



**INTERNSHIP REPORT FOR DATA SCIENCE MASTER'S DEGREE**

# EXTENDED GENERATIVE UNIFIED MODEL

**Spring 2020**

*By:*
Katherine MORALES
katymq19@gmail.com

*Internship supervisor:*
Yohan Petetin
yohan.petetin@telecom-sudparis.eu

*Performed at:*
CITI Department - Télécom SudParis

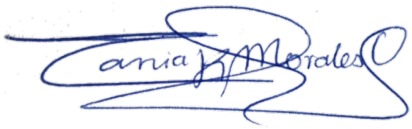
September 2020

# Contents

# List of Figures

# Anti-plagiarism declaration

I hereby declare that all material in this assignment is my own work except where there is clear acknowledgement or reference to the work of others.



Katherine Morales

# Acknowledgment

First of all, I would particularly like to thank my internship supervisor, Mr Yohan Petetin, for his availability, understanding and relevant and constructive advice.

I would like to thank my mother, *Martha*, my brother and sister, *Christopher and Diana*, my dear friends *Adrián and Any*, my family and all my friends in Ecuador who encouraged me during this year of study.

I am grateful to École Polytechnique, and their highly skilled teachers for providing me with a solid technical background and knowledge.

# About the internship

As part of my studies at Ecole Polytechnique, I will end my master studies with a 6 month internship.

This report presents my work of **4 months** and a half that I did at Communications, Images and Information Processing Department (CITI) - Télécom SudParis. I work in Information Processing for Images and Communications (TIPIC) team - SAMOVAR laboratory. Moreover, my internship tutor is Yohan Petetin, an Associate Professor at Télécom SudParis, CITI department and I will continue my internship until **October**.

CITI coordinates teaching and research in statistical signal processing: "wireless" digital communications, system identification and equalization, electromagnetism, antennas, propagation, image modeling: pattern recognition, segmentation, optimization algorithms. The emphasis is on solving open problem solving and/or advancing knowledge in these areas, in addition to traditional application-oriented development efforts. Projects are funded by various sources, including the European Commission, the Institut Télécom, as well as funding of PhD and post-doctoral students through industry support.

On the other hand, the areas of expertise of TIPIC team are at the interface of physics (optical technologies for communication networks), statistical signal processing (information transmission and processing, indoor localization) and applied mathematics (time series, Bayesian inference). Research in statistical signal and image processing focuses on Bayesian inference methods in Hidden Markovian Models (and extensions) and on the analysis of theoretical statistical performances of sequential and iterative algorithms. In the field of optical technologies, research focuses on the generation, transport, control, detection, and processing of optical signals, and more specifically on improving the spectral efficiency of multi-carrier communication techniques. These activities are complemented by contributions in indoor location and numerical calculations of electromagnetic fields.

# Introduction

Sequential data are data in which order matters, for example data from text documents, DNA sequences. This type of data is present in several areas of science and engineering, including for example, signal processing and control, bioinformatics, speech recognition, econometrics, among others. There are several tools to model this type of data, in particular we have the Hidden Markov Models and the Recurrent Neural Networks.

Artificial neural networks can be seen as a mapping from an input space to an output space with connections between neurons. Recurrent Neural Network (RNN) is a class of neural networks that has demonstrated its ability to process sequential data [DiPietro and Hager, 2020, Liu et al., 2019]. RNN has memory, in other words, the current output is considered as an input for the next output. The most common RNN are Long Short-Term Memory (LSTM), which makes it easier to remember past data in memory [Manaswi, 2018], and Gated Recurrent Units (GRU), which are similar to a LSTM with a forget gate.[Shewalkar et al., 2019]. The principal applications of RNN are in natural language processing [Thakker et al., 2019], speech recognition [Graves et al., 2013], music generation [Goel et al., 2014], sentiment classification [Tang et al., 2015] [Zhang et al., 2016] and machine translation [Liu et al., 2014].

A Markov chain can be used to calculate probabilities for a sequence of observed events. However, these events may be hidden, ie not directly observable [Keselj, 2009]. For example, in the part-of-speech tagging task, we have a sequence of words but the objective is to infer the tags from this sequence, in this case the tags are hidden because they are not observable. One of the models that incorporates both observed and hidden events, is the hidden Markov model (HMM). HMMs are statistical models that were proposed by Baum and Petrie [1966] and are used for modeling sequential data [Fine et al., 1998, Nag et al., 1986]. HMMs are used in different applications in artificial intelligence [Bengio, 1999] such as speech and handwriting recognition [Kupiec, 1992, Yamato et al., 1992], part-of-speech tagging [Keselj, 2009]. It has also used in bioinformatics and fault recognition [Boatwright et al., 1985].

Generative models can be used to construct a probability distribution of observations using latent variables (variables not observed). RNN and HMM can be seen as generative models [Cappé et al., 2006, Chung et al., 2014, Hopfield, 1982, Rabiner, 1989]. Both models are useful for processing sequential data, however, they have an important difference: in RNN the latent variables are deterministic, and are obtained from the current observation and the previous latent variable. On the other hand, in the HMM the latent variables are random. Due to the similarities between the RNN and HMM models, Salaün et al. [2019] presented a model that integrates HMM and RNN which they called Generative Unified Model (GUM).

Salaün et al. [2019] analyzed the Gaussian and linear GUM case, ie where the objective is to model a sequence of observations in which each observation follows a Gaussian distribution. In order to make a theoretical comparison between RNN and HMM models, seen as particular cases of GUM, they define a certain type of equivalence classes. These equivalence classes allow the study of the type of distributions that can be modeled and the consequences of the construction of the latent variables in each model. This comparison between the models by using equivalence classes is what they define as 'modelling power comparison'. One of the results was that the Guassian and linear GUM subclass allows to model a class of stationary multivariate Gaussian distributions with a geometric covariance sequence. Furthermore, it was shown that none of the HMM and RNN sets are included in each other (into the other one). However, a further analysis could be done with respect to the number of parameters, i.e. an even more general model. Therefore, an open question is to analyze the cost of adding new parameters to this model, in addition to analyzing the non-linear case.

The purpose of this project is to take as a reference the GUM model proposed by Salaün et al. [2019] and generalize it, this will be done by adding dependencies between observed and latent variables of the current time and its immediate past. That is, two extra parameters will be added to the model. In order to study these new dependencies it is necessary to use tools similar to those applied in [Salaün et al., 2019], but with the difference that in our case we will use several results in the multivariate context, taking as a random variable the pair composed of the latent and observed variables. Since a study will be analyzed from a theoretical point of view, we consider the linear and Gaussian case.

This project is organized in the following way: in the first chapter, RNN, HMM will be presented. The second chapter consists of my research work during the internship, *ie.* to understand the Generative Unified Model and to present a generalization of this model called Extendend Generative Unified Model. In this chapter, we present the structure of each model three particular cases where some parameters are fixed in order to obtain a simpler expression of the covariance sequence. The results and a discussion of the results are presented in chapter 3 and the conclusions are presented in chapter 4.

# Chapter 1

# Preliminaries

## 1.1 Recurrent Neural Network (RNN)

Recurrent Neural Network is a class of artificial neural networks which has been used to learn from sequential data in different domains. RNN encodes the temporal context in its connections, which are then capable of capturing the time-varying dynamics of a system. Moreover, RNN has a variety of application, such as financial prediction [Giles et al., 1997], predictive head tracking for virtual reality systems [Saad et al., 1999], among others.

There are more complex variants of RNN such as Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU). They have been proposed in order to control the vanish gradient problem. RNNs and their variants have been used in many context where the temporal dependency of the data is an important implicit feature of the model design. Other applications of RNNs and its variants are learning word embeddings [Mikolov et al., 2013], audio modeling [Oord et al., 2016], handwrinting recognition [Graves et al., 2008].

The basic idea of the RNN is that its output, at each time $t$, depends on previous inputs and past computations. This dependency allows to develop a memory of previous events, which is implicitly encoded in its hidden state variables. Furthermore, a simple RNN can be trained using backpropagation.

RNN can be seen as a generative model [Choe et al., 2017]. In other words, RNNs can be used to predict the next observation in a data sequence. Assuming a single layer, RNN is described by a set of parameters $\bar{\theta} = (\bar{\theta}_0,\ \bar{\theta}_1,\ \bar{\theta}_2)$. The output of a hidden state $\bar{h}$ depends on the previous time, this dependency allows to manage the memory of the observed sequence. Hence, the update for a hidden state at time $t$ is fully deterministic and is represented as follows:

$$\begin{aligned} \bar{h_t} &= f_{\bar{\theta_1}}(\bar{h}_{t-1}, x_t) \\ x_{t+1} &= g_{\bar{\theta_2}}(\bar{h_t}) \end{aligned} \tag{1.1}$$

where

- $x_t$: input vector
- $\bar{h}_t$: hidden state
- $\bar{\theta_1} = (W_{hh}, W_{xh}, l)$: parameter matrices and vector.

- $f, g$: activation functions.

Generally the activation functions are non-linear such as the sigmoid function, the hyperbolic tangent function (tanh) or the rectified linear unit function (ReLU).

In fact, the distribution of the observations is obtained from hidden units $\bar{h}_t$ as follows:

$$\begin{aligned} p_{\bar{\theta}}(x_{0:T}) &= p_{\bar{\theta}}(x_0) \prod_{i=1}^{T} p_{\bar{\theta}}(x_i|x_0) \\ &= p_{\bar{\theta}_0}(x_0) \prod_{i=1}^{T} p_{\bar{\theta}_2}(x_i|\bar{h}_{i-1}), \end{aligned}$$

where $p_{\bar{\theta}_0}$ and $p_{\bar{\theta}_2}$ are given parametrized distributions. In addition, the likelihood $p_{\bar{\theta}}(x_{0:T})$ is computed by construction, it can be calculated by applying a gradient ascent method.

### 1.1.1 Definition of the structure

A new $x_{t+1}$ can be generated depending on $\bar{\theta}_2 = W_{hx}$, *ie.* $W_{hx}\bar{h}_t$. This step is not necessarily deterministic. However, one can imagine that the parameters of an specific distribution of $W_{hx}\bar{h}_t$ are set, so that from such distribution a new $x_{t+1}$ is drawn. At the beginning, the latent state $\bar{h}_0$ is initially null and an external output $x_0$ is provided. This input represents the first element of a sequence we want to complete.

An index translation of the hidden units is set, that is, $h_t = \bar{h}_{t-1}$. The RNN structure is shown in Figure 1.1, the dotted arrow indicates the deterministic transition.

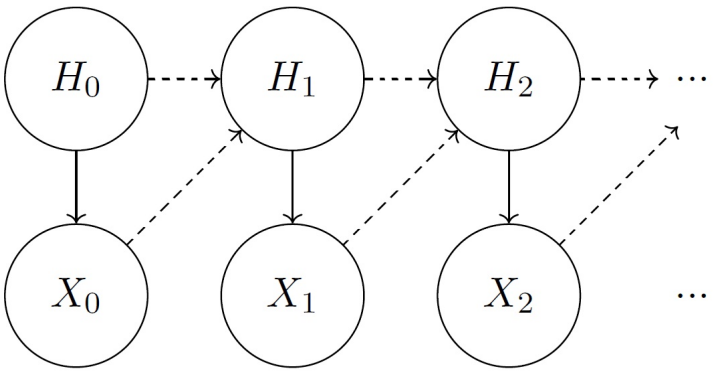



Figure 1.1: Graphical representation of the dependence structure of a Recurrent Neural Network.

## 1.2 Hidden Markov Model (HMM)

Hidden Markov Models are probabilistic models used to analyze sequential data which have been applied in different areas of the industry. HMM can be seen as an extension of Markov chains, where the states are not observables, *ie.* in HMM the observed variables depend on unobservable hidden states[1], which obey the Markov property, that is the conditional probability of the immediate next stage depends only on the present state. Moreover, HMM can also be a tool for representing probability distributions over sequences of observations [Ghahramani, 2002].

Let $X_t$ be the observable variable at time $t$, this variable can be discrete or continuous. Generally, it is considered an HMM in which the state space of the hidden variables is discrete, while the observations can be discrete or continuous (usually generated from a Gaussian distribution). HMMs can be generalized to a continuous state space, for example, when the observed and hidden variables follow a Gaussian distribution.

Discrete state HMMs are used for their simple way of representing probabilities using matrix algebra [Turin, 2012]. On the other hand, continuous state HMMs are used in different areas such as image and speech recognition [Jiang, 2011], finance [Zhang, 2004], etc. In general, a continuous state HMM is more difficult to handle than discrete state HMM because in the former one deals with integrals instead of matrices. In the continuous state HMM, there are two types of models: Gaussian and non-Gaussian HMM. In the Gaussian HMM, one tool used for calculating integrals is the Kalman Filter. On the non-Gaussian continuous state HMMs, the integrals have to be calculated numerically using numerical methods. These numerical methods can be seen as an approximation of the continuous HMM by a discrete state HMM.

There is a generalization of the HMMs known as Markov-switching models [Cappé et al., 2006]. In these types of models the conditional distribution of the variable observed at time $t$ given its past depends not only on the variable hidden at $t$ but also on the variable observed at time $t-1$.

In this project, we consider the continuous state HMM because we seek to built a generalize model that allow us to compare the HMM and the RNN.

The distribution of a random observable sequence $p_{\bar{\theta}}(x_{0:T})$ is the marginal distribution of the joint distribution of latent and observable variables $(H_{0:T},\ X_{0:T})$,

$$\begin{aligned} p_{\bar{\theta}}(h_{0:T}, x_{0:T}) &= p_{\bar{\theta}}(h_{0:T}) \times p_{\bar{\theta}}(x_{0:T}|h_{0:T}) \\ &= p(h_0) \times \prod_{i=1}^{T} p(h_i|h_{i-1}) \times \prod_{i=0}^{T} p(x_i|h_i), \end{aligned}$$

where $p_{\bar{\theta}_1}(h_i|h_{i-1})$ is the transition probability at time $i$ and $p_{\bar{\theta}_2}(x_i|h_i)$ is the conditional

[1] A latent variable is also called as hidden variable, hidden state or hidden state variable.

probability at time $i$.

### 1.2.1 Definition of the structure

In this case, the usual structure of an HMM can be seen in Figure 1.2, and will be the one used in this work.

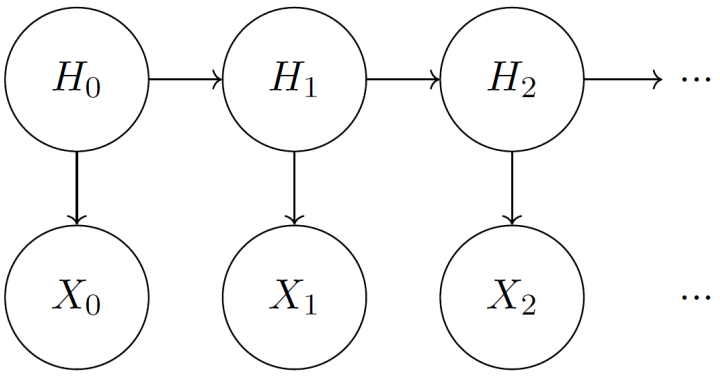



Figure 1.2: Graphical representation of the dependence structure of a Hidden Markov Model, where $(X_t)_{t\in\mathbb{N}}$ is the observable sequence and $(H_t)_{t\in\mathbb{N}}$ is the hidden sequence.

Here, both generative models are considered. These models allow to built a probability distribution functions of the observations by means of the latent variables. In the case of the RNN, every latent variable can be obtain in a deterministic way. In the case of the HMM, the distributions of the observations represent the marginal distribution of the joint distribution of both the observed and latent variables. The theoretical analysis is going to presented on the next chapters.

# Chapter 2

# Research work

## 2.1 Generative Unified Model (GUM)

RNN and HMM can be seen as generative models in order to use their similarities with a perspective of their structure. To obtain the GUM, it is necessary to consider that in both RNN and HMM, the modelling of the joint distribution of the observations depends on a sequence of hidden states. It is also necessary to take into account that in the case of the RNN, a translation of the temporal indices of the hidden units is carried out. Thus it is possible to see the RNN and the HMM as particular cases of the GUM.

### 2.1.1 Definition of the structure

In Figure 2.1, the probabilistic model $\mathcal{M}$ is considered. This model considers two sequences of random variables $(H_t)_{t\in\mathbb{N}}$ and $(X_t)_{t\in\mathbb{N}}$, where the sequence $(H_t)_{t\in\mathbb{N}}$ is called the hidden states sequence and $(X_t)_{t\in\mathbb{N}}$ is the observation states sequence.

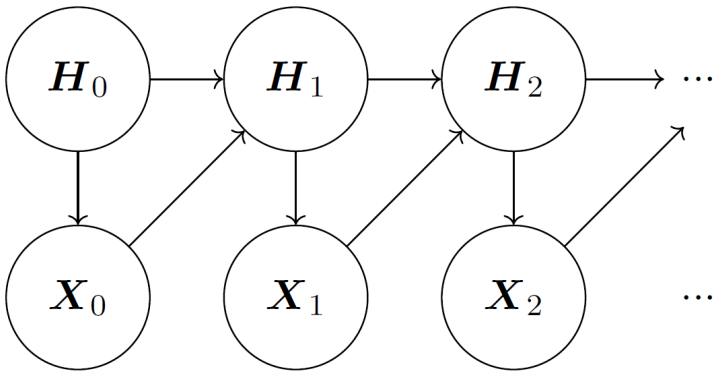



Figure 2.1: Graphical representation of the dependence structure of a Generative Unified Model.

### 2.1.2 Case of study

The objective of the GUM is to present a unified framework that is able to theoretically compare the modelling power of RNN and HMM. In particular, it focus on the case where the observations are realizations of a Gaussian distribution and the activation functions are lineal.

Let $a$, $b$, and $c$ be three real numbers and $\alpha$, $\beta$, and $\eta$ be three positive values. Salaün et al. [2019] considered that the evolution of the model $\mathcal{M}$ is described by:

$$\begin{aligned} \forall t \in \mathbb{N}^* \; p(h_t|h_{t-1}, x_{t-1}) &= \mathcal{N}(h_t; ah_{t-1} + cx_{t-1}, \alpha) \\ \forall t \in \mathbb{N} \; p(x_t|h_t) &= \mathcal{N}(x_t; bh_t, \beta) \\ p(h_0) &= \mathcal{N}(h_0; 0, \eta) \end{aligned}$$

Salaün et al. [2019] considered the following constraint on the distribution of the observations:

$$\forall t \in \mathbb{N} \; p(x_t) = \mathcal{N}(x_t; 0, 1)$$

Therefore, the parameters in the GUM are $a$, $b$, $c$, $\alpha, \eta$, and $\beta$.

In what follows, it is shown how both the HMM and the RNN can be seen as particular cases of the GUM. The graphical structures of the three models are sumarized in Figure 2.2.

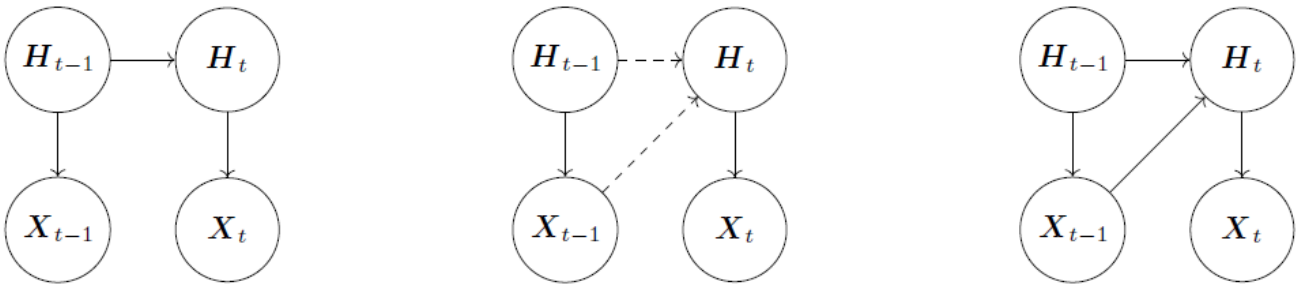



Figure 2.2: Conditional dependencies in HMM, RNN, and GUM. RNN and HMM are particular cases of the GUM. (Source: Salaün et al. [2019])

**RNN**

For the theoretical study of both models, the linear and Gaussian GUM case is considered. Thus, it can be seen that the structure defined for the RNN is obtained by setting the following parameters $\alpha = 0$ and $\eta = c^2$ as follows:

$$\begin{aligned} \forall t \in \mathbb{N}^* \; h_t &= ah_{t-1} + cx_{t-1} \\ \forall t \in \mathbb{N} \; p(x_t|h_t) &= \mathcal{N}(x_t; bh_t, \beta) \\ p(h_0) &= \mathcal{N}(h_0; 0, c^2) \end{aligned}$$

By setting $\alpha = 0$ in the hypotheses considered in the case of the linear and Gaussian GUM, a deterministic and linear function is obtained to determine $x_t$. Therefore,

RNN is a particular case of GUM as shown in Figure 2.2(a). Note that the dashed arrows in the RNN are to differentiate that the dependency is deterministic.

#### HMM

In the same way as above, the HMM can be obtained if $c = 0$. This means that the GUM structure, compared to the HMM, adds a dependency from $X_{t-1}$ to $H_t$. In this way, the hypothesis of the linear and Gaussian case is recovered:

$$\begin{aligned} \forall t \in \mathbb{N}^* \; p(h_t|h_{t-1}, x_{t-1}) &= \mathcal{N}(h_t; ah_{t-1}, \alpha) \\ \forall t \in \mathbb{N} \; p(x_t|h_t) &= \mathcal{N}(x_t; bh_t, \beta) \\ p(h_0) &= \mathcal{N}(h_0; 0, \eta) \end{aligned}$$

Thus, the HMM (Figure 2.2(b)) can be seen as a particular case of the GUM.

### 2.1.3 Results

The objective of the GUM is to present a unified framework that is able to theoretically compare the modelling power of both the RNN and the HMM. In order to do so, a theoretical study of a simple case was considered: the linear and Gaussian case. Additionally, the constraint that the distribution of each observation is a standard normal distribution is considered.

The calculation of $p(x_0)$ and $p(x_1)$ gives the parameters $\beta = 1 - b^2\eta$ and $\alpha = (1 - a^2 - 2abc)\eta - c^2$ as functions of $a$, $b$, $c$, $\eta$. That is, any linear Gaussian GUM under the constraint that each observation follows a standard normal distribution is described by these four parameters. On the other hand, $\forall T \in \mathbb{N}^*, \; p(x_{0:T}) = \mathcal{N}(x_{0:T}; \; O_T, \; \Sigma_T)$, where $\Sigma_T$ is the covariance matrix with one in the diagonal and the covariances are defined as follows:

$$\forall t \in \mathbb{N}, \forall \tau \in \mathbb{N}^*, \; cov(X_t, X_{t+\tau}) = (a + bc)^{\tau-1}(bc + ab^2\eta).$$

Therefore, for any linear and Gaussian GUM under the constraint on the observations, $cov(X_t, X_{t+\tau})$ is geometrical *i.e.* $cov(X_t, X_{t+\tau}) = A^{\tau-1}B$ for some $A$ and $B$.

## 2.2 Extended Generative Unified Model (EGUM)

In this section, the Extended Generative Unified Model (EGUM) is presented. A pairwise Markov Model $(H_t,\ X_t)_{t\in\mathbb{N}}$ is considered, where the sequence $(H_t)_{t\in\mathbb{N}}$ represents the hidden states sequence and $(X_t)_{t\in\mathbb{N}}$ is the observation states sequence. This model is called EGUM because it generalizes the GUM.

### 2.2.1 Definition of the structure

In what follows, the structure of the EGUM is described. The EGUM adds two new dependencies with respect to the GUM. Figure 2.3 shows the general structure of this model. Thus, it is given a dependence of the observation in the current time $t$ given the hidden state at previous time $t-1$ and the dependence of the hidden state in $t$ given the observation at $t-1$.

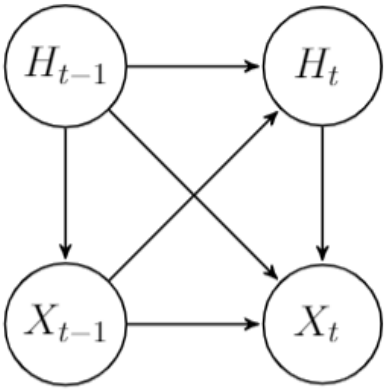



Figure 2.3: Graphical representation of EGUM.

The general pairwise Markov chain structure is described by the following transitions:

$$
\begin{aligned}
H_0 &\sim p(h_0) \\
X_0 &\sim p(x_0|h_0) \\
\forall t \in \mathbb{N}^*\ H_{t+1} &\sim p(h_{t+1}|h_t, x_t) \\
\forall t \in \mathbb{N}^*\ X_{t+1} &\sim p(x_{t+1}|h_{t+1}, h_t, x_t)
\end{aligned}
$$

This model generalizes the GUM.

### 2.2.2 Case of study

The aim of this research is to analyze the contribution of adding these new dependencies (see Figure 2.3) from a theoretical point of view. Hence, for this analysis a simple case is considered: the linear and Gaussian case. Salaün et al. [2019] analyzed the case in which the RNN, the HMM and the GUM consider that each observation

follows a standard normal distribution. Here, we will consider the same constraint in order to compare the contribution of this new model with respect to GUM. However, the probability distribution function of the observations differs only in the covariance matrix.

Particularly in the linear Gaussian case, the pairwise Markov chain structure is described by the following transitions:

$$\forall t \in \mathbb{N}^* \; p(x_t|h_t, h_{t-1}, x_{t-1}) = \mathcal{N}(x_t; bh_t + eh_{t-1} + fx_{t-1}, \beta) \tag{2.1}$$
$$\forall t \in \mathbb{N}^* \; p(h_t|h_{t-1}, x_{t-1}) = \mathcal{N}(h_t; ah_{t-1} + cx_{t-1}, \alpha) \tag{2.2}$$
$$p(h_0) = \mathcal{N}(h_0; 0, \eta) \tag{2.3}$$
$$p(x_0|h_0) = \mathcal{N}(x_0; bh_0, \beta) \tag{2.4}$$

where the parameters $a,\ b,\ c,\ e,\ f \in \mathbb{R}$ and $\alpha,\ \beta,\ \eta \in \mathbb{R}^+$.

From these information the evolution of the model (see more detail in Annex A) can be expressed. Once $Z_t = \begin{bmatrix} H_t \\ X_t \end{bmatrix}$ is set, then the transition of the pairwise $Z_t$ is described by:

$$\forall t \in \mathbb{N}^* \; p(z_t|z_{t-1}) = \mathcal{N}\Big(h_t, x_t;\ Mz_{t-1},\ \Sigma_{z_t|z_{t-1}}\Big)$$

where

$$M = \begin{bmatrix} a & c \\ ba + e & bc + f \end{bmatrix} \tag{2.5}$$
$$\Sigma_{z_t|z_{t-1}} = \begin{bmatrix} \alpha & b\alpha \\ b\alpha & \beta + b^2\alpha \end{bmatrix} \tag{2.6}$$

Additionally, a constraint of stationarity is considered, that is, the distribution of $Z_t$ at each instant $t$ is the same. Hence, the following constraint holds

$$\forall t \in \mathbb{N} \; p(z_t) = \mathcal{N}\Big(z_t;\ \begin{bmatrix} 0 \\ 0 \end{bmatrix};\ \Sigma_{z_t}\Big), \tag{2.7}$$

with

$$\Sigma_{z_t} = \begin{bmatrix} \eta & \gamma\eta \\ \gamma\eta & 1 \end{bmatrix}, \tag{2.8}$$

where $\gamma \in \mathbf{R}$.

### 2.2.3 Objective

The objective is to characterize all models for which the probability distribution of the pair $Z_t$ satisfies the constraint in Equation (2.7). First, the probability distribution function of $Z_0$ is considered:

$$\begin{aligned} p(z_0) &= \mathcal{N}\big(z_0;\ \mu_0;\ \Sigma_{z_0}\big) \\ &= \mathcal{N}\Big(\begin{bmatrix} h_0 \\ x_0 \end{bmatrix};\ \begin{bmatrix} 0 \\ 0 \end{bmatrix},\ \begin{bmatrix} \eta & \gamma\eta \\ \gamma\eta & 1 \end{bmatrix}\Big). \end{aligned}$$

Using the above expression and Proposition 1, the probability distribution function of $Z_1$ can be obtained as follows

$$\begin{aligned} p(z_1) &= \int p(z_1|z_0)p(z_0)dz_0 \\ &= \int \mathcal{N}\big(z_1;\ Mz_0;\ \Sigma_{z_1|z_0}\big)\mathcal{N}\big(z_0;\ \overrightarrow{0};\ \Sigma_{z_0}\big)dz_0 \\ &= \mathcal{N}\big(z_1;\ \overrightarrow{0};\ \Sigma_{z_1|z_0} + M\Sigma_{z_0}M^\top\big) \end{aligned}$$

Due to the fact that stationary constraint in Equation (2.7) needs to be satisfied, *ie.* that $\mathrm{Var}(Z_t)$ does not depend on $t$, it follows that $\Sigma_{z_1} = \Sigma_{z_1|z_0} + M\Sigma_{z_0}M^\top$, where for all $t \in \mathbb{N}$ $\Sigma_{z_t} = \begin{bmatrix} \eta & \gamma\eta \\ \gamma\eta & 1 \end{bmatrix}$. Consequently, the following relation has to be satisfied:

$$\begin{bmatrix} \eta & \gamma\eta \\ \gamma\eta & 1 \end{bmatrix} = \begin{bmatrix} \alpha & b\alpha \\ b\alpha & \beta + b^2\alpha. \end{bmatrix} + \begin{bmatrix} a & c \\ ab+e & bc+f \end{bmatrix} \begin{bmatrix} \eta & \gamma\eta \\ \gamma\eta & 1 \end{bmatrix} \begin{bmatrix} a & ab+e \\ c & bc+f \end{bmatrix}$$

From the previous relation the following restrictions on the parameters are obtained:

$$\eta = \alpha + a\eta(a + 2c\gamma) + c^2 \tag{2.9}$$

$$\gamma\eta = b\alpha + (a^2b + 2abc\gamma + ae + af\gamma + ce\gamma)\eta + (bc^2 + cf) \tag{2.10}$$

$$1 = \beta + b^2\alpha + ((ab+e)^2 + 2\gamma(ab+e)(bc+f))\eta + (bc+f)^2 \tag{2.11}$$

Inserting Equation (2.9) in (2.10) and (2.11), it follows:

$$\begin{aligned} \gamma\eta &= b\eta + (ae + af\gamma + ce\gamma)\eta + fc \\ 1 &= \beta + b^2\eta + (e^2 + 2ab(e + f\gamma) + 2e\gamma(bc+f))\eta + 2bcf + f^2 \end{aligned}$$

The following expressions are then gotten for the parameters $\alpha$ and $\beta$:

$$\alpha = (1 - a^2 - 2ac\gamma)\eta - c^2 \tag{2.12}$$
$$\beta = 1 - b^2\eta - 2b\eta(\gamma - b) - e\eta(e + 2f\gamma) - f^2 \tag{2.13}$$

### 2.2.4 Covariance Matrix

To do the theoretical study, it is necessary to obtain an expression for $\mathrm{Cov}(X_t, X_{t+\tau})$. In order to do so, one needs to first find an expression for $\mathrm{Cov}(Z_t, Z_{t+\tau})$. Hence, $\mathrm{Cov}(X_t, X_{t+\tau})$ is the element in the position $(2,2)$ of such matrix.

$$\begin{aligned} cov(Z_t, Z_{t+\tau}) &= \mathrm{Cov}((H_t, X_t), (H_{t+t}, X_{t+\tau})) \\ &= \begin{bmatrix} \mathrm{Cov}(H_t, H_{t+\tau}) & \mathrm{Cov}(H_t, X_{t+\tau}) \\ \mathrm{Cov}(X_t, H_{t+\tau}) & \mathrm{Cov}(X_t, X_{t+\tau}) \end{bmatrix} \end{aligned}$$

Since $Z_{t+1}|Z_t = z_t \sim \mathcal{N}(z_{t+1};\ Mz_t, \Sigma_{z_t|z_{t-1}})$, then $\mathrm{Cov}(Z_t, Z_{t+\tau})$ is just $\Sigma_{z_t}(M^\tau)^\top$ where $M$ and $\Sigma_{z_t}$ are defined in Equation (2.5) and (2.8), respectively.

To be able to make a theoretical comparison, it is assumed that the matrix $M$ is diagonalizable. The justification of this assumption is based on the fact that the EGUM is a more general model, and in the case of the GUM, the RNN and the HMM, the matrix $M$ is diagonalizable. Hence, it is reasonable to assume that $M$ is also diagonalizable in this case. Moreover, a more detailed justification will be developed throughout the rest of the internship.

$M$ can be written as $M = PDP^{-1}$ since $M$ is a diagonalizable matrix, then $M^\tau = PD^\tau P^{-1}$, where $P$, $D$ and $P^{-1}$ are written as follows:

$$\begin{aligned} P &= \begin{bmatrix} -\frac{-a+bc+f+K}{2(ab+e)} & \frac{a-bc-f+K}{2(ab+e)} \\ 1 & 1 \end{bmatrix} \\ D &= \begin{bmatrix} \frac{1}{2}(a+bc+f-K) & 0 \\ 0 & \frac{1}{2}(a+bc+f+K) \end{bmatrix} \\ P^{-1} &= \begin{bmatrix} -\frac{ab+e}{K} & \frac{a-bc-f+K}{2K} \\ \frac{ab+e}{K} & \frac{-a+bc+f+K}{2K} \end{bmatrix} \end{aligned}$$

with $K = \sqrt{(a+bc+f)^2 - 4(af-ce)}$.

Since $\mathrm{Cov}(Z_t, Z_{t+\tau}) = \Sigma_{z_t}(PD^\tau P^{-1})^\top$, the following expression for $\mathrm{Cov}(X_t, X_{t+\tau})$ reads

$$\begin{aligned}
&\forall t \in \mathbb{N}, \forall \tau \in \mathbb{N}^*, \\
&\mathrm{Cov}(X_t, X_{t+\tau}) = \Big(\frac{1}{2}(a+bc+f-K)\Big)^\tau \Big(\frac{a-bc-f-2\gamma\eta(ab+e)+K}{2K}\Big) \\
&\qquad - \Big(\frac{1}{2}(a+bc+f+K)\Big)^\tau \Big(\frac{a-bc-f-2\gamma\eta(ab+e)-K}{2K}\Big).
\end{aligned} \tag{2.14}$$

**Case GUM**

If $e = f = 0$ and $\gamma = b$, the result for the GUM model of Salaün et al. [2019] is obtained, that is, the following covariances holds:

$$\forall t \in \mathbb{N}, \forall \tau \in \mathbb{N}^*,\ cov(X_t, X_{t+\tau}) = (a+bc)^{\tau-1}(bc+ab^2\eta).$$

## 2.3 Positivity constraints on the covariance parameters

In this seccion, the analysis of three different cases is done: The first case is the EGUM with parameters $f = 0$ and $\gamma = b$, the second case is the EGUM with parameters $f = -a - bc$ and $\gamma = b$; and the third case is the GUM. Different cases are presented because the expression for $\mathrm{Cov}(X_t, X_{t+\tau})$ (Equation (2.14)) is hard to analyse in the general case. Therefore, setting some of the parameters as functions of the others parameters allows to have a simpler expression for $\mathrm{Cov}(X_t, X_{t+\tau})$ given in Equation (2.14).

### 2.3.1 Case $f = 0$ and $\gamma = b$

The case of $f = 0$ and $\gamma = b$ is analyzed. From Equation (2.9), (2.10) and (2.11), the following expressions can be obtain:

$$\begin{aligned}
a + bc &= 0 \\
\alpha &= \eta - c^2(1-b^2\eta) \\
\beta &= 1 - b^2\eta - e^2\eta
\end{aligned}$$

Thus, Equation (2.14) can be easily simplified as follows

$$\forall t \in \mathbb{N}, \forall \tau \in \mathbb{N}^*\ \mathrm{Cov}(X_t, X_{t+\tau}) = \begin{cases} \Big(\frac{K}{2}\Big)^\tau & \text{if } \tau \text{ is even} \\ \Big(\frac{K}{2}\Big)^{\tau-1} b\Big(c(1-b^2\eta)+e\eta\Big) & \text{otherwise} \end{cases} \tag{2.15}$$

where $K = \sqrt{4ce}$.

Hence, the following Toeplitz covariance matrix[1] is analysed (for some even number $k$ and $k < \tau - 1$)

$$\forall \tau \in \mathbb{N}^*,\ M_\tau(A,B) \overset{def}{=} \mathcal{T}\big([1,\ B,\ A^2,\ A^2B,\ A^4,\ A^4B,\ \dots]\big)$$

$$= \begin{bmatrix} 1 & B & \dots & A^k & A^{k-1}B & \dots \\ B & 1 & B & \dots & A^k & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \dots \\ A^k & & B & 1 & B & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \dots & A^{k-1}B & A^k & \dots & B & 1 \end{bmatrix} \tag{2.16}$$

where:

$$A = \sqrt{ce}$$
$$B = b\Big(c(1-b^2\eta) + e\eta\Big)$$

The objective in this part is to find the values of $A$ and $B$, for which the previous Toeplitz matrix defines a covariance matrix. The covariance matrix $\mathrm{Cov}(X_0, X_1) = \begin{bmatrix} 1 & B \\ B & 1 \end{bmatrix}$ has to be a positive semi-definite matrix, a necessary condition is obtained when $1 - B^2 \geq 0 \iff B \in [-1,1]$. Similarly, if $\mathrm{Cov}(X_0, X_2) = \begin{bmatrix} 1 & A^2 \\ A^2 & 1 \end{bmatrix}$, then $A \in [-1,1]$ is another necessary condition.

In fact, $A, B \in [-1,1]$ are necessary conditions for $M_\tau(A,B)$ to be covariances matrices. To obtain sufficient conditions for $A$ and $B$, it is necessary to apply the Caratheodory-Toeplitz Theorem (Theorem 1, see more details in Appendix A.1). Consequently, Toeplitz matrices, defined by Equation (2.16), are covariances matrices if and only if the condition $-\frac{A^2+1}{2} \leq B \leq \frac{A^2+1}{2}$ is verified.

**Remark:** The details of this result can be found in Annex B.

[1] **Observation:** the last term of the first row and first column of the matrix depends on whether the value of $\tau$ whether is even or odd.

### 2.3.2 Case $f = -a - bc$ and $\gamma = b$

Now, the case where $f = -a - bc$ and $\gamma = b$ is considered. In this case, the following expressions is gotten:

$$\begin{aligned}\alpha &= (1 - a^2 - 2abc)\eta - c^2\\ \beta &= 1 - b^2\eta - e\eta(e - 2(a + bc)b) - (a + bc)^2\\ 0 &= \underbrace{-(a + bc)}_{f}(e\eta - ab\eta - c)\end{aligned}$$

Without loss of generality, it can be considered $f \neq 0$ so that $c = e\eta - ab\eta$. So that, expressions for $\alpha$ and $\beta$ as functions of $a, b, e, \eta$ are found.

$$\begin{aligned}\alpha &= 1 - a^2(1 - b^2\eta) - e^2\eta\\ \beta &= 1 - a^2(1 - b^2\eta) - e^2\eta\end{aligned}$$

Equation (2.14) can then be simplified as follows:

$$\forall t \in \mathbb{N}, \forall \tau \in \mathbb{N}^* \ \mathrm{Cov}(X_t, X_{t+\tau}) = \begin{cases} \left(\frac{K}{2}\right)^{\tau} & \text{if } \tau \text{ is even} \\ \left(\frac{K}{2}\right)^{\tau-1}\Big(b\eta(ab + e) - a\Big) & \text{otherwise} \end{cases} \tag{2.17}$$

where $K = \sqrt{4(ce + a^2 + abc)}$.

In the same way as in the previous case, the following Toeplitz covariance matrix is analyzed:

$$\begin{aligned}\tau \in \mathbb{N}^*, \ \hat{M}_\tau(\hat{A}, \hat{B}) &\overset{def}{=} \mathcal{T}\big([1, \ \hat{B}, \ \hat{A}^2, \ \hat{A}^2\hat{B}, \ \hat{A}^4, \ \hat{A}^4\hat{B}, \ \dots]\big)\\ &= \begin{bmatrix} 1 & \hat{B} & \dots & \hat{A}^k & \hat{A}^{k-1}\hat{B} & \dots \\ \hat{B} & 1 & \hat{B} & \dots & \hat{A}^k & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \dots \\ \hat{A}^k & & \hat{B} & 1 & \hat{B} & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \dots & \hat{A}^{k-1}\hat{B} & \hat{A}^k & \dots & \hat{B} & 1 \end{bmatrix}\end{aligned} \tag{2.18}$$

where:

$$\hat{A} = \sqrt{e^2\eta + a^2(1 - b^2\eta)}$$
$$\hat{B} = be\eta - a(1 - b^2\eta)$$

Since the form of this matrix is the same as the previous case, it can be concluded that $\hat{A}, \hat{B} \in [-1, 1]$ are necessary conditions to satisfies that the Toeplitz matrices defined by Equation (2.18) are covariances matrices. Additionally, this is true if and only if the condition $-\frac{\hat{A}^2+1}{2} \leq \hat{B} \leq \frac{\hat{A}^2+1}{2}$ is verified.

**Remark:** Since the matrix defined in (2.18) is similar to (2.16), the same result is concluded. This is because the proof in Annex B considers $A$ and $B$ as arbitrary values.

### 2.3.3 Case $e = f = 0$

By setting $e = f = 0$, the condition $\gamma = b$ is obtained, then the GUM is recovered. The equations, presented by Salaün et al. [2019], are also recovered:

$$\beta = 1 - b^2\eta$$
$$\alpha = \big(1 - a^2 - 2abc\big)\eta - c^2.$$

An expression for the covariance sequence reads:

$$\forall t \in \mathbb{N}, \forall \tau \in \mathbb{N}^*,\ cov(X_t, X_{t+\tau}) = (a + bc)^{\tau-1}(bc + ab^2\eta).$$

In this case, the covariance matrix is given by

$$\tau \in \mathbb{N}^*,\ \bar{M}_\tau(\bar{A}, \bar{B}) \overset{def}{=} \mathcal{T}\big([1,\ \bar{B},\ \bar{A}\bar{B},\ \bar{A}^2\bar{B},\ \bar{A}^3\bar{B},\ \ldots,\ \bar{A}^{\tau-2}\bar{B}]\big)$$

$$= \begin{bmatrix} 1 & \bar{B} & \ldots & \bar{A}^k\bar{B} & \ldots & \bar{A}^{\tau-2}\bar{B} \\ \bar{B} & 1 & \bar{B} & \ldots & \ldots & \bar{A}^{\tau-3}\bar{B} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \bar{A}^k\bar{B} & & \bar{B} & 1 & \bar{B} & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \bar{B} \\ \bar{A}^{\tau-2}\bar{B} & \ldots & \bar{A}^k\bar{B} & \ldots & \bar{B} & 1 \end{bmatrix} \tag{2.19}$$

where:

$$\bar{A} = a + bc$$
$$\bar{B} = bc + ab^2\eta$$

This matrix has a different form compared to previous cases. By the way, according to the Carathéodory-Toeplitz Theorem and to similar arguments as in the previous cases, Salaün et al. [2019] concluded that the Toeplitz matrix $\bar{M}_\tau(\bar{A}, \bar{B})$ given by Equation (2.19) is a covariance matrix for all $\tau \in \mathbb{N}$ if and only if $\bar{A} \in [-1, 1]$ and $\frac{\bar{A}-1}{2} \leq \bar{B} \leq \frac{\bar{A}+1}{2}$.

**Remark:** The details of this result can be found in Section III.B of [Salaün et al., 2019].

**Observation:** The analysis of the general case is still in progress because of its complexity (Appendix C presents the work done so far).

# Chapter 3

# Results and discussion

## 3.1 Results

The main result from the previous chapter is that $\mathrm{Cov}(X_t, X_{t+\tau})$ has a particular form depending on the set of parameters (case). Three cases were considered: the first is the one with $f = 0$ and $\gamma = b$, the second one considers $f = 0$ and $\gamma = b$, and the third case is the GUM, *ie.* when $e = f = 0$ and $\gamma = b$. In fact, for the three particular cases, a linear and Gaussian EGUM under the constraint in Equation (2.7) was considered, and the distributions are Gaussian and stationary with theirs respective covariance structure.

Now, the objective is to know if any such probability distribution function can be modeled by some EGUM. For this reason, we take, as a starting point, the idea presented in [Salaün et al., 2019], that is, to study the inverse mapping of:

$$\phi : \theta \mapsto \big(A = A(\theta), B = B(\theta)\big), \tag{3.1}$$

where $\theta$ represent the set of parameters of the model. Additionally, the notations $A(\theta)$ and $B(\theta)$ mean that $A$ and $B$ are functions of the parameters $\theta$.

### 3.1.1 Case $f = 0$ and $\gamma = b$

In this case, the following parameters are considered $f = 0$, $\gamma = b$ and $a = -bc$, implying that the set of parameters is $b, c, e, \eta$ since $a, f$ and $\gamma$ are functions of such parameters.

Thus, $\phi$ can be written as:

$$\phi : (b, c, e, \eta) \mapsto \big(A = \sqrt{ce}, B = b(c(1 - b^2\eta) + e\eta)\big). \tag{3.2}$$

The domain $(A,\ B)$ have been characterized in order to obtain a covariance matrix. In fact, The Toeplitz symmetric matrix $M_\tau(A, B)$, with first row $[1,\ B,\ A^2,\ A^2B,\ A^4,\ A^4B,\ \dots]$, is a covariance matrix for all $\tau \in \mathbb{N}^*$ if and only if $A \in [-1,\ 1]$ and $-\frac{A^2+1}{2} \leq B \leq \frac{A^2+1}{2}$.

Let $\mathcal{S}$ be the surface defined by $A \in [-1,\ 1]$ and $-\frac{A^2+1}{2} \leq B \leq \frac{A^2+1}{2}$ (See Figure 3.1). In this case, the inverse mapping $\phi^{-1}$ of Equation (3.2) is the application for which, for some $(A, B) \in \mathcal{S}$, is as follows:

- if $0 < A \leq B \leq \frac{A^2+1}{2}$,

$$\phi^{-1}(A,B) = \Big\{\Big(\frac{1}{\sqrt{2\eta}}, \frac{A^2}{\sqrt{2\eta}(B-\sqrt{B^2-A^2})}, \sqrt{2\eta}(B-\sqrt{B^2-A^2}), \eta > 0\Big)\Big\} \bigcup \Big\{\Big(\frac{1}{\sqrt{2\eta}}, \frac{A^2}{\sqrt{2\eta}(B+\sqrt{B^2-A^2})}, \sqrt{2\eta}(B+\sqrt{B^2-A^2}), \eta > 0\Big)\Big\}$$

- if $0 < B \leq A$,

$$\phi^{-1}(A,B) = \Big\{\Big(\sqrt{\frac{B}{2A\eta}}, \frac{A^2}{\sqrt{\frac{AB\eta}{2}}}, \sqrt{\frac{AB\eta}{2}}, \eta > 0\Big)\Big\}$$

Consequently, the mapping $\phi$ defined by Equation (3.2) is not injective since different EGUMs can have the same observation's probability distribution. Moreover, it has not been yet proven whether the the mapping $\phi$ is surjective or not. However, for some $(A,B) \in \mathcal{S}$ there exists at least one EGUM which yields an observations probability distribution.

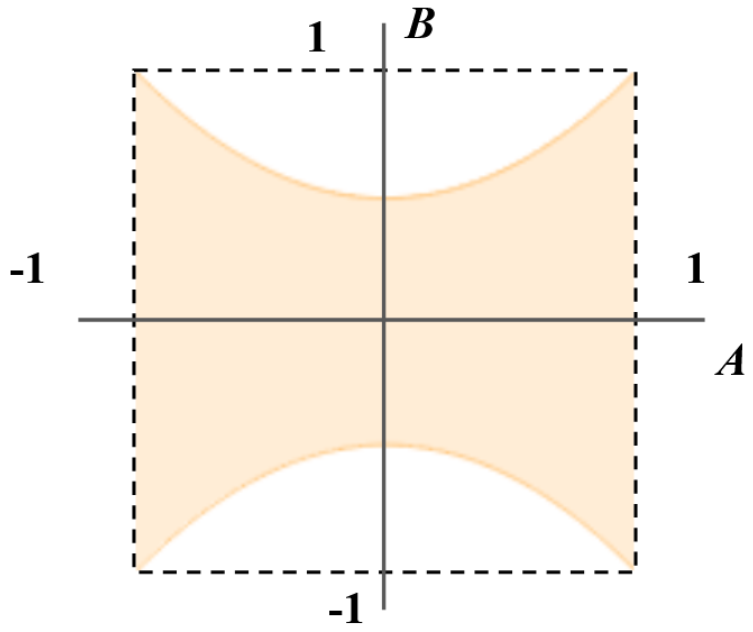



Figure 3.1: The domain $(A,\ B)$ (yellow) for which $M_\tau(A,B)$ is a covariance matrix.

**Remark:** The details of this result can be found in Annex D.

### 3.1.2 Case $f = -a - bc$ and $\gamma = b$

Given that $f = -a - bc$ and $\gamma = b$, then $c = e\eta - ab\eta$. So that, the mapping $\phi$ is as follows:

$$\phi : (a, b, e, \eta) \mapsto \left(A = \sqrt{e^2\eta + a^2(1 - b^2\eta)}, B = be\eta - a(1 - b^2\eta)\right), \tag{3.3}$$

*ie.* there are four parameters $a, b, e$ and $\eta$.

Let $\hat{M}_\tau(A, B)$ be a Toeplitz symmetric matrix with first row $[1,\ B,\ A^2,\ A^2B,\ A^4,\ A^4B,\ \dots]$ and $\hat{\mathcal{S}}$ the surface given by $A \in [-1,\ 1]$ and $-\frac{A^2+1}{2} \leq B \leq \frac{A^2+1}{2}$ (See Figure 3.2). In the previous chapter, it was shown that $\hat{M}_\tau(A, B)$ is a covariance matrix for all $\tau \in \mathbb{N}^*$ if only if $(A, B) \in \hat{\mathcal{S}}$.

Then, the inverse mapping associated to Equation (3.3) is as follows:

- if $(A, B) \in \hat{\mathcal{S}}$,

$$\begin{aligned}\phi^{-1}(A, B) = &\left\{\left(\frac{e^2\eta - A^2}{B - e\sqrt{\frac{\eta}{2}}}, \frac{1}{\sqrt{2\eta}}, e \in \mathbb{R}, \eta > 0\right)\right\} \\ &\bigcup \left\{\left(\frac{e^2\eta - A^2}{B + e\sqrt{\frac{\eta}{2}}}, -\frac{1}{\sqrt{2\eta}}, e \in \mathbb{R}, \eta > 0\right)\right\}.\end{aligned}$$

In this case, the mapping in Equation (3.3) is not injective either. Moreover, for some $(A, B) \in \hat{\mathcal{S}}$, there exists at least a EGUM provided a pdf of the observations.

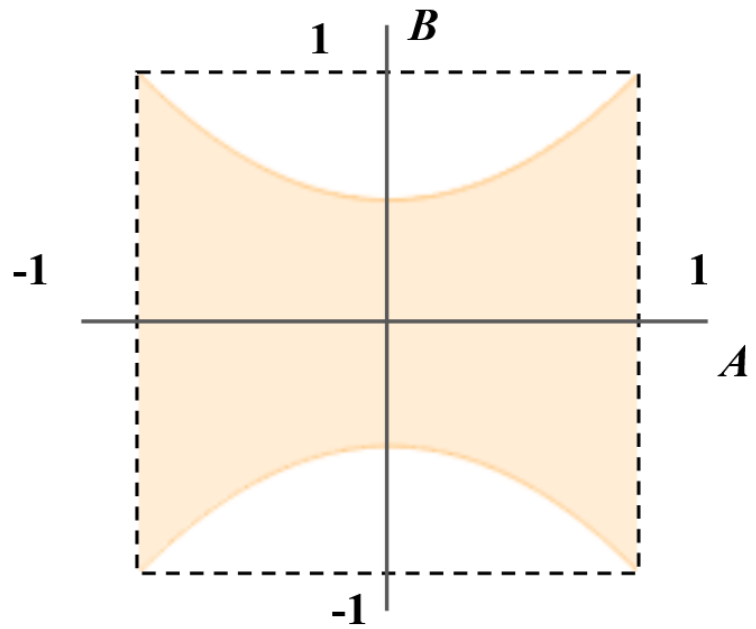



Figure 3.2: The domain $(A,\ B)$ (yellow) for which $\hat{M}_\tau(A, B)$ is a covariance matrix.

**Remark:** The details of this result can be found in Annex E.

### 3.1.3 Case $e = f = 0$

By setting $e = f = 0$, the condition $\gamma = b$ is obtained and so the GUM is recovered. In this case, there are four parameters $a, b, c, \eta$ and the mapping $\phi$ is defined as follows:

$$\phi : (a, b, c, \eta) \mapsto \big(A = a + bc, B = bc + ab^2\eta\big). \tag{3.4}$$

Let $\bar{\mathcal{S}}$ be the parallelogram defined by $A \in [-1,\ 1]$ and $\frac{A-1}{2} \leq B \leq \frac{A+1}{2}$ and $\bar{M}_\tau(A, B)$ a Toeplitz symmetric matrix with a first row $[1,\ B,\ AB,\ \ldots,\ A^{\tau-2}B]$. $\bar{M}_\tau(A, B)$ is a covariance matrix for all $\tau \in \mathbb{N}^*$ if and only if $(A, B)$ belongs to the parallelogram $\bar{\mathcal{S}}$ (see Figure 3.3).

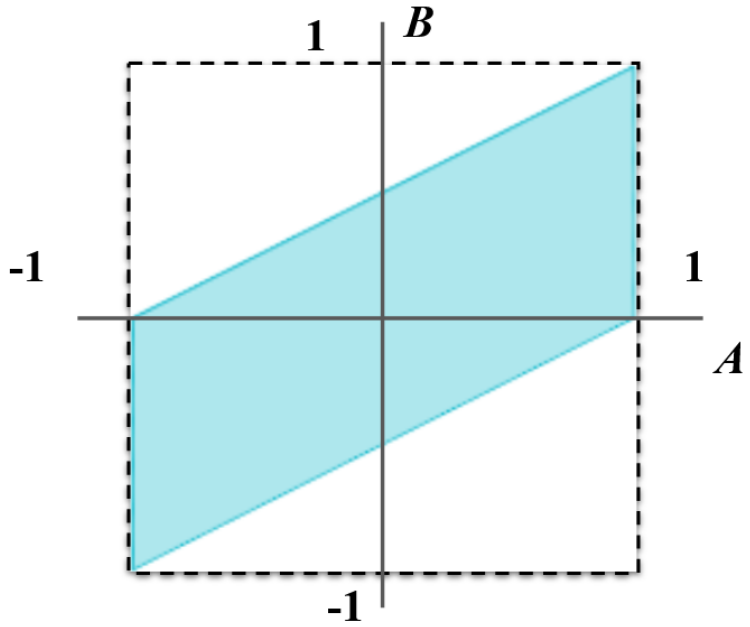



Figure 3.3: The domain $(A,\ B)$ (light blue) for which $\bar{M}_\tau(A, B)$ is a covariance matrix.

Since this case represents a GUM, *Proposition 1* of Salaün et al. [2019] is presented. Hence, for all $(A, B) \in \bar{\mathcal{S}}$ the inverse mapping $\phi^{-1}$ of Equation (3.4) is the application which associates:

- if $A \in \mathbb{R}, B = 0$.

$$\begin{aligned}\phi^{-1}(A, B) = &\Big\{\Big(a, 0, c, \eta\Big); a \in [-1, 1], c \in [-\sqrt{(1-a^2)\eta}, \sqrt{(1-a^2)\eta}], \eta > 0\Big\} \\ &\bigcup \Big\{\Big(a, b, -ab\eta, \eta\Big); a \in [-1, 1], b \in [-\frac{1}{\sqrt{\eta}}, \frac{1}{\sqrt{\eta}}], \eta > 0\Big\}\end{aligned}$$

- if $A \neq B, B \neq 0$,

$$\phi^{-1}(A, B) = \Big\{\Big(\frac{A-B}{1-b^2\eta}, b, \frac{B - Ab^2\eta}{b(1-b^2\eta)}, \eta\Big); b \in [-x_2, -x_1] \cup [x_1, x2], \eta > 0\Big\}$$

where $x_j = \sqrt{\frac{1+2AB-A^2+(-1)^j\sqrt{(A^2-1)((2B-A)^2-1)}}{2\eta}}$ for $j = 1, 2$

- if $A = B, B \neq 0$.

$$\begin{aligned}\phi^{-1}(A,B) = &\left\{\left(0, b, \frac{A}{b}, \eta\right); b \in \left[-\frac{1}{\sqrt{\eta}}, -\frac{|A|}{\sqrt{\eta}}\right] \cup \left[\frac{|A|}{\sqrt{\eta}}, \frac{1}{\sqrt{\eta}}\right], \eta > 0\right\} \\ &\bigcup \left\{\left(a, \frac{1}{\sqrt{\eta}}, (A-a)\sqrt{\eta}, \eta\right); a \in \mathbb{R}, \eta > 0\right\} \\ &\bigcup \left\{\left(a, -\frac{1}{\sqrt{\eta}}, -(A-a)\sqrt{\eta}, \eta\right); a \in \mathbb{R}, \eta > 0\right\}\end{aligned}$$

Salaün et al. [2019] concluded that this function is not injective. However, the function is surjective, i.e. for any $(A, B) \in \bar{\mathcal{S}}$ there is at least one GUM that allows to obtain an observations distribution. Additionally, they performed the study of which distributions $p_{A,B}(x_{0:t})$ can be obtained by an RNN or HMM, for $(A, B) \in \bar{\mathcal{S}}$.

In Figure 3.4, the blue areas coincide with the value of $A$ and $B$ which can be taken by the HHM. On the other hand, the orange curves coincide with the value of $A$ and $B$ which can be taken by the RNN. Therefore, the modeling power of the GUM is larger than that of the HMM and the RNN. Additionally, the modeling power of the HMM is larger than thet of the RNN.

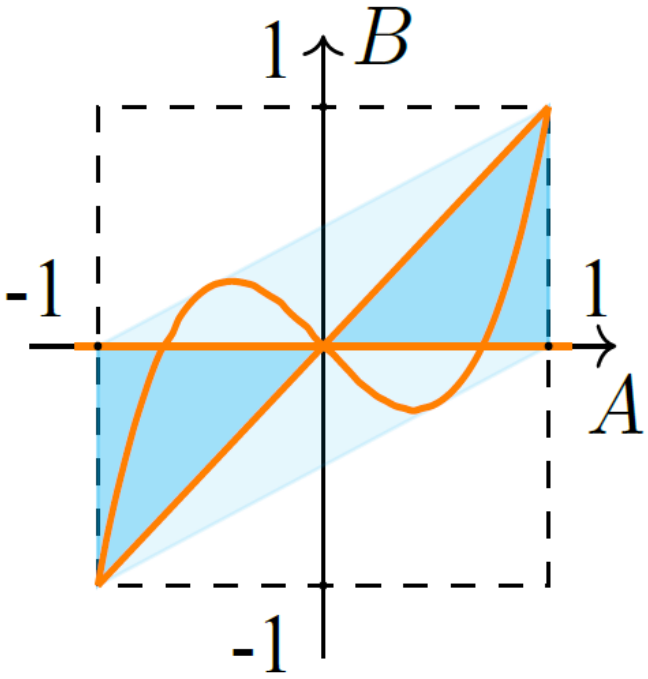



Figure 3.4: Modeling powers of RNN, HMM and GUM with regards to $A$ and $B$. A distribution can be modeled by an RNN (orange), an HMM (blue), a GUM (light blue). (Source: Salaün et al. [2019])

## 3.2 Discussion

In the three analyzed case from the linear and Gaussian EGUM under the constraint given by Equation (2.7), it can be concluded that the pdf of the observations is Gaussian stationary, and that they only differ in the covariance matrix. The question is then whether any pdf of the observations with a given covariance matrix can be modeled by the model we proposed. In this way, the modeling power of each case of the linear and Gaussian EGUM can be analyzed. In order to answer the above question, we have considered and have analyzed the mapping $\phi : \theta\big(A(\theta), B(\theta)\big)$, where $\theta$ represents the set of parameters.

The set of parameters are not the same in all the cases since in order to simplify the calculations some parameters have to be fixed as functions of others. Now, the main interest is to study the inverse mapping of $\phi$, in such a way that it allows to know if given $(A, B)$ some EGUM model can be obtained. $(A, B)$ belongs to a domain which satisfies that the Toeplitz symmetric matrix (defined in each case) is a covariance matrix.

First, the three analyzed cases have four parameters: in the first case, we have $\theta_1 = (b, c, e, \eta)$, in the second one $\theta_2 = (a, b, e, \eta)$, and the third one $\theta_2 = (a, b, c, \eta)$. However, the first case considers $f = 0$, which implies that the observed variable $X_t$ at time $t$ is no longer dependent on the observation at time $t - 1$. In the second case, all the dependencies between the variables are preserved, since the parameters $f, c$ and $\gamma$ are functions of $\theta_2$. In the third case, by setting $e = f = 0$, it is obtained that $\gamma$ is equal to $b$, which means that the GUM is recovered. That means that the dependencies of the observed variable at time $t$ given the observation and the hidden state at time $t - 1$ are no longer considered. Therefore, the second model is more general than the other ones, and has the same number of parameters to estimate.

On the other hand, for the first and second case, the (different) Toeplitz symmetric matrices are covariance matrices in the same domain $(A, B)$. However, the domain of GUM is not the same. Figure 3.5 shows that some GUM covariance sequences are not more covariance sequences in the two other cases and vice versa.

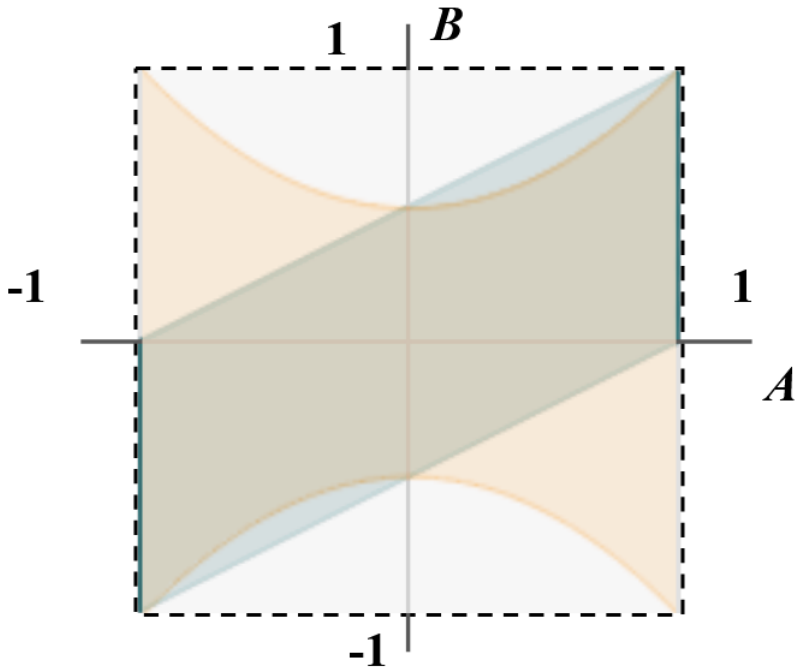



Figure 3.5: The domain $(A, B)$ of the first and second case (yellow) and of the third case, GUM (light blue) for which the Toeplitz symmetric matrix are covariance matrices.

Additionally, the mapping defined in Equation (3.1), is not injective in all cases. Nevertheless, this function is surjective in the GUM. In the other two cases, it is not necessarily true, because $A$ is the positive root of an expression which is a function of the parameters in each case. That is, in the first and second cases it was only considered the case $A \in [0, 1]$. Despite this factor, in both cases it was demonstrated that at least one EGUM generates a observations distribution for some $(A, B)$ belonging to the domain presented in Figure 3.1.

# Chapter 4

# Conclusion

In this project, the EGUM has been presented as a model generalization to the GUM presented in Salaün et al. [2019]. The GUM generalizes both the RNN and the HMM, since they are particular instances of GUM. In order to compare in a theoretical way the proposed model, the linear and Gaussian case was considered. In particular, we considered three specific cases of linear and Gaussian EGUM, one of them being the GUM. Thus, we demonstrated that the three cases considered can model a large class of stationary multivariate Gaussian distributions with an specific covariance sequence. However, applying the Carathéodory theorem in the general case where $\theta = (a, b, c, d, f, \eta)$ is still a wotk in progress because of its difficulty. In fact, there are four parameters in all considered cases. However, the case $f = -a - bc$ and $\gamma = b$ is the more general one since it holds all the dependencies between the hidden and observable variables.

Bear in mind that at the moment we only have considered the linear and Gaussian case, during the rest of the internship the general case will be further analyzed, *ie.*, the linear and Gaussian EGUM which considers all the parameters. In addition, the non-linear EGUM case will also be analyzed. The evaluation of this case will be based on the Stochastic Recurrent Networks (STORNS) presented in Bayer and Osendorfer [2014] which uses stochastic gradient variational Bayes [Rezende et al., 2014] as an estimator. Finally, it can be stated that STORNS are particular cases of GUM, therefore the extension of STORNS is a natural question.

Regarding the internship, it was a great opportunity to work on a research project on the field of Probability Theory on a research laboratory. During my internship, I was able to develop my skills on so many levels. In fact, this experience allowed me to have a different point of view regarding the use of Recurrent Neural Networks and Hidden Markov Models.

First of all, on a technical level, I was able to develop my knowledge in Probability Theory. On a personal level, I became much more comfortable when dealing with a theoretical projects. Also, I am very grateful to my tutor with whom I keep contant contact. I have learnt a lot of great thing from him, and moreover, I could share my opinions to develop this project.

My experience with the CITI Department at Télécom SudParis is a great success and I'm very satisfied. Additionally, I am very happy that I decided to join their teams and do my internship with them.

# Chapter 5

# Annexes

## Annexe A

The transition of $(H_t, X_t)$ is described by:

$$p(h_t,\ x_t|h_{t-1},\ x_{t-1}) = \mathcal{N}\Big(h_t, x_t;\ \begin{bmatrix}\mu_{H_t|H_{t-1},\ X_{t-1}=h_{t-1},\ x_{t-1}}\\ \mu_{X_t|H_{t-1},\ X_{t-1}=h_{t-1},\ x_{t-1}}\end{bmatrix},;\Sigma_{h_t,\ x_t|h_{t-1},\ x_{t-1}}\Big)$$

where

$$\Sigma_{h_t,\ x_t|h_{t-1},\ x_{t-1}} = \begin{bmatrix}\mathrm{Var}(H_t|H_{t-1},\ X_{t-1}=h_{t-1},\ x_{t-1}) & \mathrm{Cov}(H_t, X_t|H_{t-1},\ X_{t-1}=h_{t-1},\ x_{t-1}))\\ \mathrm{Cov}(H_t, X_t|H_{t-1},\ X_{t-1}=h_{t-1},\ x_{t-1})) & \mathrm{Var}(X_t|H_{t-1},\ X_{t-1}=h_{t-1},\ x_{t-1}))\end{bmatrix}$$

Let $a,\ b,\ c,\ e,\ f$ be real numbers and $\alpha,\ \beta,\ \eta$ positive real numbers. We use the hypothesis (2.2) and (2.4) in order to obtain the following expression:

$$\begin{aligned}\mu_{H_t|h_{t-1},x_{t-1}} &= ah_{t-1} + cx_{t-1}\\ \mathrm{Var}(H_t|H_{t-1},\ X_{t-1}=h_{t-1}, x_{t-1}) &= \alpha.\end{aligned}$$

Moreover, the distribution of $X_t|Z_{t-1} = z_{t-1}$ is deduced from Equation (2.1) and (2.2) and with help of Proposition 1.

$$\begin{aligned}p(x_t|h_{t-1}, x_{t-1}) &= \int p(x_t, h_t|h_{t-1}, x_{t-1})dh_t\\ &= \int p(x_t|h_t, h_{t-1}, x_{t-1})p(h_t|h_{t-1}, x_{t-1})dh_t\\ &= \int \mathcal{N}(x_t; bh_t + eh_{t-1} + fx_{t-1}; \beta)\mathcal{N}(h_t; ah_{t-1} + cx_{t-1}; \alpha)dh_t\\ &= \mathcal{N}(x_t;\ (ab+e)h_{t-1} + (bc+f)x_{t-1};\ \beta + b^2\alpha).\end{aligned}$$

Then $\mu_{z_t|z_{t-1}}$ can be written as the product of a $2\times 2$ matrix with the vector $z_{t-1}$:

$$\begin{aligned}\mu_{z_t|z_{t-1}} &= Mz_{t-1}\\ &= \begin{bmatrix}a & c\\ ab+e & bc+f\end{bmatrix}\begin{bmatrix}h_{t-1}\\ x_{t-1},\end{bmatrix}\end{aligned}$$

and the covariance matrix $\Sigma_{z_t|z_{t-1}}$ is written as follows :

$$\Sigma_{z_t|z_{t-1}} = \begin{bmatrix}\alpha & b\alpha\\ b\alpha & \beta + b^2\alpha.\end{bmatrix}$$

# Annex B

We have the following Toeplitz covariance matrix:

$$T \in \mathbb{N}^*,\ \Sigma_\tau \overset{def}{=} \mathcal{T}\big([1,\ B,\ A^2,\ A^2B,\ A^4,\ A^4B,\ \dots]\big) \tag{5.1}$$

In order to have the sufficient conditions it's necessary to apply Theorem 1: Carathéodory-Toeplitz Theorem.

$$T \in \mathbb{N}^*,\ \Sigma_T \geq 0 \iff z \in \{u \in \mathbb{C};\ |u| < 1\}, \Re\Big(1 + 2(Bz + A^2z^2) \sum_{\tau=0}^{\infty} (A^2z^2)^\tau \Big) \geq 0$$

Since

$$\begin{aligned}
&1 + 2(Bz + A^2z^2 + A^2Bz^3 + A^4z^4 + A^4Bz^5 + \dots) \\
=&1 + 2\big[Bz((A^2z^2)^0 + (A^2z^2)^1 + (A^2z^2)^2 + (A^2z^2)^3 + \dots) \\
&+ A^2z^2((A^2z^2)^0 + (A^2z^2)^1 + (A^2z^2)^2 + (A^2z^2)^3 + \dots)\big] \\
=&1 + 2(Bz + A^2z^2) \sum_{\tau=0}^{\infty} (A^2z^2)^\tau
\end{aligned}$$

For all $z \in \{u \in \mathbb{C};\ |u| < 1\}, |A^2z^2| < 1$ then $\sum_{\tau=0}^{\infty}(A^2z^2)^\tau = \frac{1}{1-A^2z^2}$

$$\begin{aligned}
&\Re\Big(1 + 2(Bz + A^2z^2) \sum_{\tau=0}^{\infty} (A^2z^2)^\tau \Big) \\
&\overset{(i)}{=} \Re\Big(1 + 2\ \frac{Bz + A^2z^2}{1 - A^2z^2}\ \Big) \\
&= \Re\Big(\frac{1 + 2Bz + A^2z^2}{1 - A^2z^2}\ \Big) \\
&\overset{(ii)}{=} \Re\Big(\frac{1 + 2Bre^{i\theta} + A^2r^2e^{2i\theta}}{1 - A^2r^2e^{2i\theta}}\ \Big) \\
&= \Re\Big(\frac{(1 + 2Bre^{i\theta} + A^2r^2e^{2i\theta})(1 - A^2r^2e^{-2i\theta})}{|1 - A^2r^2e^{2i\theta}|^2}\ \Big) \geq 0 \\
\iff\ &\Re\Big((1 + 2Bre^{i\theta} + A^2r^2e^{2i\theta})(1 - A^2r^2e^{-2i\theta})\Big) \geq 0 \\
\iff\ &1 + 2Br\cos(\theta) - 2A^2Br^3\cos(-\theta) - A^4r^4 \geq 0 \\
\overset{(iii)}{\iff}\ &1 + 2Br\cos(\theta) - 2A^2Br^3\cos(\theta) - A^4r^4 \geq 0 \\
\iff\ &1 + 2Br\cos(\theta)(1 - A^2r^2) - A^4r^4 \geq 0
\end{aligned}$$

We've used the following arguments:

(i) $|A^2z^2| < 1$ since $A \in [-1, 1]$ and $|z| < 1$

(ii) Writing $z = re^{i\theta}$, for all $r \in [0,1)$ and $\theta \in [-\pi, \pi]$.

(iii) Cosine is an even function.

Therefore, we have to analyze the following expression:

$$1 + 2Br\cos(\theta)(1 - A^2r^2) - A^4r^4 \geq 0 \tag{5.2}$$

**Cases:**

1. Case $A = 0$: let first consider the case where $A = 0$, (5.2) is written as: $1 + 2Br\cos(\theta) \geq 1 - 2|B| \geq 0$ then $|B| \leq \frac{1}{2}$.
2. Case $B = 0$: we have the condition $|A| \leq 1$, it's true.
3. Case $B > 0$:

$$\begin{aligned} &1 + 2Br\cos(\theta)(1 - A^2r^2) - A^4r^4 \\ &\geq 1 - 2B(1 - A^2) - A^4. \end{aligned}$$

Note that $1 - A^2r^2 \geq 0$ and $A^4r^4 \geq 0$, then

$$\begin{aligned} &1 + 2Br\cos(\theta)(1 - A^2r^2) - A^4r^4 \geq 0 \\ \iff\ & B \leq \frac{A^2 + 1}{2} \end{aligned}$$

4. Case $B < 0$:

$$\begin{aligned} &1 + 2Br\cos(\theta)(1 - A^2r^2) - A^4r^4 \\ &\geq 1 + 2B(1 - A^2) - A^4. \end{aligned}$$

Note that $1 - A^2r^2 \geq 0$ and $A^4r^4 \geq 0$, then

$$\begin{aligned} &1 + 2Br\cos(\theta)(1 - A^2r^2) - A^4r^4 \geq 0 \\ \iff\ & B \geq -\frac{A^2 + 1}{2} \end{aligned}$$

# Annex C

## General case

The expression of the covariance sequence defined in (2.14) can be rewrite as follows:

$$\forall t \in \mathbb{N}, \forall \tau \in \mathbb{N}^*,\ \mathrm{Cov}(X_t, X_{t+\tau}) = A^\tau\big(B+\frac{1}{2}\big) - C^\tau\big(B-\frac{1}{2}\big)$$

where:

$$\begin{aligned} A &= \frac{a+bc+f-K}{2} \\ C &= \frac{a+bc+f+K}{2} \\ B &= \frac{a-bc-f-2\gamma\eta(ab+e)}{2K} \\ K &= \sqrt{(a+bc+f)^2-4(af-ce)} \end{aligned}$$

We have the following Toeplitz[1] covariance matrix, $\forall \tau \in \mathbb{N}^*$

$$\Sigma_\tau \overset{def}{=} \mathcal{T}\big([1, A\big(B+\frac{1}{2}\big) - C\big(B-\frac{1}{2}\big), \ldots, A^{\tau-1}\big(B+\frac{1}{2}\big) - C^{\tau-1}\big(B-\frac{1}{2}\big)]\big)$$

$$= \begin{bmatrix} 1 & A\big(B+\frac{1}{2}\big) - C\big(B-\frac{1}{2}\big) & \ldots & A^{\tau-1}\big(B+\frac{1}{2}\big) - C^{\tau-1}\big(B-\frac{1}{2}\big) \\ A\big(B+\frac{1}{2}\big) - C\big(B-\frac{1}{2}\big) & 1 & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ A^{\tau-1}\big(B+\frac{1}{2}\big) - C^{\tau-1}\big(B-\frac{1}{2}\big) & \ldots & \ldots & 1 \end{bmatrix}$$

The objective in this part is to find the values of $A$, $B$ and $C$, for which the previous Toeplitz matrix defines a covariance matrix. From the covariance matrix of $(X_0,\ X_1)$ and $(X_0, X_\tau)$, necessary conditions on $A$ and $B$ can be obtained. The covariance matrix of $(X_0,\ X_1)$ is as follows:

$$\mathrm{Cov}(X_0,\ X_1) = \begin{bmatrix} 1 & A\big(B+\frac{1}{2}\big) - C\big(B-\frac{1}{2}\big) \\ A\big(B+\frac{1}{2}\big) - C\big(B-\frac{1}{2}\big) & 1 \end{bmatrix}$$

This matrix must be semi-definite matrix, then we obtain a necessary condition.

[1] More detail in Appendix A

$$
\begin{aligned}
&1-\Big(A\big(B+\frac{1}{2}\big)-C\big(B-\frac{1}{2}\big)\Big)^2\geq 0\\
&A\big(B+\frac{1}{2}\big)-C\big(B-\frac{1}{2}\big)\in[-1,\ 1]\\
\iff &\frac{A+C}{2}-B(C-A)\in[-1,\ 1]
\end{aligned}
$$

Moreover, the condition $\frac{A^\tau+C^\tau}{2}-B(C^\tau-A^\tau)\in[-1,1]$ is obtained from $cov(X_0,X_\tau)$,

$$
\mathrm{Cov}(X_0,\ X_\tau)=\begin{bmatrix}1 & A^\tau\big(B+\frac{1}{2}\big)-C^\tau\big(B-\frac{1}{2}\big)\\ A^\tau\big(B+\frac{1}{2}\big)-C^\tau\big(B-\frac{1}{2}\big) & 1\end{bmatrix}.
$$

In order to obtain sufficient conditions for $A$ and $B$, it is necessary to apply Caratheodory-Toeplitz Theorem. Therefore, we have the following condition:

$$
\begin{aligned}
\forall\tau\in\mathbb{N}^*,\ \Sigma_\tau\geq 0\iff & z\in\{u\in\mathbb{C};\ |u|<1\},\\
&\Re\Big(1+2\big(B+\frac{1}{2}\big)Az\sum_{\tau=0}^{\infty}(Az)^\tau-2\big(B-\frac{1}{2}\big)Cz\sum_{\tau=0}^{\infty}(Cz)^\tau\Big)\geq 0\\
\iff &\Re\Big(\frac{[1+2B(A-C)re^{i\theta}-ACr^2e^{i2\theta}](1-Are^{-i\theta})(1-Cre^{-i\theta})}{|1-Are^{i\theta}|^2|1-Cre^{i\theta}|^2}\Big)\geq 0
\end{aligned}
$$

Therefore, it's necessary to analyze the following expression .

$$
\begin{aligned}
\iff\ & 1+A^2C^2r^4-2B(A-C)(A+C)r^2+(2B(A-C)-(A+C))rcos(\theta)\\
&+AC(2B(A-C)+(A+C))r^3cos(\theta)\geq 0
\end{aligned}
\tag{5.3}
$$

**Calculus:**
According to Caratheodory-Toeplitz Theorem:

$$
\begin{aligned}
T\in\mathbb{N}^*,\ \Sigma_T\geq 0\iff & z\in\{u\in\mathbb{C};\ |u|<1\},\\
&\Re\Big(1+2\big(B+\frac{1}{2}\big)Az\sum_{\tau=0}^{\infty}(Az)^\tau-2\big(B-\frac{1}{2}\big)Cz\sum_{\tau=0}^{\infty}(Cz)^\tau\Big)\geq 0
\end{aligned}
$$

We assume the following conditions $A\in[-1,1]$ and $C\in[-1,1]$ to ensure convergence.

From these conditions and $z = re^{i\theta}$, for all $r \in [0,1)$ and $\theta \in [-\pi, \pi]$, we will obtain the following result:

$$\begin{aligned}
&\Re\Big(1 + 2\big(B + \frac{1}{2}\big)Az \sum_{\tau=0}^{\infty}(Az)^{\tau} - 2\big(B - \frac{1}{2}\big)Cz \sum_{\tau=0}^{\infty}(Cz)^{\tau}\Big) \\
&= \Re\Big(1 + \frac{2\big(B + \frac{1}{2}\big)Az}{1 - Az} - \frac{2\big(B - \frac{1}{2}\big)Cz}{1 - Cz}\Big) \\
&= \Re\Big(\frac{1 + 2B(A - C)z - ACz^2}{(1 - Az)(1 - Cz)}\Big) \\
&= \Re\Big(\frac{1 + 2B(A - C)re^{i\theta} - ACr^2e^{i2\theta}}{(1 - Are^{i\theta})(1 - Cre^{i\theta})}\Big) \\
&= \Re\Big(\frac{[1 + 2B(A - C)re^{i\theta} - ACr^2e^{i2\theta}](1 - Are^{-i\theta})(1 - Cre^{-i\theta})}{|1 - Are^{i\theta}|^2|1 - Cre^{i\theta}|^2}\Big)
\end{aligned}$$

Therefore, it's necessary to analyze the following expression

$$\begin{aligned}
&\Re\Big(\frac{[1 + 2B(A - C)re^{i\theta} - ACr^2e^{i2\theta}](1 - Are^{-i\theta})(1 - Cre^{-i\theta})}{|1 - Are^{i\theta}|^2|1 - Cre^{i\theta}|^2}\Big) \geq 0 \\
\iff\ & 1 + A^2C^2r^4 - 2B(A - C)(A + C)r^2 + (2B(A - C) - (A + C))rcos(\theta) \\
& + AC(2B(A - C) + (A + C))r^3cos(\theta) \geq 0 \qquad (5.4)
\end{aligned}$$

**Observation:** The analysis of this expression is still in progress because of its complexity.

# Annexe D

The case of $f = 0$ and $\gamma = b$ is analyzed. From Equation (2.9), (2.10) and (2.11), the following expressions can be obtain:

$$\begin{aligned} a + bc &= 0 \\ \alpha &= \eta - c^2(1 - b^2\eta) \\ \beta &= 1 - b^2\eta - e^2\eta \end{aligned}$$

In addition, $A$ and $B$ are defined as follows:

$$\begin{aligned} A &= \sqrt{ce} \\ B &= b\Big(c(1 - b^2\eta) + e\eta\Big) \end{aligned}$$

There are constraints on the parameters $\alpha$ and $\beta$ since they are variances. Therefore, from previous expressions of $\alpha$ and $\beta$ the following conditions on the parameters $c$ and $a$ are obtained:

$$\begin{aligned} \alpha \geq 0 \quad c \quad &\in \left[-\sqrt{\frac{\eta}{1 - b^2\eta}}, \sqrt{\frac{\eta}{1 - b^2\eta}}\right] \\ \beta \geq 0 \quad e \quad &\in \left[-\sqrt{\frac{1 - b^2\eta}{\eta}}, \sqrt{\frac{1 - b^2\eta}{\eta}}\right] \end{aligned}$$

Let $A$ and $B$ two fixed values such that $A \in [-1,\ 1]$ and $-\frac{A^2+1}{2} \leq B \leq \frac{A^2+1}{2}$, then:

$$\begin{aligned} A = \sqrt{c \cdot e} \quad &\Rightarrow \quad c = \frac{A^2}{e} \\ &\Rightarrow c \geq 0 \\ &\Rightarrow e \in \left[0, \sqrt{\frac{1 - b^2\eta}{\eta}}\right] \end{aligned}$$

We want to holds $c \in \Big[0, \sqrt{\frac{\eta}{1 - b^2\eta}}\Big]$, then

$$\frac{A^2}{e} \leq \sqrt{\frac{\eta}{1 - b^2\eta}} \quad \Rightarrow \quad e \leq A^2\sqrt{\frac{1 - b^2\eta}{\eta}}$$

.

Thus, in order to satisfy the condition $c \in \left[0, \sqrt{\frac{\eta}{1 - b^2\eta}}\right]$, we have that $e \in \left[A^2\sqrt{\frac{1 - b^2\eta}{\eta}}, \sqrt{\frac{1 - b^2\eta}{\eta}}\right]$ and it ensures the existence of $c$.

The next step is to replace $c$ in $B = b[c(1-b^2\eta)+e\eta]$

$$\begin{aligned} eB &= bA^2(1-b^2\eta)+b\eta e^2 \\ \iff b\eta e^2 - Be + bA^2(1-b^2\eta) &= 0 \end{aligned}$$

A polynomial in $e$ is obtained, then its discriminant $\Delta_e$ is equals to $B^2 - 4b^2\eta A^2(1-b^2\eta)$. Since the condition $\Delta e \geq 0$ have to be satisfied, the following expression is obtained:

$$b^2\eta(1-b^2\eta) \leq \frac{B^2}{4A^2}.$$

Moreover, the condition $b \in \left[-\frac{1}{\sqrt{\eta}}, \frac{1}{\sqrt{\eta}}\right]$ is satisfied since the matrix defined in (2.8) is a covariance matrix, *ie.* its determinant is non negative. This condition implies $b^2\eta(1-b^2\eta) \in \left[0, \frac{1}{4}\right]$. In addition, if $\frac{B^2}{4A^2} \geq \frac{1}{4}$, we have that $B \geq A$.

We set $b = \frac{1}{\sqrt{2\eta}}$ and then $b^2\eta(1-b^2\eta) = \frac{1}{4}$. Thus, the root $e_1$ of the polynomial in $e$ is given by $e_1 = \frac{B - \sqrt{B^2 - A^2}}{2b\eta}$. $e_1$ needs to be in the interval $\left[A^2\sqrt{\frac{1-b^2\eta}{\eta}}, \sqrt{\frac{1-b^2\eta}{\eta}}\right]$ and it is verified since

$$\begin{aligned} & e_1 \geq A^2\sqrt{\frac{1-b^2\eta}{\eta}} \\ \iff & \frac{B-\sqrt{B^2-A^2}}{2b\eta} \geq A^2\sqrt{\frac{1-b^2\eta}{\eta}} \\ \iff & A^2 - 2B + 1 \geq 0 \\ \iff & B \leq \frac{A^2+1}{2} \quad \textit{it is true} \end{aligned}$$

Now, the second root is $e_2 = \frac{B + \sqrt{B^2 - A^2}}{2b\eta}$ and it is in the interval $\left[A^2\sqrt{\frac{1-b^2\eta}{\eta}}, \sqrt{\frac{1-b^2\eta}{\eta}}\right]$:

$$\begin{aligned} & e_2 \leq \sqrt{\frac{1-b^2\eta}{\eta}} \\ \iff & \frac{B+\sqrt{B^2-A^2}}{2b\eta} \leq \sqrt{\frac{1-b^2\eta}{\eta}} \\ \iff & B \leq \frac{A^2+1}{2} \textit{ it is true} \end{aligned}$$

Therefore, if $0 \leq A \leq B \leq \dfrac{A^2+1}{2}$, we have the following values:

$$
\begin{aligned}
b &= \frac{1}{\sqrt{2\eta}} \\
c &= \frac{A^2}{\sqrt{2\eta}(B - \sqrt{B^2 - A^2})} \\
e &= \sqrt{2\eta}(B - \sqrt{B^2 - A^2})
\end{aligned}
$$

If $0 \leq B \leq A$ then $\frac{B^2}{4A^2} \leq \frac{1}{4}$, we can take $b$ such that $b^2\eta(1 - b^2\eta) = \dfrac{B^2}{4A^2}$ then $e_1 = \dfrac{B}{2b\eta}$ and $e_1 \in \left[A^2\sqrt{\dfrac{1-b^2\eta}{\eta}}, \sqrt{\dfrac{1-b^2\eta}{\eta}}\right]$

$$
\begin{aligned}
& e_1 \geq A^2\sqrt{\frac{1-b^2\eta}{\eta}} \\
\iff & \frac{B}{2b\eta} \geq A^2\sqrt{\frac{1-b^2\eta}{\eta}} \\
\iff & B \geq 2A^2\sqrt{\frac{B^2}{4A^2}} \\
\iff & A \leq 1 \textit{ it is true}
\end{aligned}
$$

$$
\begin{aligned}
& e_1 \leq \sqrt{\frac{1-b^2\eta}{\eta}} \\
\iff & \frac{B}{2b\eta} \geq \sqrt{\frac{1-b^2\eta}{\eta}} \\
\iff & B \geq 2\sqrt{\frac{B^2}{4A^2}} \\
\iff & A \leq 1 \textit{ it is true}
\end{aligned}
$$

Therefore, if $A \geq B \geq 0$, we have the following values:

$$
\begin{aligned}
b &= \sqrt{\frac{B}{2A\eta}} \\
e &= \sqrt{\frac{AB\eta}{2}} \\
c &= \frac{A^2}{\sqrt{\frac{AB\eta}{2}}}
\end{aligned}
$$

# Annexe E

If we have the case $f = -a - bc$ and $\gamma = b$. Here, it can be considered $f \neq 0$ so that $c = e\eta - ab\eta$. So that, expressions for $\alpha$ and $\beta$ as functions of $a, b, e, \eta$ are found.

$$\alpha = 1 - a^2(1 - b^2\eta) - e^2\eta$$
$$\beta = 1 - a^2(1 - b^2\eta) - e^2\eta$$

In addition, $A$ and $B$ are defined as follows:

$$A = \sqrt{e^2\eta + a^2(1 - b^2\eta)}$$
$$B = be\eta - a(1 - b^2\eta)$$

Let $A$ and $B$ two fixed values such that $A \in [-1,\ 1]$ and $-\frac{A^2+1}{2} \leq B \leq \frac{A^2+1}{2}$. On the other hand, there are constraints on the parameters $\alpha$ and $\beta$ since they are variances. In fact, we take $e^2\eta = A^2 - a^2(1 - b^2\eta)$ the constraints $\alpha \geq 0$ and $\beta \geq 0$ are satisfied.

Moreover, the condition $b \in \left[-\frac{1}{\sqrt{\eta}}, \frac{1}{\sqrt{\eta}}\right]$ is satisfied since the matrix defined in (2.8) is a covariance matrix, *ie.* its determinant is non negative. Therefore, we can set $b = \frac{1}{\sqrt{2\eta}}$. Additionally, using $A$ and $B$ defined above, we have $b = \frac{A^2 - e^2\eta + aB}{ae\eta}$ thus $a = \frac{e^2\eta - A^2}{B - e\sqrt{\frac{\eta}{2}}}$.

In conclusion, we have the following values:

$$b = \frac{1}{\sqrt{2\eta}}$$
$$a = \frac{e^2\eta - A^2}{B - e\sqrt{\frac{\eta}{2}}}$$

A similar analysis can be don for $b = -\frac{1}{\sqrt{2\eta}}$ which gives the following values:

$$b = -\frac{1}{\sqrt{2\eta}}$$
$$a = \frac{e^2\eta - A^2}{B + e\sqrt{\frac{\eta}{2}}}$$

# Appendices

# Appendix A

# Review

## A.1 Matrices

**Definition.** ***Mathematical objects***
***Scalar:*** *A scalar is a single number*
***Vector:*** *A vector is an array of numbers*
***Matrix:*** *A matrix is a 2-D array of numbers.*

**Definition.** ***Positive definite matrix [Strang and Borre, 1997]***

*Let $A$ be a $n \times n$ symmetric matrix is positive definite if one of the following properties is satisfied:*

- $x^T Mx > 0$ *except at* $x = 0$, *for* $x \in \mathbb{R}^n$
- *All the eigenvalues are positive.*
- *All the upper determinants are positive.*
- *All the pivots are positive.*

*where the upper left determinants of a $n \times n$ matrix are 1 by 1, 2 by 2, $n$ by $n$.*

**Definition.** ***Diagonalizable Matrix***

*A squared matrix $A$ is a diagonalizable matrix if there exists an invertible matrix $P$ an a diagonal matrix $D$ such that $A = PDP^{-1}$.*

The diagonalization process of a matrix $A$ usually consists of finding its eigenvalues and eigenvectors. Diagonalization can be useful to efficiently compute the powers of a matrix $A$. If $A$ can be written as $A = PDP^{-1}$ then $A^k = PD^kP^{-1}$ for $k \in \mathbb{N}$.

**Definition.** ***Toepliz matrix***

*A Toeplitz matrix is an $n \times n$ matrix $\mathcal{T} = \left[t_{kj};\ k, j = 0, 1, \ldots, n-1\right]$ where $t_{kj} = t_{k-j}$, i.e., a matrix of the form:*

$$\mathcal{T} = \begin{bmatrix} t_0 & t_1 & \ldots & t_{n-1} \\ t_1 & t_0 & t_1 & t_{n-2} \\ \vdots & & \ddots & \vdots \\ t_{n-1} & \ldots & t_1 & t_0 \end{bmatrix}$$

The following theorem allows us to verify if a Toeplitz matrix is a covariance matrix.

**Theorem 1.** ***Carathéodory-Toeplitz Theorem*** [1]

*$\{a_0, a_1, a_2, \dots\}$ is a covariance sequence if and only if $S(z) = a_0 + 2\sum_{i=1}^{+\infty} a_i z^i$ is a function of the Carathéodory class, i.e., $S(z)$ has a positive real part for $z$ in the open unit disk.*

## A.2 Multivariate Concepts

**Definition.** ***Random variable [Beaumont, 2005]***

*A random variable $X$ is a function on a sample space with two properties:*

1. *the values are real numbers; and*
2. *for every real number $x$, the probability that the value of the function is less than or equal to $x$ can be calculated.*

**Definition.** ***Random vector***

*A random vector is a finite collection on random variables defined on a common probability space.*

A random variable (or vector) is represented by capital letters, for example $X$, and its realizations are written in lower case, for example $x$.

**Definition.** ***Expectation Value.***

*The expected value (mean) of a random vector $\mathbf{X} = [X_1, \dots, X_n]$ is a vector $\mathbb{E}[X]$, where each element is the expected value of the respective random variable.*

$$\mu_x = \mathbb{E}[X] = [\mathbb{E}[X_1], \dots, \mathbb{E}[X_n]]^T$$

**Definition.** ***Covariance matrix.***

*The covariance matrix of a random vector $\mathbf{X} = [X_1, \dots, X_n]$ is a $n \times n$ matrix, where the $(i, j)$ element is the convariance between the $i^{th}$ and $j^{th}$ random values.*

$$\Sigma_x = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T$$

**Definition.** ***Cross-covariance matrix***

*Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^p$ be two random vectors, the covariance matriz is a $n \times p$ matrix.*

$$Cov[XY] = \mathbb{E}[XY^T] - \mathbb{E}[X]\mathbb{E}[Y]^T$$

The covariance matrix is a symmetric matrix and a positive semidefinite matrix.

**Properties.** *Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^p$ be two random vectors; and $a$, $A$, $B$ other vector and matrices of appropriate dimensions.*

- $\mathbb{E}[a^T X] = a^T \mu_x$
- $Cov(AX, BY) = ACov(X, Y)B^T$

[1] NIU Department of Mathematical Sciences

## A.3 Gaussian vectors and properties

**Definition.** *A vector random variable* $\mathbf{X} = [X_1, \ldots, X_n]$ *has a multivariable normal distribution,* $X \sim \mathcal{N}(\mu, \Sigma)$*, with mean* $\mu \in \mathcal{R}^n$ *and covariance matrix* $\Sigma \in S^n_{++}$*, i.e., the space of symmetric positive definite* $n \times n$ *matrices; if it has a probability density function of the form:*

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\Big( -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\Big).$$

**Remark:** $\mathcal{N}(x; \mu, \Sigma)$ stands for the Gaussian distribution with mean $\mu$ and variance $\Sigma$, taken at point x.

We now present a classic result on Gaussian vector conditioning that will be used in our derivations [Rao et al., 1973].

**Proposition 1.** ***[Rao et al., 1973]***
*Let* $x, \in \mathbb{R}^p$*,* $y \in \mathbb{R}^q$*,* $\Sigma_1$ *and* $\Sigma_2$ *be* $p \times p$ *and* $q \times q$ *positive definite matrices, respectively, then*

$$\int \mathcal{N}(x;\ Fy + d, \Sigma_1)\mathcal{N}(y;\ m, \Sigma_2)dy = \mathcal{N}(x;\ Fm + d, \Sigma_1 + F\Sigma_2 F^T),$$

*where* $F, d, m$ *and other vectors and matrices, are of appropriate dimensions.*

# Bibliography


Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

Justin Bayer and Christian Osendorfer. Learning stochastic recurrent networks. *arXiv preprint arXiv:1411.7610*, 2014.

Geoffrey P Beaumont. *Probability and random variables*. Elsevier, 2005.

Yoshua Bengio. Markovian models for sequential data. *Neural computing surveys*, 2 (199):129–162, 1999.

GO Boatwright, FW Ravet, and TW Taylor. Development of early warning models. *ARS-United States Department of Agriculture, Agricultural Research Service (USA)*, 1985.

Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer Science & Business Media, 2006.

Yo Joong Choe, Jaehyeok Shin, and Neil Spencer. Probabilistic interpretations of recurrent neural networks. *Probabilistic Graphical Models*, 2017.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Robert DiPietro and Gregory D Hager. Deep learning: Rnns and lstm. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, pages 503–519. Elsevier, 2020.

Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32(1):41–62, 1998.

Z Ghahramani. Hidden markov models: applications in computer vision, 2002.

C Lee Giles, Steve Lawrence, and Ah Chung Tsoi. Rule inference for financial prediction using recurrent neural networks. In *Proceedings of the IEEE/IAFE 1997 Computational Intelligence for Financial Engineering (CIFEr)*, pages 253–259. IEEE, 1997.

Kratarth Goel, Raunaq Vohra, and Jajati Keshari Sahoo. Polyphonic music generation by modeling temporal dependencies using a rnn-dbn. In *International Conference on Artificial Neural Networks*, pages 217–224. Springer, 2014.

Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31 (5):855–868, 2008.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Xuefeng Jiang. A facial expression recognition model based on hmm. In *Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology*, volume 6, pages 3054–3057. IEEE, 2011.

Vlado Keselj. Speech and language processing daniel jurafsky and james h. martin (stanford university and university of colorado at boulder) pearson prentice hall, 2009, xxxi+ 988 pp; hardbound, isbn 978-0-13-187321-6, 115.00, 2009.

Julian Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer speech & language*, 6(3):225–242, 1992.

Larkin Liu, Yu-Chung Lin, and Joshua Reid. Comparing the performance of the lstm and hmm language models via structural similarity. *arXiv*, pages arXiv–1907, 2019.

Shujie Liu, Nan Yang, Mu Li, and Ming Zhou. A recursive recurrent neural network for statistical machine translation. 2014.

Navin Kumar Manaswi. Rnn and lstm. In *Deep Learning with Applications Using Python*, pages 115–126. Springer, 2018.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

R Nag, K Wong, and Frank Fallside. Script recognition using hidden markov models. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 2071–2074. IEEE, 1986.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, and Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and variational inference in deep latent gaussian models. In *International Conference on Machine Learning*, volume 2, 2014.

Emad W Saad, Thomas P Caudell, and Donald C Wunsch. Predictive head tracking for virtual reality. In *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, volume 6, pages 3933–3936. IEEE, 1999.

Achille Salaün, Yohan Petetin, and François Desbouvries. Comparing the modeling powers of rnn and hmm. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1496–1499. IEEE, 2019.

Apeksha Shewalkar, Deepika Nyavanandi, and Simone A Ludwig. Performance evaluation of deep neural networks applied to speech recognition: Rnn, lstm and gru. *Journal of Artificial Intelligence and Soft Computing Research*, 9(4):235–245, 2019.

Gilbert Strang and Kai Borre. *Linear algebra, geodesy, and GPS*. Siam, 1997.

Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432, 2015.

Urmish Thakker, Ganesh Dasika, Jesse Beu, and Matthew Mattina. Measuring scheduling efficiency of rnns for nlp applications. *arXiv preprint arXiv:1904.03302*, 2019.

William Turin. *Performance Analysis and modeling of digital transmission systems*. Springer Science & Business Media, 2012.

Junji Yamato, Jun Ohya, and Kenichiro Ishii. Recognizing human action in time-sequential images using hidden markov model. In *CVPR*, volume 92, pages 379–385, 1992.

Yangsen Zhang, Yuru Jiang, and Yixuan Tong. Study of sentiment classification for chinese microblog based on recurrent neural network. *Chinese Journal of Electronics*, 25(4):601–607, 2016.

Yingjian Zhang. *Prediction of financial time series with Hidden Markov Models*. PhD thesis, Applied Sciences: School of Computing Science, 2004.