# A Multivariate Generalization of the Markov Switching Model

## with an application to volatility clusters

BY MOHAMAD KHALED

Paris School of Economics and University of Paris I Panthéon-Sorbonne

Job Market Paper

*September 2008*

**Abstract**

I present a multivariate generalization of the simple markov-switching model. I allow for the introduction of several latent processes that have a simple parametric distribution. The matrix-variate bernoulli distribution yields a flexible yet parsimonious pattern of dependence between the different latent processes while preserving the markovian property. I derive several analytic results and show how to compute quantities such marginal and conditional distributions. I also show how to estimate the model in the bayesian framework and give several examples.

I then apply the approach to multivariate volatility clustering models. In the usual approaches to the problem, volatility clusters need either occur simultaneously in different series or be completely independent across those series. Contrary to those approaches, the framework in the paper allows for a rich pattern of dependence in the volatility clusters taking place across different variables.

# 1 Introduction

The markov-switching regression model has proved to be a useful tool in econometrics over the past two decades. Generalizations of the model to the multivariate case have been undertaken in several papers (see for instance [Krolzig, 1997] and [Sims and Zha, 2006].) However, there are several shortcomings of the current literature of multivariate markov-switching models. For instance, the usual generalizations either assume the existence of a single rudimentary latent process underlying the model or suppose some simple way of combining different latent processes such as assuming they are independent.

Before discussing the econometric aspects of those details, I shall give some economic motivation that justifies addressing those shortcomings. Let us consider a cornerstone of markov-switching models in econometrics and study its implications in the multivariate setting. The GDP growth model of the seminal paper

[Hamilton, 1989] formulates a markov-switching model in which the mean of the GDP growth autoregressive process is time-varying. There are two states, each of which corresponds respectively to each one of the two-values $\mu_1$ and $\mu_2$ assumed by the time-varying mean with $\mu_1 < \mu_2$. The standard economic interpretation is that one state (say that corresponding to $\mu_1$) represents economic recessions and the other one corresponds to economic expansions. One observes that the duration of the recession state is smaller than that of the expansion state. Now consider going to the multivariate setting and taking a vector of country GDP growth series, say, the U.S., Canada and France. On one hand, an economist thinking about each individual country's expansion and recession states might think of them as occurring simultaneously with high probability due to the inter-connection of their economies, that is, due to close economic ties between say the U.S. and Canada, recessions are likely to hit both countries at the same time. On the other hand, the same economist can not exclude that some recessions might occur in some countries and not others, that they might last less in one country than another or that there might exist some delay effects. For instance, one might think that recessions occur more often simultaneously in the U.S. and Canada than in the U.S. and France. Turning back to the multivariate markov switching model, it is difficult to think of an easy way of incorporating those features in it. If one wants to be flexible, one should consider a binary latent variable for each country. This will create a markov-switching model with a total of $2^3 = 8$ states (that is, respectively for the U.S.-Canada-France, the states expansion-expansion-recession, expansion-recession-expansion, expansion-expansion-expansion etc...)). This is quite unwieldy and moreover there is a total of $2^{2\times 3} - 2^3 = 56$ transition probability parameters to estimate. Those parameters are very difficult to interpret and it is their estimation that will characterize the simultaneity and duration features of those states. One common assumption used in multivariate markov-switching models but that is completely absurd from an economic point of view is to assume those latent variables to be independent. That will greatly reduce the number of parameters, make the interpretation easier, but to say that a recession in the U.S. is completely independent of a recession in Canada will be immediately rejected by any economist with a common sense. A third solution that is often used is to simplify the model by supposing a common latent process and therefore forcing recessions to occur all at the same time in all countries. This economic framework is described in more technical terms in appendix C.

The model that I propose will solve all problems exposed above with no loss in flexibility and with great parsimony. First, the number of parameters in my framework grows at the rate $O(m^2)$ and not at the prohitive rate $O\left(2^{2m}\right)$ mentioned above. The economist will get parameters that are meaningful and easy to interpret and further utilize from an economic perspective, that is, there will be parameters that describe the strength of simultaneity of recession occurrences between each pairs of country. In other words, he will not have to rely on the complex transition parameters mentioned earlier to study this phenomenon but rather refer to formulating some immediately interpretable parameters characterizing the dependence structure of the latent variables. This will be explored in more details in the next sections.

As another possible economic application, consider modeling financial returns series. A financial economist might consider a markov-switching model for the returns series of a given financial asset whereby he studies the time-varying nature of the series variance. Volatility clusters are a natural phenomenon for returns series. The variance seems to take a jump up suddenly, stays high for a while, takes a jump down, stays low for a while and so forth. I shall call those forms volatility clusters. This could naturally be described by a markov-switching model on the returns series with a time-varying volatility. The episodes of sustained peaks of volatility are often attributed to a more volatile economic environment. I shall simply refer to those states as ones of "high speculation" in an obvious abuse of language. When going to the multivariate setting, that is considering several financial assets at the same time, a financial economist might be interested in describing the nature of occurrence of volatility clusters across different assets. Volatility clusters occurring simultaneously in different markets can describe spill-over effects and financial contamination as investors behave similarly in those markets. At the same time, very different financial assets might display different behaviors when it comes to volatility clusters. This is a natural application where traditional multivariate markov-switching models usually face great difficulties and where my multivariate frameworks will find a natural simple illustrative application. I consider this in more details in section 4 and illustrate with real data in section 5.

The main contributions of the paper are the following. I formulate a new multivariate markov-switching model that is flexible and parsimonious and fill a gap in the multivariate-switching literature. The model relies on a parametric structure of the latent variables that is difficult to handle and that is known as the matrix-variate bernoulli distribution. I derive new results concerning the constant of integration of the matrix-variate bernoulli distribution and several marginal and conditional distributions in closed form which will make possible inference both from the frequentist and bayesian perspectives. I illustrate with an application to multivariate volatility clusters and concentrate on bayesian inference.

# 2 The Model

I consider here a multivariate generalization of the markov-switching model taking the following form

$$\begin{cases} \boldsymbol{y}_t|\boldsymbol{x}_t & \sim & f(\boldsymbol{\phi}_{\boldsymbol{z}_t}) \\ \boldsymbol{Z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_T)' & \sim & \mathcal{MB}(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) \end{cases}$$

In the first equation $\boldsymbol{y}_t$ is a vector of dependent variables, $\boldsymbol{x}_t$ is a vector of covariates and $\boldsymbol{z}_t$ is an $m \times 1$ vector of binary latent variables. $\boldsymbol{\phi}$ is a $p \times 1$ vector of parameters. Notice that $\boldsymbol{\phi}$ is indexed by $\boldsymbol{z}_t$, which means here that $\boldsymbol{\phi}$ takes as many different values as $\boldsymbol{z}_t$ can take. Since $\boldsymbol{z}_t$ is an $m$-dimensional vector of binary variables, there are therefore $2^m$ different vectors of parameters $\boldsymbol{\theta}_{\boldsymbol{z}_t}$. The first equation can be considered as the measurement or observation equation in a state-space model setting.

The second equation, $\boldsymbol{Z}$ is a $T \times m$ matrix where the $t$-th row contains $\boldsymbol{z}_t$. $\boldsymbol{Z}$ is the matrix of latent variables and is distributed as first order markov matrix-variate bernoulli with parameter matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}$. More details on this will come in the next section. The parameter matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}$ characterize the dependence structure of the latent variables. In a state-space model setting, the second equation can be considered as the law of motion or state equation.

The econometrician who uses that model start with a given parametric modeling framework $\boldsymbol{y}_t|\boldsymbol{x}_t \sim f(\boldsymbol{\phi})$ and generalizes it by making some or all of the parameters in $\boldsymbol{\phi}$ time-varying. An interesting feature is that, due to $\boldsymbol{z}_t$ being multivariate, subsets of the parameters vector $\boldsymbol{\phi}$ can be made to be time-varying differently from other subsets. We shall illustrate this unique feature with several examples.

**Example 1.** As a simple illustrative example that should clarify the notation, take $m = 2$, $p = 3$ and $\boldsymbol{\phi} = (\psi, \xi, \zeta)'$. $\boldsymbol{z}_t$ can take here four different values for each time period $t$, that is $(0, 0)'$, $(0, 1)'$, $(1, 0)'$ and $(1, 1)'$. In this example, we decide to make the latent variables index $\xi$ and $\zeta$ separately without indexing $\psi$. $z_{t,1}$ and $z_{t,2}$ indexes $\xi$ and $\zeta$. Here, there are the different values allowed here

| $\boldsymbol{z}_t$ | $\boldsymbol{\phi}_{\boldsymbol{z}_t}$ |
|---|---|
| $(0,0)'$ | $(\psi, \xi_1, \zeta_1)'$ |
| $(0,1)'$ | $(\psi, \xi_1, \zeta_2)'$ |
| $(1,0)'$ | $(\psi, \xi_2, \zeta_1)'$ |
| $(1,1)'$ | $(\psi, \xi_2, \zeta_1)'$ |

The previous example shows the flexibility that is allowed by our model. Other models could have been also allowed such that making $z_{t,1}$ index both $\psi$ and $\zeta$ and $z_{t,2}$ index $\xi$, or making $z_{t,1}$ index both $\psi$ and $\xi$ and $z_{t,2}$ index $\zeta$.

$m$ and the different possible configurations are chosen by the econometrician to conform with his modeling decisions.

An important special case multivariate markov-switching model is the multivariate regressions special case. This takes the following form

$$\begin{cases} \boldsymbol{y}_t = \boldsymbol{x}_t.\boldsymbol{\beta}_{\boldsymbol{z}_t} + \boldsymbol{u}_t \\ \\ \boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{z}_t}) \end{cases}$$

where in that case $\boldsymbol{\phi} = (\boldsymbol{\beta}, \text{vech}(\boldsymbol{\Sigma}))$ and $\boldsymbol{u}_t$ is an error term defined as the difference between $\boldsymbol{y}_t$ and $\boldsymbol{x}_t.\boldsymbol{\beta}_{\boldsymbol{z}_t}$. Two important examples that belong to that special case of models are the following.

**Example 2. Regression coefficients and error variance that are separately switching**

This is a very simple illustrative problem

$$y_t = \boldsymbol{x}_t.\boldsymbol{\beta}_{z_{t,1}} + \sigma_{z_{t,2}}.\varepsilon_t$$

$$\varepsilon_t \sim \mathcal{N}(0,1)$$

$\boldsymbol{\beta}$ and $\sigma$ depend on two different latent chains $z_{1,t}$ and $z_{2,t}$ that are dependent.

Conditional on $\boldsymbol{Z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_T)'$, the problem simplifies to the heteroscedastic regression model

$$y_t = \boldsymbol{x}_t.(1 - z_{t,1}).\boldsymbol{\beta}_1 + \boldsymbol{x}_t.z_{t,1}.\boldsymbol{\beta}_2 + u_t$$

$$\boldsymbol{u} = (u_1, ..., u_T)' \sim \mathcal{N}(0, \boldsymbol{\Omega})$$

$$\boldsymbol{\Omega} = \text{diag}(\omega_1, ..., \omega_T)$$

$$\omega_t = h(z_{t,2}, \sigma_1^2, \sigma_2^2) = \sigma_1^2.(1 - z_{t,2}) + \sigma_2^2.z_{t,2}$$

To see how to implement the part of the Gibbs sampler that conditions on $\boldsymbol{Z}$, see e.g. [Bauwens et al., 1999] and [Koop, 2003].

In the next example, the latent variables configuration in the sense that each latent underlies a single equation in the system of equations of the multivariate regression model.

**Example 3. A multivariate regression model with a different chain underlying each equation**.

$$\begin{pmatrix} y_{t,1} & y_{t,2} & y_{t,3} \end{pmatrix} = \boldsymbol{x}_t.\begin{pmatrix} \boldsymbol{\beta}_{1,z_{t,1}} & \boldsymbol{\beta}_{2,z_{t,2}} & \boldsymbol{\beta}_{3,z_{t,3}} \end{pmatrix} + \boldsymbol{u}_t$$

$$\boldsymbol{u}_t \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} \sigma_{1,z_{t,1}}^2 & & \\ 0 & \sigma_{2,z_{t,2}}^2 & \\ 0 & 0 & \sigma_{3,z_{t,3}}^2 \end{pmatrix}\right)$$

one can think of the system as a collection of univariate markov-switching regressions depending each on a distinct latent process $z_{i,t}$ for $i = 1, ..., 3$. The latent processes are correlated each one with another and moreover, the residuals from the different regressions are not independent.

I can also consider a model where the covariances are not zeros.

$$\boldsymbol{u}_t \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} \sigma_{1,z_{t,1}}^2 & & \\ \sigma_{21} & \sigma_{2,z_{t,2}}^2 & \\ \sigma_{31} & \sigma_{32} & \sigma_{3,z_{t,3}}^2 \end{pmatrix}\right)$$

In the next section, I will study the first-order markov matrix-variate bernoulli distribution and its use for describing the structure of the latent variables. We shall also derive several results that will prove essential for inference purposes.

# 3 The probability structure of the latent processes

In this section, I describe in detail the way to construct the joint distribution of the latent processes. Subsection 3.1 reviews the matrix-variate bernoulli distribution as introduced in [Lovison, 2006]. In subsection 3.2, I introduce our concept of "augmented kernel vector". That concept is crucial since it is going to allow us to easily prove several of our results here and it will allow us the possibility of compactly and neatly writing several of our results.

## 3.1 The matrix-variate bernoulli distribution

This subsection quickly reviews the paper of [Lovison, 2006] in which the matrix-variate bernoulli distribution was introduced. We then present those aspects of the distribution that we will require for our modeling of the markov-switching multivariate regression. Throughout the section, we will attempt to be as close as possible to [Lovison, 2006]'s initial notation.

As described earlier, we want to model $m$ different binary latent processes consisting each of $T$ observation. Therefore, the typical random variable considered is a matrix $\boldsymbol{Z}$ of size $T \times m$ where each column of the matrix corresponds to a different latent process. Throughout the paper, we will denote the $t$-th row of $\boldsymbol{Z}$ by $\boldsymbol{z}_t$ for $t = 1, ..., T$ and we will denote the $j$-th column of $\boldsymbol{Z}$ by $\boldsymbol{z}_{(:,j)}$ for $j = 1, ..., m$. Equally, we will denote the $(i, j)$-th entry of $\boldsymbol{Z}$ by $z_{i,j}$.

The matrix-variate bernoulli distribution allows for different patterns of dependence. Those include dependence between different variables (simple column-wise dependence between, say, $\boldsymbol{z}_{(j,:)}$ and $\boldsymbol{z}_{(j',:)}$ for $j' \neq j$), dependence between different periods in time (or, put differently, observational unit dependence between, say, $z_{t,j}$ and $z_{s,j}$ for $s \neq t$) and finally mixed variable-unit dependence (say between $z_{t,j}$ and $z_{s,j'}$). In this paper, I will not address mixed variable-unit dependence.

### 3.1.1 The density and dependence parameters

The density of $\boldsymbol{Z}$ is equal to

$$p(\boldsymbol{Z}\,|\boldsymbol{\Psi}) = \frac{1}{K_T(\boldsymbol{\Psi})}.\exp\left\{\text{vec}(\boldsymbol{Z}')'.\boldsymbol{\Psi}.\text{vec}(\boldsymbol{Z}')\right\}$$

where $\boldsymbol{\Psi}$ is a $T\,m \times T\,m$ matrix that contains the density parameters. $K_T(\boldsymbol{\Psi})$ is the integration constant that depends on $\boldsymbol{\Psi}$ and that is given by the following formula

$$K_T(\boldsymbol{\Psi}) = \left(\sum_{k=1}^{2^{Tm}} \exp\left\{\text{vec}(\boldsymbol{Z}_k')'.\boldsymbol{\Psi}.\text{vec}(\boldsymbol{Z}_k')\right\}\right)$$

where each $\boldsymbol{Z}_k$ represent one possible $\boldsymbol{Z}$ matrix among all the $2^{Tm}$ possible such matrices.

The matrix $\boldsymbol{\Psi}$ groups all the parameters that describe the dependence or association structure of the different $z_{t,j}$ random variables (for $t = 1, ..., T$ and $j = 1, ..., m$). Therefore, it has a very special structure that will be explicitly defined after describing the different dependence patterns allowed in the matrix-variate bernoulli distribution.

The patterns of dependence allowed can be classified in three categories. The three categories of parameters associated with each pattern will be given a distinct notation each in order to distinguish between them

- The parameters describing pure variable association

$$\theta_t^{j,j'} = \theta^{j,j'}, \forall t$$

  Those parameters describe the dependence between two different variables, i.e. between two different columns of $\boldsymbol{Z}$.

- The parameters describing pure unit (or observational) association

$$\lambda_{t,t'}^{j}$$

  Within a single column of $\boldsymbol{Z}$, say the $j$-th column, i.e the $j$-th variable, those parameters describe the dependence pattern between two different rows $t$ and $t$, i.e. two different observations.

- The parameters describing mixed unit-variable association

$$\phi_{t,t'}^{j,j'}$$

  The parameter $\phi_{t,t'}^{j,j'}$ describes the dependence between the $t$-th observation of the $j$-th variable and the $t'$-th observation of the $j'$-th variable. I.e., it describes the dependence between the $(t,j)$-th entry of $\boldsymbol{Z}$ and its $(t',j')$-th entry.

In addition to the parameters describing patterns of dependence or association, we need some parameters that describe the within probabilistic structure of each variable or column of $\boldsymbol{Z}$. I.e., more concretely, we need parameters that describe the $1-0$ frequency whiting each variable. For that, we use the parameters

$$\theta_t^j = \theta^j, \forall t$$

The difference between parameters $\theta^j$ (a single index) and parameters $\theta^{j,j'}$ (two indeces) is that $\theta^j$ characterizes the overall $0-1$ pattern within variable $j$ and that $\theta^{j,j'}$ characterizes the overall contemporaneous (i.e for the same observation $t$) dependence between variables $j$ and $j'$.

It is possible to tidy the presentation up by putting the parameters into matrices. Let us introduce the following symmetric $m \times m$ matrices $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}_{t,t'}$

$$\boldsymbol{\Theta} = \begin{pmatrix} \theta^1 & \theta^{1,2} & \cdots & \theta^{1,m} \\ & \theta^2 & & \\ & & \ddots & \\ & & & \theta^m \end{pmatrix}$$

$$\boldsymbol{\Lambda}_{t,t'} = \boldsymbol{\Lambda}'_{t',t} = \begin{pmatrix} \lambda_{t,t'}^1 & \phi_{t,t'}^{1,2} & \cdots & \phi_{t,t'}^{1,m} \\ & \lambda_{t,t'}^2 & & \vdots \\ & & \ddots & \vdots \\ & & & \lambda_{t,t'}^m \end{pmatrix}$$

Now, it is possible to define the parameter matrix $\boldsymbol{\Psi}$ of the matrix-variate bernoulli distribution as a function of $\boldsymbol{\Theta}$ and the $\boldsymbol{\Lambda}_{t,t'}$s

$$\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Theta} & \boldsymbol{\Lambda}_{1,2} & \cdots & \boldsymbol{\Lambda}_{1,T-1} & \boldsymbol{\Lambda}_{1,T} \\ \boldsymbol{\Lambda}_{2,1} & \boldsymbol{\Theta} & & \vdots & \boldsymbol{\Lambda}_{2,T} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ \boldsymbol{\Lambda}_{T-1,1} & \vdots & & \boldsymbol{\Theta} & \boldsymbol{\Lambda}_{T-1,T} \\ \boldsymbol{\Lambda}_{T,1} & \boldsymbol{\Lambda}_{T,2} & \cdots & \boldsymbol{\Lambda}_{T,T-1} & \boldsymbol{\Theta} \end{pmatrix} = \boldsymbol{I}_T \otimes \boldsymbol{\Theta} + \sum_{t=1,t'\neq t}^{T} \boldsymbol{E}_{t,t'} \otimes \boldsymbol{\Lambda}_{t,t'}$$

The $\boldsymbol{E}_{t,t'}$ matrices are such that all of their entries contain zeros except for the ($t$, $t'$)th ones which contain a one.

One can factor the likelihood $p(\boldsymbol{Z}\,|\boldsymbol{\Theta}, ..., \boldsymbol{\Lambda}_{t,t'}, ...)$ so as to write it in the following form

$$\frac{1}{K_T(\boldsymbol{\Theta}, ..., \boldsymbol{\Lambda}_{t,t'}, ...)} \exp\{\mathrm{tr}[\boldsymbol{Z}'.\boldsymbol{Z}.\boldsymbol{\Theta}]\} \prod_{t=1,t'\neq t}^{T} \exp\{\mathrm{tr}[\boldsymbol{Z}'.\boldsymbol{E}_{t,t'}.\boldsymbol{Z}.\boldsymbol{\Lambda}_{t,t'}]\}$$

After writing the likelihood in that form, we can immediately see that the quantities

$$\boldsymbol{Z}'.\boldsymbol{Z}$$

and each one of

$$\boldsymbol{Z}'.\boldsymbol{E}_{t,t'}.\boldsymbol{Z}$$

are jointly sufficient statistics for $\boldsymbol{\Theta}$ and each one of the $\boldsymbol{\Lambda}_{t,t'}$ respectively.

As a matter of fact, we can describe the parameters by sufficiency. We are going to explain in further detail in subsection 3.1.4. We delay that interpretation because we are more interested in the special case of the first-order markov case and in that case, the interpretation is simpler.

Another possible interpretation is through the use of "conditional" log-odds ratios. This is quickly resumed in subsection 3.1.2.

In this paper, I will only address the special case of the matrix-variate bernoulli distribution, that is the first-order markov case (that case was give as an example in the paper of [Lovison, 2006]). Subsection 3.1.3 is totally devoted to that task.

For further details on those general parameterizations, see [Lovison, 2006] (and e.g. [Cox, 1972], [Zhao and Prentice, 1990] or [Cox and Wermuth, 1994]).

### 3.1.2 "Conditional" log-odds ratio interpretation of the dependence parameters

One sees from the definition of the dependence parameters that the distribution allows for pairwise interactions only. That might prove a formidable restriction in certain applications, but in our case, the distribution offers exactly what we need.

The way each of those parameters is described is through the use of odds ratios. For instance $\phi_{t,t'}^{j,j'}$ can be written as

$$\phi_{t,t'}^{j,j'} = \log\left\{ \frac{P\{z_{t,j}=1, z_{t',j'}=1\}.P\{z_{t,j}=0, z_{t',j'}=0\}}{P\{z_{t,j}=1, z_{t',j'}=0\}.P\{z_{t,j}=0, z_{t',j'}=1\}} \right\}$$

The probabilities in the log-odds ratios are conditional on the rest being zero.

The $\theta^j$s are written as

$$\theta_t^j = \theta^j = \log\left\{ \frac{P\{z_{t,j}=1\}}{P\{z_{t,j}=0\}} \right\}, \forall t$$

and again the probabilities appearing in the fractions are conditional on the rest being zero.

Although conditional log odds ratios constitute a neat and elegant way of interpreting the dependence parameters, we prefer to use sufficiency for that purpose. In particular, since the distribution used by those log odds ratios is conditional on all the other entries in the matrix $\boldsymbol{Z}$ being set to zero, it might prove less useful in economic applications than the interpretation that relies on the concept sufficiency. We will explore that idea in more detail in subsection 3.1.4.

### 3.1.3 The first-order markov case

Now I am going to illustrate the first-order markov special case. The density of the matrix-variate bernoulli distribution simplifies greatly since, in order to account for temporal dependence of the first order, one will only need a single parameter within each latent process only. That is, $z_{t,j}$ and $z_{s,j}$ will dependent if and only if $s = t + 1$ for $s \geqslant t$. Therefore, the parameters reflecting markovian dependence between units in each latent process are

$$\lambda_{t,s}^j = \begin{cases} \lambda^j & \text{if } s = t+1 \quad \text{for } s \geqslant t \\ 0 & \text{otherwise} \end{cases}$$

On the other hand, the contemporaneous dependence between latent process $j$ and latent process $j'$ will be captured by $\theta^{j,j'}$.

Mixed dependence of the first order can also be allowed (i.e. dependence in one chain on past values of other chains)

$$\phi_{t,t'}^{j,j'} = \begin{cases} \phi^{j,j'} & \text{if } t' = t+1 \\ 0 & \text{otherwise} \end{cases}$$

However, for the purposes of our paper, I will set all mixed dependence parameters to zero.

In matrix notation

$$\boldsymbol{\Lambda}_{i,i+1} = \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1^1 & \phi_1^{1,2} & \cdots & \phi_1^{1,m} \\ & \lambda_1^2 & & \vdots \\ & & \ddots & \vdots \\ & & & \lambda_1^m \end{pmatrix}$$

and since I set all mixed dependence parameters zero, the matrix $\mathbf{\Lambda}$ will be diagonal

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1^1 & 0 & \cdots & 0 \\ & \lambda_1^2 & & \vdots \\ & & \ddots & \vdots \\ & & & \lambda_1^m \end{pmatrix}$$

Similarly, the matrix $\mathbf{\Theta}$ is defined as in the general case

$$\mathbf{\Theta} = \begin{pmatrix} \theta^1 & \theta^{1,2} & \dots & \theta^{1,m} \\ & \theta^2 & & \\ & & \ddots & \\ & & & \theta^m \end{pmatrix}$$

Therefore, for the general notations of the matrix-variate bernoulli distribution, the general matrix containing all parameters can, as a result, been written as

$$\mathbf{\Psi} = \boldsymbol{I}_T \otimes \mathbf{\Theta} + \boldsymbol{L}_1 \otimes \mathbf{\Lambda}$$

where $\boldsymbol{L}_1$ is a matrix containing ones on the first right off-diagonal

$$\boldsymbol{L}_1 = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}$$

The likelihood of the first order matrix-variate bernoulli distribution is therefore

$$p(\boldsymbol{Z}|\mathbf{\Theta}, \mathbf{\Lambda}) = \frac{1}{K_T(\mathbf{\Theta}, \mathbf{\Lambda})}.\exp\{\mathrm{tr}[\boldsymbol{Z}'.\boldsymbol{Z}.\mathbf{\Theta}] + \mathrm{tr}[\boldsymbol{Z}'.\boldsymbol{L}_1.\boldsymbol{Z}.\mathbf{\Lambda}_1]\}$$

where $\boldsymbol{Z}'.\boldsymbol{Z}$ and $\boldsymbol{Z}'.\boldsymbol{L}_1.\boldsymbol{Z}$ are the sufficient statistics for $\mathbf{\Theta}$ and $\mathbf{\Lambda}$ respectively.

I will hereafter refer to the first-order markov matrix-variate bernoulli distribution through the abbreviation FOMMVB.

### 3.1.4 Interpretation of the dependence parameters

The formula for the joint density is all that one needs to interpret the parameters. Instead of interpreting the general case with $m$ variables, I will explicitly write the univariate and bivariate cases so as to shed some light on the distribution.

In the univariate case, $\boldsymbol{Z}$ is of size $T \times 1$ and its probability density function is given by

$$p(\boldsymbol{Z}|\theta, \lambda) = \frac{1}{K_T(\theta, \lambda)}.\exp\left(\left[\sum_{t=1}^{T} z_t\right].\theta + \left[\sum_{t=2}^{T} z_t.z_{t-1}\right].\lambda\right)$$

where one commonly writes $z_{t,1}$ as $z_t$ and where $\sum_{t=1}^{T} z_t^2 = \sum_{t=1}^{T} z_t$ since $z_t$ is binary.

Here we see how $\theta$ and $\lambda$ are defined by sufficiency. To estimate $\theta$, a sufficient statistic ( $\sum_{t=1}^{T} z_t$) is the number of times the state labeled by 1 has occurred. Similarly, to estimate $\lambda$ a sufficient statistic ( $\sum_{t=2}^{T} z_t.z_{t-1}$) is the number of times the state labeled as one was followed in time by the same state, i.e. the number of times that 1 occurs consecutively in the column of the matrix is sufficient to estimate $\lambda$.

For the univariate case, the $T \times 2$ matrix $\boldsymbol{Z}$ has the following density

$$\begin{aligned} p(\boldsymbol{Z}|\theta^1,\theta^2,\theta^{1,2},\lambda^1,\lambda^2) &= \frac{1}{K_T(\theta^1,\theta^2,\theta^{1,2},\lambda^1,\lambda^2)} \\ &\times \exp\left(\left[\sum_{t=1}^{T} z_{t,1}\right].\theta^1 + \left[\sum_{t=2}^{T} z_{t,1}.z_{t-1,1}\right].\lambda^1\right. \\ &+ \left[\sum_{t=1}^{T} z_{t,2}\right].\theta^2 + \left[\sum_{t=2}^{T} z_{t,2}.z_{t-1,2}\right].\lambda^2 \\ &+ \left.\left[\sum_{t=1}^{T} z_{t,1}.z_{t,2}\right].\theta^{1,2}\right) \end{aligned}$$

Here, the interpretation of $(\theta^1, \theta^2, \lambda^1, \lambda^2)$ is the same as in the univariate case. As for $\theta^{1,2}$, the interpretation is also given by sufficiency. The sufficient statistic for estimation $\theta^{1,2}$ is $\sum_{t=1}^{T} z_{t,1}.z_{t,2}$ which the number of times that the state labeled by 1 occurred simultaneously for both latent processes.

## 3.2 Augmented kernel vector

The main point is that, in order to compute several important quantities such as marginal densities and such as the constant of integration, one needs to integrate over several elements of $\boldsymbol{Z}$. Integration with respect to several elements of $\boldsymbol{Z}$, e.g. with respect to $\boldsymbol{z}_t$, will yield a functional expression that is different from the kernel of matrix-variate bernoulli distribution (i.e. the density without dividing by the constant of integration.) It turns out that it is possible to write, for each date $t$, a vector of size $2^m$ which has the unique property that, if one somehow integrates one of its entries with respect to $\boldsymbol{z}_t$, then we will obtain a linear combination of the same vector at date $t-1$. That is, those vectors of size $2^m$ are somehow "closed" to the integration operation. I will dub those vectors as "augmented kernel vectors".

Before writing the main result of the paper, I will try to justify the use of the term kernel vector.

The kernel (i.e. the density function without the normalizing constant) of the FOMMVB distribution is equal to

$$\exp\{\mathrm{tr}[\boldsymbol{Z}'.\boldsymbol{Z}.\boldsymbol{\Theta}] + \mathrm{tr}[\boldsymbol{Z}'.\boldsymbol{L}_1.\boldsymbol{Z}.\boldsymbol{\Lambda}_1]\}$$

In less compact notation but one that might prove helpful for the purpose of introducing the concept, we can write the kernel as

$$\exp\left[\sum_{i=1}^{m} y_i^T.\theta^i + \sum_{i=1}^{m} w_i^T.\lambda^i + 2\sum_{i=1, j\neq i}^{m} u_{i,j}^T.\theta^{i,j}\right]$$

where I introduced the following notation

**Notation 1.** *Define the following quantities*

$$y_i^t \triangleq \sum_{\tau=1}^{t} z_{\tau,i}$$

$$w_i^t \triangleq \sum_{\tau=2}^{t} z_{\tau,i}.z_{\tau-1,i}$$

$$u_{i,j}^t \triangleq \sum_{\tau=1}^{t} z_{\tau,i}.z_{\tau,j}$$

*Obviously*

$$\boldsymbol{Z}'_{1:t}.\boldsymbol{Z}'_{1:t} = \begin{pmatrix} y_1^t & \cdots & u_{1,i}^t & \cdots & u_{1,m}^t \\ \vdots & \ddots & & & \vdots \\ \vdots & & y_i^t & & \vdots \\ \vdots & & & \ddots & \vdots \\ u_{m,1}^t & \cdots & \cdots & \cdots & y_m^t \end{pmatrix}$$

*where $\boldsymbol{Z}_{1:t}$ denote the first $t$ rows of $\boldsymbol{Z}$.*

*We also have obviously that*

$$\boldsymbol{Z}'_{1:t}.\boldsymbol{L}_1.\boldsymbol{Z}_{1:t} = \begin{pmatrix} w_1^t & * & \cdots & \cdots & * \\ * & \ddots & \ddots & * & \vdots \\ \vdots & * & w_i^t & * & \vdots \\ \vdots & * & \ddots & \ddots & * \\ * & * & \cdots & * & w_m^t \end{pmatrix}$$

*where of course $\boldsymbol{L}_1$ is of dimension $t \times t$ here and where the blanks $*$ refers to terms that we would not need to define since $\boldsymbol{\Lambda}$ is a diagonal matrix and we only care about the trace of $\boldsymbol{Z}'_{1:t}.\boldsymbol{L}_1.\boldsymbol{Z}_{1:t}.\boldsymbol{\Lambda}$.*

If one integrates the kernel with respect to $\boldsymbol{z}_T$, then we will obtain a linear combination of exponential terms that do not have the same functional form as the kernel (which it is itself an exponential term). The only difference are the multiplicands of the different $\lambda^i$ for $i = 1, ..., m$, that is, we will have expressions such as

$$(z_{t,i} + w_i^t).\lambda^i \text{ for } i = 1, ..., M$$

inside the exponential terms.

In particular, we will prove that we will have $2^m$ different such terms.

Moreover, there exists a one-to-one and onto mapping from the power set of $\mathcal{M}$ into those terms. I will begin by providing the following definition.

**Definition 1.** *Let us define the set*

$$\mathcal{M} \triangleq \{1, ..., m\}$$

*and its power set as $\mathcal{P}(\mathcal{M})$.*

*Define the index function that maps each element of* $\mathcal{P}(\mathcal{M})$ *into the integer set* $\{1, ..., 2^m\}$

$$\varrho: \mathcal{P}(\mathcal{M}) \longrightarrow \{1, ..., 2^m\}$$

*For instance, one can map the null set into the integer 1, i.e.* $\varrho(\phi) = 1$, *the set* $\{1\}$ *into the integer 2, i.e.* $\varrho(\phi) = 2$ *etc...*

*Also, consider the row vector* $\boldsymbol{z}_t$ *of size* $1 \times m$. *There are* $2^m$ *different possible such vectors. Define that set as* $\mathcal{Z}$. *Therefore, there exists a one-to-one and onto mapping from either the index set* $\{1, ..., 2^m\}$ *or the power set* $\mathcal{P}(\mathcal{M})$ *into the set* $\mathcal{Z}$. *Define that mapping as* $\mathcal{C}$.

$$\mathcal{C}: \mathcal{Z} \longrightarrow \{1, ..., 2^m\}$$

*As an example, suppose* $m = 2$ *and* $\boldsymbol{z}_t = (0, 0)$, *then we can consider that* $\mathcal{C}(\boldsymbol{z}_t) = 1$.

*Moreover, we can easily use* $\varrho^{-1} \circ \mathcal{C}$ *and* $\mathcal{C}^{-1} \circ \varrho$ *as one-to-one and onto mappings between* $\mathcal{Z}$ *and* $\mathcal{P}(\mathcal{M})$.

It is important to be able to efficiently construct one such mapping as $\mathcal{C}$. We will show one such way in appendix B. Moreover, the algorithm given in appendix B will implicitly show how to undertake several of the numerical computations associated with the analytic ones exposed in the paper.

Using the previous notation, we can, for instance, write the kernel of the FOMMVB as

$$\exp\left[\sum_{i \in \mathcal{M}} y_i^T.\theta^i + w_i^T.\lambda^i + 2.\sum_{i \in \mathcal{M}, j \neq i \in \mathcal{M}} u_{i,j}^T.\theta^{i,j}\right]$$

Now, let us create a $2^m \times 1$ vector where each entry will be equal to one of the functional form that we introduced earlier.

**Definition 2.** *Define the* $t$*-th (for* $t = 1, ..., T$*)* ***augmented kernel vector*** $\boldsymbol{\gamma}_t(\boldsymbol{Z})$, *(simply referred to hereafter as* $\boldsymbol{\gamma}_t$*) as a* $2^m \times 1$ *vector* $\boldsymbol{\gamma}_t$ *where the* $j$*-th entry (for* $k = 1, ..., 2^m$*) is given by*

$$\exp\left[\sum_{i \in \mathcal{M}} y_i^t.\theta^i + \sum_{j \in \mathcal{K}} \left(z_{t,j} + w_j^t\right).\lambda^j + \sum_{j \in \mathcal{M} \backslash \mathcal{K}} w_j^t.\lambda^j + 2 \sum_{i \in \mathcal{M}, j \neq i \in \mathcal{M}} u_{i,j}^t.\theta^{i,j}\right]$$

*where* $k = \varrho(\mathcal{K})$ *and* $\mathcal{M} \backslash \mathcal{K}$ *is the complement of* $\mathcal{K}$ *in* $\mathcal{M}$.

Equivalently, we could have introduced an alternative way of writing $\boldsymbol{\gamma}_t(\varrho(\mathcal{K}))$ that is useful in the rest of paper and that is identical to the previous one

$$\boldsymbol{\gamma}_t(\varrho(\mathcal{K})) = \exp\left[\sum_{i \in \mathcal{M}} y_i^t.\theta^i + w_i^t.\lambda + \sum_{j \in \mathcal{K}} z_{t,j}.\lambda^j + 2.\sum_{i \in \mathcal{M}, j \neq i \in \mathcal{M}} u_{i,j}^t.\theta^{i,j}\right]$$

I will reintroduce some matrix notation here, which will make the subsequent exposition more compact.

**Definition 3.** *Define a selector matrix* $\boldsymbol{S}_j$ *for* $j = 1, ..., 2^m$ *as an* $m \times m$ *matrix where the* $(i, k)$*-th entry is defined as*

$$\delta_{i,k}.\mathcal{I}_{i \in \mathcal{K}}$$

*where* $\mathcal{K} = \varrho^{-1}(j)$, $\delta_{i,k}$ *is kronecker's delta function and* $\mathcal{I}_{i\in\mathcal{K}}$ *is an indicator function that is equal to 1 if* $i \in \mathcal{K}$.

$\boldsymbol{S}_j$ *can also be defined as being*

$$\boldsymbol{S}_j = \text{diag}(\mathcal{C}^{-1}(j))$$

As an example of a selector matrix, if $m = 2$, $\mathcal{K} = \{1\}$ and $j = \varrho(\mathcal{K}) = 2$ then

$$\boldsymbol{S}_3 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

Now, in matrix notation, the $j$-th term of $\boldsymbol{\gamma}_t$ will be given by

$$\exp(\text{tr}[\boldsymbol{Z}'_{1:t}.\boldsymbol{Z}_{1:t}.\boldsymbol{\Theta}] + \text{tr}[\boldsymbol{Z}'_{1:t}.\boldsymbol{L}_1.\boldsymbol{Z}_{1:t}.\boldsymbol{\Lambda}] + \text{tr}[\text{diag}(\boldsymbol{z}_t).\boldsymbol{S}_j.\boldsymbol{\Lambda}])$$

where $j = \varrho(\mathcal{K})$ for a given element $\mathcal{K}$ of $\mathcal{P}(\mathcal{M})$ and where $\boldsymbol{Z}_{1:t}$ is the matrix formed by the first $t$ rows of $\boldsymbol{Z}$. Of course, the size of $\boldsymbol{L}_1$ was changed here because we only consider the first $t$ rows of $\boldsymbol{Z}$ and the whole matrix $\boldsymbol{Z}$. Therefore, $\boldsymbol{L}_1$ is now of dimension $t \times t$ instead of $T \times T$.

Using these notation, it is easy to see that the kernel of the distribution is the first element of $\boldsymbol{\gamma}_T$ where $1 = \varrho(\phi)$ and $\phi$ is the null set (see appendix B to see how $\phi$ is mapped into 1). Moreover, the FOMMVB density is equal to

$$p(\boldsymbol{Z}\,|\boldsymbol{\Theta}, \boldsymbol{\Lambda}) = \frac{1}{K_T(\boldsymbol{\Theta}, \boldsymbol{\Lambda})}.\boldsymbol{\gamma}_T(\varrho(\phi))$$

Now, a central result of the paper.

**Theorem 1.** *If one integrates any element of* $\boldsymbol{\gamma}_t$ *with respect to* $\boldsymbol{z}_t$, *one will obtain a linear combination of* $\boldsymbol{\gamma}_{t-1}$.

**Corollary 1.**

$$\sum_{\boldsymbol{z}_t} \boldsymbol{\gamma}_t = \boldsymbol{A}.\boldsymbol{\gamma}_{t-1}$$

*where* $\boldsymbol{A} = [\boldsymbol{a}'_1, ..., \boldsymbol{a}'_j, ..., \boldsymbol{a}'_{2^m}]'$ *and where we refer to coefficients of the linear combination of* $\boldsymbol{\gamma}_t(j)$ *as* $\boldsymbol{a}_j$ *with* $j = \varrho(\mathcal{K})$ *for some* $\mathcal{K} \in \mathcal{M}(\mathcal{P})$.

*Moreover the* $(p, q)$*-th entry of* $\boldsymbol{A}$ *is*

$$\boldsymbol{A}(p, q) = \exp\left[\sum_{i\in\mathcal{J}} \theta^i + \sum_{j\in\mathcal{K}\cap\mathcal{J}} \lambda^j + 2.\sum_{i\in\mathcal{J}, j\neq i\in\mathcal{J}} \theta^{i,j}\right]$$

*where* $p = \varrho(\mathcal{K})$ *and* $q = \varrho(\mathcal{J})$.

As an illustrative simple example that will clarify some of the ideas, let us first consider the univariate FOMMVB case. Since $m = 1$, the augmented kernel vector is of size $2 \times 1$

$$\boldsymbol{\gamma}_t = \begin{pmatrix} \exp\Big(\Big[\sum_{\tau=1}^t z_\tau\Big].\theta + \Big[\sum_{\tau=2}^t z_\tau.z_{\tau-1}\Big].\lambda\Big) \\ \exp\Big(\Big[\sum_{\tau=1}^t z_\tau\Big].\theta + \Big[z_t + \sum_{\tau=2}^t z_\tau.z_{\tau-1}\Big].\lambda\Big) \end{pmatrix}$$

by using some earlier obvious notation, we can write $\boldsymbol{\gamma}_t$ more compactly as

$$\boldsymbol{\gamma}_t = \begin{pmatrix} \exp(y^t.\theta + w^t.\lambda) \\ \exp(y^t.\theta + [z_t + w^t].\lambda) \end{pmatrix}$$

Now, let us integrate $\boldsymbol{\gamma}_t$ with respect to $z_t$

$$\begin{aligned}
\sum_{\boldsymbol{z}_t} \boldsymbol{\gamma}_t &= \sum_{\boldsymbol{z}_t} \begin{pmatrix} \exp(y^t.\theta + w^t.\lambda) \\ \exp(y^t.\theta + [z_t + w^t].\lambda) \end{pmatrix} \\
&= \sum_{\boldsymbol{z}_t} \begin{pmatrix} \exp\big(\big[z_t + y^{t-1}\big].\theta + \big[z_t.z_{t-1} + w^{t-1}\big].\lambda\big) \\ \exp\big(\big[z_t + y^{t-1}\big].\theta + \big[z_t + z_t.z_{t-1} + w^{t-1}\big].\lambda\big) \end{pmatrix} \\
&= \begin{pmatrix} \exp\big(\big[1 + y^{t-1}\big].\theta + \big[z_{t-1} + w^{t-1}\big].\lambda\big) + \exp\big(y^{t-1}.\theta + w^{t-1}.\lambda\big) \\ \exp\big(\big[1 + y^{t-1}\big].\theta + \big[1 + z_{t-1} + w^{t-1}\big].\lambda\big) + \exp\big(y^{t-1}.\theta + w^{t-1}.\lambda\big) \end{pmatrix} \\
&= \begin{pmatrix} \mathrm{e}^{\theta}.\exp\big(y^{t-1}.\theta + \big[z_{t-1} + w^{t-1}\big].\lambda\big) + \exp\big(y^{t-1}.\theta + w^{t-1}.\lambda\big) \\ \mathrm{e}^{\theta+\lambda}.\exp\big(y^{t-1}.\theta + \big[z_{t-1} + w^{t-1}\big].\lambda\big) + \exp\big(y^{t-1}.\theta + w^{t-1}.\lambda\big) \end{pmatrix} \\
&= \begin{pmatrix} 1 & \mathrm{e}^{\theta} \\ 1 & \mathrm{e}^{\theta+\lambda} \end{pmatrix}.\begin{pmatrix} \exp\big(y^{t-1}.\theta + w^{t-1}.\lambda\big) \\ \exp\big(y^{t-1}.\theta + \big[z_{t-1} + w^{t-1}\big].\lambda\big) \end{pmatrix} \\
&= \boldsymbol{A}.\boldsymbol{\gamma}_{t-1}
\end{aligned}$$

Therefore

$$\boldsymbol{A} = \begin{pmatrix} 1 & \mathrm{e}^{\theta} \\ 1 & \mathrm{e}^{\theta+\lambda} \end{pmatrix}$$

This simple example shows that even in the univariate case, the analytic computations might be quite complex.

Now we specifically saw how the integration operation is carried out in practice.

To clarify the framework even further, we shall give the analytic results for the bivariate case. However, we will not show the details of the computations for problem of space.

Here $m = 2$ and therefore each $\boldsymbol{\gamma}_t$ is of dimension $4 \times 1$.

$$\boldsymbol{\gamma}_t = \begin{pmatrix} \exp(y_1^t.\theta^1 + y_2^t.\theta^2 + w_1^t.\lambda^1 + w_2^t.\lambda^2 + 2u^t.\theta^{12}) \\ \exp(y_1^t.\theta^1 + y_2^t.\theta^2 + \big[z_{t,1} + w_1^t\big].\lambda^1 + w_2^t.\lambda^2 + 2u^t.\theta^{12}) \\ \exp(y_1^t.\theta^1 + y_2^t.\theta^2 + w_1^t.\lambda^1 + \big[z_{t,2} + w_2^t\big].\lambda^2 + 2u^t.\theta^{12}) \\ \exp(y_1^t.\theta^1 + y_2^t.\theta^2 + \big[z_{t,1} + w_1^t\big].\lambda^1 + \big[z_{t,2} + w_2^t\big].\lambda^2 + 2u^t.\theta^{12}) \end{pmatrix}$$

where, in an obvious notation, $u^t$ is actually $u_{1,2}^t$.

As earlier, $\sum_{\boldsymbol{z}_t} \boldsymbol{\gamma}_t = \boldsymbol{A}.\boldsymbol{\gamma}_{t-1}$ where the reader can verify that the matrix containing the coefficients of the linear combination is given by

$$\begin{pmatrix} 1 & \mathrm{e}^{\theta^1} & \mathrm{e}^{\theta^2} & \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}} \\ 1 & \mathrm{e}^{\theta^1+\lambda^1} & \mathrm{e}^{\theta^2} & \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}+\lambda^1} \\ 1 & \mathrm{e}^{\theta^1} & \mathrm{e}^{\theta^2+\lambda^2} & \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}+\lambda^2} \\ 1 & \mathrm{e}^{\theta^1+\lambda^1} & \mathrm{e}^{\theta^2+\lambda^2} & \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}+\lambda^1+\lambda^2} \end{pmatrix}$$

## 3.3 Marginal distributions

In this subsection, we are interested about marginal densities of the form $p(\boldsymbol{z}_1, ..., \boldsymbol{z}_t)$ for $t = 1, ..., T$. Each one of those densities is obtained by integration over $\boldsymbol{z}_{t+1}, ..., \boldsymbol{z}_T$.

We are interested about densities of that form because our goal is to describe the conditional distribution $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$. In subsection 3.5, we will show the markov property, i.e. $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, ..., \boldsymbol{z}_1) = p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$. However, here, we will concentrate on the quantity

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, ..., \boldsymbol{z}_1) = \frac{p(\boldsymbol{z}_1, ..., \boldsymbol{z}_{t-1}, \boldsymbol{z}_t)}{p(\boldsymbol{z}_1, ..., \boldsymbol{z}_{t-1})}$$

We see that the conditional distribution can be obtained by the ratios of two marginal densities of the form $p(\boldsymbol{z}_1, ..., \boldsymbol{z}_t)$ for $t = 1, ..., T$. Being able to write those quantities in closed form will be helpful for the rest of the paper.

**Proposition 1.** *Each marginal density of the form $p(\boldsymbol{z}_1, ..., \boldsymbol{z}_t)$ can be written as a linear combination of the augmented kernel vector $\boldsymbol{\gamma}_t(\boldsymbol{Z})$ or $\boldsymbol{\gamma}_t$.*

$$p(\boldsymbol{z}_1, ..., \boldsymbol{z}_t) = \boldsymbol{b}_t'.\boldsymbol{\gamma}_t/K_T$$

*where $K_T \triangleq K_T(\boldsymbol{\Theta}, \boldsymbol{\Lambda})$ is the constant of integration of $\boldsymbol{Z}$.*

*Moreover, the coefficients of the linear combination are given by the following recursion*

$$\boldsymbol{b}_{t-1} = \boldsymbol{A}'.\boldsymbol{b}_t$$

*with the boundary condition*

$$\boldsymbol{b}_T = \boldsymbol{e}_1$$

*where $\boldsymbol{e}_1$ is the first column of the identity matrix $\boldsymbol{I}_{2^m}$.*

## 3.4 The constant of integration

Let $\boldsymbol{Z}_k$ denote one possible $\boldsymbol{Z}$ matrix among the $2^{Tm}$ possible $\boldsymbol{Z}$ matrices that could have occurred. The constant of integration is then, as the reader remembers, will be equal to the sum of the kernel over all those $2^{Tm}$ possibilities.

$$K_T(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) = \sum_{k=1}^{2^{Tm}} \exp\{\mathrm{tr}[\boldsymbol{Z}_k'.\boldsymbol{Z}_k.\boldsymbol{\Theta}] + \mathrm{tr}[\boldsymbol{Z}_k'.\boldsymbol{L}_1.\boldsymbol{Z}_k.\boldsymbol{\Lambda}_1]\}$$

Doing the integration over all $2^{Tm}$ possibilities is intractable in most practical situations. However, since this integration can be carried out by integrating successively over $\boldsymbol{z}_t$, we can resort to the augmented kernel vector in order to carry out that integration.

**Proposition 2.** *The constant of integration can be obtained from the relation*

$$K_T(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) = \boldsymbol{e}_1'.\boldsymbol{A}^{T-1}.\boldsymbol{\kappa}_1$$

*with $\boldsymbol{\kappa}_1 = \sum_k^{2^m} \boldsymbol{\gamma}_1(k)$ with $k = \varrho(\mathcal{K})$ for some $\mathcal{K} \in \mathcal{M}(\mathcal{P})$*

**Corollary 2.** *The vector $\boldsymbol{\kappa}_t$ follows a first-order difference equation*

$$\boldsymbol{\kappa}_t = \boldsymbol{A}.\boldsymbol{\kappa}_{t-1}$$

*with the initial condition* $\boldsymbol{\kappa}_1 = \sum_k^{2^m} \boldsymbol{\gamma}_1(k) = \sum_{k=1}^{2^m} A_{(:,k)}$ *(i.e. the sum of columns of* $\boldsymbol{A}$*).*

*The constant of integration of different sample size is just the first element of that vector*

$$K_t(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) = \boldsymbol{e}_1'.\boldsymbol{\kappa}_t$$

Now we will illustrate with some simple examples.

In the univariate case

$$\begin{aligned}
\boldsymbol{\kappa}_1 &= \sum_{z_1} \boldsymbol{\gamma}_1 \\
&= \sum_{z_1} \begin{pmatrix} \exp(z_1.\theta) \\ \exp(z_1.\theta + z_1.\lambda) \end{pmatrix} \\
&= \begin{pmatrix} \mathrm{e}^{\theta} + 1 \\ \mathrm{e}^{\theta+\lambda} + 1 \end{pmatrix}
\end{aligned}$$

and therefore a sample of size 1 has the following integration constant

$$\begin{aligned}
K_1(\theta, \lambda) &= \boldsymbol{e}_1'.\boldsymbol{\kappa}_1 \\
&= \mathrm{e}^{\theta+1}
\end{aligned}$$

A sample of size two has the following integration constant

$$\begin{aligned}
K_2(\theta, \lambda) &= \boldsymbol{e}_1'.\begin{pmatrix} 1 & \mathrm{e}^{\theta} \\ 1 & \mathrm{e}^{\theta+\lambda} \end{pmatrix}\begin{pmatrix} \mathrm{e}^{\theta} + 1 \\ \mathrm{e}^{\theta+\lambda} + 1 \end{pmatrix} \\
&= \boldsymbol{e}_1'.\begin{pmatrix} \mathrm{e}^{2\theta+\lambda} + 2\mathrm{e}^{\theta} + 1 \\ \mathrm{e}^{2\theta+2\lambda} + \mathrm{e}^{\theta+\lambda} + \mathrm{e}^{\theta} + 1 \end{pmatrix} \\
&= \mathrm{e}^{2\theta+\lambda} + 2\mathrm{e}^{\theta} + 1
\end{aligned}$$

For instance $K_5(\theta, \lambda) = \mathrm{e}^{5\theta+4\lambda} + 2\mathrm{e}^{4\theta+3\lambda} + 3\mathrm{e}^{4\theta+2\lambda} + 3\mathrm{e}^{3\theta+2\lambda} + 6\mathrm{e}^{3\theta+\lambda} + 4\mathrm{e}^{2\theta+\lambda} + \mathrm{e}^{3\theta} + 6\mathrm{e}^{2\theta} + 5\mathrm{e}^{\theta} + 1$.

This gives an idea of the complexity of the computations and at the rate at which $K_t(\theta, \lambda)$. (This can be given by the spectral radius of $\boldsymbol{A}$).

For the bivariate case

$$\begin{aligned}
\boldsymbol{\kappa}_1 &= \sum_{\boldsymbol{z}_1} \boldsymbol{\gamma}_1 \\
&= \sum_{\boldsymbol{z}_1} \begin{pmatrix} \exp\left(z_{1,1}.\theta^1 + z_{1,2}.\theta^2 + 2z_{1,1}.z_{1,2}.\theta^{12}\right) \\ \exp(z_{1,1}.\theta^1 + z_{1,2}.\theta^2 + z_{1,1}.\lambda^1 + 2z_{1,1}.z_{1,2}.\theta^{12}) \\ \exp(z_{1,1}.\theta^1 + z_{1,2}.\theta^2 + z_{1,2}.\lambda^2 + 2z_{1,1}.z_{1,2}.\theta^{12}) \\ \exp(z_{1,1}.\theta^1 + z_{1,2}.\theta^2 + z_{1,1}.\lambda^1 + z_{1,2}.\lambda^2 + 2z_{1,1}.z_{1,2}.\theta^{12}) \end{pmatrix} \\
&= \begin{pmatrix} \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}} + \mathrm{e}^{\theta^1} + \mathrm{e}^{\theta^2} + 1 \\ \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}+\lambda^1} + \mathrm{e}^{\theta^1+\lambda^1} + \mathrm{e}^{\theta^2} + 1 \\ \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}+\lambda^2} + \mathrm{e}^{\theta^1} + \mathrm{e}^{\theta^2+\lambda^2} + 1 \\ \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}+\lambda^1+\lambda^2} + \mathrm{e}^{\theta^1+\lambda^1} + \mathrm{e}^{\theta^2+\lambda^2} + 1 \end{pmatrix}
\end{aligned}$$

Remember that the integration over $\boldsymbol{z}_1$ is actually over (1, 1), (1, 0), (0,1) and (0, 0). Therefore $K_1(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) = \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}} + \mathrm{e}^{\theta^1} + \mathrm{e}^{\theta^2} + 1$.

Similarly

$$\begin{aligned}
K_2(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) &= \boldsymbol{e}_1'.\boldsymbol{A}.\boldsymbol{\kappa}_1 \\
&= \boldsymbol{e}_1'.\begin{pmatrix} 1 & \mathrm{e}^{\theta^1} & \mathrm{e}^{\theta^2} & \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}} \\ 1 & \mathrm{e}^{\theta^1+\lambda^1} & \mathrm{e}^{\theta^2} & \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}+\lambda^1} \\ 1 & \mathrm{e}^{\theta^1} & \mathrm{e}^{\theta^2+\lambda^2} & \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}+\lambda^2} \\ 1 & \mathrm{e}^{\theta^1+\lambda^1} & \mathrm{e}^{\theta^2+\lambda^2} & \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}+\lambda^1+\lambda^2} \end{pmatrix} \\
&\times \begin{pmatrix} \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}} + \mathrm{e}^{\theta^1} + \mathrm{e}^{\theta^2} + 1 \\ \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}+\lambda^1} + \mathrm{e}^{\theta^1+\lambda^1} + \mathrm{e}^{\theta^2} + 1 \\ \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}+\lambda^2} + \mathrm{e}^{\theta^1} + \mathrm{e}^{\theta^2+\lambda^2} + 1 \\ \mathrm{e}^{\theta^1+\theta^2+2\theta^{12}+\lambda^1+\lambda^2} + \mathrm{e}^{\theta^1+\lambda^1} + \mathrm{e}^{\theta^2+\lambda^2} + 1 \end{pmatrix} \\
&= \mathrm{e}^{2\theta^1+2\theta^2+4\theta^{12}+\lambda^1+\lambda^2} + 2\mathrm{e}^{2\theta^1+\theta^2+2\theta^{12}+\lambda^1} + 2\mathrm{e}^{\theta^1+2\theta^2+2\theta^{12}+\lambda^2} \\
&+ \mathrm{e}^{2\theta^1+\lambda^1} + \mathrm{e}^{2\theta^2+\lambda^2} + 2\mathrm{e}^{\theta^1+\theta^2+2\theta^{12}} + 2\mathrm{e}^{\theta^1+\theta^2} + 2\mathrm{e}^{\theta^1} + 2\mathrm{e}^{\theta^2} + 1
\end{aligned}$$

We will not give the result for $K_5(\boldsymbol{\Theta}, \boldsymbol{\Lambda})$ since it will take too many lines to write it down.

Now that we have the constant of integration in closed form, it is straightforward to write the likelihood of the model in closed form.

## 3.5 Transition probabilities

As already explained in subsection 3.3, computing marginal densities of the form $p(\boldsymbol{z}_1, ..., \boldsymbol{z}_t)$ will easily yield the conditional distribution

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, ..., \boldsymbol{z}_1) = \frac{p(\boldsymbol{z}_1, ..., \boldsymbol{z}_{t-1}, \boldsymbol{z}_t)}{p(\boldsymbol{z}_1, ..., \boldsymbol{z}_{t-1})}$$

From that, we easily see that

$$\begin{aligned}
p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, ..., \boldsymbol{z}_1) &= \frac{\boldsymbol{b}_t'.\boldsymbol{\gamma}_t/K_T}{\boldsymbol{b}_{t-1}'.\boldsymbol{\gamma}_{t-1}/K_T} \\
&= \frac{\boldsymbol{b}_t'.\boldsymbol{\gamma}_t}{\boldsymbol{b}_{t-1}'.\boldsymbol{\gamma}_{t-1}}
\end{aligned}$$

And we already mentioned that $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, ..., \boldsymbol{z}_1) = p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$. We shall easily prove it!

**Proposition 3.** *$\boldsymbol{z}_t$ is markovian, i.e. $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}, ..., \boldsymbol{z}_1) = p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$. Moreover, the markov process formed by $\boldsymbol{z}_t$ is non-homogeneous.*

In the previous proof, we explicitly computed the formula for the conditional density $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$. We will summarize that in a proposition.

**Proposition 4.** *Define $\boldsymbol{\zeta}_t$ as a $2^m \times 1$ vector whose $j$-th entry is given by*

$$\boldsymbol{\zeta}_t(j) = \exp(\mathrm{tr}[\boldsymbol{z}_t'.\boldsymbol{z}_t.\boldsymbol{\theta}] + \mathrm{tr}[\boldsymbol{z}_t'.\boldsymbol{z}_{t-1}.\boldsymbol{\Lambda}] + \mathrm{tr}[\mathrm{diag}(\boldsymbol{z}_t).\boldsymbol{S}_j.\boldsymbol{\Lambda}])$$

*where $j = \varrho(\mathcal{K})$ for the set $\mathcal{K}$ that is mapped into $\boldsymbol{z}_t$ through $\varrho^{-1} \circ \mathcal{C}$.*

*$\boldsymbol{\zeta}_t$ is a function of $\boldsymbol{z}_t$ and $\boldsymbol{z}_{t-1}$ and can therefore be written as $\boldsymbol{\zeta}_t(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t)$.*
*Also, define $\boldsymbol{\xi}_t$ as a $2^m \times 1$ vector whose $j$-th entry is given by*

$$\boldsymbol{\xi}_t(j) = \exp(\mathrm{tr}[\mathrm{diag}(\boldsymbol{z}_t).\boldsymbol{S}_j.\boldsymbol{\Lambda}])$$

*where, again, $j = \varrho(\mathcal{K})$ for the set $\mathcal{K}$ that is mapped into $\boldsymbol{z}_t$ through $\varrho^{-1} \circ \mathcal{C}$.*
*$\boldsymbol{\xi}_t$ is a function of $\boldsymbol{z}_t$ and can therefore be written as $\boldsymbol{\xi}_t(\boldsymbol{z}_t)$.*
*Then the conditional density $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ is given by the following formula*

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \boldsymbol{b}_t'.\boldsymbol{\zeta}_t / \boldsymbol{b}_{t-1}'.\boldsymbol{\xi}_{t-1}$$

*Moreover, the $2^m \times 2^m$ matrix of transition probabilities $\boldsymbol{P}_t$ contains at its $(i, j)$-th entry*

$$\boldsymbol{P}_t(i, j) = \boldsymbol{b}_t'.\boldsymbol{\zeta}_t(\mathcal{C}^{-1}(i), \mathcal{C}^{-1}(j)) / \boldsymbol{b}_{t-1}'.\boldsymbol{\xi}_{t-1}(\mathcal{C}^{-1}(i))$$

As a simple example, let us consider the univariate case.

The two vectors $\boldsymbol{\zeta}_t$ and $\boldsymbol{\xi}_t$ are respectively

$$\boldsymbol{\zeta}_t = \begin{pmatrix} \exp(z_t.\theta + z_t.z_{t-1}.\lambda) \\ \exp(z_t.\theta + (z_t + z_t.z_{t-1}).\lambda) \end{pmatrix}$$

$$\boldsymbol{\xi}_t = \begin{pmatrix} 1 \\ \exp(z_t.\lambda) \end{pmatrix}$$

Using index notation over $k = 1, ..., 2^m$, $\boldsymbol{\xi}_t(1) = 1$ and $\boldsymbol{\xi}_t(2) = \exp(z_t.\lambda)$. Using the notation over $\mathcal{Z}$, we see that $\boldsymbol{\xi}_t(z_t) = (1, \exp(z_t.\lambda))'$ and $\boldsymbol{\zeta}_t(z_{t-1}, z_{t-1}) = (\exp(z_t.\theta + z_t.z_{t-1}.\lambda), \exp(z_t.\theta + (z_t + z_t.z_{t-1}).\lambda)'$.

The conditional density $p(z_t|z_{t-1})$ is now given by the following formula

$$p(z_t|z_{t-1}) = \frac{b_t(1).\exp(z_t.\theta + z_t.z_{t-1}.\lambda) + b_t(2).\exp(z_t.\theta + (z_t + z_t.z_{t-1}).\lambda)}{b_{t-1}(1) + b_{t-1}(2).\exp(z_{t-1}.\lambda)}$$

Now, replacing $z_t$ and $z_{t-1}$ by their possible binary values, we obtain a matrix of transition probabilities of the following form

$$\boldsymbol{P}_t = \begin{pmatrix} \frac{b_t(1) + b_t(2)}{b_{t-1}(1) + b_{t-1}(2)} & \frac{b_t(1) + b_t(2)}{b_{t-1}(1) + b_{t-1}(2).\mathrm{e}^{\lambda}} \\ \\ \frac{b_t(1).\mathrm{e}^{\theta} + b_t(2).\mathrm{e}^{\theta+\lambda}}{b_{t-1}(1) + b_{t-1}(2)} & \frac{b_t(1).\mathrm{e}^{\theta+\lambda} + b_t(2).\mathrm{e}^{\theta+2\lambda}}{b_{t-1}(1) + b_{t-1}(2).\mathrm{e}^{\lambda}} \end{pmatrix}$$

where the configuration in $\boldsymbol{P}_t$ is through our usual mappings, i.e. $\mathcal{C}(0) = 1$ and $\mathcal{C}(1) = 2$.

We can easily see that the rows sum to one after replacing the $b_{t-1}(.)$ coefficients as the corresponding linear combination of the $b_t(.)$ coefficients.

We can also give formulas for the bivariate case. We begin by writing the formulas for $\boldsymbol{\zeta}_t$ and $\boldsymbol{\xi}_t$

$$\boldsymbol{\zeta}_t = \begin{pmatrix} \exp(z_{t,1}.\theta^1 + z_{t,2}.\theta^2 + z_{t,1}.z_{t-1,1}.\lambda^1 + z_{t,2}.z_{t-1,2}.\lambda^2 + 2z_{t,1}.z_{t,2}.\theta^{12}) \\ \exp(z_{t,1}.\theta^1 + z_{t,2}.\theta^2 + [z_{t,1} + z_{t,1}.z_{t-1,1}].\lambda^1 + z_{t,2}.z_{t-1,2}.\lambda^2 + 2z_{t,1}.z_{t,2}.\theta^{12}) \\ \exp(z_{t,1}.\theta^1 + z_{t,2}.\theta^2 + z_{t,1}.z_{t-1,1}.\lambda^1 + [z_{t,2} + z_{t,2}.z_{t-1,2}].\lambda^2 + 2z_{t,1}.z_{t,2}.\theta^{12}) \\ \exp(z_{t,1}.\theta^1 + z_{t,2}.\theta^2 + [z_{t,1} + z_{t,1}.z_{t-1,1}].\lambda^1 + [z_{t,2} + z_{t,2}.z_{t-1,2}].\lambda^2 + 2z_{t,1}.z_{t,2}.\theta^{12}) \end{pmatrix}$$

and

$$\boldsymbol{\xi}_t = \begin{pmatrix} 1 \\ \exp(z_{t,1}.\lambda^1) \\ \exp(z_{t,2}.\lambda^2) \\ \exp(z_{t,1}.\lambda^1 + z_{t,2}.\lambda^2) \end{pmatrix}$$

We will not give here explicitly the formulas for $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ and $\boldsymbol{P}_t$ but it suffices to say that $\boldsymbol{P}_t$ will have the following form

$$\boldsymbol{P}_t = \begin{array}{c|cccc} \boldsymbol{z}_{t-1} & (0,0) & (1,0) & (0,1) & (1,1) \\ \boldsymbol{z}_t & & & & \\ (0,0) & p_{t,1,1} & \cdots & \cdots & p_{t,1,4} \\ (1,0) & \vdots & \ddots & & \vdots \\ (0,1) & \vdots & & \ddots & \vdots \\ (1,1) & p_{t,4,1} & \cdots & \cdots & p_{t,4,1} \end{array}$$

where to get each of the $p_{t,i,j}$, we replace $\boldsymbol{z}_t$ and $\boldsymbol{z}_{t-1}$ by their values in the formula of $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$.

For the actual computation on a computer, it is more straightforward to use the formulas in proposition 4 for the general case.

# 4 Volatility clustering

## 4.1 Motivation

A wide array of time series data is characterized by volatility clustering. At a given date, the series shows an increase in volatility that remains for a given number of periods and then subsides, forming a "volatility cluster". Those clusters could manifest themselves several times in a given series. Such phenomena created the need to model the underlying volatility of time series in a dynamic (autoregressive) way which was an impetus behind the creation of volatility models such as GARCH and stochastic volatility. For more details, consult [Bauwens et al., 2006] which is a survey of multivariate GARCH models. If volatility can be thought to be showing jumps and varying in a non-smooth way, then markov switching (also know as hidden markov or markov mixture) models could be used to model the phenomenon.

This section concerns itself with the case when volatility clustering is thought of as a discrete latent phenomenon, i.e. in the markov switching case. The objective is to show how to generalize markov switching volatility models to the multivariate case. In a previous section, I introduced a general framework for extending the markov switching model to a multivariate setting. Here, we present an illustration of that methodology in the multivariate volatility modeling framework.

I will begin by a simple illustration that will show the difficulties and problems and that will pave the way for the rest of the paper.

**Motivating Example**

For instance, a quite rudimentary model is the following. Let $y_t$ be the variable of interest,

$$y_t = \sigma_{z_t}.\varepsilon_t$$

$$\varepsilon_t \sim \mathcal{N}(0,1)$$

and $z_t$ is a binary markov with the matrix of transition probabilities $\boldsymbol{P}$ and where state 1 stands for low volatility and 2 stands for high volatility, i.e. $\sigma_1 < \sigma_2$.

This model is a simple markov switching model that can be estimated in a standard way (see e.g. [Hamilton, 1989], [Kim and Nelson, 1999]).

It is unclear how to generalize the previous model to the multivariate case. Let $\boldsymbol{y}_t$ be a $p \times 1$ vector.

- Strategy 1

The most straightforward (and perhaps the less attractive) way is the following

$$\boldsymbol{y}_t = \boldsymbol{\Sigma}_{z_t}^{1/2}.\boldsymbol{\varepsilon}_t$$

$$\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$$

and $z_t$ is a binary markov with the matrix of transition probabilities $\boldsymbol{P}$. $\boldsymbol{\Sigma}_{z_t}^{1/2}$ is the cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma}_{z_t}$. The classification of high and low volatility clustering is unclear in this case because the two matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are not readily ordered.

One way to get a clearer picture is to use the special decomposition of a covariance matrix into a diagonal matrix $\boldsymbol{S}$ containing the standard deviations and a symmetric matrix $\boldsymbol{R}$ containing the correlation coefficients as in e.g. [Barnard et al., 2000] and [Pelletier, 2006]

$$\boldsymbol{\Sigma} = \boldsymbol{S}.\boldsymbol{R}.\boldsymbol{S}$$

Another shortcoming of that approach is that the change of volatility must occur in all variables at the same time (we shall refer to that phenomenon as *concomitant clustering*).

- Strategy 2

A second modeling strategy is the following

$$y_{t,i} = \sigma_{z_{t,i}}.\varepsilon_{t,i}$$

$$\varepsilon_{t,i} \sim \mathcal{N}(0,1)$$

and for $i \neq j$

$$\operatorname{Corr}(\varepsilon_{t,i}, \varepsilon_{t',j}) = \delta_{t,t'}.\rho_{i,j}$$

where $\delta_{t,t'} = \mathbb{1}\{t = t'\}$ is kronecker's delta and $\rho_{i,j}$ is the $(i,j)$th correlation coefficient.

Each $z_{t,i}$ is a binary markov chain with the matrix of transition probabilities $\boldsymbol{P}_i$ and each two chains $z_{t,i}$ and $z_{t,j}$ are independent for $i \neq j$.

I concatenate all the chains $z_{t,i}$ in a global chain $z_t$ with $2^p$ states in the following manner.

$$\boldsymbol{P} = \bigotimes_{i=1}^{p} \boldsymbol{P}_i$$

This framework solve th concomitant clustering problem in the sense that if a variable $i$ experiences a surge in volatility, it does not necessarily mean that variable $j \neq i$ is experiencing the same surge. On the other hand the model assumes that surges in volatility in different variables are necessarily independent. This assumption is quite restrictive in many cases. For instance, in financial markets, a surge in volatility can be interpreted as an increase in speculation and that could occur simultaneously in different variables. Hence the necessity to relax the independence assumption.

- Strategy 3

We need to relax the assumption of independence, yet we need a parsimonious model that is tractable. We apply the strategy of the previous chapter.

$$y_{t,i} = \sigma_{z_{t,i}} . \varepsilon_{t,i}$$

$$\varepsilon_{t,i} \sim \mathcal{N}(0,1)$$

and for $i \neq j$

$$\text{Corr}(\varepsilon_{t,i}, \varepsilon_{t',j}) = \delta_{t,t'} . \rho_{i,j}$$

where $\delta_{t,t'} = \mathbb{1}\{t = t'\}$ is kronecker's delta and $\rho_{i,j}$ is the $(i,j)$th correlation coefficient.

The matrix $\boldsymbol{Z}$ of latent variables (i.e. the markov chains in the previous strategy)

$$\boldsymbol{Z} = \begin{pmatrix} z_{1,1} & \cdots & z_{1,p} \\ \vdots & \vdots & \vdots \\ \vdots & z_{t,i} & \vdots \\ \vdots & \vdots & \vdots \\ z_{T,1} & \cdots & z_{T,p} \end{pmatrix}$$

is a first-order markov matrix-variate bernoulli with parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}$ of dimensions $p \times p$ respectively.

$$\boldsymbol{Z} \sim \mathcal{MB}(\boldsymbol{\Theta}, \boldsymbol{\Lambda})$$

The advantages of that representation is that the dependence between the latent variables is taken into account in a very straightforward manner. For instance

- $\boldsymbol{\Theta}(i,\, j)$ describes the dependence between the respective latent volatilities in variables $i$ and $j$, i.e. $\boldsymbol{\Theta}(i,\, j)$ tells us how likely a surge in volatility in variable $i$ be accompanied by a surge in volatility in variable $j$.
- $\boldsymbol{\Theta}(i,\, i)$ describes the probability structure of the latent volatility in each variable $i$, i.e. $\boldsymbol{\Theta}(i,i)$ tells us how likely a surge in volatility in variable $i$ is to occur.

- $\mathbf{\Lambda}(i,\ i)$ describes the dynamic structure (conditional heteroscedasticity structure) within each latent volatility in variable $i$, i.e. $\mathbf{\Lambda}(i,i)$ tells us how likely a surge in volatility in variable $i$ is to be followed by hight volatility in that same variable.

The matrix $\mathbf{\Lambda}$ is diagonal whereas $\mathbf{\Theta}$ is just symmetric (if it is diagonal, then the latent volatilities are independent of one another.)

## 4.2 Model and Inference

Inference in multivariate volatility clustering model can be considered from both the classical and bayesian point of views.

From the classical point of view, estimation would typically require use of the EM algorithm of [Dempster et al., 1977]. However, in this paper, I shall limit myself to the explicit description of the bayesian approach. I shall describe a hybrid gibbs/metropolis sampler for drawing from the posterior of the model. I shall also address some computational issues relating to the matrix-variate bernoulli distribution.

### 4.2.1 The model

The observables of the model are a length $T$ time series of $m \times 1$ vectors $\boldsymbol{y}_t$, for $t = 1, ..., T$. Put the observables in the matrix $\boldsymbol{Y}$ of dimension $T \times m$.

The unobservables are

1. A matrix $\boldsymbol{Z}$ of latent variables of dimension $T \times m$.
2. The parameters of the model which comprise
    a. The $m + \frac{m(m+1)}{2}$ parameters of the matrix-variate bernoulli distribution which are included in the matrices $m \times m$ matrices $\mathbf{\Theta}$ and $\mathbf{\Lambda}$.
    b. The $2m + \frac{m(m-1)}{2}$ parameters describing the covariance structure. Those include $2m$ parameters describing the different variances in each volatility state for each variable and $\frac{m(m-1)}{2}$ parameters which are the entries of the unique correlation matrices across all the states.

Therefore, the model contains a total of $m^2 + 3m$ parameters which grow at a polynomial rate with the number of equations. (The traditional approach of multivariate markov switching models exhibits an exponential growth of the number of parameters if the hypothesis of independence is dropped. The advantage of our approach is the exhibition of a polynomial rate with an arbitrary pattern of dependence across the different equations.)

The priors that will used are the following. For each one of the different parameters in the matrices $\mathbf{\Theta}$ and $\mathbf{\Lambda}$, a normal prior is assigned that is centered and has a huge standard deviation. For instance

$$\theta^i \sim \mathcal{N}(0, \sigma^2_{\theta^i})$$

$$\theta^{i,j} \sim \mathcal{N}(0, \sigma^2_{\theta^{i,j}})$$

$$\lambda^i \sim \mathcal{N}(0, \sigma^2_{\lambda^i})$$

The variances in those priors should be quite large for them to be non-informative. Typically, they should be chosen to be $10^4$ or higher.

There corresponds a set of two variances for each variable $y_{t,i}$ which are respectively the variance of the high volatility state and that of the low one.

$$\boldsymbol{s}_k = \begin{pmatrix} \sigma^2_{k,1} \\ \vdots \\ \sigma^2_{k,i} \\ \vdots \\ \sigma^2_{k,m} \end{pmatrix}$$

for $k = 1, 2$ (state index) and $i = 1, ..., m$ (variable index).

We shall pick an inverse gamma prior for each one of those variances

$$\sigma^2_{k,i} \sim \mathcal{IG}\left(\frac{\nu^0_{k,i}}{2}, \frac{s^0_{k,i}}{2}\right)$$

The superscript 0 indicates that the corresponding quantity is a hyperparameter (i.e. a parameter of the prior distribution).

For a detailed discussion of inverse gamma priors see e.g. [Robert, 2007] [Bauwens et al., 1999] and [Gelman et al., 1995]. $s \sim \mathcal{IG}(\alpha,\ \beta)$ if its density is $\frac{\beta^\alpha}{\Gamma(\alpha)}.s^{-(\alpha+1)}.\exp\big(-\beta.s^{-1}\big)$.

Finally, for the prior of the correlation matrix, we adopt a strategy suggested in [Barnard et al., 2000] and use a uniform prior.

$$p(\boldsymbol{R}) \propto 1$$

The uniform prior here not cause the posterior to be improper because the space of correlation parameters is a compact subset of the hyper cube $[-1,\ 1]^{\frac{m(m-1)}{2}}$ which is itself a compact subset of the Euclidean space $\mathbb{R}^{\frac{m(m-1)}{2}}$. See [Rousseeuw and Molenberghs, 1994] for a discussion.

It is not necessary to use those same priors for the bayesian inference approach. They seem to be the ones to give the most straightforward way to sample from the posterior. Inverse Wishart posteriors for $\boldsymbol{R}$ are preferred when they can be obtained.

## 4.3 The MCMC Algorithm

In this subsection, an MCMC algorithm is described for taking draws form the posterior of the model. It is a gibbs sampler that contains a metropolis-hastings sweep. See [Chen et al., 2000] or [Robert and Casella, 2004] for a review of the MCMC methodology.

The algorithm is describe in the following way. Repeat the following iteration for $M$ times. Drop $B$ iterations at the beginning (burn-in period). Then the remaining $(M - B)$ iterations are the draws from the posterior. In the following, we denote the $n$-th draw of, say, parameter $\psi$ as $\psi^{(n)}$.

**Algorithm 1.** ***The Gibbs Sampler***

1. *Initialize the sampler with*

$$\boldsymbol{Z}^{(0)},\ \boldsymbol{\Theta}^{(0)},\ \boldsymbol{\Lambda}^{(0)},\ \boldsymbol{s}_1^{(0)},\ \boldsymbol{s}_2^{(0)},\ \boldsymbol{R}^{(0)}$$

2. *At iteration $n$, draw*

$$\boldsymbol{Z}^{(n)},\ \boldsymbol{\Theta}^{(n)},\ \boldsymbol{\Lambda}^{(n)},\ \boldsymbol{s}_1^{(n)},\ \boldsymbol{s}_2^{(n)},\ \boldsymbol{R}^{(n)}$$

*consecutively in the following order*

   a. *Draw a matrix of latent variable $\boldsymbol{Z}^{(n)}$ from the distribution*

$$p(\boldsymbol{Z}^{(n)}|\boldsymbol{\Theta}^{(n-1)}, \boldsymbol{\Lambda}^{(n-1)}, \boldsymbol{s}_1^{(n-1)}, \boldsymbol{s}_2^{(n-1)}, \boldsymbol{R}^{(n-1)}, \boldsymbol{Y})$$

   b. *Draw the parameters of the matrix-variate bernoulli from*

$$p(\boldsymbol{\Theta}^{(n)}, \boldsymbol{\Lambda}^{(n)}|\boldsymbol{Z}^{(n)})$$

   c. *Draw the variances for $k = 1, 2$ from*

$$p\left(\boldsymbol{s}_k^{(n)}\middle|\boldsymbol{Z}^{(n)}, \boldsymbol{Y}\right)$$

   d. *Draw the correlations from*

$$p(\boldsymbol{R}^{(n)}|\boldsymbol{s}_1^{(n)}, \boldsymbol{s}_2^{(n)}, \boldsymbol{Z}^{(n)}, \boldsymbol{Y})$$

I shall investigate the different sweeps in the Gibbs sampler in more detail.

## 4.4 Computations for the FOMMVB distribution

Taking a draw of $\boldsymbol{Z}^{(n)}$ is done in the following way. Since the rows $\boldsymbol{z}_t^{(n)}$ of $\boldsymbol{Z}^{(n)}$ constitute a markov chain, they could be sampled in the following way. Using the closed form formulas for $p(\boldsymbol{z}_1)$ and $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$, those densities can be readily computed from $\boldsymbol{\Theta}^{(n)}$ and $\boldsymbol{\Lambda}^n$.

Therefore, using the transition probabilities and the previous draws of $\boldsymbol{s}_k$, a filter as in [Hamilton, 1989] can be readily constructed. That filter[1] will give the *a posteriori* probabilities $p^\star(\boldsymbol{z}_1)$ and $p^\star(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$. Sampling from $p^\star(\boldsymbol{z}_1)$ and $p^\star(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ gives $\boldsymbol{Z}^{(n)}$. Computing the *a posteriori* $p^\star()$ densities from the $p()$ densities shall not described in detail because it is standard in markov switching models. (See e.g. the last chapter of [Hamilton, 1994] or the detailed exposition of [Kim and Nelson, 1999].)

1. Standard filtering software should be modified in our case here because the markov chain is non-homogeneous and therefore, at each period $t$, the updating uses different set of probabilities and not the same ones.

On the other hand computing $p(\boldsymbol{z}_1)$ and $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$ is not standard and it uses the closed form results exposed earlier. They should be done in the following way.

First we need to explicitly be able to do computations with the mappings $\mathcal{C}$ and $\varrho$ described earlier. Those mappings should be coded (and stored once and for all throughout the whole MCMC iterations) in the following way.

**Algorithm 2.** ***Computation of Mappings $\mathcal{C}$ and $\varrho$***

*As an input to this step, give $m$, the number of latent variables.*

*The output is a matrix $\boldsymbol{X}$ of dimensions $2^m \times m$. (This is shown in detail in appendix B). The matrix $\boldsymbol{X}$ allows us to compute $\varrho$ and $\mathcal{C}$ because it has the property that $\mathcal{C}(\boldsymbol{X}(j,:)) = j$, i.e. to compute the inverse image ($\mathcal{C}^{-1}$) of the integer $j$ (for $j = 1, ..., 2^m$), simply take the $j$th row of $\boldsymbol{X}$. Also, to compute the inverse image ($\varrho^{-1}$) of the integer $j$, simply construct the set of the columns corresponding to the non null entry of the $j$th row of $\boldsymbol{X}$.*

Now that we know how to do calculations with the mappings $\mathcal{C}$ and $\varrho$, we can show how to compute the *a priori* transition probabilities necessary in the filtering step of the gibbs sampler.

**Algorithm 3.** ***a priori Transition Probabilities in the Gibbs Sampler***

*As an input to this step, give the matrices $\boldsymbol{\Theta}^{(n)}$ and $\boldsymbol{\Lambda}^{(n)}$.*

- *Compute the matrix $\boldsymbol{A}^{(n)}$ from $\boldsymbol{\Theta}^{(n)}$ and $\boldsymbol{\Lambda}^{(n)}$ using the mapping $\varrho$ using the following formula*

$$\boldsymbol{A}^{(n)}(\varrho(\mathcal{K}), \varrho(\mathcal{J})) = \exp\left[\sum_{i\in\mathcal{J}} \theta^i + \sum_{j\in\mathcal{K}\cap\mathcal{J}} \lambda^j + 2.\sum_{i\in\mathcal{J}, j\neq i\in\mathcal{J}} \theta^{i,j}\right]$$

- *Compute the sequence of vectors $\boldsymbol{b}_t$ using the following recursion*

$$\boldsymbol{b}_{t-1} = \boldsymbol{A}^{(n)\prime}.\boldsymbol{b}_t$$

*where the recursion is initialized by $\boldsymbol{b}_T = \boldsymbol{e}_1$ the first column of the identity matrix $\boldsymbol{I}_{2^m}$.*

- *Compute the constant of integration $K_T^{(n)}$*

$$K_T = \boldsymbol{e}_1'.\boldsymbol{A}^{T-1}.\boldsymbol{\kappa}_1$$

*where $\boldsymbol{\kappa}_1$ is the sum of the columns of $\boldsymbol{A}^{(n)}$ and $\boldsymbol{e}_1$ is the first column of the identity matrix $\boldsymbol{I}_{2^m}$.*

- *Compute the two sequences of $2^m \times 1$ vectors $\boldsymbol{\xi}_t$ and $\boldsymbol{\zeta}_t$ given by the formulas*

$$\begin{aligned}\boldsymbol{\zeta}_t(j,1) &= \exp(\mathrm{tr}[\boldsymbol{z}_t'.\boldsymbol{z}_t.\boldsymbol{\theta}] + \mathrm{tr}[\boldsymbol{z}_t'.\boldsymbol{z}_{t-1}.\boldsymbol{\Lambda}] + \mathrm{tr}[\mathrm{diag}(\boldsymbol{z}_t).\mathfrak{S}_j.\boldsymbol{\Lambda}]) \\ \boldsymbol{\xi}_t(j,1) &= \exp(\mathrm{tr}[\mathrm{diag}(\boldsymbol{z}_t).\mathfrak{S}_j.\boldsymbol{\Lambda}])\end{aligned}$$

*where $\mathfrak{S}_j = \mathrm{diag}(\mathcal{C}^{-1}(j))$.*

*Using the previous computations, the a priori transition probabilities are*

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \boldsymbol{b}_t'.\boldsymbol{\zeta}_t/\boldsymbol{b}_{t-1}'.\boldsymbol{\xi}_{t-1}$$

*and the initial density is*

$$p(\boldsymbol{z}_1) = \boldsymbol{b}_1'.\boldsymbol{\xi}_1/K_T^{(n)}?$$

Now we turn to drawings from $p(\boldsymbol{\Theta}^{(n)}, \boldsymbol{\Lambda}^{(n)}|\boldsymbol{Z}^{(n)})$. This conditional density is proportional to the product of the matrix-variate bernoulli density $p(\boldsymbol{Z}^{(n)}|\boldsymbol{\Theta}^{(n)}, \boldsymbol{\Lambda}^{(n)})$ and the product of the normal priors on the entries of $\boldsymbol{\Theta}^{(n)}$ and $\boldsymbol{\Lambda}^{(n)}$. A Metropolis-Hastings step is used to draw from this density.

## 4.5 An Illustration in detail: The Bivariate Case

### 4.5.1 The covariance structure in detail

In the bivariate case, $\boldsymbol{y}_t = (y_{t,1}, y_{t,2})'$ and $\boldsymbol{z}_t = (z_{t,1}, z_{t,2})'$.

$$\begin{cases} \boldsymbol{y}_t = (\boldsymbol{\Sigma}_{\boldsymbol{z}_t})^{1/2}.\boldsymbol{\varepsilon}_t \\ \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_2) \end{cases}$$

where $(\boldsymbol{\Sigma}_{\boldsymbol{z}_t})^{\frac{1}{2}}$ is the cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{z}_t}$

$$\boldsymbol{\Sigma}_{\boldsymbol{z}_t} = \begin{pmatrix} \sigma^2_{z_{t,1}} & \rho_{1,2}.\sigma_{z_{t,1}}.\sigma_{z_{t,2}} \\ & \sigma^2_{z_{t,2}} \end{pmatrix}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{z}_t} = \boldsymbol{S}_{\boldsymbol{z}_t}.\boldsymbol{R}.\boldsymbol{S}_{\boldsymbol{z}_t}$$

Therefore, there are four possible matrices $\boldsymbol{\Sigma}_{\boldsymbol{z}_t}$ (In the general case, there would be $2^m$ possible matrices $\boldsymbol{\Sigma}_{\boldsymbol{z}_t}$ because there are $2^m$ possible vectors $\boldsymbol{z}_t$.)

$$\begin{aligned} \boldsymbol{\Sigma}_{(0,0)} &= \begin{pmatrix} \sigma^2_{1,1} & \rho_{1,2}.\sigma_{1,1}.\sigma_{1,2} \\ & \sigma^2_{1,2} \end{pmatrix} \\ \boldsymbol{\Sigma}_{(0,1)} &= \begin{pmatrix} \sigma^2_{1,1} & \rho_{1,2}.\sigma_{1,1}.\sigma_{2,2} \\ & \sigma^2_{2,2} \end{pmatrix} \\ \boldsymbol{\Sigma}_{(1,0)} &= \begin{pmatrix} \sigma^2_{2,1} & \rho_{1,2}.\sigma_{2,1}.\sigma_{1,2} \\ & \sigma^2_{1,2} \end{pmatrix} \\ \boldsymbol{\Sigma}_{(1,1)} &= \begin{pmatrix} \sigma^2_{2,1} & \rho_{1,2}.\sigma_{2,1}.\sigma_{2,2} \\ & \sigma^2_{2,2} \end{pmatrix} \end{aligned}$$

In the previous four matrices $\sigma^2_{k,i}$ stands for the variance of the $i$-th variable in $\boldsymbol{y}_t$ (i.e. $i = 1, ..., m$) and the $k$-th state of that same variable. Because there are two states for each variance, i.e. state 1 and state 2, then $k = 1$ or $k = 2$.[2]

The variances for each state are put in two vectors

$$\boldsymbol{s}_1 = \begin{pmatrix} \sigma^2_{1,1} \\ \sigma^2_{1,2} \end{pmatrix} \quad , \quad \boldsymbol{s}_2 = \begin{pmatrix} \sigma^2_{2,1} \\ \sigma^2_{2,2} \end{pmatrix}$$

where $\boldsymbol{s}_1$ corresponds to the state coded as 0 (for the whole vector $\boldsymbol{y}_t$) and $\boldsymbol{s}_2$ corresponds to the state coded as 1.

Note also that

$$\begin{aligned} \boldsymbol{S}_{(0,0)} &= \begin{pmatrix} \sigma^2_{1,1} & 0 \\ & \sigma^2_{1,2} \end{pmatrix} \\ \boldsymbol{S}_{(0,1)} &= \begin{pmatrix} \sigma^2_{1,1} & 0 \\ & \sigma^2_{2,2} \end{pmatrix} \\ \boldsymbol{S}_{(1,0)} &= \begin{pmatrix} \sigma^2_{2,1} & 0 \\ & \sigma^2_{1,2} \end{pmatrix} \\ \boldsymbol{S}_{(1,1)} &= \begin{pmatrix} \sigma^2_{2,1} & 0 \\ & \sigma^2_{2,2} \end{pmatrix} \end{aligned}$$

and that the correlation matrix is

$$\boldsymbol{R} = \begin{pmatrix} 1 & \rho_{1,2} \\ & 1 \end{pmatrix}$$

**Note 1. Identification issues for the covariances matrices**

I chose here the more parsimonious way where the correlations between the different entries of $\boldsymbol{y}_t$ is fixed.

In that setting, either of two methodologies can be adopted. We will illustrate those two methodologies in the bivariate case since it is clearer in that special case.

First, either fix the correlation between $y_{t,1}$ and $y_{t,2}$, or fix the covariance. The covariance and correlation cannot be fixed at the same time because the following equalities will be impossible

$$\rho_{\boldsymbol{y}_1,\boldsymbol{y}_2} = \frac{\sigma_{\boldsymbol{y}_1,\boldsymbol{y}_2}}{\sigma_{1,1}.\sigma_{1,2}} = \frac{\sigma_{\boldsymbol{y}_1,\boldsymbol{y}_2}}{\sigma_{1,1}.\sigma_{2,2}} = \frac{\sigma_{\boldsymbol{y}_1,\boldsymbol{y}_2}}{\sigma_{2,1}.\sigma_{1,2}} = \frac{\sigma_{\boldsymbol{y}_1,\boldsymbol{y}_2}}{\sigma_{2,1}.\sigma_{2,2}}$$

where the denominator correspond to the standard deviations in different states ($\sigma_{k,i}$ is the standard deviation of $y_{t,i}$ when the state is $k$ and $k = 1, 2$) and where $\rho_{\boldsymbol{y}_1,\boldsymbol{y}_2}$ and $\sigma_{\boldsymbol{y}_1,\boldsymbol{y}_2}$ are respectively the correlation and covariance between $y_{t,1}$ and $y_{t,2}$.

2. Notice that here there is a slight abuse of notation because if $\boldsymbol{z}_t = (a,\ b)$ then $\boldsymbol{\Sigma}_{(a,b)} = \begin{pmatrix} \sigma^2_{a+1,1} & \sigma_{1,2} \\ & \sigma^2_{b+1,2} \end{pmatrix}$ and not $\begin{pmatrix} \sigma^2_{a,1} & \sigma_{1,2} \\ & \sigma^2_{b,2} \end{pmatrix}$, i.e. we are coding the state 0 by 1 and the state 1 by 2.

### 4.5.2 The priors and conditional posteriors of the covariance parameters

The prior on $\rho_{1,2}$ is the uniform distribution on $[-1,1]$, i.e. $p(\rho_{1,2})=\frac{1}{2}$.

The priors on $\sigma^2_{i,k}$ are $\mathcal{IG}\left(\frac{\nu^0_{k,i}}{2},\frac{s^0_{k,i}}{2}\right)$, i.e. they have the densities

$$p(\sigma^2_{i,k})=\frac{\left(\frac{s^0_{k,i}}{2}\right)^{\frac{\nu^0_{k,i}}{2}}}{\Gamma\left(\frac{\nu^0_{k,i}}{2}\right)}.\left(\sigma^2_{i,k}\right)^{-\left(\frac{\nu^0_{k,i}}{2}+1\right)}.\exp\left(-\frac{s^0_{k,i}}{2\sigma^2_{i,k}}\right)$$

From an inferential point of view, conditioning on $\boldsymbol{Z}$ is equivalent to dividing the sample for each series into two series.

For instance, the first variable $\boldsymbol{y}_1$ is divided into two subsamples and $y_{t,1}$ is put in one of those samples if the $z_{t,1}$ is 0 or 1. Let $n_{1,1}$ and $n_{2,1}$ be the sizes of those two subsamples. Therefore $n_{k,i}$ corresponds to the size of the $k$-th subsample of variable $i$ ($k=1,2$ and $i=1,...,m$)

Let $s_{k,i}=\sum_{j=1}^{n_{k,i}} y^2_{j,i}$ (i.e the sum of squares of the observation in the $k$-th subsample of variable $i$).

Note that, in the case of zero correlations, the posterior of $\sigma^2_{k,i}$ conditional on everything else is

$$\mathcal{IG}\left(\frac{n_{k,i}+\nu^0_{k,i}}{2},\frac{s_{k,i}+s^0_{k,i}}{2}\right)$$

we can define

$$\begin{aligned} n^1_{k,i} &= n_{k,i}+\nu^0_{k,i} \\ s^1_{k,i} &= s_{k,i}+s^0_{k,i} \end{aligned}$$

$$p(\sigma^2_{k,i}|\text{everything else})=\frac{\left(\frac{s^1_{k,i}}{2}\right)^{\frac{\nu^1_{k,i}}{2}}}{\Gamma\left(\frac{\nu^1_{k,i}}{2}\right)}.\left(\sigma^2_{i,k}\right)^{-\left(\frac{\nu^1_{k,i}}{2}+1\right)}.\exp\left(-\frac{s^1_{k,i}}{2\sigma^2_{i,k}}\right)$$

This shall be proved in the next paragraph.

**A Proof for the posterior formulation conditional on the states and its parameters**

We are going to attempt to write the likelihood conditional on $\boldsymbol{Z}$ (and its parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}$) up to a constant of proportionality. Remember that if the state realizations are known, then the likelihood is equivalent to that of a linear models on different subsamples.

As earlier $\mathcal{I}$ is the index set $\{1,...,2^m\}$. I.e., in this case, $\mathcal{I}=\{1,2,3,4\}$ that corresponds, by the mapping $\mathcal{C}^{-1}$ to $\mathcal{Z}=\{(0,0),(0,1),(1,0),(1,1)\}$. Let $n_i$ for $i\in\mathcal{I}$ be the number of occurrences of each vector in $\mathcal{Z}$. That is $T=\sum_{i\in\mathcal{I}} n_i$. Let $\boldsymbol{y}^{(i)}$ be $n_i$ occurrences of vectors $\boldsymbol{y}_t$.

Then the likelihood of each one of the samples indexed by $i$ is given

$$p(\boldsymbol{y}^{(i)}|.) \propto \prod_{j=1}^{n_i} |\boldsymbol{\Sigma}_{\mathcal{C}^{-1}(i)}|^{-\frac{1}{2}}.\exp\left( -\frac{1}{2}\boldsymbol{y}^{(i)\prime}.\boldsymbol{\Sigma}_{\mathcal{C}^{-1}(i)}^{-1}.\boldsymbol{y}^{(i)}\right)$$

and therefore $p(\boldsymbol{Y}|.) = \prod_{i\in\mathcal{I}} p(\boldsymbol{y}^{(i)}|.)$.

Multiplying by the priors we obtain the posterior

$$p\left(\left\{\boldsymbol{\Sigma}_{\mathcal{C}^{-1}(i)}\right\}_{i\in\mathcal{I}}\middle|\boldsymbol{Z},\boldsymbol{Y},\boldsymbol{\Theta},\boldsymbol{\Lambda}\right) \propto p(\boldsymbol{R}).\prod_{i'\in\mathcal{M}}\prod_{k\in\{1,2\}} p(\sigma_{i',k}^2).\prod_{i\in\mathcal{I}} p(\boldsymbol{y}^{(i)}|.)$$

**A Note on the quadratic form $\boldsymbol{x}'.\boldsymbol{\Sigma}^{-1}.\boldsymbol{x}$**

$$\begin{aligned}\boldsymbol{x}'.\boldsymbol{\Sigma}^{-1}.\boldsymbol{x} &= \boldsymbol{x}'.(\boldsymbol{S}.\boldsymbol{R}.\boldsymbol{S})^{-1}.\boldsymbol{x}\\ &= \boldsymbol{x}'.\boldsymbol{S}^{-1}.\boldsymbol{R}^{-1}.\boldsymbol{S}^{-1}.\boldsymbol{x}\\ &= ...\\ &= \frac{1}{|\boldsymbol{R}|}\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{x_i.x_j}{\sigma_i.\sigma_j}M_{i,j}(\boldsymbol{R})\end{aligned}$$

where $|\boldsymbol{R}|$ is the determinant of $\boldsymbol{R}$ and $M_{i,j}(\boldsymbol{R})$ is the $(i,\,j)$-th minor of $\boldsymbol{R}$. Also $x_i$ is the $i$-th entry of the vector $\boldsymbol{x}$.

This signifies that, when the quadratic form is considered as a function of $\boldsymbol{x}$, the coefficient of $x_i^2$ is $\frac{M_{i,i}(\boldsymbol{R})}{\sigma_i^2.|\boldsymbol{R}|}$ and the coefficient of the cross products $x_i.x_j$ is $2\frac{M_{i,j}(\boldsymbol{R})}{\sigma_i.\sigma_j.|\boldsymbol{R}|}$. When, the quadratic form is considered as a function of each one of the standard deviations $\sigma_i$ for $i = 1, ..., m$, then the coefficient of $1/\sigma_i^2$ and $1/\sigma_i$ are respectively

| Variable | Coefficient |
|---|---|
| $\frac{1}{\sigma_i^2}$ | $x_i^2.\frac{M_{i,i}(\boldsymbol{R})}{\lvert\boldsymbol{R}\rvert}$ |
| $\frac{1}{\sigma_i}$ | $2x_i.\sum_{j\neq i}\frac{x_j}{\sigma_j}.\frac{M_{i,j}(\boldsymbol{R})}{\lvert\boldsymbol{R}\rvert}$ |

This proves that in the case of non-zero correlations, the posteriors for $\sigma_{k,i}^2$ are no longer inverse gamma and in that case, a metropolis-hastings step is utilized.

# 5 An empirical illustration

I consider here a part of the dataset used in [Audrino and Trojani, 2006] and it was download from the Journal of Applied Econometrics data archive. I use daily returns obtained from three stock market indices that respectively the French CAC40 index, the Swiss SMI Index and US S&P500 Index. The daily price data from the return series were constructed span the period from January 1, 1990 to November 4, 2002. Our sample consists of three variables with 3350 observations each.

In the following figures, the return series is showed with the filtered probabilities of the state of lower volatility estimated from the multivariate markov-switching model considered earlier with diagonal covariance matrices.

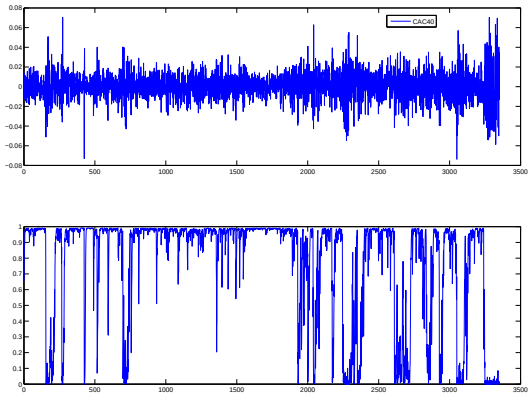


**Figure 1.**

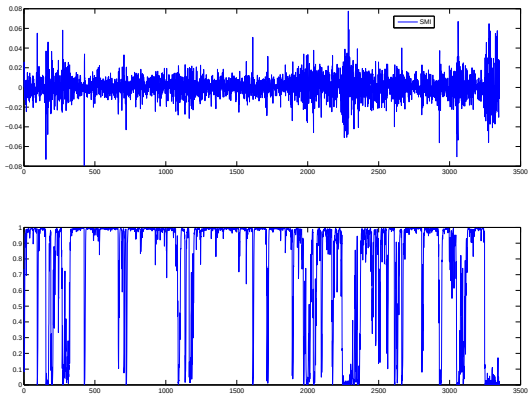


**Figure 2.**

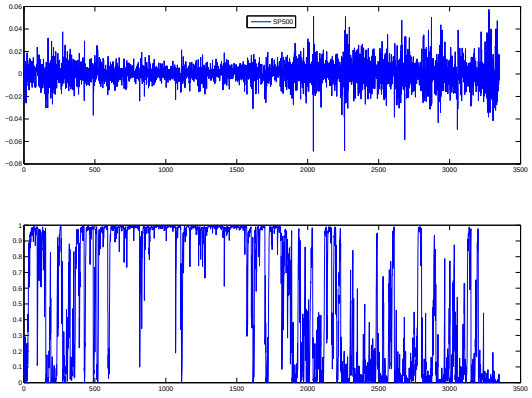


**Figure 3.**

By plotting the three filtered probabilities next to each other, we can see that this seems to indicate that a single latent binary variable for modeling the three-dimensional returns vector is inadequate at best. For instance, the mean absolute deviation between the filtered probabilities series between the Swiss and US is 0.3, not high enough to warrant independence and not low enough to assume a single latent variable. (It is 0.15 between the French and the Swiss and 0.28 between the French and the US).

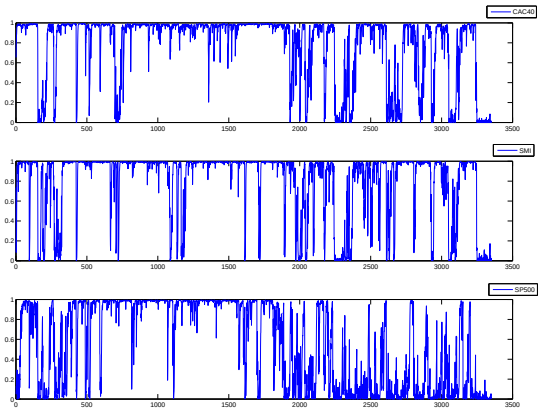


**Figure 4.**

# 6 Conclusion

The basic setup of the paper can be extended into several directions. One topic for future research is relaxing the restriction to the two-state model. What is needed is a matrix-variate multinomial distribution for a more general use of the methodology in this paper. This seems promising for future research. Since a natural representation of a vector of multinomial variables can be a matrix of indicator variables coding the states, the matrix-variate bernoulli distribution can in fact be used to model multinomial variables. The ideas are going to be further explored by the author in future work.

Another investigation are is model selection. Due to the big number of possible alternative modeling possibilities, automatic selection procedures are of great interest and very difficult to carry out. This might prove a quite intriguing and definitely challenging problem for future research.

On the purely computational side, a study of the efficiency of different MCMC algorithms constitutes an interesting complementary study to be performed in the future. One intriguing theoretical aspect for further study is the non-homogeneity of the latent markov chain underlying the markov-swithcing model. It should be investigated in several directions.

On the application side, the multivariate volatility cluster special case could be augmented in order to incorporate more complex mean dynamics.

# Bibliography

**[Audrino and Trojani, 2006]** Audrino, F. and Trojani, F. (2006). Estimating and predicting multivariate volatility thresholds in global stock markets. *Journal of Applied Econometrics*, 21:345–369.

**[Barnard et al., 2000]** Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10:1281–1311.

**[Bauwens et al., 2006]** Bauwens, L., Laurent, S., and Rombouts, J. V. K. (2006). Multivariate garch models: a survey. *Journal of Applied Econometrics*, 21:79–109.

**[Bauwens et al., 1999]** Bauwens, L., Lubrano, M., and Richard, J.-F. (1999). *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press.

**[Chen et al., 2000]** Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer.

**[Cox, 1972]** Cox, D. R. (1972). The analysis of multivariate binary data. *Applied Statistics*, 21:113–120.

**[Cox and Wermuth, 1994]** Cox, D. R. and Wermuth, N. (1994). A note on the quadratic exponential binary distribution. *Biometrika*, 81:403–408.

**[Dempster et al., 1977]** Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38.

**[Gelman et al., 1995]** Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman and Hall.

**[Hamilton, 1994]** Hamilton, J. (1994). *Time Series Analysis*. Princeton University Press.

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

**[Kim and Nelson, 1999]** Kim, C.-J. and Nelson, C. (1999). *State-Space Models with Regime-Switching: Classical and Gibbs-Sampling Approaches with Applications*. MIT Press.

**[Koop, 2003]** Koop, G. (2003). *Bayesian Econometrics*. Wiley.

**[Krolzig, 1997]** Krolzig, H. (1997). *Markov-Switching Vector Autoregressions*. Springer.

**[Lovison, 2006]** Lovison, G. (2006). A matrix-valued bernoulli distribution. *Journal of Multivariate Analysis*, 97:1573–1585.

**[Pelletier, 2006]** Pelletier, D. (2006). Regime switching for dynamic correlations. *Journal of Econometrics*, 131:445–473.

**[Robert and Casella, 2004]** Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Models*. Springer-Verlag, second edition.

**[Robert, 2007]** Robert, C. P. (2007). *The Bayesian Choice. From Decision-Theoretic Foundations to Computational Implementation*. Springer, second edition.

**[Rousseeuw and Molenberghs, 1994]** Rousseeuw, P. J. and Molenberghs, G. (1994). The shape of correlation matrices. *The American Statistician*, 48:276–279.

**[Sims and Zha, 2006]** Sims, C. and Zha, T. (2006). Were there regime switches in US monetary policy? *American Economic Review*, 96:54–81.

**[Zhao and Prentice, 1990]** Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, 77:642–648.

# Appendix A Proofs

**Proof. Theorem 1**

Let us integrate some element of $\boldsymbol{\gamma}_t$ with respect to $\boldsymbol{z}_t$, say $\boldsymbol{\gamma}_t(\varrho(\mathcal{K}))$ for some $\mathcal{K} \in \mathcal{P}(\mathcal{M})$. If we obtain a linear combination of $\boldsymbol{\gamma}_{t-1}$, then the rest of the proof follows by induction.

$\boldsymbol{\gamma}_t(\varrho(\mathcal{K}))$ is equal to

$$\exp\left[\sum_{i=1}^{m} y_i^t.\theta^i + \sum_{j\in\mathcal{K}} \left(z_{t,j} + w_j^t\right).\lambda^j + \sum_{j\in\mathcal{M}\setminus\mathcal{K}} w_j^t.\lambda^j + 2\sum_{i=1,j\neq i}^{m} u_{i,j}^t.\theta^{i,j}\right]$$

which is also equal to

$$\exp\left[\sum_{i=1}^{m}\left[z_{t,i} + y_i^{t-1}\right].\theta^i + \sum_{j\in\mathcal{K}}\left[z_{t,j} + z_{t,j}.z_{t-1,j} + w_j^{t-1}\right].\lambda^j + \sum_{j\in\mathcal{M}\setminus\mathcal{K}}\left[z_{t,j}.z_{t-1,j} + w_j^{t-1}\right].\lambda^j + 2\sum_{i=1,j\neq i}^{m}\left[z_{t,i}.z_{t,j} + u_{i,j}^{t-1}\right].\theta^{i,j}\right]$$

or equivalently

$$\exp\left[\sum_{i\in\mathcal{M}}\left[z_{t,i} + y_i^{t-1}\right].\theta^i + \left[z_{t,i}.z_{t-1,i} + w_i^{t-1}\right].\lambda^i + \sum_{j\in\mathcal{K}} z_{t,j}.\lambda^j + 2\sum_{i\in\mathcal{M},j\neq i\in\mathcal{M}}\left[z_{t,i}.z_{t,j} + u_{i,j}^{t-1}\right].\theta^{i,j}\right]$$

Now, integrating over $\boldsymbol{z}_t$ is done by summing over all $2^m$ different vectors $\boldsymbol{z}_t$ that can occur. Every such $\boldsymbol{z}_t$ is given by $\mathcal{C}^{-1}\circ\varrho(\mathcal{J})\ \forall\mathcal{J}\in\mathcal{P}(\mathcal{M})$.

As a preliminary step, let us replace one $\boldsymbol{z}_t$ (say the one corresponding to $\mathcal{C}^{-1}\circ\varrho(\mathcal{J})$ for a given $\mathcal{J}$) by its value. Such $\boldsymbol{z}_t$ contains 1 over $i\in\mathcal{J}$ and 0 over over $i\in\mathcal{M}\setminus\mathcal{J}$. Therefore, the kernel entry corresponding to $\mathcal{K}$ evaluated at the $\boldsymbol{z}_t$ corresponding to $\mathcal{J}$, i.e. $\boldsymbol{\gamma}_t(\varrho(\mathcal{K}))|_{\boldsymbol{z}_t=\mathcal{C}^{-1}\circ\varrho(\mathcal{J})}$ can be written in the following way

$$\exp\left[\sum_{k\in\mathcal{J}}\left\{\sum_{i\in\mathcal{M}\cap\mathcal{J}}\left[1 + y_i^{t-1}\right].\theta^i + \left[z_{t-1,i} + w_i^{t-1}\right].\lambda^i + \sum_{j\in\mathcal{K}\cap\mathcal{J}}\lambda^j + 2\sum_{i\in\mathcal{M}\cap\mathcal{J},j\neq i\in\mathcal{J}}\left[1 + u_{i,j}^{t-1}\right].\theta^{i,j}\right\} + \sum_{k\in\mathcal{M}\setminus\mathcal{J}}\left\{\sum_{i\in\mathcal{M}\cap\mathcal{M}\setminus\mathcal{J}} y_i^{t-1}.\theta^i + w_i^{t-1}.\lambda^i + 2\sum_{i\in\mathcal{M}\cap\mathcal{M}\setminus\mathcal{J},j\neq i\in\mathcal{M}\setminus\mathcal{J}} u_{i,j}^{t-1}.\theta^{i,j}\right\}\right]$$

Now, of course $\mathcal{M} \cap \mathcal{J} = \mathcal{J}$ and $\mathcal{M} \cap (\mathcal{M} \backslash \mathcal{J}) = \mathcal{M} \backslash \mathcal{J}$ because $\mathcal{J} \subset \mathcal{M}$. And also replace redundant summation operators such as $\sum_{k \in \mathcal{J}} \sum_{i \in \mathcal{J}}$ by simply $\sum_{i \in \mathcal{J}}$, we rewrite $\boldsymbol{\gamma}_t(\varrho(\mathcal{K}))|_{\boldsymbol{z}_t = \mathcal{C}^{-1} \circ \varrho(\mathcal{J})}$ as

$$\exp\left[\sum_{i \in \mathcal{J}} \theta^i + \sum_{j \in \mathcal{K} \cap \mathcal{J}} \lambda^j + 2. \sum_{i \in \mathcal{J}, j \neq i \in \mathcal{J}} \theta^{i,j}\right].\exp\left[\sum_{i \in \mathcal{J}} y_i^{t-1}.\theta^i + \left[z_{t-1,i} + w_i^{t-1}\right].\lambda^i + 2 \sum_{i \in \mathcal{J}, j \neq i \in \mathcal{J}} u_{i,j}^{t-1}.\theta^{i,j} + \sum_{i \in \mathcal{M} \backslash \mathcal{J}} y_i^{t-1}.\theta^i + w_i^{t-1}.\lambda^i + 2 \sum_{i \in \mathcal{M} \backslash \mathcal{J}, j \neq i \in \mathcal{M} \backslash \mathcal{J}} u_{i,j}^{t-1}.\theta^{i,j}\right]$$

where we have factored out the therm $\exp\left[\sum_{i \in \mathcal{J}} \theta^i + \sum_{j \in \mathcal{K} \cap \mathcal{J}} \lambda^j + 2. \sum_{i \in \mathcal{J}, j \neq i \in \mathcal{J}} \theta^{i,j}\right]$ because it does not depend on the data $\boldsymbol{Z}$. Now, grouping the terms over $\mathcal{J}$ and those over $\mathcal{M} \backslash \mathcal{J}$, we rewrite $\boldsymbol{\gamma}_t(\varrho(\mathcal{K}))|_{\boldsymbol{z}_t = \mathcal{C}^{-1} \circ \varrho(\mathcal{J})}$ as

$$\exp\left[\sum_{i \in \mathcal{J}} \theta^i + \sum_{j \in \mathcal{K} \cap \mathcal{J}} \lambda^j + 2. \sum_{i \in \mathcal{J}, j \neq i \in \mathcal{J}} \theta^{i,j}\right].\exp\left[\sum_{i \in \mathcal{M}} y_i^{t-1}.\theta^i + w_i^{t-1}.\lambda^i + \sum_{j \in \mathcal{J}} z_{t-1,j}.\lambda^j + 2 \sum_{i \in \mathcal{M}, j \neq i \in \mathcal{M}} u_{i,j}^{t-1}.\theta^{i,j}\right]$$

Now the formula is very clear. On the right, the exponential term is simply $\boldsymbol{\gamma}_{t-1}(\varrho(\mathcal{J}))$. On the the left, the exponential term is simply a multiplicative form that does not depend on the data.

Therefore, replacing $\boldsymbol{z}_t = \mathcal{C}^{-1} \circ \varrho(\mathcal{J})$ inside $\boldsymbol{\gamma}_t(\varrho(\mathcal{K}))$ yields the following formula

$$\boldsymbol{\gamma}_t(\varrho(\mathcal{K}))|_{\boldsymbol{z}_t = \mathcal{C}^{-1} \circ \varrho(\mathcal{J})} = \exp\left[\sum_{i \in \mathcal{J}} \theta^i + \sum_{j \in \mathcal{K} \cap \mathcal{J}} \lambda^j + 2. \sum_{i \in \mathcal{J}, j \neq i \in \mathcal{J}} \theta^{i,j}\right].\boldsymbol{\gamma}_{t-1}(\varrho(\mathcal{J}))$$

Now, summing over $\boldsymbol{z}_t$ inside $\boldsymbol{\gamma}_t(\varrho(\mathcal{K}))$ is equivalent to summing over all $\boldsymbol{z}_t = \mathcal{C}^{-1} \circ \varrho(\mathcal{J})\ \forall \mathcal{J} \in \mathcal{P}(\mathcal{M})$.

$$\begin{aligned}\sum_{\boldsymbol{z}_t} \boldsymbol{\gamma}_t(\varrho(\mathcal{K})) &= \sum_{\mathcal{J} \in \mathcal{P}(\mathcal{M})} \boldsymbol{\gamma}_t(\varrho(\mathcal{K})) \\ &= \sum_{\mathcal{J} \in \mathcal{P}(\mathcal{M})} \exp\left[\sum_{i \in \mathcal{J}} \theta^i + \sum_{j \in \mathcal{K} \cap \mathcal{J}} \lambda^j + 2. \sum_{i \in \mathcal{J}, j \neq i \in \mathcal{J}} \theta^{i,j}\right].\boldsymbol{\gamma}_{t-1}(\varrho(\mathcal{J}))\end{aligned}$$

Therefore, we see that integrating any element of $\boldsymbol{\gamma}_t$, say $\boldsymbol{\gamma}_t(\varrho(\mathcal{K}))$, will yield a linear combination of $\boldsymbol{\gamma}_{t-1}$ where the coefficients are equal to $\exp\left[\sum_{i \in \mathcal{J}} \theta^i + \sum_{j \in \mathcal{K} \cap \mathcal{J}} \lambda^j\right]$

□

**Proof. Corollary 1**

The proof follows easily from that of theorem 1.

All we have to do is write the identity

$$\sum_{\boldsymbol{z}_t} \boldsymbol{\gamma}_t(\varrho(\mathcal{K})) = \sum_{\mathcal{J} \in \mathcal{P}(\mathcal{M})} \exp\left[\sum_{i \in \mathcal{J}} \theta^i + \sum_{j \in \mathcal{K} \cap \mathcal{J}} \lambda^j + 2. \sum_{i \in \mathcal{J}, j \neq i \in \mathcal{J}} \theta^{i,j}\right].\boldsymbol{\gamma}_{t-1}(\varrho(\mathcal{J}))$$

in matrix form.

□

**Proof. Proposition 1**

The proof is carried out by induction.

Begin by writing $p(\boldsymbol{z}_1, ..., \boldsymbol{z}_T)$ as $\boldsymbol{e}_1'.\boldsymbol{\gamma}_T/K_T$. This simply true because the kernel of the density of $\boldsymbol{Z}$ is equal to the first entry in $\boldsymbol{\gamma}_T$.

The first marginal density $p(\boldsymbol{z}_1, ..., \boldsymbol{z}_T)$ can be written in the form $\boldsymbol{b}_T'.\boldsymbol{\gamma}_T/K_T$ with $\boldsymbol{b}_T = \boldsymbol{e}_1$.

Let us assume that $p(\boldsymbol{z}_1, ..., \boldsymbol{z}_t)$ can be written as $\boldsymbol{b}_t'.\boldsymbol{\gamma}_t/K_T$. We would like to deduce now that $p(\boldsymbol{z}_1, ..., \boldsymbol{z}_{t-1})$ can be written in the form $\boldsymbol{b}_{t-1}'.\boldsymbol{\gamma}_{t-1}/K_T$ and then deduce the relation between $\boldsymbol{b}_t$ and $\boldsymbol{b}_{t-1}$.

Carry out the integration operation over $\boldsymbol{z}_t$

$$\begin{aligned} p(\boldsymbol{z}_1, ..., \boldsymbol{z}_{t-1}) &= \sum_{\boldsymbol{z}_t} p(\boldsymbol{z}_1, ..., \boldsymbol{z}_{t-1}, \boldsymbol{z}_t) \\ &= \\ &= \sum_{\boldsymbol{z}_t} \boldsymbol{b}_t'.\boldsymbol{\gamma}_t/K_T \\ &= K_T^{-1}.\boldsymbol{b}_t' \sum_{\boldsymbol{z}_t} \boldsymbol{\gamma}_t \\ &= K_T^{-1}.\boldsymbol{b}_t'.\boldsymbol{A}.\boldsymbol{\gamma}_{t-1} \\ &\quad \text{now, replace } \boldsymbol{b}_t'.\boldsymbol{A} \text{ by } \boldsymbol{b}_{t-1}' \\ &= K_T^{-1}.\boldsymbol{b}_{t-1}'.\boldsymbol{\gamma}_{t-1} \end{aligned}$$

which proves that we can deduce the formula $p(\boldsymbol{z}_1, ..., \boldsymbol{z}_{t-1}) = \boldsymbol{b}_{t-1}'.\boldsymbol{\gamma}_{t-1}/K_T$ and the recursion $\boldsymbol{b}_{t-1} = \boldsymbol{A}'.\boldsymbol{b}_t$. □

**Proof. Proposition 2**

Remember that we can obtain $p(\boldsymbol{z}_1)$ from $p(\boldsymbol{z}_1, ..., \boldsymbol{z}_T)$ by way of proposition 1, i.e. $p(\boldsymbol{z}_1) = \boldsymbol{b}_1'.\boldsymbol{\gamma}_1/K_T$

Integrating over $\boldsymbol{z}_1$ will yield 1. $\sum_{\boldsymbol{z}_1} p(\boldsymbol{z}_1) = 1$ and therefore

$$\begin{aligned} \sum_{\boldsymbol{z}_1} p(\boldsymbol{z}_1) &= 1 \\ &= K_T^{-1}.\boldsymbol{b}_1'. \sum_{\boldsymbol{z}_1} \boldsymbol{\gamma}_1 \end{aligned}$$

And therefore $K_T = \boldsymbol{b}_1' \sum_{\boldsymbol{z}_1} \boldsymbol{\gamma}_1$ □

**Proof. Corollary 2**

The proof follows easily by induction from the application of the formula

$$K_T(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) = \boldsymbol{e}_1'.\boldsymbol{A}^{T-1}.\boldsymbol{\kappa}_1$$

□

**Proof. Proposition 3**

We will analyze the ratio $\boldsymbol{b}_t'.\boldsymbol{\gamma}_t/\boldsymbol{b}_{t-1}'.\boldsymbol{\gamma}_{t-1}$

Write $\boldsymbol{b}'_{t-1}.\boldsymbol{\gamma}_{t-1}$ as $\sum_{k=1}^{2^m} b_{t-1}(k).\boldsymbol{\gamma}_{t-1}(k)$ and then divide it by the scalar $\boldsymbol{\gamma}_{t-1}(\varrho^{-1}(\phi))$ (i.e. the kernel of a density that has $t-1$ observations only)

$$\begin{aligned}\boldsymbol{b}'_{t-1}.\boldsymbol{\gamma}_{t-1} &= \frac{\sum_{k=1}^{2^m} b_{t-1}(k).\boldsymbol{\gamma}_{t-1}(k)}{\boldsymbol{\gamma}_{t-1}(\varrho(\phi))} \\ &= \sum_{k=1}^{2^m} b_{t-1}(k).\frac{\boldsymbol{\gamma}_{t-1}(k)}{\boldsymbol{\gamma}_{t-1}(\varrho(\phi))}\end{aligned}$$

(note here that the set of indices $k$ for $k = 1, ..., 2^m$ are given by the mapping $k = \varrho(\mathcal{K})$).

The ratio $\boldsymbol{\gamma}_{t-1}(k)/\boldsymbol{\gamma}_{t-1}(\varrho(\phi))$ can be easily proved to be equal to

$$\begin{aligned}\frac{\boldsymbol{\gamma}_{t-1}(k)}{\boldsymbol{\gamma}_{t-1}(\varrho(\phi))} &= \exp\Bigg[\sum_{i=1}^{M} y_i^{t-1}.\theta^i + \sum_{j\in\mathcal{K}}\left(z_{t-1,j} + w_j^{t-1}\right).\lambda^j + \sum_{j\in\mathcal{M}\setminus\mathcal{K}} w_j^{t-1}.\lambda^j + \\ &\quad 2\sum_{i=1,j\neq i}^{M} u_{i,j}^{t-1}.\theta^{i,j} - \sum_{i=1}^{M} y_i^{t-1}.\theta^i - \sum_{j\in\phi}\left(z_{t-1,j} + w_j^{t-1}\right).\lambda^j - \\ &\quad \sum_{j\in\mathcal{M}\setminus\phi} w_j^{t-1}.\lambda^j - 2\sum_{i=1,j\neq i}^{M} u_{i,j}^{t-1}.\theta^{i,j}\Bigg] \\ &= \exp\left[\sum_{j\in\mathcal{K}}\left(z_{t-1,j} + w_j^{t-1}\right).\lambda^j + \sum_{j\in\mathcal{M}\setminus\mathcal{K}} w_j^{t-1}.\lambda^j - \sum_{j\in\mathcal{M}} w_j^{t-1}.\lambda^j\right] \\ &= \exp\left[\sum_{j\in\mathcal{K}} z_{t-1,j}.\lambda^j\right]\end{aligned}$$

Similarly, we have

$$\boldsymbol{b}'_t.\boldsymbol{\gamma}_t = \sum_{k=1}^{2^m} b_t(k).\frac{\boldsymbol{\gamma}_t(k)}{\boldsymbol{\gamma}_t(\varrho(\phi))}$$

and the ratio $\boldsymbol{\gamma}_t(k)/\boldsymbol{\gamma}_t(\varrho(\phi))$ can be computed accordingly

$$\begin{aligned}\frac{\boldsymbol{\gamma}_t(k)}{\boldsymbol{\gamma}_t(\varrho(\phi))} &= \exp\Bigg[\sum_{i=1}^{m} y_i^t.\theta^i + \sum_{j\in\mathcal{K}}\left(z_{t,j} + w_j^t\right).\lambda^j + \sum_{j\in\mathcal{M}\setminus\mathcal{K}} w_j^t.\lambda^j + 2\sum_{i=1,j\neq i}^{m} u_{i,j}^t.\theta^{i,j} - \\ &\quad \sum_{i=1}^{m} y_i^{t-1}.\theta^i - \sum_{j\in\phi}\left(z_{t-1,j} + w_j^{t-1}\right).\lambda^j - \sum_{j\in\mathcal{M}\setminus\phi} w_j^{t-1}.\lambda^j - 2\sum_{i=1,j\neq i}^{m} \\ &\quad u_{i,j}^{t-1}.\theta^{i,j}\Bigg] \\ &= \exp\Bigg[\sum_{i=1}^{m}\left(y_i^t - y_i^{t-1}\right).\theta^i + \sum_{j\in\mathcal{K}} z_{t,j}.\lambda^j + \sum_{j\in\mathcal{M}}\left(w_j^t - w_j^{t-1}\right).\lambda^j + 2\sum_{i=1,j\neq i}^{m} \\ &\quad \left(u_{i,j}^t - u_{i,j}^{t-1}\right).\theta^{i,j}\Bigg] \\ &= \exp\left[\sum_{i=1}^{m} z_{t,i}.\theta^i + \sum_{j\in\mathcal{K}} z_{t,j}.\lambda^j + \sum_{j\in\mathcal{M}} z_{t,j}.z_{t-1,j}.\lambda^j + 2\sum_{i=1,j\neq i}^{m} z_{t,i}.z_{t,j}.\theta^{i,j}\right]\end{aligned}$$

we see that neither $\boldsymbol{\gamma}_{t-1}(k)/\boldsymbol{\gamma}_{t-1}(\varrho(\phi))$ nor $\boldsymbol{\gamma}_t(k)/\boldsymbol{\gamma}_t(\varrho(\phi))$ do contain $\boldsymbol{z}_{t-j}$, $\forall j \leqslant 2$ which proves our claim.

To quickly resume, write the density

$$p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \frac{\sum_{k=1}^{2^m} b_t(k).\exp\Big[\sum_{i=1}^{m} (z_{t,i}.\theta^i + z_{t,i}.z_{t-1,i}.\lambda^i) + \sum_{j\in\mathcal{K}} z_{t,j}.\lambda^j + 2\sum_{i=1,j\neq i}^{m} z_{t,i}.z_{t,j}.\theta^{i,j}\Big]}{\sum_{k=1}^{2^m} b_{t-1}(k).\exp\Big[\sum_{j\in\mathcal{K}} z_{t-1,j}.\lambda^j\Big]}$$

□

**Proof. Proposition 4**

The proof follows easily but tediously from dividing both numerator and denominator of the conditional density by $\boldsymbol{\gamma}_t(\varrho(\phi))$ and simplifying the expression.

□

# Appendix B An algorithm for constructing the mapping $\mathcal{C}$

There are numerous ways of constructing a mapping $\mathcal{C}$ that uniquely pinpoints an integer in the index set $\{1, ..., 2^m\}$ to one possible occurrence of the vector $\boldsymbol{z}_t$. Therefore, it is important to explicitly describe a way of doing so.

We propose the following extremely simple algorithm; Construct a $2^m \times m$ matrix $\boldsymbol{X}$ such that each of its columns is equal to $\boldsymbol{\iota}_{2^{m-i}} \otimes \begin{pmatrix} \mathbf{0}_{2^{i-1}} \\ \boldsymbol{\iota}_{2^{i-1}} \end{pmatrix}$, i.e.

$$\boldsymbol{X}(:,i) = \boldsymbol{\iota}_{2^{m-i}} \otimes \begin{pmatrix} \mathbf{0}_{2^{i-1}} \\ \boldsymbol{\iota}_{2^{i-1}} \end{pmatrix}$$

where $\mathbf{0}_k$ is a column vector of size $k$ that only contains zeros and where $\boldsymbol{\iota}_k$ is a column vector of size $k$ that only only contains ones.

Each row of $\boldsymbol{X}$ will correspond to a unique possible $\boldsymbol{z}_t$ among all possible unique 0-1 permutations of the entries of $\boldsymbol{z}_t$.

And now define $\mathcal{C}$ as being given by the following identity

$$\mathcal{C}(\boldsymbol{X}(j,:)) = j$$

We will give a simple example to clarify the ideas.

Let $m = 3$, then there are $2^m$ different possible vectors $\boldsymbol{z}_t$. The set $\mathcal{Z}$ is therefore

$$\mathcal{Z} = \{(0,0,0),(1,0,0),(0,1,0),(1,1,0),(0,0,1),(1,0,1),(0,1,1),(1,1,1)\}$$

The different columns of $\boldsymbol{X}$ are computed in the following way

$$\begin{aligned} \boldsymbol{X}(:,1) &= \boldsymbol{\iota}_{2^2} \otimes \begin{pmatrix} \mathbf{0}_{2^0} \\ \boldsymbol{\iota}_{2^0} \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
&= \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \\
\boldsymbol{X}(:,2) &= \boldsymbol{\iota}_{2^1} \otimes \begin{pmatrix} \mathbf{0}_{2^1} \\ \boldsymbol{\iota}_{2^1} \end{pmatrix} \\
&= \begin{pmatrix} 1 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \\
\boldsymbol{X}(:,3) &= \boldsymbol{\iota}_{2^0} \otimes \begin{pmatrix} \mathbf{0}_{2^2} \\ \boldsymbol{\iota}_{2^2} \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}
\end{aligned}
$$

and therefore

$$
\boldsymbol{X} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}
$$

Each row of $\boldsymbol{X}$ is a unique $\boldsymbol{z}_t$. The reader can easily check that $\mathcal{C}((0,\,0,\,0)) = 1$, $\mathcal{C}(1,0,0) = 2$, $\mathcal{C}(0,1,0) = 3$ etc...

To quickly show the connection between $\mathcal{C}$ and $\varrho$, we only need to note that $\varrho^{-1} \circ \mathcal{C}$ maps a row of $\boldsymbol{X}$ (where there are $k$ integers 1) into a set $\{i_1, ..., i_k\}$ where each element of $\{i_1, ..., i_k\}$ is the column index referring to where the 1 occurred. E.g.

$$\mathcal{C}\left(\begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{pmatrix} = \varrho\left(\begin{pmatrix} \phi \\ \{1\} \\ \{2\} \\ \{1,2\} \\ \{3\} \\ \{1,3\} \\ \{2,3\} \\ \{1,2,3\} \end{pmatrix}\right)$$

Similarly, if $m = 2$, then

$$\boldsymbol{X} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$$

Here, we can easily see the relation between $\mathcal{C}$ and $\varrho$.

$$\mathcal{C}\left(\begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} = \varrho\left(\begin{pmatrix} \phi \\ \{1\} \\ \{2\} \\ \{1,2\} \end{pmatrix}\right)$$

# Appendix C An Economic Example

**Example 4.** Let $\boldsymbol{y}_t$ be a vector of GDP growth data where each component of the vector corresponds to one country

| variable | country | latent variable |
|---|---|---|
| $y_{t,1}$ | U.S. | $z_{t,1}$ |
| $y_{t,2}$ | Canada | $z_{t,2}$ |
| $y_{t,3}$ | France | $z_{t,3}$ |

The model described in the introduction is

$$\boldsymbol{\Phi}(L).\boldsymbol{y}_t = \boldsymbol{\mu}_{\boldsymbol{z}_t} + \boldsymbol{\varepsilon}_t$$

$$\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu}_{\boldsymbol{z}_t} = \begin{pmatrix} \mu_{1_{z_{t,1}}} \\ \mu_{2_{z_{t,2}}} \\ \mu_{3_{z_{t,3}}} \end{pmatrix}$$

where $\boldsymbol{\Phi}(L)$ is multivariate lag polynomial.

Let $i$ be the country index. $z_{t,i}$ is the latent variable describing expansions and recessions in country $i$. Since $z_{t,i}$ indexes $\mu_i$, the mean of the autoregressive process of GDP growth of country $i$, $\mu_i$ takes two different values $\mu_{i,j}$ for $j = 1, 2$, that is $\mu_{i,1}$ if $z_{t,1} = 0$ and $\mu_{i,2}$ if $z_{t,2} = 1$. If $\mu_{i,1} < \mu_{i,2}$, we can say that the state 0 of latent variable $z_{t,i}$ corresponds to a recession for country $i$. We see that the model is quite flexible and corresponds to the economic discussion in the introduction.