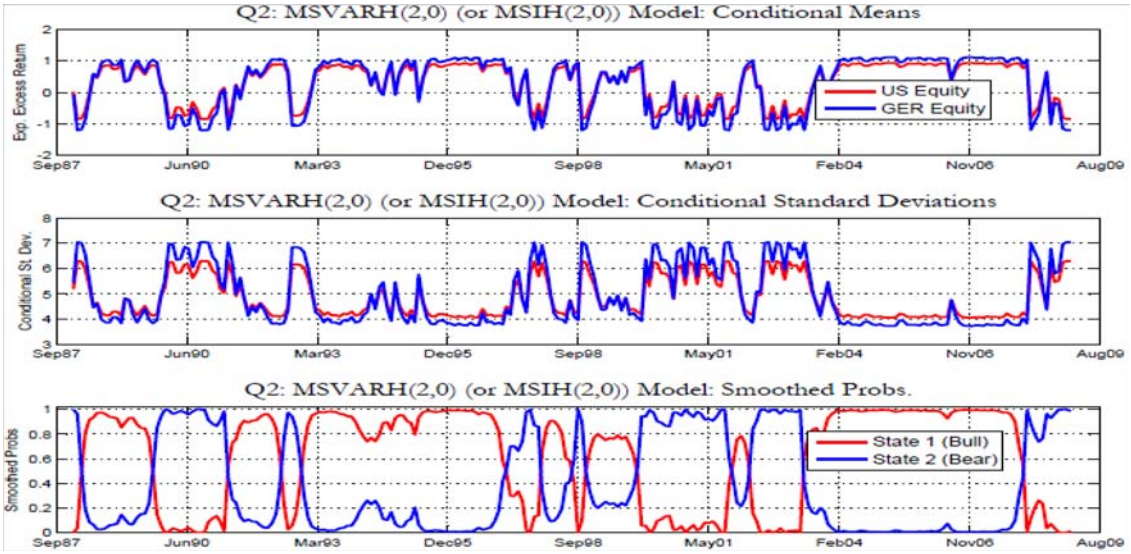


Figure A4:Implied conditional means, volatilities, and smoothed probabilities from MSIH(2,0)

In computing dynamic correlations, we have made sure to adjust for the effects on both variances and covariance of the joint presence of switches in expected excess returns, which is accomplised by the following lines of code:

%Extracts transition matrix from parameter vector

p11=Spec_Out_1.param(11);

p22=Spec_Out_1.param(14);

%Computes ergodic probabilities; notice the numerator, (1-p22) to compute ergodc1

ergodic1=(1-p22)/(2-p11-p22);

ergodic2=(1-p11)/(2-p11-p22);

%Computes unconditional mean estimates

mubar1=ergodic1*Spec_Out_1.param(7)+ergodic2*Spec_Out_1.param(8);

mubar2=ergodic1*Spec_Out_1.param(9)+ergodic2*Spec_Out_1.param(10);

%Transforms structure that contains covariance matrix into matrix

Covarmat=cell2mat(Spec_Out_1.Coeff.covMat);

cov_reg1=Covarmat(1,2); cov_reg2=Covarmat(1,4);

%Extracts standard deviations and covariances

Sigma1_reg1=sqrt(Covarmat(1,1)); Sigma1_reg2=sqrt(Covarmat(1,3));

Sigma2_reg1=sqrt(Covarmat(2,2)); Sigma2_reg2=sqrt(Covarmat(2,4));

%Computes filtered covariance over time

cov_dyna=Spec_Out_1.filtProb(:,1).*cov_reg1+Spec_Out_1.filtProb(:,2).*cov_reg2;

%Adjusts filtered covariance to take into account the effects of regime switches

[sizevec,cc]=size(Spec_Out_1.condMean(:,1));

```
cov_dyna=cov_dyna+Spec_Out_1.filtProb(:,1).*Spec_Out_1.filtProb(:,2)
.*(Spec_Out_1.condMean(:,1)-mubar1*ones(sizevec,1))
.*(Spec_Out_1.condMean(:,2)-mubar2*ones(sizevec,1));
%Computes filtered standard deviations over time
Sigma1_dyna=Spec_Out_1.filtProb(:,1).*Sigma1_reg1^2+
Spec_Out_1.filtProb(:,2).*Sigma1_reg2^2;
Sigma2_dyna=Spec_Out_1.filtProb(:,1).*Sigma2_reg1^2
+Spec_Out_1.filtProb(:,2).*Sigma2_reg2^2;
%Adjusts filtered variance to take into account the effects of regime switches
Sigma1_dyna=Sigma1_dyna+Spec_Out_1.filtProb(:,1)
.*Spec_Out_1.filtProb(:,2).*(Spec_Out_1.condMean(:,1)-mubar1*ones(sizevec,1))
.*(Spec_Out_1.condMean(:,1)-mubar1*ones(sizevec,1));
Sigma2_dyna=Sigma2_dyna+Spec_Out_1.filtProb(:,1).*Spec_Out_1.filtProb(:,2)
.*(Spec_Out_1.condMean(:,2)-mubar2*ones(sizevec,1)).*
(Spec_Out_1.condMean(:,2)-mubar2*ones(sizevec,1));
Sigma1_dyna=sqrt(Sigma1_dyna);
Sigma2_dyna=sqrt(Sigma2_dyna);
%Computes filtered correlation over time
cor_dyna=cov_dyna./(Sigma1_dyna.*Sigma2_dyna);

figure(3);
plot(date(ss:se)',cor_dyna,'b', 'LineWidth',2);
dateaxis('x',12)
set(gca,'fontname','garamond','fontsize',13);
ylabel('Dynamic Markov Switching Correlations');
title('Q2: MSVARH(2,0) (or MSIH(2,0)) Model: Dynamic Correlations',
'fontname','Garamond','fontsize',16);
```

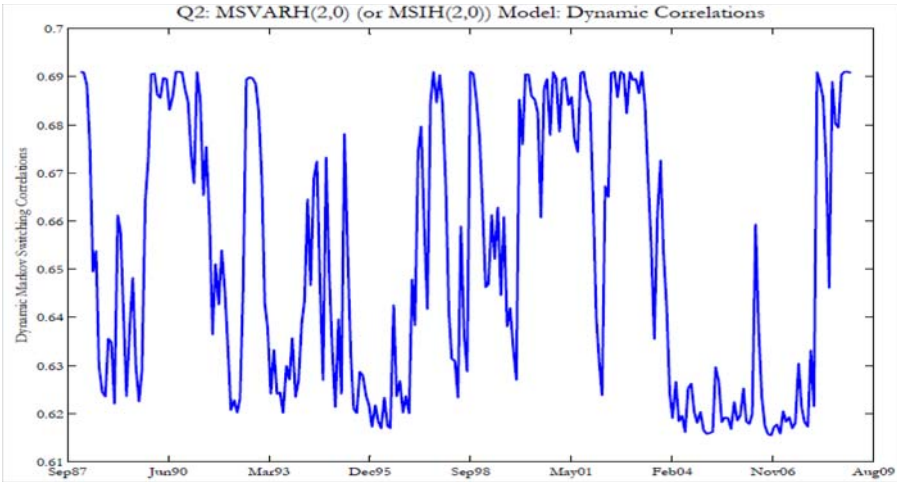


Figure A5:Implied conditional correlations from MSIH(2,0)

3. We now use the dynamic variance-covariance matrix and the dynamic conditional means filtered from question 2 to build an in-sample, recursive dynamic Markowitz portfolio based on the simple expression

$$\mathbf{w}_t^{Markow} = \frac{1}{\lambda}[\hat{\mathbf{\Omega}}_t]^{-1}\hat{\boldsymbol{\mu}}_t,$$

where $\hat{\boldsymbol{\mu}}_t \equiv [\hat{\mu}_{us} \ \hat{\mu}_{ger}]'$ and $\lambda = 0.2$ (this is of course a measure of aversion to risk). These calculations are performed by the lines of code:

```
%Extracts filtered vectors of expected returns and regime-specific covariance matrices
filtProb1 = Spec_Out_1.filtProb;
Sigma_s1 = Spec_Out_1.Coeff.covMat{1};
Sigma_s2 = Spec_Out_1.Coeff.covMat{2};
%These instructions build column vectors of mean parameter estimates
mu_s1 = [Spec_Out_1.Coeff.S_Param{1,1}(1);Spec_Out_1.Coeff.S_Param{1,2}(1)];
mu_s2 = [Spec_Out_1.Coeff.S_Param{1,1}(2);Spec_Out_1.Coeff.S_Param{1,2}(2)];
%Builds recursive filtered expected returns based on filtered probabilities
FIT_1 = repmat(filtProb1(:,1),1,2).*(repmat(mu_s1',n+1,1))...
+ repmat(filtProb1(:,2),1,2).*(repmat(mu_s2',n+1,1));
%Computes filtered matrices of covariances of returns
SIGMA_1 = zeros(2,2,n);  rho_1 = zeros(1,n);
vol_1_1 = zeros(1,n);  vol_1_2 = zeros(1,n);  vol_rt_1 = zeros(1,n);
for t=1:n+1
Sigma_vec = filtProb1(t,1)*vec(Sigma_s1)+filtProb1(t,2)*vec(Sigma_s2)...
+ filtProb1(t,1)*filtProb1(t,2)* vec((mu_s1-FIT_1(t,:)')*(mu_s2-FIT_1(t,:)')');
SIGMA_1(:,:,t) = reshape(Sigma_vec,2,2);
rho_1(1,t) = SIGMA_1(1,2,t)/sqrt(SIGMA_1(1,1,t)*SIGMA_1(2,2,t));
vol_1_1(1,t) = sqrt(SIGMA_1(1,1,t));
vol_1_2(1,t) = sqrt(SIGMA_1(2,2,t));
vol_rt_1(1,t) = vol_1_1(1,t)/vol_1_2(1,t);
At_1(:,t) = inv(0.2*SIGMA_1(:,:,t))*FIT_1(t,:)';
Wt_1(t,1) = At_1(1,t);
Wt_1(t,2) = At_1(2,t);
Wt_1(t,3) = 1-At_1(1,t)-At_1(2,t);
end
```

The corresponding recursive, real-time portfolio weights (that include a reminder allocated to the

riskless asset, here a short-term euro-denominated bond) are plotted in Figure A6.

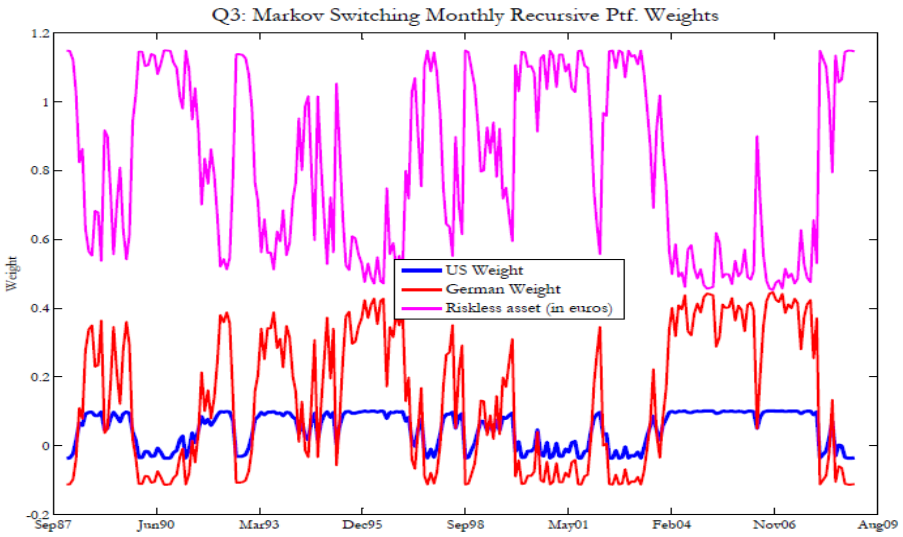


Figure A6:Recursive (filtered probs-based) mean-variance weights computed under $\lambda = 0.2$ from MSIH(2,0)

Even though on average almost 80% of the portfolio is allocated to cash, during bull markets both stock weights shoot up (especially the German portfolio share) to exceed 30 and 10 percent, respectively, thus lowering the percentage allocated to cash to less than 60%. However, during bear markets, the weights on both stock markets turn (slightly negative), as it is sensible, to indicate that one ought to short them in order to allocate more than 100% into safe cash. For instance, this would have been the optimal strategy suggested with reference to late 2008 and early 2009, clearly a rather attractive one.

4. At this point, we simply repeat point 2 with reference to a full-MSIVARH(2,1) in which however correlations are imposed to equal zero in both regimes. Figure A7 plots the dynamics of (i) expected excess returns, (ii) standard deviations, and (iii) the full-sample, ex-post smoothed probabilities implied by the two-state Markov chain for this case.

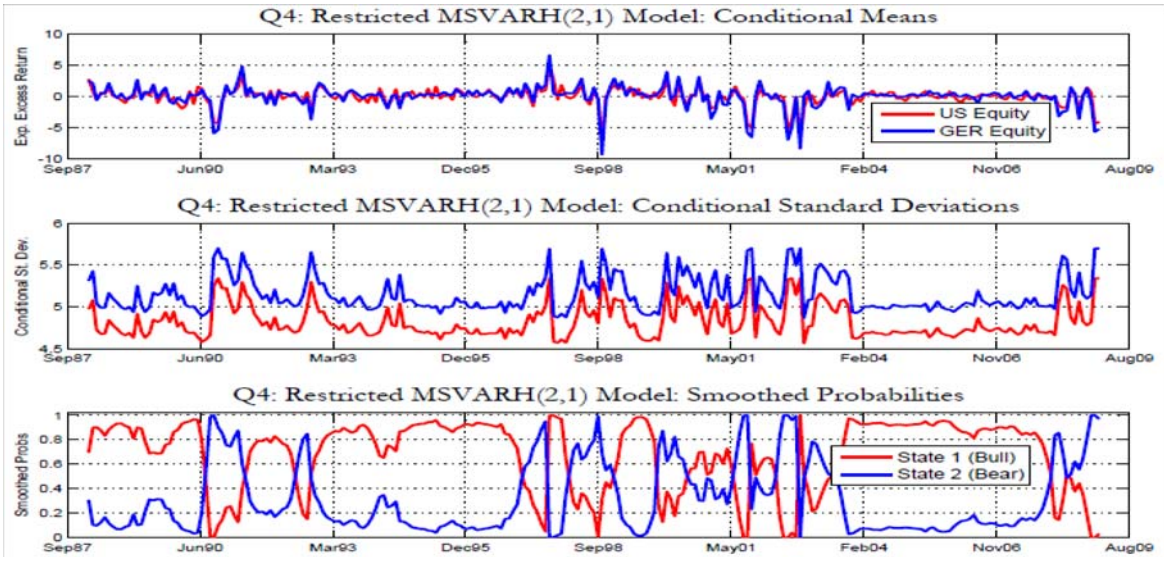


Figure A7:Implied conditional means, volatilities, and smoothed probs from zero-correlation MSIVARH(2,1)

We simply comment the key differences vs. Figure A4. The plot of smoothed probabilities tells a story that is similar to the one emphasized before, but now the state probabilities are much frequently removed from both 0 and 1.[74] For instance, in 2000-2001 the existence of substantial uncertainty is obvious. In this case, the conditional standard deviations of the shocks to US returns is always, systematically lower than that for German returns. However, you need to recall that in the case of a MSVARH model, the regime-specific covariance matrix that you estimate simply concerns $\boldsymbol{\epsilon}_{t+1}$ and not the vector of excess returns, as these are also affected by the state-specific VAR components. In fact, the first plot in Figure A7 shows that in such a model, the conditional expectations of excess returns become considerably volatile over time, as a result of lagged stock returns forecasting subsequent ones. In fact, the corresponding estimates concerning the first row of the model (i.e., the equation for U.S. excess returns) as printed on the Matlab screen are:

```
--->   Switching Parameters (Distribution Parameters)  <---

State 1
    Model's Variance:      19.964199
    Std Error (p. value):  2.3973 (0.00)
State 2
    Model's Variance:      29.999443
    Std Error (p. value):  4.3956 (0.00)

--->   Switching Parameters (Regressors)  <---

Switching Parameters for Equation #1 - Indep column 1

State 1
   Value:                 0.4709
   Std Error (p. value): 0.3765 (0.21)
State 2
   Value:                 -0.3209
   Std Error (p. value): 0.7896 (0.68)

Switching Parameters for Equation #1 - Indep column 2

State 1
   Value:                 0.1146
   Std Error (p. value): 0.1744 (0.51)
State 2
   Value:                 0.3259
   Std Error (p. value): 0.2590 (0.21)

Switching Parameters for Equation #1 - Indep column 3

State 1
   Value:                 -0.3441
   Std Error (p. value): 0.1212 (0.00)
State 2
   Value:                 0.1879

 Std Error (p. value): 0.2764 (0.50)
```

Figure A8 : EM estimates from a zero-correlations MSIAH(2,1) model

In this case, $\hat{\mu}^{us}_{bear} = -0.32$ and $\hat{\mu}^{us}_{bull} = 0.47$, but none of these intercepts is statistically significant; moreover, $\hat{a}^{us,us}_{bear} = 0.33$ and $\hat{a}^{us,us}_{bull} = 0.12$ are again not significant; finally, $\hat{a}^{us,ger}_{bear} = 0.19$ and $\hat{a}^{us,ger}_{bull} = -0.34$, with the latter coefficient being highly statistically significant (this may indicate that with a one-month lag, the two markets substitute for each other). Please make sure to

[74] One idea would be to compute RCM_1 for the models in questions 2 and now 4 and compare them. The plots suggest that while the first model will have a RCM_1 close to 100, the second will not.

visualize and interpret the remaining estimates, although it remains the case that most of them fail to be significant, which may an indication that while a MSIH(2,0) is an interesting model, a MSIVARH(2,1) is not.

Also note that in this case the estimation has been performed using the command *MS_Regress_Fit* with the commands:

% Defining a constant vector in mean equation

constVec=ones(length(dep),1);

% Defining explanatory variables in the two equations (one lag of both returns)

indep{1}=[constVec R_eq(1:end-1,1) R_eq(1:end-1,2)];

indep{2}=[constVec R_eq(1:end-1,1) R_eq(1:end-1,2)];

% Defining which parts of the two equations will switch states

S{1}=[1 1 1 1];

S{2}=[1 1 1 1];

advOpt.distrib='Normal';

advOpt.std_method=1;

advOpt.diagCovMat = 1; % This means that we will NOT estimate by MLE also MS covariances

advOpt.doPlots = 0;

advOpt.printIter = 0; % When set to 0, does not print iterations to the screeen

[Spec_Out_2]=MS_Regress_Fit(dep,indep,k,S,advOpt);

The line before the last specifies that no lengthy iteration information is to be printed on the Matlab screen. MS_Regress fit is a toolbox function that is fit to estimate MS regressions, besides MS VAR models like in this specific application. Figure A9 computes and plots the dynamics of the conditional correlations implied by the two-state model using only real time information (i.e., using filtered and not smoothed probabilities).

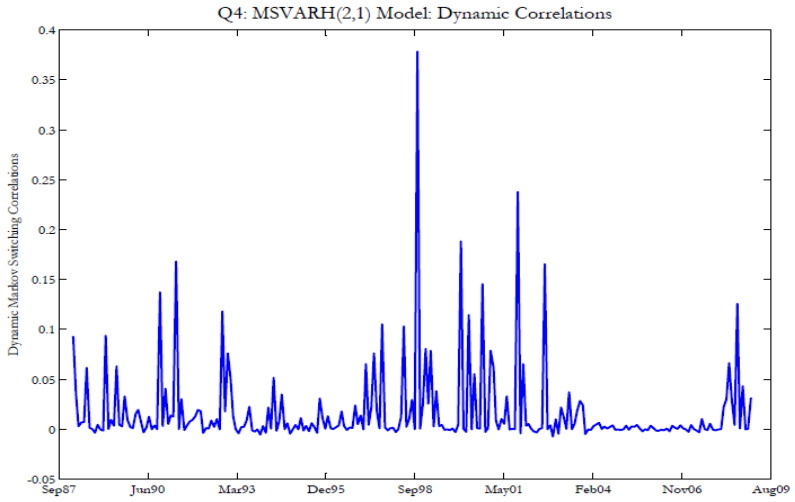


Figure A9:Implied conditional correlations from zero-correlations MSIAH(2,1)

5. At this point, we replicate question 3 and use the dynamic variance-covariance matrix and the dynamic conditional means filtered from question 4 to build an in-sample, recursive dynamic Markowitz portfolio when $\lambda = 0.2$. Figure A10 shows such optimal weights.

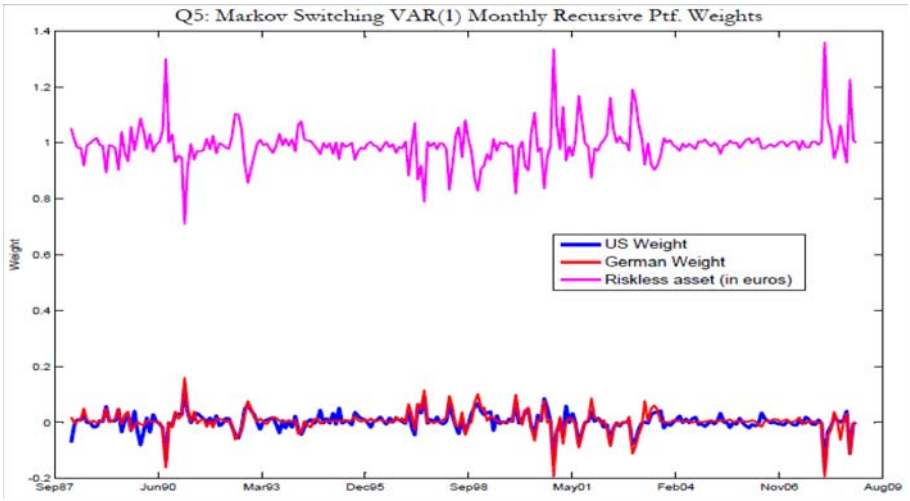


Figure A10:Recursive (filtered-based) MV weights under $\lambda = 0.2$ from zero-correlation MSIAH(2,1)

These are clearly different than those in Figure A6. On the one hand, all these weights are very stable over time, and oscillate around zero in the case of the stock allocations. On the other hand, visibly, an investor using filtered probabilities from the model in Figure A8, would end up always investing close to 100% in cash, which is probably due to the imprecise estimates of most of the parameters.

6. With reference to the out-of-sample period January 2009 - December 2012, we proceed to compute optimal weights for the two-state Markov switching model in questions 4-5. We use the same estimated conditional mean parameters and the regime-dependent covariance matrix parameters estimated in question 4 but compute the dynamic means and covariance matrix on the basis of those parameter performing the updating on the basis of the out-of-sample forecast errors over the out-of-sample period. Figure A11 shows the results.

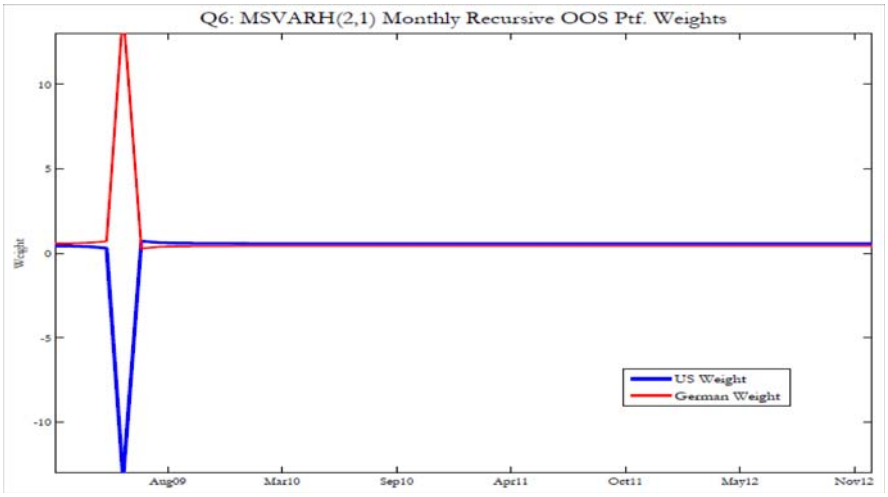


Figure A11:Recursive out-of-sample MV weights from zero-correlation MSIAH(2,1)

The weights behave in a way similar to what already found in the workout example of chapter 6: the financial crisis in early 2009 is marked by a shock to U.S. excess returns that, even though only temporarily draws an investor away from U.S. markets and towards the European ones. After obtaining the weights, we havel compute the realized Sharpe ratios (for the pure equity, risky portfolio) over the out of sample period, obtaining:

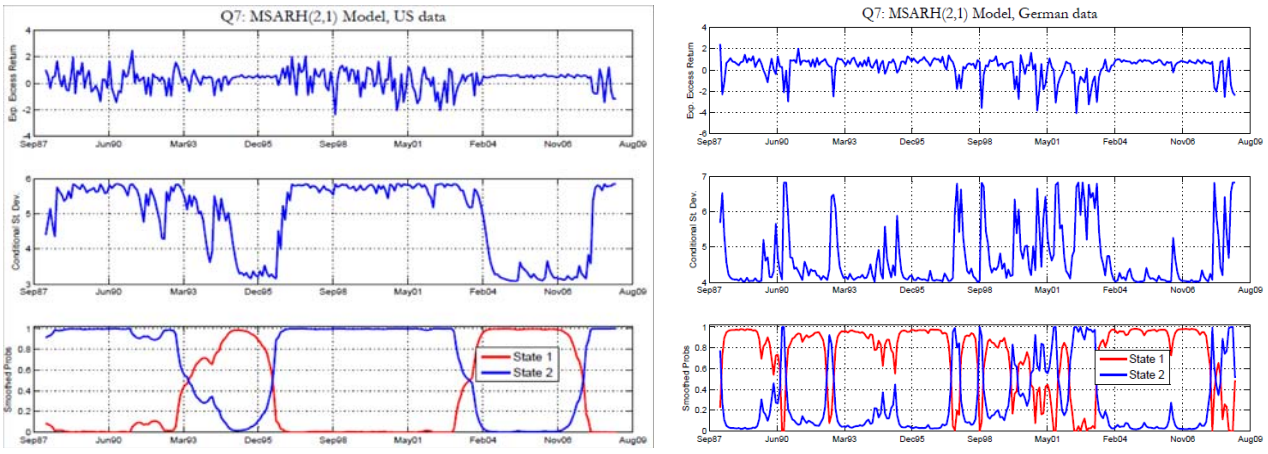
```
Sharpe Ratio
1. MSVARH(2,1) Markowitz Portfolio out-of-sample Sharpe ratio:
    0.2621

2. US Equity Sharpe ratio during the forecasting period:
    0.3232

2. German Equity Sharpe ratio during the forecasting period:
    0.1601
```

Although the MSIAH(2,1) model does not yield the highest of the Sharpe ratios, our earlier concerns on the specification of the model make the chances of this model outperforming all other models rather slim.

7. We now estimate two distinct, univariate MSARH(2,1) models with regime-dependent variance for excess stock returns on the US and the German index:

$$r_{i,t+1} = \mu^i_{S^i_{t+1}} + \phi^i_{S^i_{t+1}} r_{i,t} + \phi^{-i}_{S^i_{t+1}} r_{-i,t} + \epsilon_{us,t+1} \qquad \epsilon_{i,t+1} \sim N\left(0, \sigma^2_{i,S^i_{t+1}}\right),$$

where i = US, Germany, and $-i$ means Germany if i = US, and $-i$ means US if i = Germany. The Markov chain S^i_{t+1} driving the switching dynamics in the two models is now country-specific, i.e., S^{us}_{t+1} follows a chain that is potentially different (possibly, independent) of S^{ger}_{t+1}. For each of the two countries, Figure A12 plots the dynamics of (i) expected excess returns, (ii) standard deviations, and (iii) the full-sample, ex-post smoothed probabilities implied by the two-state Markov chain.



U.S. data German data

Figure A12:EM estimation outputs from two univariate MSIAH(2,1) models

Although parameter estimates and their standard errors are not reported here, please make sure

to visualize and study them on your Matlab output screen. Figure A12 makes it clear that the dynamics over time of S_{t+1}^{us} and S_{t+1}^{ger} are rather different.

8. Next, we have estimated the same bivariate two-state MSVARH(2,1)) model as in question 4; just to make this repetition interesting, we have now used the function MS_VAR_Fit from the toolbox. This model is however interesting in connection to question 7 because it is a restricted version of the pair of univariate models for US and Germany estimated in question 7 in which $S_{t+1} = S_{t+1}^{us} = S_{t+1}^{ger}$, i.e., a unique Markov chain is assumed to drive switches in both US and German data. In particular, the model of this question may be obtained from the model in question 7 when (i) the mean and variance parameters are set to be identical; (ii) $p_{11} = p_{11}^{us} = p_{11}^{ger}$ and $p_{22} = p_{22}^{us} = p_{22}^{ger}$ which amounts to imposing 4 equality restrictions.[75] Because of the diagonal structure of the covariance matrix, the total log-likelihood for the pair of univariate models for US and Germany estimated in question 7 can be simply computed as the sum of the individually maximized log-likelihood functions:

dep=R_eq;

nLag=1;

k=2;

S{1}=[1 1 1]; S{2}=[1 1 1];

doIntercept=1;

advOpt.distrib='Normal';

advOpt.std_method=1;

% THIS IS IMPORTANT: this MSVARH(2,1) model obtains as a restriction of the two univariate models in question 7 only when the simultaneous covariance coefficient is restricted to be zero

advOpt.diagCovMat = 1;

advOpt.doPlots = 0;

advOpt.prinIter = 1;

advOpt.printOut = 1;

advOpt.constCoeff.nS_Param{1}={'e'};

advOpt.constCoeff.nS_Param{2}={'e'};

advOpt.constCoeff.S_Param{1}=

{mu_us(1), mu_us(2); var_us1(1), var_us1(2); var_us2(1), var_us2(2)};

advOpt.constCoeff.S_Param{2}=

{mu_ger(1), mu_ger(2); var_ger1(1), var_ger1(2); var_ger2(1), var_ger2(2)};

advOpt.constCoeff.covMat{1}(1,1)={variance_us(1)};

advOpt.constCoeff.covMat{1}(2,2)={variance_us(2)};

[75] Technically, one also needs the two country-specific regimes to be initialized to be identical at the beginning of the sample.

advOpt.constCoeff.covMat{2}(1,1)={variance_ger(1)};

advOpt.constCoeff.covMat{2}(2,2)={variance_ger(2)};

advOpt.constCoeff.p={'e','e'; 'e','e'};

[Spec_Out_3]=MS_VAR_Fit(dep,nLag,k,doIntercept,advOpt);

% Performs Likelihood Ratio test:

% Log-likelihood of restricted case, MSVARH(2,1)

LL_A = Spec_Out_3.LL;

% Log likelihood of general case with separate regime process (from univariate estimations)

LL_B = Spec_Out_us.LL + Spec_Out_ger.LL;

% Likelihood ratio test has structure 2*(Unrestricted Log-lik - Restricted Log-Lik)

LRT = 2*(LL_B-LL_A);

We have then used a likeihood ratio test (LRT) to assess the null hypothesis that this restriction (formulated as a null hypothesis) cannot be rejected based on the available data, finding:

```
Likelihood Ratio Test:
Likelihood Ratio Statistic:  251.9871
chi_sq(4) 5 percent critical value:  9.4877
chi_sq(4) 1 percent critical value:  13.2767
```

The LRT of almost 255 under 4 restrictions leads to a rejection because the $\chi^2_{1,0.99} = 13.28$. This means that the restriction can be rejected, i.e., $S^{us}_{t+1} \neq S^{ger}_{t+1}$, the regime driving U.S. and German data are different and therefore, at least at this non-linear level, there is no evidence of contagion.

9. Finally, we have repeated point 7 above when the marginal distribution of the errors is assumed to follow a t-student distribution, i.e., the model is

$$r_{i,t+1} = \mu^i_{S^i_{t+1}} + \phi^i_{S^i_{t+1}} r_{i,t} + \phi^{-i}_{S^i_{t+1}} r_{-i,t} + \epsilon_{us,t+1} \qquad \epsilon_{i,t+1} \sim t\left(0, \sigma^2_{i,S^i_{t+1}}; d\right),$$

where i = US, Germany, and -i means Germany if i = US, and -i means US if i = Germany. In this case, for instance for the German data, we obtain the outputs in Figure A13. Interestingly, the degrees of freedom of the t-Student ($\hat{d}$) are strongly time-varying. Changing the marginal error distribution does affect parameter estimates and ends up making both regimes considerably

more persistent than what found in question 7, to the point that $\hat{p}_{22} \simeq 1$.

```
Final log Likelihood: -764.5544
Number of estimated parameters: 14
Type of Switching Model: Univariate
Distribution Assumption -> t
Method SE calculation -> 1

***** Final Parameters for Equation #1 *****

---> Non Switching Parameters <---


--->   Switching Parameters (Distribution Parameters)  <---

State 1
    Model's Variance:      26.504545
    Std Error (p. value):  711.0397 (0.97)
    Degrees of Freedom (t dist): 13.9991
    Std Error (p. value):        453.8011 (0.98)
State 2
    Model's Variance:      15.638331
    Std Error (p. value):  2.5438 (0.00)
    Degrees of Freedom (t dist): 3.9769
    Std Error (p. value):        1.0727 (0.00)

--->   Switching Parameters (Regressors)  <---

Switching Parameters for Equation #1 - Indep column 1

State 1
   Value:                4.4286
   Std Error (p. value): 138.0050 (0.97)
State 2
   Value:                0.7535
   Std Error (p. value): 0.3124 (0.02)
```

```
Switching Parameters for Equation #1 - Indep column 2

State 1
   Value:                -7.6378
   Std Error (p. value): 82.8754 (0.93)
State 2
   Value:                0.0933
   Std Error (p. value): 0.0796 (0.24)

Switching Parameters for Equation #1 - Indep column 3

State 1
   Value:                8.6387
   Std Error (p. value): Inf (1.00)
State 2
   Value:                -0.0140
   Std Error (p. value): 0.0845 (0.87)

---> Transition Probabilities Matrix (std. error, p-value) <---

     0.92 (42.74,0.98)   0.00 (0.00,1.00)
     0.08 (1.93,0.96)    1.00 (0.06,0.00)

---> Expected Duration of Regimes <---

    Expected duration of Regime #1: 11.81 time periods
    Expected duration of Regime #2:  Inf time periods
```

Figure A13:EM estimation outputs from two univariate t-Student MSIAH(2,1) models

References

[1] Ang, A., and G., Bekaert, 2002. "International Asset Allocation With Regime Shifts", *Review of Financial Studies*, 15, 1137-1187.

[2] Baum, L., T., Petrie, G., Soules, and N., Weiss, 1970, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", *Annals of Mathematical Statistics*, 41, 164-171.

[3] Bollerslev, T., 1990. "Modelling the Coherence in Short-Run Nominal Exchange Rates: A Multivariate Generalized ARCH Model", *Review of Economics and Statistics*, 72, 498-505.

[4] Cai, J., 1994. "A Markov Model of Switching-Regime ARCH", *Journal of Business and Economic Statistics*, 12, 309-316.

[5] Davidson, R., and J., MacKinnon, 1981. "Several Tests for Model Specification in the Presence of Alternative Hypothesis", *Econometrica*, 49, 781-793.

[6] Davies, R., 1977. "Hypothesis Testing When a Nuisance Parameter Is Present Only Under the Alternative", *Biometrika*, 64, 247-254.

[7] Doan, T., R., Littermann, and C., Sims, 1984. "Forecasting and Conditional Projection Using Realistic Prior Distributions", *Econometric Reviews*, 3, 1-14.

[8] French, K., W., Schwert, and R., Stambaugh, 1987. "Expected Stock Returns and Volatility", *Journal of Financial Economics*, 19, 3-29.

[9] Glosten, L., R., Jagannathan, and D., Runkle, 1993. "On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks", *Journal of Finance*, 48, 1779-1801.

[10] Gray, S., 1996. "Modeling the Conditional Distribution of Interest Rates as a Regime-Switching Process", *Journal of Financial Economics*, 42, 27-62.

[11] Guidolin, M., 2009. "Detecting and Exploiting Regime Switching ARCH Dynamics in US Stock and Bond Returns", in *Stock Market Volatility* (G. Gregoriou editor), Chapman Hall, London, pp. 92-133.

[12] Guidolin M., 2012. "Markov Switching Models in Empirical Finance", in *Advances in Econometrics* (D. Drukker et al., eds.), Emerald Publishers Ltd., London, pp. 1-86.

[13] Guidolin, M., and S., Ono, 2006. "Are the Dynamic Linkages Between the Macroeconomy and Asset Prices Time-Varying?" *Journal of Economics and Business*, 58, 480-518.

[14] Guidolin, M., F., Ravazzolo, and A., Tortora, 2013. "Econometric Implementations of Multi-Factor Models of the U.S. Financial Markets", *Quarterly Review of Economics and Finance*, 53, 87-111.

[15] Guidolin, M., and F. Ria, 2010, "Regime Shifts in Mean-Variance Efficient Frontiers: Some International Evidence", Federal Reserve Bank of St. Louis working paper 2010-040B. [also published in *Journal of Asset Management*, 2011, 12, 322-349]

[16] Guidolin, M., and A., Timmermann, 2005, "Economic Implications of Bull and Bear Regimes in UK Stock and Bond Returns", *Economic Journal*, 115, 111-143.

[17] Guidolin, M., and A., Timmermann, 2006a. "An Econometric Model of Nonlinear Dynamics in the Joint Distribution of Stock and Bond Returns", *Journal of Applied Econometrics*, 21, 1-22.

[18] Guidolin, M., and A., Timmermann, 2006b. "Term Structure of Risk under Alternative Econometric Specifications", *Journal of Econometrics*, 131, 285-308.

[19] Guidolin,M., and A., Timmermann, 2007. "Asset Allocation under Multivariate Regime Switching", *Journal of Economic Dynamics and Control*, 31, 3503-3544.

[20] Hamilton J., 1988. "Rational-Expectations Econometric Analysis of Changes in Regime: An Investigation of the Term Structure of Interest Rates", *Journal of Economic Dynamics and Control*, 12, 385-423.

[21] Hamilton J., 1994. *Time Series Analysis*. Princeton University Press, chapter 22.

[22] Hamilton, J., and R., Susmel, 199., "Autoregressive Conditional Heteroskedasticity and Changes in Regime", *Journal of Econometrics*, 64, 307-333.

[23] Hansen, B., 1992. "The Likelihood Ratio Test Under Non-Standard Conditions: Testing the Markov Switching Model of GNP", *Journal of Applied Econometrics*, 7, S61-S82.

[24] Karlsen, H., 1990. "Existence of Moments in a Stationary Stochastic Difference Equation", *Advances in Applied Probability*, 22, 129-146.

[25] Kim, C.-J., 1994. "Dynamic Linear Models with Markov-Switching", *Journal of Econometrics*, 60, 1-22.

[26] Leroux, B., 1992. "Maximum Likelihood Estimation for Hidden Markov Models", *Stochastic Processes and their Applications*, 40, 127-143.

[27] Lettau, M., and S., Ludvigson, 2001. "Resurrecting the (C)CAPM: a Cross-Sectional Test when Risk Premia are Time-Varying", *Journal of Political Economy*, 109, 1238-1287.

[28] Krolzig, H.-M., 1997. *Markov-Switching Vector Autoregressions*, Berlin, Springer-Verlag.

[29] Magli, M., L., 2013. *Modelli con Regimi Markoviani e l'Illusorio Pricing del Rischio Idiosincratico nella Cross Section: Un'Analisi Empirica.* MSc. Finance dissertation, Bocconi University.

[30] Magnani, C., A., 2012. *Decomposing the Great Real Estate Bubble: Evidence from Commercial and Residential REIT Data*. MSc. Finance dissertation, Bocconi University.

[31] Pelletier, D., 2006. "Regime Switching for Dynamic Correlations", *Journal of Econometrics*, 131, 445-473.

[32] Turner, C., R., Startz, and C., Nelson, 1989. "A Markov Model of Heteroskedasticity, Risk, and Learning in the Stock Market", *Journal of Financial Economics*, 25, 3-22.

[33] Tong, H., 1983. *Threshold Models in Non-linear Time Series Analysis*, New York, Springer-Verlag.

[34] White, H., 1982. "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1-25.

[35] Wolfe, J., 1971. *A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixture of Multinormal Distributions*. San Diego, NITS Research Laboratory.