Lecture 10: Further Topics*

1 GARCH Model Families

There are a few key features of financial data. First, the variance seems to be varying from time to time, and usually one large movement tends to be followed by another large movement. In other words, large movements tend to cluster. This can be seen from Figure 1.

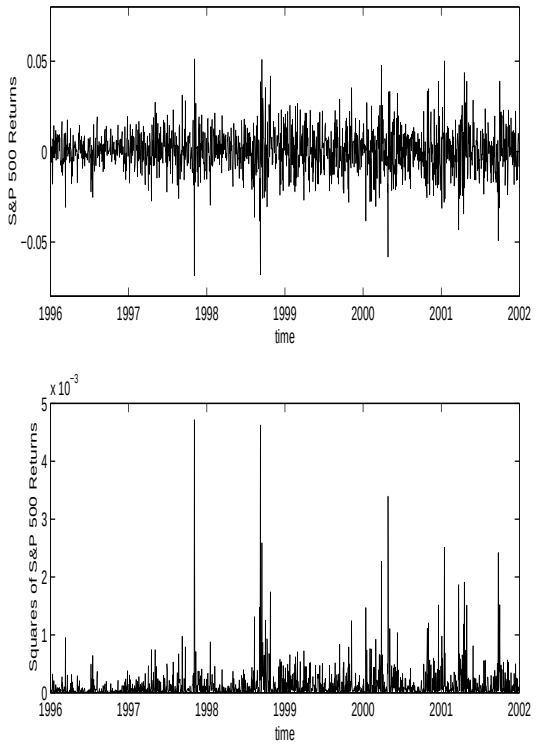

Figure 1: The S&P 500 Return 1996-2001

Figure 1 plots the S&P 500 returns from Jan. 1996 to Dec. 2001. The upper figure plots the returns and the lower figure plots the squares of the returns. From both graphs, you could see the clustering of large movements.

Second, the distributions of financial data have heavy tails (heavier than Gaussian). For the same data described above, I plot the empirical density and the normal density (with mean zero and standard deviation equal to the standard deviation of the data) in Figure 2.

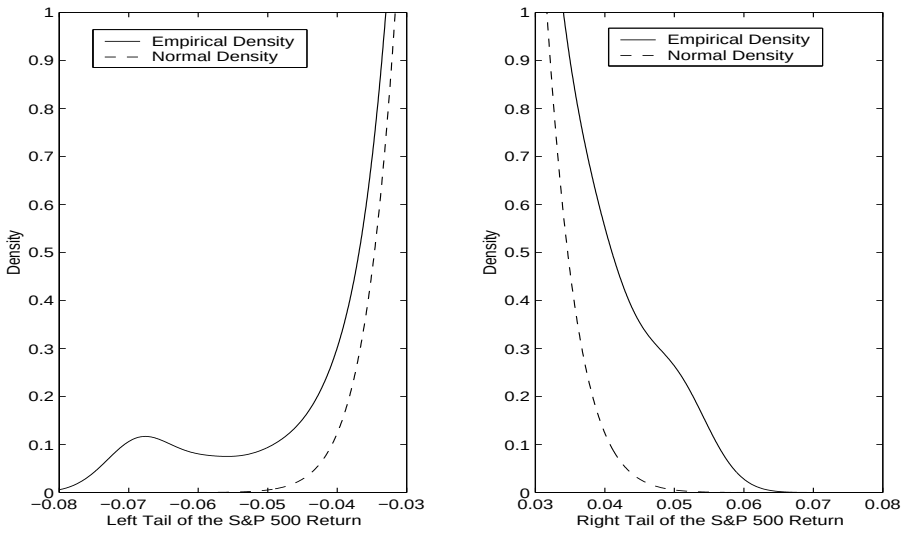

Figure 2: The Tails of S&P 500 Return 1996-2001

Another way to measure tail thickness is to use kurtosis statistics. We know that the kurtosis of Gaussian is 3. But for the data we have here, the empirical kurtosis is 5.8867.

Both time-varying variance and fat tails are important in finance applications. The ARCH-GARCH family models have been constructed to capture these features of financial data.

1.1 Autoregressive Conditional Heteroskedasticity (ARCH)

ARCH model was introduced by Engle (1982). For an AR(p) process

$$x_t = \sum_{k=1}^{p} \phi_k x_{t-k} + \epsilon_t,$$

where $E(\epsilon_t) = 0$, $E(\epsilon_t^2) = \sigma^2 > 0$ and $E(\epsilon_t \epsilon_s) = 0$ for $t \neq s$.

The idea of ARCH model is that the variance of ϵ_t, denoted by σ_t^2, follows an autoregressive process.

$$\sigma_t^2 = c + \sum_{i=1}^{m} \beta_i \epsilon_{t-i}^2 + u_t, \tag{1}$$

where $u_t \sim WN(0, \eta^2)$. Then we say ϵ_t follows an ARCH(m) process.

Note: first, we must have that $\sigma_t^2 > 0$. A sufficient condition is that $c \geq 0$ and $\beta_i \geq 0$ for $i = 1, \ldots, m$.

Second, we are modeling a time-varying conditional variance for x_t, but we'd still like to restrict our discussion to covariance-stationary process, therefore, we want that the unconditional variance

of x_t, therefore the unconditional variance of ϵ_t, is constant. For the AR(m) process σ_t^2 to be stationary, we must have that all the roots of

$$1 - \beta_1 z - \ldots - \beta_m z^m = 0$$

lie outside the unit circle. Combine this with the condition that all β_is are nonnegative, we need to impose the following condition on the coefficient,

$$\sum_{i=1}^{m} \beta_i < 1.$$

Then the unconditional variance of ϵ_t is given by

$$\sigma^2 = E(\sigma_t^2) = c/(1 - \sum_{i=1}^{m} \beta_i).$$

Another way to specify an ARCH(m) process for ϵ_t is to let

$$\epsilon_t = \sqrt{h_t}\nu_t$$

where $E(\nu_t) = 0$ and $Var(\nu_t) = 1$, and

$$h_t = c + \sum_{i=1}^{m} \beta_i \epsilon_{t-i}^2.$$

It is easy to see that $E(\epsilon_t) = 0$ and $E_{t-1}(\epsilon_t^2) = h_t$.

To estimate the parameters in an ARCH model, we can specify a distribution for ν_t and use maximum likelihood estimation, or use GMM estimation based on some orthogonalization conditions.

Recall that we motivate this section by describing two features of financial data: clustering of large movements and heavy tails. An ARCH model can capture both of these two features. First, it is easy to see that we got dependence between σ_t and $\sigma_{t-1}, \ldots$, therefore can produce clustered large movements. Second,the distribution of ϵ_t will have heavy tails. We can show this by computing kurtosis, denoted by ks, for a simple ARCH(1) process,

$$\epsilon_t = \sqrt{h_t}\nu_t$$

where $\nu_t \sim i.i.d.N(0,1)$, and

$$h_t = c + \beta \epsilon_{t-1}^2.$$

The moments for ϵ_t are

$$E(\epsilon_t) = 0,$$
$$E(\epsilon_t^2) = E(h_t \nu_t^2) = \sigma^2 = c/(1-\beta)$$
$$E(\epsilon_t^4) = 3\frac{c^2 + 2c\beta\sigma^2}{1 - 3\beta^2},$$

so the kurtosis is

$$ks = \frac{E(\epsilon_t^4)}{(E(\epsilon_t^2))^2} = 3\frac{1-\beta^2}{1-3\beta^2} > 3 \quad \text{for} \quad \beta^2 < 1/3.$$

Therefore, an ARCH model can produce heavier than normal tails.

1.2 Generalized Autoregressive Conditional Heteroskedasticity (GARCH)

Bollerslev (1986) extends the ARCH model and let σ_t^2 in (1) not only depend on the lagged values of ϵ_t^2, but also the lagged values of σ_t^2,

$$\sigma_t^2 = c + \sum_{i=1}^{p} \alpha_i \sigma_{t-i}^2 + \sum_{i=1}^{q} \beta_i \epsilon_{t-i}^2 + u_t. \tag{2}$$

We use GARCH(p, q) to denote such a process. The sufficient condition for $\sigma_t^2 > 0$ and that the process is stationary is that $\alpha_i \geq 0$ for $i = 1, \ldots, p$, $\beta_j \geq 0$ for $j = 1, \ldots, q$, and

$$\sum_{i=1}^{p} \alpha_i + \sum_{i=1}^{q} \beta_i < 1.$$

Finally, the unconditional variance is

$$\sigma^2 = E(\sigma_t^2) = c/(1 - \sum_{i=1}^{p} \alpha_i - \sum_{i=1}^{q} \beta_i).$$

For the S&P 500 data we have displayed, we estimate a GARCH(1, 1) process for the return r_t and the estimates we got are:

$$\sigma_t^2 = 0.0021^2 + 0.876\sigma_{t-1}^2 + 0.097\epsilon_{t-1}^2 + \hat{u}_t.$$

The unconditional variance is

$$\sigma^2 = E(\sigma^2) = 0.0021^2/(1 - 0.876 - 0.097) = 0.0127^2.$$

Figure 3 plots the estimated σ_t^2 (solid line in the lower graph) and the unconditional variance σ^2 (dashed line in the lower graph).

1.3 Multivariate GARCH

Consider a k-vector VAR(h) process,

$$\mathbf{x}_t = \sum_{i=1}^{h} \mathbf{\Phi}_i \mathbf{x}_{t-i} + \boldsymbol{\epsilon}_t$$

where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $E(\boldsymbol{\epsilon_t}'\boldsymbol{\epsilon_s}) = \Omega$ for $t = s$ and zero otherwise.

Recall that a univariate GARCH model produces time-varying variance σ_t^2. Using the same idea, we could construct a multivariate GARCH model to produce time-varying covariance matrix Ω_t.

A simple extension from univariate GARCH(p, q) processes to multivariate GARCH(p, q) is to let each element of Ω_t, say $\sigma_{ij,t}^2$, follows a GARCH(p, q) process,

$$\sigma_{ij,t}^2 = c_{ij} + \sum_{l=1}^{p} \alpha_{ij,l} \sigma_{ij,t-l}^2 + \sum_{k=1}^{q} \beta_{ij,k} \epsilon_{ij,t-k}^2 + u_{ij,t}.$$

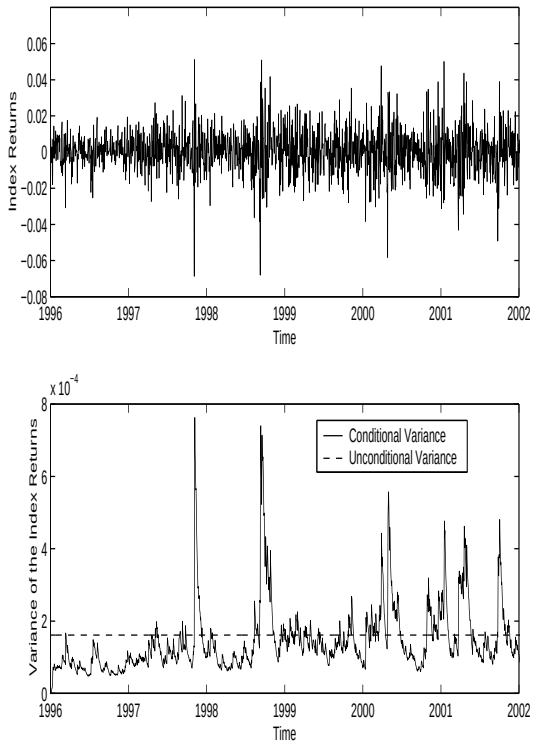

Figure 3: The Estimated Variance of S&P 500 Returns

The problem of this approach is the number of parameters may get too big when k is large. For instance, even if we assume a GARCH(1, 1) process, when $k = 10$, we need to estimate $3 \times 10 \times 11/2 = 165$ parameters.

To solve this problem, we can impose some structures of Ω_t. For instance, Bollerslev (1990) suggested that the conditional correlations are constant over time. Then $\sigma_{ij,t} = \rho_{ij}\sigma_{i,t}\sigma_{j,t}$, with only one parameter, ρ_{ij}, instead of c_{ij}, α_{ij} and β_{ij}.

1.4 Variants of GARCH Models

We will briefly introduce a few other members in the GARCH model family.
IGARCH In a GARCH(p, q) model, when the coefficient satisfy

$$\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j = 1,$$

Engle and Bollerslev (1986) referred to it as integrated GARCH process. In this case, the unconditional variance of ϵ_t is infinite so the process is no longer covariance (weakly) stationary but still strictly stationary.
EGARCH Recall that we let the innovation take the form $\epsilon_t = h_t\nu_t$ with ν_t is i.i.d. with mean zero and unit variance. Nelson (1991) proposed the following specification for h_t.

$$\log h_t = c + \sum_{i=1}^{\infty} \gamma_i(|\nu_{t-i}| - E|\nu_{t-i}| + \theta\nu_{t-i})$$

.

The parameter γ_i captures the effects of the deviation of $|\nu_t|$ from its expectations. A more interesting specification in an EGARCH model is the parameter θ.

We have discussed two features of financial data: dependence in volatility and heavy tails. There is another feature of financial return data: negative skewness. For the normal distribution, we have zero skewness. But for the data set we have used as example, the empirical skewness is -0.1806. In other words, in the financial return data, negative shock tends to have larger volatility than positive shocks. The parameter θ in the EGARCH model can capture this effect. If $\theta = 0$ then the volatilities for positive and negative shocks are symmetric; if $\theta < 0$, then negative shocks tend to have larger effect to the volatility.

There are still other models that belongs to GARCH family, GARCH with threshold (Zakoian 1990), GARCH with regime-switching (Cai 1994), etc.

Readings: Hamilton Ch 21

2 Hidden Markov Chains and Regime Switching Models

2.1 Markov chains

Let s_t be a random variable that only take integer value. If the probability of s_t takes a particular value j depends on the past only through the most recent value s_{t-1}:

$$P\{s_t = j|s_{t-1} = i, s_{t-2} = ...\} = P\{s_t = j|s_{t-1} = i\} = P_{ij}.$$

This process is called a *Markov Chain*, and P_{ij} is called transition probability: the probability that state i will be followed by state j. Suppose there are N states, then we must have

$$\sum_{j=1}^{N} P_{ij} = 1.$$

We can collect the transition probabilities in an $N \times N$ matrix, denoted by P, and it is known as the *transition matrix*:

$$P = \begin{bmatrix} P_{11} & P_{21} & \dots & P_{N1} \\ \vdots & \vdots & \vdots & \vdots \\ P_{1N} & P_{2N} & \dots & P_{NN} \end{bmatrix}.$$

For example, suppose there is a squirrel, who may stay inside a house (in the roof) or stay in the tree (tree by the house). We can specify the transition matrix (P') of this squirrel as:

		t	
		House	Tree
$t-1$	House	0.7	0.3
	Tree	0.1	0.9

To study the forecast of a Markov chain, using our two state example, we can assign integer 1 and 2 to the two states, and we can define

$$\xi_t = \begin{cases} (1,0)' & \text{when} \quad s_t = 1 \\ (0,1)' & \text{when} \quad s_t = 2 \end{cases}$$

Conditional on $s_t = 1$, the expected value of ξ_{t+1} is $(p_{11}, p_{12})'$. Hence we can write one period forecast for the Markov chain as

$$E(\xi_{t+1}|\xi_t, \xi_{t-1}, \dots) = P\xi_t.$$

Hence we can express a Markov chain using a VAR(1) representation

$$\xi_{t+1} = P\xi_t + v_{t+1} \tag{3}$$

where

$$v_{t+1} = \xi_{t+1} - E(\xi_{t+1}|\xi_t, \xi_{t-1}, \dots).$$

It is easy to see that v_t is a martingale difference sequence.

Similarly, the m-period forecast of a Markov chain is

$$E(\xi_{t+m}|\xi_t, \xi_{t-1}, \dots) = P^m\xi_t. \tag{4}$$

Now, suppose $p_{11} = 1$ instead of 0.7, then the squirrel will stay in the roof forever. In this case, the state 'House' is an *absorbing* state and that the Markov chain is *reducible*. On the other hand, if the Markov chain is not reducible, we say it is *irreducible*. For our two-state example, this requires that $P_{11} < 1$, and $P_{22} < 1$.

For a transition matrix P, suppose that one of the eigenvalues is unity and that all other eigenvalues of P are inside the unit circle. Then the Markov chain is said to be *ergodic*. The vector corresponding to the unit eigenvalue is the ergodic probablity (after rescaled so that its elements sum to unity ($\mathbf{1}'\boldsymbol{\pi} = 1$). It can be shown that (Hamilton, page 681)

$$\lim_{m\to\infty} P^m = \boldsymbol{\pi}\cdot\mathbf{1}'.$$

From (4), we can write that

$$E(\xi_{t+m}|\xi_t, \xi_{t-1}, \ldots) = P^m\xi_t \to \boldsymbol{\pi}\cdot\mathbf{1}'\xi_t = \boldsymbol{\pi}.$$

Hence the forecast of ξ_{t+m} converge to $\boldsymbol{\pi}$ no matter what is ξ_t. So, we can see that this $\boldsymbol{\pi}$ is the unconditional probability for the process (the matrix P gives the conditional probability).

In our example of the squirrels, we can compute that $\boldsymbol{\pi} = [0.25, 0.75]$, or, the squirrel stays in the house with about one fourth of the time.

In general, for a two-state Markov chain to be ergodic, besides the conditions for irreducible, which is $P_{11} < 1$, $P_{22} < 1$, we also require $P_{11} + P_{22} > 0$, which means at least one of these two probability is positive. If both probability is zero, then in our example with squirrel, we got that the squirrel jump from the house to the tree and jump from the tree to the house, then at time $t + m$, the position of the squirrel depends on its position at time t. If the squirrel is in the house at time t and m is even number, we know that the squirrel is in the house at time $t + m$. Hence, no matter how large m is, we can always tell where is the squirrel given the position of the squirrel at time t.

2.2 The Hidden Markov Chain and *i.i.d.* mixture distributions

Let s_t be a Markov chain and there are N possible states. Let x_t denote another sequence, and the distribution of x_t at time t depends on s_t. For example, suppose there are two states and s_t take values 1 or 2. When $s_t = 1$, x_t equals 0 with probability 0.9 and equals 1 with probabilty 0.1; while when $s_t = 2$, x_t equals 0 with probability 0.1 and equals 1 with proabbilty 0.9. Further assume that we could not observe s_t and we can only observe x_t, this is a simple example of a *hidden Markov chain*. (draw a picture here)

x_t can also be drawn from a continuous distribution, such as a normal distribution. For example, when $s_t = 1$, x_t is drawn from $N(0, 1)$, and when $s_t = 2$, x_t is drawn from $N(2, 4)$. We write the density of x_t conditional on s_t as follows

$$\begin{aligned} f(x_t|s_t = 1) &= \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{x_t^2}{2}\right], \\ f(x_t|s_t = 2) &= \frac{1}{2\sqrt{2\pi}}\exp\left[-\frac{(x_t-2)^2}{8}\right]. \end{aligned}$$

To compute the unconditional distributions, we need to know the distribution of s_t. For example, if s_t is *i.i.d.*, and $P(s_t = 1) = 1/3$, then the unconditional distribution of x_t is

$$f(x_t) = (1/3)f(x_t|s_t = 1) + (2/3)f(x_t|s_t = 2).$$

Figure (4) plots the density of this mixture distribution as well as those two normal distributions.

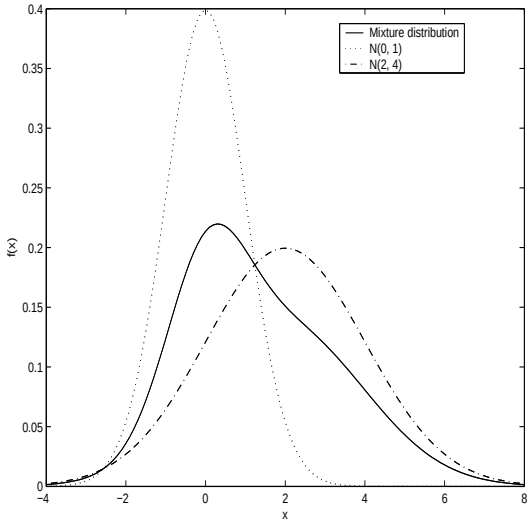

Figure 4: The plot of the mixture density

Now, although we could not observe s_t, we can make an inference for s_t based on x_t. In our example,

$$P(s_t = 1|x_t) = \frac{p(x_t, s_t = 1)}{f(x_t)} = \frac{(1/3) \cdot f(x_t|s_t = 1)}{f(x_t)}. \tag{5}$$

Similarly, we can write this for $P(s_t = 2|x_t)$. From this expression, we see that two factors jointly determine this probability: one is the unconditional probability of s_t, and the other is the probability that each component generate x_t. Consider some numerical examples. Suppose we observe that $x_t = 3$, then we know that $N(2, 4)$ is more likely to generate this observation and also we know that the unconditional probability of $s_t = 2$ is larger. Hence we believe that s_t is much more likely to be 2. Using (5), we can compute that $P(s_t = 1|x_t = 3) = 0.01$, which supports our hypothesis. On the other hand, if we observe that $x_t = -1$, then things are not that clear. Although s_t has larger unconditional probability to be 2, from figure (4) we can see that N(0, 1) has a much larger probability to generate $x_t = -1$ than $N(2, 4)$ distribution. Using (5) we can then compute that the probability that $s_t = 1$ conditional on $x_t = -1$ is about 0.65.

Above is a simple example, where we assume that we know all coefficient, and it largely illustrate how to work with an *i.i.d.* mixture models. In general, if there are N states (N number of individual distributons in the mixture), and if we assume x_t is drawn from $N(\mu_i, \sigma_i^2)$ when $s_t = i$, then we can write

$$f(x_t|s_t = i; \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i^2} \exp\left[-\frac{(x_t - \mu_i)^2}{2\sigma_i^2}\right]. \tag{6}$$

Let π_i denote the probability that $s_t = i$, and let $\theta = (\mu_1, \ldots, \mu_N, \sigma_1^2, \ldots, \sigma_N^2, \pi_1, \ldots, \pi_N)$, then the joint probability density for x_t and $s_t = i$ is

$$p(x_t, s_t = i; \theta) = \pi_i f(x_t|s_t = i; \mu_i, \sigma_i^2),$$

hence the unconditional distribution for x_t is just

$$f(x_t;\theta) = \sum_{i=1}^{N} \pi_i f(x_t|s_t = i; \mu_i, \sigma_i^2).$$

If s_t is *i.i.d.*, the likelihood for observations $\{x_1, \ldots, x_T\}$ is

$$l(\theta) = \sum_{t=1}^{T} \log f(x_t;\theta).$$

We can then solve for the maximum likelihood estimator for θ with the restrictions that $\pi_i \geq 0$ and $\sum_{i=1}^{N} \pi_i = 1$.

Note that to maximize this function, we first take sum over different component and then take log, hence it is not possible to solve them analytically for $\hat{\theta}$ as a function of the data. In empirical studies, MLE of mixture models is computed using the EM algorithm. E represent expectation, and M represent maximization. This is an iterative method, and the likelihood is guaranteed to increase in each iteration.

The MLE estimator for the system can be shown as (P699-701 in Hamilton)

$$\begin{aligned}
\hat{\mu}_i &= \frac{x_t P(s_t = i|x_t;\hat{\theta})}{\sum_{t=1}^{T} P(s_t = i|x_t;\hat{\theta})} \\
\hat{\sigma}_i^2 &= \frac{(x_t - \hat{\mu}_i)^2 \cdot P(s_t = i|x_t;\hat{\theta})}{\sum_{t=1}^{T} P(s_t = i|x_t;\hat{\theta})} \\
\hat{\pi}_i &= T^{-1} \sum_{t=1}^{T} P(s_i = i|x_t;\hat{\theta})
\end{aligned}$$

EM algorithm was originally designed to solve estimation with missing data. In a mixture model, if we know what each observation x_t is drawn from which regime (state), then the problem is much easier. $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ are just the mean and variance computed using the data that from regime i. And $\hat{\pi}_i$ is just the proportion of data from regime i. Since we don't know this information, we can use an iterative algorithm.

We can start with an arbitrary value for θ, denote it θ^0, plug this θ^0 to the right hand side of the above equations, we can obtain a new estimate for θ, denoted by θ^1. We can continue this iteration and stop till θ^m and θ^{m+1} are close.

2.3 Time series regime switching model

Next, we apply the hidden Markov model to time series studies which allow time varying parameters. The idea is that under different regimes, the parameters, which represent the level or relationships, maybe different. For simplicity, we assume that there are two regimes: $s_t = 1$ or $s_t = 2$ and let P denote the transition matrix. We continue to assume that we could not observe s_t, and we can only observe y_t, whose process is specified as

$$y_t = c_{s_t} + \phi y_{t-1} + u_t \tag{7}$$

where $u_t \sim i.i.d.N(0, \sigma^2)$. Then the conditional density of y_t can be written as

$$\boldsymbol{\eta}_t = \begin{bmatrix} f(y_t|s_t = 1, y_{t-1}; \theta) \\ f(y_t|s_t = 2, y_{t-1}; \theta) \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_t - c_1 - \phi y_{t-1})^2}{2\sigma^2}\right] \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_t - c_2 - \phi y_{t-1})^2}{2\sigma^2}\right] \end{bmatrix}. \tag{8}$$

To analyze this system, let's first assume that we know all parameters $\theta = (c_1, c_2, \phi, \sigma^2, p_{11}, p_{22})$. Unlike in the *i.i.d.* case, now we can draw inference about s_t based on all observations. Let Y_t denote observations up to time t, then we can write the conditional probability about s_t at time t as

$$\hat{\xi}_{t|t} = \begin{bmatrix} P(s_t = 1|Y_t; \theta) \\ P(s_t = 2|Y_t; \theta) \end{bmatrix}.$$

Let's first consider how to compute density of y_t conditional on Y_{t-1}. If we take point by point product of $\xi_{t|t-1}$ and $\boldsymbol{\eta}_t$, which can be written as $(\xi_{t|t-1} \odot \boldsymbol{\eta}_t)$, we get

$$\begin{aligned} p(y_t, s_t = 1|Y_{t-1}; \theta) &= P(s_t = 1|Y_{t-1}; \theta) \times f(y_t|s_t = 1, y_{t-1}; \theta) \\ p(y_t, s_t = 2|Y_{t-1}; \theta) &= P(s_t = 2|Y_{t-1}; \theta) \times f(y_t|s_t = 2, y_{t-1}; \theta). \end{aligned}$$

If we add these two elements, we just got the density of y_t conditional on Y_{t-1}, i.e.

$$f(y_t|Y_{t-1}, \theta) = \mathbf{1}'(\xi_{t|t-1} \odot \boldsymbol{\eta}_t). \tag{9}$$

Then the likelihood function can be written as

$$l(\theta) = \sum_{t=1}^{T} \log f(y_t|Y_{t-1}, \theta). \tag{10}$$

To derive a rule in updating the forecast and optimal inference about s_t, note that if we divide each element in $(\xi_{t|t-1} \odot \boldsymbol{\eta}_t)$ by $f(y_t|Y_{t-1}, \theta) = \mathbf{1}'(\xi_{t|t-1} \odot \boldsymbol{\eta}_t)$., we have

$$\frac{p(y_t, s_t = i|Y_{t-1}; \theta)}{\mathbf{1}'(\xi_{t|t-1} \odot \boldsymbol{\eta}_t)} = \frac{p(y_t, s_t = i|Y_{t-1}; \theta)}{f(y_t|Y_{t-1}, \theta)} = P(s_t = i|y_t, Y_{t-1}; \theta) = P(s_t = i|Y_t; \theta).$$

We can do this for each element in the vector and obtain that

$$\hat{\xi}_{t|t} = \frac{(\xi_{t|t-1} \odot \boldsymbol{\eta}_t)}{\mathbf{1}'(\xi_{t|t-1} \odot \boldsymbol{\eta}_t)}. \tag{11}$$

Finally, if we take expectation of (3) conditional on Y_t, we have

$$\hat{\xi}_{t+1|t} = P \cdot \hat{\xi}_{t|t}. \tag{12}$$

These two equations, (11) and (12) compose an iterating algorithm to compute the optimal inference for s_t. This iteration starts from $\xi_{1|0}$, which can be specified in several ways (see page 693 in Hamilton).

So far when we draw inference about s_t, we base on information up to time t. While, as we obtain more information and look back, we may have different ideas about what happened at time t. Such an inference, say, $P(s_t = i|Y_\tau; \theta)$ for $\tau > t$, is called the *smoothed inference*.

Above we assume that we know the parameter θ. To estimate θ, we can find the estimator that maximizes the likelihood (10), using some numerical optimization techniques.