Mixture models for ROC curve and spatio-temporal clustering

McMaster
University

MIXTURE MODELS FOR ROC CURVE AND SPATIO-TEMPORAL CLUSTERING

BY

AMAY S.M. CHEAM

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Doctor of Philosophy

Science Doctor of Philosophy (2016) McMaster University
(Department of Mathematics and Statistics) Hamilton, Ontario, Canada

TITLE:	Mixture models for ROC curve and spatio-temporal clustering
AUTHOR:	Amay S.M. Cheam Ph.D., (Mathematics and Statistics) McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	xiii, 104

To my beloved family

"Don't be afraid to dream big and take risks. If you succeed you'll be happy, and if you fail you'll be smarter."

Abstract

Finite mixture models have had a profound impact on the history of statistics, contributing to modelling heterogeneous populations, generalizing distributional assumptions, and lately, presenting a convenient framework for classification and clustering.

A novel approach, via Gaussian mixture distribution, is introduced for modelling receiver operating characteristic curves. The absence of a closed-form for a functional form leads to employing the Monte Carlo method. This approach performs excellently compared to the existing methods when applied to real data.

In practice, the data are often non-normal, atypical, or skewed. It is apparent that non-Gaussian distributions be introduced in order to better fit these data. Two non-Gaussian mixtures, i.e., t distribution and skew t distribution, are proposed and applied to real data.

A novel mixture is presented to cluster spatial and temporal data. The proposed model defines each mixture component as a mixture of autoregressive polynomial with logistic links. The new model performs significantly better compared to the most well known model-based clustering techniques when applied to real data.

Acknowledgements

Foremost, I would like to express my gratitude toward my advisor, Dr. Paul D. McNicholas, for his guidance, continuous encouragement, understanding and great patience during my PhD. His perpetual energy and enthusiasm in research motivated me throughout the years. I would also like to thank Dr. Matthieu Marbac for his insightful advice, support and generous friendship. I would not be able to finish my thesis without his valuable guidance. "*Merci mille fois pour tout et aussi à Cléo.*"

I would like to acknowledge my friends from University of Guelph and McMaster University for their support, interesting discussions, advices and friendship throughout this journey. A special thank you to my dearest friends: Noémi, Marie-Pascal and Anna. I feel very fortunate to have all of you by my side. Thank you for your support, encouragement, extremely undeniable patience and constant availability.

Finally a big THANK YOU to my family. Thank you for your ongoing support and love. But mostly for teaching me that "challenges are masked opportunities to better things, a stepping stones to greater experience and perhaps, in the future, I will be thankful to those temporary failures or difficulties". Thank you for bringing me to the realization that when one door closes, another always opens.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Why Modelling?

David Hilbert, a German mathematician, once said

> *"Mathematics knows no races or geographic boundaries; for mathematics, the cultural world is one country."*

This universal language allows us to transcribe quantitative problems into reality, namely modelling. A model serves as an abstraction or approximate representation of reality to better understand and interpret what is happening. Modelling can be accomplished through various techniques such as analytics, numerics, statistics and many more.

Mixture models have had a profound impact on the history of statistics, contributing to modelling heterogeneous populations, generalizing distributional assumptions, and lately, presenting a convenient framework for classification and clustering. This thesis focuses on the versatility of the finite mixtures in modelling, specifically in two areas: the performance of a binary system and model-based clustering. Note that though both topics evolved from finite mixture modelling, they have distinct goals: one is typically associated with inference on the model and its parameters while the other's aim is to provide a partition of the data into groups of homogeneous observations. Thus, clustering requires an additional step after model fitting, i.e., assigning each observation to a group according to some pre-specified rule.

Since its emergence in the signal detection theory, the receiver operating characteristic (ROC) curve has remained a useful method of describing the intrinsic accuracy of a diagnostic test. This method overcomes the limitations of single sensitivity and specificity pairs by including all of the decision thresholds. The empirical curve is not smooth and does not respect certain theoretical properties, which makes it unattractive. Modelling the diagnostic test can circumvent this issue. A novel approach to modelling the ROC curve is proposed using finite mixture models. These models are then applied to simulated and real data, and perform favourably when compared to existing methods.

In practice, it is often useful to separate data into meaningful groups, where the similarity within groups and the dissimilarity between groups are maximized. Such a method is called clustering. The availability of cheap sensor devices and remarkable development of computer power has created complex data such as spatio-temporal and functional data which emerge as a challenge in clustering. Consequently, researchers must turn their attention toward methods that can define homogeneous partitions and facilitate their interpretation, such as model-based clustering. Likewise, we introduce a novel model-based clustering for spatio-temporal data, employing finite mixture models. In addition, the proposed model can be applied for functional data under some conditions. To illustrate the benefits of this new multi-functionality model, a simulation study along with two challenging applications is conducted.

1.2 Thesis Structure

1.2.1 Chapter 2

Principal concepts surrounding finite mixture models are introduced along with a literature review. Some useful definitions about identifiability are presented, followed by inference in finite mixtures, more precisely parameter estimation and model selection.

1.2.2 Chapter 3

An overview on the literature on the ROC curve is discussed in this chapter. Some basic definitions and properties are given, along with two performance measures.

1.2.3 Chapter 4

In the literature on the ROC curve modelling, the most well-known model is the binormal model. This chapter outlines the most popular model, called LABROC. As an alternative, a novel method is proposed using Gaussian and non-Gaussian mixtures, in conjunction with the Monte Carlo method. Details of its parameter estimation via EM algorithm are discussed. Two applications on pancreatic cancer are considered to demonstrate the flexibility and smoothness of the proposed method.

1.2.4 Chapter 5

This chapter marks the change in emphasis from univariate data to functional or spatio-temporal data and thus the switch from the receiver operating characteristic curve to model-based clustering. Thus, an overview of the literature on the functional and spatio-temporal data clustering is presented. Furthermore, the performance measures and the existing method are presented.

1.2.5 Chapter 6

A novel model for spatio-temporal clustering is introduced, employing a finite mixture model, where each component is an autoregressive polynomial regression mixture of which the logistic weights depend on the spatial and temporal dimensions. Under the maximum likelihood framework, parameter estimation is carried out via an expectation-maximization algorithm while the classical information criteria can be used for model selection. Additionally, the proposed model can be used as a functional data clustering. To illustrate the benefits of the new model, two challenging applications are conducted.

1.2.6 Chapter 7

The ideas and methods demonstrated in this thesis are summarized in this last chapter. Possible research prospects are presented, both based upon and arising from this thesis.

1.3 The Contribution of this Work

The impact of this work on the body of literature is summarized here. The principal novel features of this work are:

(i) ROC curve has been widely used in medical work for its ability to measure the accuracy of diagnostic tests. A new way to tackle the modelling of the ROC curve is introduced. Instead of following the conventional path, i.e., adopting a distribution where the closed-form for the functional form is guaranteed, the proposed method utilizes Gaussian mixtures, in conjunction with the Monte Carlo method. This method performs equivalently or outperforms the well-established existing method LABROC.

(ii) Following the positive results using Gaussian mixtures, the idea of introducing, for the first time, non-Gaussian mixture distributions in the literature of ROC curve is very appealing. The mixture of t distributions and the mixture of skew t distributions are the chosen candidates. Again, when applied to real data, this proposed methods have not disappointed us.

(iii) Clustering spatio-temporal data requires to consider both spatial and temporal aspects in order to extract useful knowledge. A new model for these complex data, adding an autoregressive polynomial regression mixture to each component, is introduced. Furthermore, this new model exhibits the feature of modelling spatial dependencies for multivariate functional data. This model performs significantly better compared to the popular model-based clustering model, when applied to real data. In addition, this model is implemented in R package, called `SpaTimeClust`[1].

[1] The latest version can be downloaded at `https://r-forge.r-project.org/R/?group_id=2163`.

Chapter 2

Finite Mixture Models

2.1 Overview

Finite mixture models made their first appearance in the statistical literature in the nineteenth century in a paper by Newcomb (1886) who used them in the context of modelling outliers. A few years later, Pearson (1894) used a mixture of two Gaussian densities, with unknown unequal variances, to analyze biological data that consisted of the ratio of forehead to body length of 1000 crabs sampled from the Bay of Naples. Because the maximum likelihood estimate for such a model does not have a tractable analytical form, Pearson employed the method of moments to estimate the parameters in the model.

A finite mixture is a convex linear combination of two or more probability density functions. Mixture models are capable of approximating any arbitrary distribution by combining the properties of each probability density function. This leads to its status as a powerful and flexible tool for a statistical-based approach to the modelling of various phenomena (see McLachlan and Basford, 1988). Consequently, finite mixture models have drawn attention in different areas of application ranging from biology to physics, economics and more (Eisen *et al.*, 1998; Vardi *et al.*, 1985; Keane and Wolpin, 1997).

Typically, finite mixture modelling is associated with inference on the model and its parameters. In the literature, Gaussian mixture models are the most popular choice (Titterington *et al.*, 1985; McLachlan and Basford, 1988; Banfield and Raftery, 1993). Later, several different mixture models have been studied such as t distribution (McLachlan and Peel, 1998; Andrews and McNicholas, 2012), skew t distribution (Vrbik and McNicholas, 2012; Lee and McLachlan, 2013b), skew normal (Azzalini, 1985; Lin *et al.*, 2007) and more (Franczak *et al.*, 2014; Browne and McNicholas, 2015). All of this contributes to its extensive uses and flexibility to model unknown distributional shapes. For instance, in clustering, these models can be applied to data where observations originate from diverse groups, and the group memberships are unknown.

An important problem in mixture modelling arises from the selection of the number of components. Thus, the trade-off in model selection problems is: (i) with too many components, the mixture may over-fit the data, while (ii) a mixture with too few components may not be flexible enough to approximate the true underlying model. However this issue can be solved via the Bayesian information criterion which is discussed later in this chapter.

2.2 Definition

In the finite mixture modelling framework, each group is characterized by a probability distribution. Let $\mathbf{x} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$ be the dataset describing n observations

$\mathbf{x}_i = (x_i^1, \ldots, x_i^p)$ by p variables. The probability distribution function of $\mathbf{x}_i$ can be written as

$$f(\mathbf{x}_i \mid \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i \mid \boldsymbol{\vartheta}_g), \tag{2.1}$$

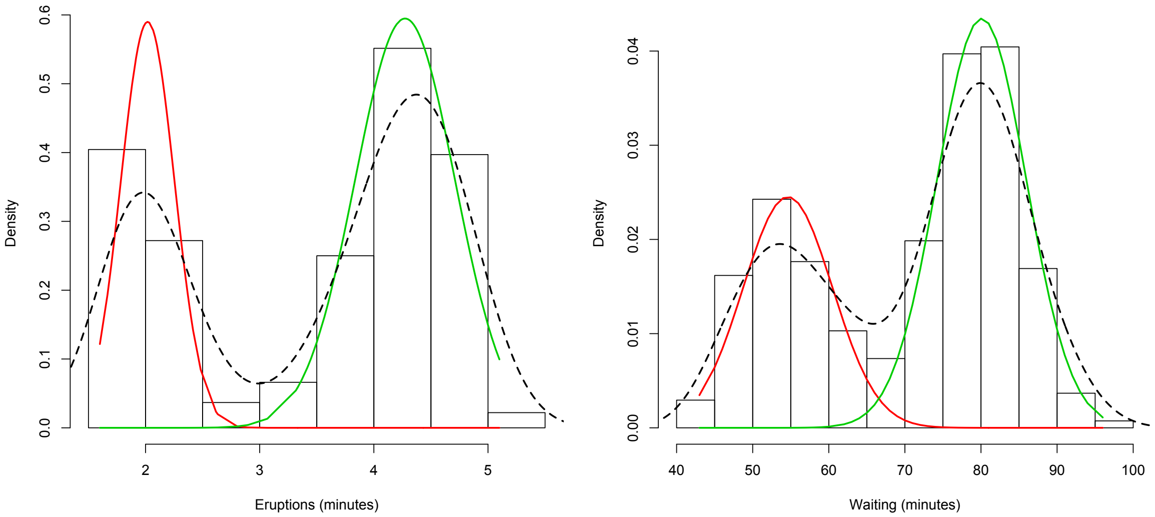
where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, are called mixing proportions, $f_g(\mathbf{x}_i \mid \boldsymbol{\vartheta}_g)$ is the gth component distribution of the mixture with $\boldsymbol{\vartheta}_g$ the set of parameters, and the set of all model parameters is $\boldsymbol{\theta} = \{\pi_g, \boldsymbol{\vartheta}_g; g = 1, \ldots, G\}$.

In the formulation of the mixture models, the number of components G is considered fixed. However, in practice, the value of G is often unknown and has to be estimated.

Figure 2.1 displays the Faithful dataset, publicly available in the R package `mass` (Venables and Ripley, 2002). This dataset contains the waiting time between 272 eruptions and the duration of the eruptions for the Old Faithful geyser in Yellowstone National Park ($p = 2$ variables). This example demonstrates the usefulness of mixture models to fit the data with two components, i.e., $G = 2$.

2.3 Identifiability

Given the expanding importance of models, it is crucial to verify the uniqueness of the parameters which allows easier interpretation of the parameters and the partitions are therefore unique. In other words, for one family of distributions, if two parameters define the same distribution then they must be equal. This prerequisite step is referred to as the identifiability of the model. The assumption of identifiability for statistical models lies at the heart of most statistical theory and practice. The interpretation is based on the parameters and the uniqueness of the partition.



(a) Eruption duration (b) Waiting time until next eruption

Figure 2.1: Histograms and marginal densities of the two-component mixture models for Old Faithful dataset.

Teicher (1963) conduced a useful study of identifiability for finite mixtures and since then numerous special cases have been proven. Thus, identifiability allows for the recovery of the mixing distribution from the mixture and for the consistency of estimation. Identifiability is a general concept that has to be carefully defined.

Definition 2.3.1. *Let $\boldsymbol{\theta}$ be the parameter value of the mode, $\mathbf{x}_i$ be the observations, and $f(\mathbf{x}_i \mid \boldsymbol{\theta})$ be the probability distribution of the data. A model is said to be identifiable if $\forall\ (\boldsymbol{\theta}, \boldsymbol{\theta}') \in \boldsymbol{\Theta}$ and $\forall\ \mathbf{x}_i \in \mathcal{X}$,*

$$f(\mathbf{x}_i \mid \boldsymbol{\theta}) = f(\mathbf{x}_i \mid \boldsymbol{\theta}') \Leftrightarrow \boldsymbol{\theta} = \boldsymbol{\theta}', \tag{2.2}$$

where $\boldsymbol{\Theta}$ represents the set of all possible parameter values, and $\mathcal{X}$ denotes the set of all possible values of the data. Note this definition is sometimes referred to as 'strict' *identifiability.*

Unfortunately, in some cases, the above mapping is not strictly injective. Consider models with discrete hidden variables, such as in finite mixture models: the latent class memberships can be relabelled without modifying the distribution of the observations, see Example 2.3.2. However, this does not disturb inference of the parameters using expectation-maximization algorithm, described in the following section. Note that, in the Bayesian framework, the relabelling problem can be rather onerous (Stephens, 2000).

Example 2.3.2. *Suppose a mixture model with two components parametrized by* $\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2)$ *and* $\boldsymbol{\theta}' = (\pi_2, \pi_1, \boldsymbol{\vartheta}_2, \boldsymbol{\vartheta}_1)$*. Hence, for all* $\mathbf{x}_i$*, we have*

$$\begin{aligned} f(\mathbf{x}_i \mid \boldsymbol{\theta}) &= \pi_1 f(\mathbf{x}_i \mid \boldsymbol{\vartheta}_1) + \pi_2 f(\mathbf{x}_i \mid \boldsymbol{\vartheta}_2) \\ &= \pi_1' f(\mathbf{x}_i \mid \boldsymbol{\vartheta}_1') + \pi_2' f(\mathbf{x}_i \mid \boldsymbol{\vartheta}_2') \\ &= f(\mathbf{x}_i \mid \boldsymbol{\theta}'), \end{aligned}$$

where $\pi_1' = \pi_2, \pi_2' = \pi_1, \boldsymbol{\vartheta}_1' = \boldsymbol{\vartheta}_2$ *and* $\boldsymbol{\vartheta}_2' = \boldsymbol{\vartheta}_1$*.*

The notion of ‘weak’ identifiability is introduced by Teicher (1963) for mixture models because the strict identifiability has too many constraints and to avoid the problem of component relabeling.

Definition 2.3.3. *The mixture model* $f(\mathbf{x}_i \mid \boldsymbol{\theta})$ *is said to be weakly identifiable if* $\forall\ \mathbf{x}_i \in \mathcal{X}$

$$f(\mathbf{x}_i \mid \boldsymbol{\theta}) = f(\mathbf{x}_i \mid \boldsymbol{\theta}') \Leftrightarrow \boldsymbol{\theta} \text{ and } \boldsymbol{\theta}' \text{ are equivalent.} \tag{2.3}$$

The introduction of this definition results in the weakly identifiable proof for Gaussian mixtures, mixtures of Gamma distributions and mixtures of Poisson distributions (Teicher, 1963, 1967; Yakowitz and Spragins, 1968). Using the conditions in Theorem 2.3.4, taken directly from Teicher (1963), one can demonstrate the weak identifiability of some univariate mixture models, such as the univariate Gaussian mixture, see Proposition 2.3.5.

Theorem 2.3.4. *Let $\mathcal{F} = \{F\}$ be a family of one-dimensional cumulative distribution functions with transforms $\phi(t)$ defined for $t \in S_\phi$ (the domain of definition of φ) such that the mapping $M : F \to \phi$ is linear and one-to-one. Suppose that there exists a total ordering ($\preceq$) of $\mathcal{F}$ such that $F_1 \prec F_2$ implies*

(i) $S_{\phi_1} \subseteq S_{\phi_2}$,

(ii) the existence of some $t_1 \in \bar{S}_{\phi_1}$ (t_1 being independent of ϕ_2) such that $\lim\limits_{t \to t_1} \frac{\phi_2(t)}{\phi_1(t)} = 0$.

Then the class of all finite mixtures of $\mathcal{F}$ is weakly identifiable.

Proposition 2.3.5. *The class of all finite mixtures of univariate Gaussian distributions is weakly identifiable.*

Proof. Let $\Phi = \Phi(\cdot \mid \mu, \sigma^2)$ denote the Gaussian cumulative distribution function with mean μ and variance $\sigma^2 > 0$. Its bilateral Laplace transform is given by $\phi(t) = \exp\{\sigma^2 t^2/2 - \mu t\}$. Order the family lexicographically by $f_1 = \Phi(\mathbf{x}_i \mid \mu_1, \sigma_1^2) \prec \Phi(\mathbf{x}_i \mid \mu_2, \sigma_2^2) = f_2$ if $\sigma_1 > \sigma_2$ or if $\sigma_1 = \sigma_2$ but $\mu_1 < \mu_2$. Then, Theorem 2.3.4 applies with $S_\phi = (-\infty, +\infty)$ and $t_1 = +\infty$. □

Sometimes mixtures are non-identifiable but are of interest because they provide meaningful results and their parameters seem identifiable. Thus, the definition of 'generic' identifiability is derived using less stringent conditions.

Definition 2.3.6. *A model is said to be generically identifiable when the parameter space, for which identifiability does not hold, has measure zero.*

Definition 2.3.6 implies, in other words, that any observed dataset has probability one of being drawn from a distribution with identifiable parameters.

2.4 Inference in Finite Mixture Models

Finite mixture models provide great flexibility in fitting models with many modes, skewness and non-standard distributional characteristics. However, this flexibility comes with the price of an increase in the number of parameters and components. Here, we address issues in estimation and model selection with regard to mixture models. We assume, in what follows, that the functional form of f_g is parametric and the same for all components.

2.4.1 Expectation-Maximization Algorithm

Conditionally on the data, maximum likelihood estimation is a method of estimating the parameters by solving the parameter values that maximize the likelihood $\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x})$. For computational reasons, the natural logarithm of the likelihood is used. The log-likelihood function can be written as follows:

$$\log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}) = \sum_{i=1}^{n} \log \left(\sum_{g=1}^{G} \pi_g f(\mathbf{x}_i \mid \boldsymbol{\vartheta}_g) \right). \tag{2.4}$$

Under certain conditions, it has been shown that the maximum likelihood estimate (MLE) is consistent (i.e., it converges in probability to the true parameters), asymptotic normality and efficiency. The MLE owes its popularity to these properties.

One practical problem related to maximum likelihood estimation in finite mixtures is troublesome optimization caused by the complicated and severely multi-modal form of the likelihood function. This issue requires laborious, and sometimes unachievable, analytical or numerical solutions. The most popular and standard procedure for finding the MLE for mixtures is the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977).

The EM algorithm is an iterative method where there is dependency upon unobserved data. Thus the data are incomplete or are treated as incomplete. Here, the component memberships are taken to be missing data. Let $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ denote the component memberships with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})$ and

$$z_{ig} = \begin{cases} 1 & \text{if observation } \mathbf{x}_i \text{ arises from component } g, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Therefore, $\mathbf{z}_i$ is the realization of the latent variable $\mathbf{Z}_i$, which follows a multinomial distribution $\mathcal{M}_G(\pi_1, \ldots, \pi_G)$. Note that f is the conditional probability distribution of $\mathbf{X}_i$ given $\mathbf{Z}_i = \mathbf{z}_i$, where $\mathbf{X}_i = (X_i^1, \ldots, X_i^p)$ is the p-variate random variable defined on the space $\mathcal{X}$. Also, (2.1) can be interpreted as the marginal distribution of $\mathbf{X}_i$.

Let $\boldsymbol{\theta}^{[r]}$ represent the value of the model parameters $\boldsymbol{\theta}$ at iteration r of the EM algorithm. Starting with an initial value of the parameters $\boldsymbol{\theta}^{[0]}$, the algorithm iterates between two processes: the expectation (E-step) and the maximization (M-step), until some convergence criterion is satisfied.

In the E-step, the expected value of the log-likelihood is computed based on the current estimates of the model parameters and the complete-data log-likelihood, i.e., data composed with observed ($\mathbf{x}_i$) and missing ($\mathbf{z}_i$) data:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{[r]}) := \mathbb{E}\left[\log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{z})\right], \tag{2.6}$$

where

$$\begin{aligned} \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{z}) &= \sum_{i=1}^{n} \log f(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta}) \qquad (2.7) \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log\left[\pi_g f(x_i \mid \boldsymbol{\vartheta}_g)\right]. \qquad (2.8) \end{aligned}$$

In the M-step, this expected value is maximized with respect to the model parameters $\boldsymbol{\theta}$ and we have

$$\boldsymbol{\theta}^{[r+1]} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\; Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{[r]}). \tag{2.9}$$

There are a variety of ways to measure convergence. One common approach is to stop the algorithm when the increase in the log-likelihood between successive iterations is smaller than a given threshold ε, i.e.,

$$\log \mathcal{L}(\boldsymbol{\theta}^{[r+1]} \mid \mathbf{x}) - \log \mathcal{L}(\boldsymbol{\theta}^{[r]} \mid \mathbf{x}) < \varepsilon. \tag{2.10}$$

2.4.2 Initialization

This algorithm is susceptible to converge to a local optimum, thus global maximization depends severely on the initial starting values for the EM algorithm (Wu, 1983). Therefore, good initialization is crucial for finding ML estimates. As a consequence, many different initialization procedures have been suggested in the literature (Biernacki *et al.*, 2003; Figueiredo and Jain, 2002). A standard way to obtain $\boldsymbol{\theta}^{[0]}$ consists of initializing it from multiple random positions. Generally this random initial position is obtained by drawing at random component means in the dataset. A random position involves randomly assigning observations into groups and estimating $\boldsymbol{\theta}^{[0]}$ based on this random partition. Then we can repeat the algorithm for a certain number of iterations or until convergence is reached. For example, we execute an algorithm from 50 starting points $\boldsymbol{\theta}_1^{[0]}, \ldots, \boldsymbol{\theta}_{50}^{[0]}$ that create the estimates $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_{50}$. The selected starting point, say $\hat{\boldsymbol{\theta}}^{[0]}$, is the one that obtained the highest log-likelihood in conformity with the trial. If a different initialization is required, depending on the problem at hand, a detailed description will be given in their respective chapters.

2.4.3 Model Selection

The logical step following the resolution of parameter estimation is to select an appropriate model from a set of competing candidate models $\mathcal{M}$, i.e.,

$$m = \{f(\mathbf{x}_i \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\} \in \mathcal{M}. \tag{2.11}$$

Thus, the model m defines the number of components and the nature of the component distributions (i.e., the family of distributions and the parsimony obtained by constraining model parameters). The philosophy behind the choice of a model is to obtain a balance between a good modelling (reducing the bias) and a reasonable number of parameters (reducing the variance).

Despite the vast literature devoted to address the ubiquitous tradeoff between goodness-of-fit and parsimony, the Bayesian information criterion, BIC (Schwarz, 1978), remains an appealing way of choosing the best model (Dasgupta and Raftery, 1998; Madison and Vermunt, 2004). For this thesis, we employed the Bayesian information criterion (BIC). The BIC is derived using a Laplace approximation and by dropping all terms that do not depend on n, i.e., the number of observations. Its principle is to select the model that minimizes the following quantity:

$$\text{BIC}(m) = -2 \log \mathcal{L}(m, \hat{\boldsymbol{\theta}}_m \mid \mathbf{x}) + \nu_m \log n, \tag{2.12}$$

where ν_m is the number of free parameters in the model $m \in \mathcal{M}$, $\log \mathcal{L}(\cdot)$ is the maximized observed log-likelihood and n is the sample size.

Note the value of BIC is a sum of two terms: the log-likelihood that that reflects the goodness-of-fit of the model and the penalty that grows with the model complexity. This criterion is consistent when the model used in the sampling scheme belongs to the set of the competing candidate models. However, it tends to overestimate the number of components, in practice, for which the true model is unknown (Biernacki *et al.*, 2000). The explanation is that real data do not emerge from the mixture densities at hand, and thus, the penalty term is not strong enough to balance the tendency of the likelihood to increase with G in order to improve the fit of the mixture model. Thus, it can provide a poor partition. Hence, we can consider the integrated completed likelihood (ICL) (Biernacki *et al.*, 2000) as an alternative

$$\text{ICL}(m) = \log \mathcal{L}(m \mid \mathbf{x}, \hat{z}_m), \tag{2.13}$$

where $\hat{z}_m = (\hat{z}_{m1}, \ldots, \hat{z}_{mn})$ is the partition given by the MAP rule evaluated at the

MLE $\hat{\boldsymbol{\theta}}_m$. Recall if $g = \underset{\ell=1,\ldots,G}{\arg\max}\ m_{i\ell}(\hat{\boldsymbol{\theta}}_m)$ then $\hat{z}_{mig} = 1$. When the mixture components belong to the exponential family, the ICL has a closed-form but it requires the prior distribution of the parameters to be specified. Otherwise, the BIC-like approximation can be used

$$\mathrm{ICL}(m) = \mathrm{BIC}(m) + 2\sum_{i=1}^{n}\sum_{g=1}^{G}\hat{z}_{mig}\log m_{ig}(\hat{\boldsymbol{\theta}}_m). \tag{2.14}$$

The ICL penalizes the BIC for uncertainty in estimating the number of components, for this reason it is considered to be more appropriate for model-based clustering.

Another option is to consider the Akaike information criterion (AIC) (Akaike, 1974) which tends to select model that minimizes the Kullback-Leibler deviance between the candidate and the selected models:

$$\mathrm{AIC} = -2\log\mathcal{L}(m, \hat{\boldsymbol{\theta}}_m \mid \mathbf{x}) - 2\nu_m. \tag{2.15}$$

When $n > 8$, the penalty of the AIC criterion is less than those of the BIC criterion. Hence, it tends to select a more complex model which results to overestimate the number of components (McLachlan and Peel, 2000).

Chapter 3

Receiver Operating Characteristic Curve

3.1 Overview

The receiver operating characteristic curve, commonly called ROC, made its initial appearance in signal detection during World War II for the analysis of radar images. Its arcane name is derived from the fact that it was developed to better understand the performance of radar operators whose difficult assignment was to observe a noisy screen on the radar receiver and to detect signals among the noise. Thus setting the criterion is a crucial step. Consider the following situation: when setting the criterion too leniently, the operator risks issuing a false alarm - announcing the intrusion of the enemy aircraft when, in fact, it was just noise - whereas, by setting it too strictly, the risk of missing actual targets is highly destructive.

The prompt expansion of ROC curve to other fields was widespread. For example, Swets (1986) introduced it in psychology to study the perceptual detection of stimuli.

Over the years, it has been widely applied in many disciplines, although the ROC curve only gained popularity after the pivotal paper published by Lusted (1971). Therein, Lusted described how ROC curve could be used to assess the accuracy of a diagnostic test. This marked the birth of ROC analysis in medicine.

Depending on different aims, the ROC analysis is useful for: (i) evaluating the ability of a continuous marker or diagnostic test to correctly assign observations into a two states classification, i.e., 'non-diseased' or 'diseased'; (ii) finding the optimal cut-off point to reduce misclassification of the two states; and (iii) comparing the efficacy of two or more markers or diagnostic tests. Two quantities are computed at each threshold of the marker: the false positive (FP) and true positive (TP) rates. The ROC curve is a graphical representation of the relationship between these two quantities. Many ROC-based methods exist for estimating and comparing test accuracy. The literature can be divided into two categories: direct and indirect.

The direct approach does not depend on any distributional assumptions as its name may suggest. Thus, the ROC curve is constructed directly from the diagnostic tests (Lloyd, 1998). Unfortunately, the empirical curve failed to respect certain theoretical properties such as the monotonicity. The non-smoothness makes the empirical curve aesthetically unappealing and suggests that it is not an efficient way to utilize the data. To solve this issue, some works have suggested nonparametric estimation of the density function of each state employing kernel smoothing methods (Hall and Hyndman, 2003; López-de Ullibarri *et al.*, 2008; Qiu and Le, 2001). The problem is translated to selecting an optimal bandwidth (Peng and Zhou, 2004; Zhou and Harezlak, 2002). The lack of a one-to-one correspondence between FP and TP values makes inference unhandy. Moreover some smoothing methods, such as splines

(Ren *et al.*, 2004; Du and Tang, 2009), may not guarantee monotonicity and their ROC curve estimates may fall beyond the range $[0, 1]$. These drawbacks led to the development of alternative modelling option.

Within the indirect approach, two methods have been proposed for estimating ROC curve, i.e., parametric and semi-parametric methods. First, pure parametric approaches are based on the assumption of parametric distributions of the diagnostic variables for the non-diseased and diseased populations. For instance, the classical binormal model assumes that both populations follow a normal distribution. Gönen (2013) assumed that the non-diseased population follows a normal distribution while the diseased follows a Gaussian mixture. Second, semi-parametric methods are highly considered. The addition of nonparametric components makes these approaches very flexible. Under the binormal framework, various works suggested different techniques to estimate its parameters (Hsieh and Turnbull, 1996; Cai and Moskowitz, 2004; Zhou and Lin, 2008). The best known extension of the binormal model is propounded by Metz *et al.* (1998) using the Dorfman and Alf (1968) maximum likelihood algorithm for ordinal data. They developed an algorithm, called LABROC, that assumes the underlying distributions to be normal under an unspecified monotone transformation. However, in practice, the data may not always follow normal distributions as Goddard and Hinberg (1990) pointed out.

In this thesis, the prime focus is on ROC curve fitting with continuous diagnostic variables with gold standard, i.e., the true disease status of each patients is known.

3.2 Definitions

In this section, medical terminology is used to discuss the ROC curve. The ROC curve is defined as follows. Let $X \sim F$ and $Y \sim H$ represent the diagnostic variables for non-diseased and diseased group, respectively. Note they are well defined because the true disease status of each patient is known beforehand, i.e., the case with the gold standard for the truth. By varying the threshold value c_t and plotting the true positive (TP) rate against the false positive (FP) rate, or sensitivity versus 1-specificity, the ROC curve is obtained, see Figure 3.1.

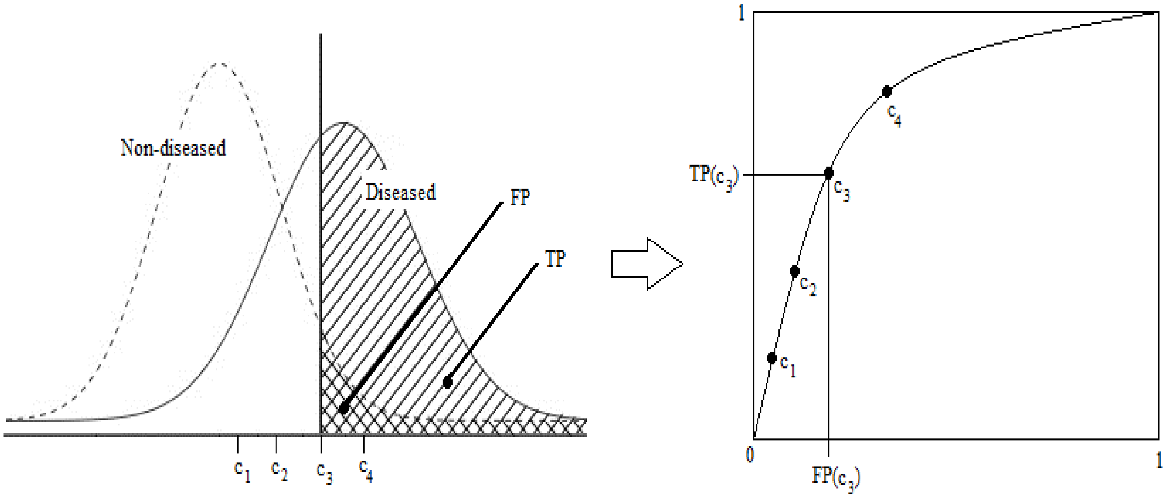


Figure 3.1: Construction of the ROC curve from non-diseased and diseased populations at different threshold, $c_t \in \mathbb{R}$.

It is given by, for $c_t \in \mathbb{R}$,

$$\{(t, R(t))\} = \{(\mathrm{FP}(c_t), \mathrm{TP}(c_t))\}, \tag{3.1}$$

where

$$\text{FP}(c_t) = \int_{-\infty}^{+\infty} f_X(x) I\left(x - c_t\right) dx = P(X > c_t), \tag{3.2}$$

$$\text{TP}(c_t) = \int_{-\infty}^{+\infty} h_Y(y) I\left(y - c_t\right) dy = P(Y > c_t), \tag{3.3}$$

and

$$I(u) = \begin{cases} 1, & \text{if } u > 0, \\ 0, & \text{if } u \leq 0. \end{cases}$$

For a given $t \in D \subset [0, 1]$, $c_t = \bar{F}^{-1}(t) = F^{-1}(1 - t)$, where $F^{-1}(\zeta) = \inf\{x : F(x) \geq \zeta\}$. Mathematically, the functional form of ROC curve can be written as follows:

$$R(t) = TP(c_t) = \bar{H}(\bar{F}^{-1}(t)) = \bar{H}(c_t) = P(Y > c_t) = P(Y > \bar{F}^{-1}(t)), \tag{3.4}$$

where $\bar{F}(u) = P(X > u)$ and $\bar{H}(u) = P(Y > u)$ are known as survival functions of X and Y, respectively. Note, by convention, we will assume that larger values of the test results are more indicative of the disease.

3.3 Properties

Two major mathematical properties of the ROC curve are presented in this section.

Property 3.3.1. *The ROC curve is a monotonically increasing function in the positive unit quadrant.*

Proof. Using (3.4) and noting that $F^{-1}(\cdot)$ is strictly increasing and $H(\cdot)$ is monotonically increasing (Casella and Berger, 2002). Let t_1 and t_2 be two false positive rates such that $t_1 < t_2$. Then $1 - t_2 < 1 - t_1$ and $F^{-1}(1 - t_2) < F^{-1}(1 - t_1)$. Because $H(\cdot)$

is monotonically increasing, we have $H(F^{-1}(1-t_2)) \leq H(F^{-1}(1-t_1))$. Thus,

$$\begin{aligned} R(t_1) &= \bar{H}(\bar{F}^{-1}(t_1)) = 1 - H(F^{-1}(1-t_1)) && (3.5) \\ &= 1 - H(F^{-1}(1-t_1)) \\ &\leq 1 - H(F^{-1}(1-t_2)) && (3.6) \\ &= \bar{H}(\bar{F}^{-1}(t_2)) = R(t_2). && (3.7) \end{aligned}$$

For a given t_1 and t_2 such that $t_1 < t_2$ implies $R(t_1) \leq R(t_2)$, and so the ROC curve is monotonically increasing with the false positive rate t. □

Property 3.3.2. *The ROC curve is unaltered under strictly increasing transformations of the diagnostic variables.*

Proof. To demonstrate this property, let $\varphi(\cdot)$ a strictly increasing transformation, i.e.,

$$b > a \quad \Longleftrightarrow \quad \varphi(b) > \varphi(a) \quad \forall\ a, b \in S,$$

where S denotes the set of diagnostic values, $S \subset \mathbb{R}$. Therefore, for the random variables X and Y from respectively the non-diseased and diseased group, $P(X > c_t) = P(\varphi(X) > \varphi(c_t))$ and $P(Y > c_t) = P(\varphi(Y) > \varphi(c_t))$. Points on the ROC curve for the transformed diagnostic values satisfy

$$t^* = P(\varphi(X) > \varphi(c_t)) = P(X > c_t) = t,$$

and

$$R(t^*) = P(\varphi(Y) > \varphi(c_t)) = P(Y > c_t) = R(t).$$

Thus, the ROC curve for the transformed values is identical to the original ROC curve. □

3.4 Performance Measures

The ROC curve is a summary of the information about the accuracy of a continuous predictor. Nevertheless, it is often useful to summarize the accuracy of a test by a single number. Moreover, they can be employed as the basis of inferential statistics for comparing ROC curves.

3.4.1 Area Under the Curve

The most commonly used summary statistic for an ROC curve is the area under the ROC curve (AUC). The ROC curve alone is beneficial, but with the addition of AUC becomes substantial in the analysis of a diagnostic test. The AUC is defined as follows:

$$\text{AUC} = \int_0^1 R(t)dt, \tag{3.8}$$

where $0 \leq \text{AUC} \leq 1$. However, because random guessing produces the diagonal line between $(0, 0)$ and $(1, 1)$, which has an $\text{AUC} = 0.5$, no diagnostic test should have an $\text{AUC} \leq 0.5$. The closer AUC is to 1, the better the overall accuracy of a diagnostic test.

Corollary 3.4.1. *The AUC can be interpreted as $AUC = P(Y > X)$ for continuous tests.*

Proof. By the definition of AUC, we have

$$\text{AUC} = \int_0^1 R(t)dt = \int_0^1 \bar{H}(\bar{F}^{-1}(t))dt \tag{3.9}$$

$$= \int_{\infty}^{-\infty} \bar{H}(y)dF(y) = \int_{-\infty}^{\infty} P(Y > y)f(y)dy = P(Y > X). \tag{3.10}$$

□

Remark 3.4.2. *If two diagnostic tests are ordered such that test B is uniformly better than test A, i.e.,*

$$R_A(t) \leq R_B(t) \quad \Longrightarrow \quad AUC_A \leq AUC_B \quad \forall\ t \in \mathbb{R}.$$

However, the reverse implication is not necessarily true, because of the possibility that two curves intersect, see Figure 3.2.

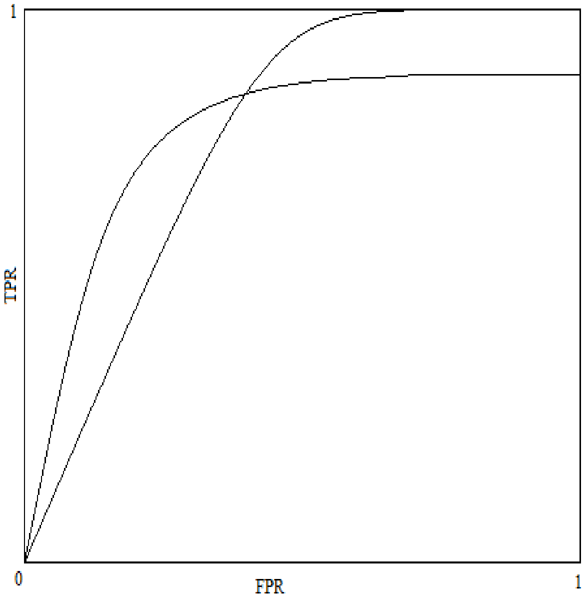


Figure 3.2: Two distinct curves with the same AUC value.

3.4.2 Mean Squared Error

While AUC is by far the most popular one used in practice, it has a limitation. For instance, two tests may have equal AUC, but the tests may differ in clinically important regions of the curve. Assessing the quality of a predictor is essential when modelling because it allows to measure the difference between the observed values and the estimated values, and therefore, to compare models. The mean squared error (MSE) can be estimated by:

$$\text{MSE} = \frac{1}{D}\sum_{d=1}^{D}(R_d(t) - \tilde{R}_d(t))^2, \tag{3.11}$$

where $R(t)$ designates the true TP, $\tilde{R}(t)$ the estimated TP and d the number of cuts on the x-axis representing FP or t.

Chapter 4

Mixture Model Approaches for ROC Curve

4.1 Binormal

The binormal is unequivocally the most explored and used parametric ROC model, and it is described in this section. Similar to how the normal distribution has long been a cornerstone of distribution functions, the binormal model has formed the foundation of ROC curve. The binormal model assumes that both populations follow a normal distribution, i.e., $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, for the non-diseased and diseased group, respectively. By convention that larger test results are more indicative of disease, we assume that $\mu_X < \mu_Y$, thus

$$t = P(X > c_t) = P\left(Z > \frac{c_t - \mu_X}{\sigma_X}\right) = P\left(Z \leq \frac{\mu_X - c_t}{\sigma_X}\right) = \Phi\left(\frac{\mu_X - c_t}{\sigma_X}\right), \quad (4.1)$$

where Z and $\Phi(\cdot)$ denote the standardized normal variable and the standard normal distribution, respectively. Let z_t be the value of Z, then

$$z_t = \Phi^{-1}(t) = \frac{\mu_X - c_t}{\sigma_X} \qquad \Longrightarrow \qquad c_t = \mu_X - z_t\sigma_X.$$

Therefore, the ROC curve at this point t is

$$R(t) = P(Y > c_t) = P\left(Z > \frac{c_t - \mu_Y}{\sigma_Y}\right) = \Phi\left(\frac{\mu_Y - c_t}{\sigma_Y}\right), \tag{4.2}$$

and by substituting the value of c_t obtained previously, we have

$$R(t) = \Phi\left(\frac{\mu_Y - \mu_X + z_t\sigma_X}{\sigma_Y}\right). \tag{4.3}$$

Thus, the functional form of the binormal curve can be summarized in the following way:

$$R(t) = \Phi(a + b\Phi^{-1}(t)) \quad \text{for } 0 \leq t \leq 1, \tag{4.4}$$

where

$$a = \frac{\mu_Y - \mu_X}{\sigma_Y}, \quad b = \frac{\sigma_X}{\sigma_Y}. \tag{4.5}$$

Note that a represents the separation (or intercept) while b represents the symmetry coefficients (or slope). Furthermore, X and Y are independent which implies that $Y - X \sim \mathcal{N}(\mu_Y - \mu_X, \sigma_X^2 + \sigma_Y^2)$. Using the Corollary (3.4.1), we have

$$\text{AUC} = P\left(Z > \frac{0 - (\mu_Y - \mu_X)}{\sqrt{\sigma_X^2 + \sigma_Y^2}}\right) = \Phi\left(\frac{\mu_Y - \mu_X}{\sqrt{\sigma_X^2 + \sigma_Y^2}}\right). \tag{4.6}$$

Then, the corresponding AUC has a closed form given by

$$\text{AUC} = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right). \tag{4.7}$$

The above expression is derived under the assumption that the diagnostic tests follow normal distributions in each of the two groups. However, the ROC curve remains invariant if a monotone increasing transformation is applied to the data. Although some monotone transformations preserve normality of the groups, others may not. Furthermore, Goddard and Hinberg (1990) pointed out that if the distribution of the observed data is non-normal, the AUC derived from a directly fitted binormal can be seriously distorted.

4.1.1 Software: LABROC

The choice of the binormal is usually justified by theoretical considerations, mathematical tractability and familiarity with the normal model. Stating that the ROC curve is binormal simply means that there exists some strictly increasing transformation that would simultaneously transform the raw data into normally distributed random variables. Thus, the problem is reduced to estimating the intercept and the slope parameters from (4.5).

Algorithms have been described (Dorfman and Alf, 1968) and computer programs have been written to implement these techniques. Perhaps the most well known is the software introduced by Metz *et al.* (1998), entitled LABROC[1].

[1] The software is available in the Department of Radiology, University of Chicago: `http://metz-roc.uchicago.edu/MetzROC/software`.

Metz *et al.* (1998) suggested a semi-parametric algorithm, LABROC4, that categorized or binned the ordered continuous test results into runs of non-diseased and diseased patients. Consider the sequence $\{n, n, d, n, n, d, d, d\}$, where n and d represent the non-diseased and diseased, respectively. Thus, that sequence is equivalent to the dataset $\{2n, 1d, 2n, 2d\}$, where rn and sd indicate a run of r non-diseased cases and s diseased cases, respectively. Thus, the ordinal category is $r = \{2, 0, 2, 0\}$ and $s = \{0, 1, 0, 2\}$. Borrowing the authors' notations, the likelihood is defined as:

$$\mathcal{L}(r, s \mid a, b, t^*) = \prod_{i=1}^{I} (P_{i|\bar{D}})^{r_i} \prod_{j=1}^{I} (P_{i|D})^{s_j}, \tag{4.8}$$

where r_i is the number of observations from the non-diseased group in ith category, s_j is the number of observations from the diseased group in jth category, and $P_{i|\bar{D}}$ and $P_{i|D}$ are the probability of one observation from the non-diseased group in the ith category and the diseased group in the jth category, respectively. Here t^* is the latent fixed boundary value generated by truth state runs, where $[t_{i-1}, t_i)$ is the range of the ith category.

The binormal parameters can then be estimated with the Dorfman and Alf (1968) maximum likelihood algorithm for ordinal test results. The LABROC4 procedure assumes that the underlying distributions of the grouped non-diseased and diseased test results can be transformed simultaneously to normal distributions by a single monotone transformation. This assumption is less strict than the assumption of explicit distributional form for the continuous data in the pure parametric approach.

4.2 Modelling Using Finite Mixture Models

Predominantly, ROC curve modelling assumes that each group follows a distribution, mostly the normal distribution. Unfortunately, the binormal model can yield fitted curves with inappropriate shapes when observations are non-normal (Metz, 1989). Mixture models have a rich history in statistics as a compelling apparatus in modelling since they were employed for clustering by Wolfe (1963). One of the major foci of this thesis is to adopt mixture models in modelling the ROC curve. In this section, we introduce Gaussian mixtures and non-Gaussian mixtures, more specifically mixtures of t distributions and mixtures of skew t distributions, as an alternative to the binormal model. Recall that X and Y represent the diagnostic variables for non-diseased and diseased groups, respectively. Thus, the corresponding diagnostic test results for the non-diseased and diseased patients are x_i, for $i = 1, \ldots, n_X$, and y_j, for $j = 1, \ldots, n_Y$, respectively.

4.2.1 Gaussian Mixtures

Mixtures of Gaussian densities are by far the most commonly used representation in statistical modelling for both theoretical and computational reasons. Gaussian mixtures allow us to model heterogeneity in data which is often observed in diseased group.

Model

Suppose X and Y follow Gaussian mixtures, therefore (2.1) can be written as

$$f(x_i \mid \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g \phi(x_i \mid \mu_g, \sigma_g), \tag{4.9}$$

where

$$\phi(x_i \mid \mu_g, \sigma_g) = \frac{1}{\sigma_g\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_g^2}(x_i - \mu_g)^2\right\} \tag{4.10}$$

is the gth Gaussian component density with mean μ_g and standard deviation σ_g, and the collection of all model parameters is $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_G, \mu_1, \ldots, \mu_G, \sigma_1, \ldots, \sigma_G)$. This model is named the MG model. The density $h(y_j \mid \boldsymbol{\psi})$ is defined similarly, where $\boldsymbol{\psi}$ is the collection of all model parameters.

Maximum Likelihood Inference of the Parameters

For simplicity, we only compute parameter estimation for the non-diseased group, X, but the process is similar for diseased group, Y. Using the mixture defined in (4.9) and the log-likelihood function in (2.4), the complete-data log-likelihood is

$$\log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n_X} \sum_{g=1}^{G} z_{ig} \left(\log \pi_g + \log \phi\left(x_i \mid \mu_g, \sigma_g^2\right)\right). \tag{4.11}$$

The iteration $[r]$ of the EM algorithm is written as follows:

1. The E-step: Computation of the conditional probability based on the current value of $\boldsymbol{\theta}^{[r]}$;

$$\eta_{ig}(\boldsymbol{\theta}^{[r]}) := \mathbb{E}\left[Z_{ig} = 1 \mid x_i, \boldsymbol{\theta}_g^{[r]}\right] = \frac{\pi_g^{[r]} f_g(x_i \mid \boldsymbol{\theta}_g^{[r]})}{\sum_{\ell=1}^{G} \pi_\ell^{[r]} f_\ell(x_i \mid \boldsymbol{\theta}_\ell^{[r]})}.$$

Recall that $Z_i \sim \mathcal{M}_G(\pi_1, \ldots, \pi_G)$ from Section 2.2. Here η_{ig} is the conditional probability that an observation x_i is drawn from component g.

2. The M-step: Maximization of the expectation of the complete-data log-likelihood;

   (i) Update the mixing weights, $\pi_g^{[r]}$, by maximizing (2.6) with respect to π_g, which leads to
$$\pi_g^{[r+1]} = \frac{n_g^{[r]}}{n_X};$$

   (ii) Update the mean, $\mu_g^{[r]}$, by maximizing (2.6) with respect to μ_g, which leads to
$$\mu_g^{[r+1]} = \frac{1}{n_g^{[r]}} \sum_{i=1}^{n_X} \eta_{ig}(\boldsymbol{\theta}^{[r]}) x_i;$$

   (iii) Update the variance, $\sigma_g^{2[r]}$, by maximizing (2.6) with respect to σ_g^2, which leads to
$$\sigma_g^{2[r+1]} = \frac{1}{n_g^{[r]}} \sum_{i=1}^{n_X} \eta_{ig}(\boldsymbol{\theta}^{[r]})(x_i - \mu_g^{[r+1]})^2,$$

where $n_g^{[r]} = \sum_{i=1}^{n_X} \eta_{ig}(\boldsymbol{\theta}^{[r]})$.

4.2.2 Non-Gaussian Mixtures

As mentioned earlier, most of the work in finite mixture modelling involves Gaussian mixtures due to their computational tractability. Recently, however, there has been growing of interest in mixtures of non-Gaussian distributions. For instance, Kotz and Nadarajah (2004) pointed out that the multivariate t distribution gives a more realistic model in applied problems. Meanwhile, Lin *et al.* (2007) stated that Gaussian mixture models have a tendency to overfit skewed observations.

An alternative to Gaussian mixtures is to use mixtures of t distributions and mixtures of skew t distributions. McLachlan and Peel (1998) had proved that mixtures of t distributions are effective for dealing with components containing outliers. Although the mixtures have proven robust, they lack the capacity to capture skewness of the observations. This section focuses on mixtures of t distributions and mixtures of skew t distributions that allow us to tackle scenarios that the normality assumption cannot properly handle.

Recall (2.1) for the finite mixture. Depending on the distributional assumption, $f_g(x_i \mid \boldsymbol{\vartheta}_g)$ will change. To lighten the reading, we will omit the density $h(y_j \mid \boldsymbol{\psi})$ that is defined similarly.

Mixtures of t Distributions

Model

Suppose X follows a mixture of t distributions, then its density is defined as

$$f_g(x_i \mid \boldsymbol{\vartheta}_g) = f_T(x_i \mid \mu_g, \sigma_g^2, \nu_g) = \frac{\Gamma(\frac{\nu_g+1}{2})}{\sqrt{\nu_g \pi}\sigma_g \Gamma(\frac{\nu_g}{2})} \left(1 + \frac{(x_i - \mu_g)^2}{\nu_g \sigma_g^2}\right)^{-\frac{\nu_g+1}{2}}, \tag{4.12}$$

where μ_g is the mean, σ_g is the standard deviation, ν_g is the number of degrees of freedom and Γ is the gamma function. The addition of the degrees of freedom parameter allows for heavy tails which give less weight to outlying points during parameter estimation. This model is called the MT model.

Maximum Likelihood Inference of the Parameters

Parameter estimation for the mixture of t distribution is performed using the R package, `teigen`, presented by Andrews and McNicholas (2015). The authors used a variant of the EM algorithm, i.e., the expectation-conditional maximization (ECM) algorithm. In the ECM algorithm, the M-step is adjusted with a set of conditional maximization steps in which each parameter is maximized individually, conditionally on the other parameters remaining fixed. More precisely, they used the multi-cycle ECM algorithm, where a cycle is defined to be one E-step followed by one CM-step. For further details, the reader is referred to Andrews and McNicholas (2012) for the package and Meng and Rubin (1993) for the multi-cycle ECM algorithm. Note for the purpose of our method, we chose the model with unconstrained variance and degree of freedom, i.e., 'univUU'.

Mixtures of Skew t Distributions

Model

Suppose X follows a mixture of skew t distributions. Some recent works have proposed different ways to represent the mixture of skew t as listed by Lee and McLachlan (2013b). For convenience, we use the representation described by Pyne *et al.* (2009). The density is then written as

$$f_g(x_i \mid \boldsymbol{\vartheta}_g) = f_{ST}(x_i \mid \mu_g, \sigma_g, \delta_g, \nu_g) = \frac{2}{\sigma_g} t_{\nu_g}(d(x_i)) T_{\nu_g+1}\left(\delta_g d(x_i)\sqrt{\frac{\nu_g+1}{\nu_g+d(x_i)^2}}\right), \tag{4.13}$$

and

$$d(x_i) = \frac{x_i - \mu_g}{\sigma_g},$$

where t_ν and T_ν, denote, respectively, the pdf and the cdf of standard t distribution with ν degrees of freedom. This model is referred to as MST model.

Maximum Likelihood Inference of the Parameters

Parameter estimation for mixture of skew t distribution is performed using the R package, `EMMIXskew`, introduced by Wang *et al.* (2009). The authors used a standard EM algorithm where a detailed description can be found in Lee and McLachlan (2013a).

4.2.3 Methodology

We propose to model ROC curves using finite mixture models. The Monte Carlo method is used to circumvent the problem of the absence of a closed-form for the functional form of the ROC curve defined in (3.4). Thanks to its properties, computation of confidence bands is feasible. Because each pair (X, Y) constructs one ROC curve, the idea of the proposed approach is to generate an ensemble of replica ROC curves by simulating many pairs $(\widetilde{X}, \widetilde{Y})$. The following details, step-by-step, the proposed method. Let $X^s = (x_1^s, \ldots, x_{n_X}^s)$ and $Y^s = (y_1^s, \ldots, y_{n_Y}^s)$ be the vector of non-diseased and diseased sample groups, respectively.

Step 1: Distributional Assumption

The choice of distributional assumption for the non-diseased and diseased group, i.e., X and Y, respectively, is the first step of the proposed method. Let $X \sim f^*(x_i \mid \boldsymbol{\theta})$ and $Y \sim h^*(y_i \mid \boldsymbol{\psi})$, where * represents either MG, MT or MST model described in Section 4.2.

Step 2: Parameter Estimation

Parameter estimation is carried out directly from the sample groups, i.e., X^s and Y^s, via an EM algorithm described in Sections 2.4.1 and 4.2. Thus, the parameter estimates are obtained, i.e., $\widehat{\boldsymbol{\theta}}^*_X$ and $\widehat{\boldsymbol{\theta}}^*_Y$. The number of components, G, is determined by the BIC described in Section 2.4.3.

Step 3: Replica Generations

This step is the core of the proposed method. With the parameter estimates, obtained previously, $(\widetilde{X}^*, \widetilde{Y}^*)$ are generated from $\widetilde{X}^* \sim F(\widehat{\boldsymbol{\theta}}^*_X)$ and $\widetilde{Y}^* \sim H(\widehat{\boldsymbol{\theta}}^*_Y)$, respectively. From the simulated ensemble, the ROC curve is computed via (3.1) in Section 3.2. Note that it gives only one ROC curve. Then, using the Monte Carlo method, this step is repeated S times. This produces S ROC curves, i.e.,

$$\{(t, \widetilde{R}^*_1(t))\}, \ldots, \{(t, \widetilde{R}^*_S(t))\}.$$

As consequence, S AUC values are obtained, i.e., $\widetilde{A}_1, \ldots, \widetilde{A}_S$. By the properties of Monte Carlo method, based on the strong law of large numbers and the central limit theorem, the distribution of these S points $\widetilde{TP}^*$ converges asymptotically to a Gaussian distribution (Graham and Talay, 2013). To ensure the properties of Monte Carlo method, we choose a large S, i.e., $S = 1000$. The beauty of this finding is that it simplifies hugely the computation of summary measures because of the absence of a closed-form.

Step 4: Averaging Curves

The model estimate, denoted as $\widehat{R}^*(t)$, is derived by averaging the random realizations of the ROC curves such that

$$\widehat{R}^*(t) = \text{mean}(\widetilde{R}^*(t)),$$

where $t \in D \subset [0,1]$. In the same way, the AUC estimate is obtained $\widehat{A}^* = \text{mean}(\widetilde{A}^*)$. Note that by averaging over the ensemble of random realizations $\widetilde{R}^*(t)$, the estimate $\widehat{R}^*(t)$ gives a much smoother curve than the empirical estimate.

Step 5: Performance Measures

For a fixed point t (or $\widetilde{\text{FP}}^*$), we obtained S points $\widetilde{R}^*(t)$ (or $\widetilde{\text{TP}}^*$). Recall that the distribution of these S points $\widetilde{R}^*(t)$ converges asymptotically to a Gaussian distribution. Thus, the $100(1-\alpha)\%$ confidence interval can be easily computed as follows, for a given t,

$$100(1-\alpha)\% \text{ CI} = \widehat{R}^*(t) \pm z_{1-\alpha}\frac{\text{sd}_t}{\sqrt{S}}, \tag{4.14}$$

where the standard deviation of $\widehat{R}^*(t)$ is computed as

$$\text{sd}_t = \sqrt{\frac{1}{S-1}\sum_{l=1}^{S}\left(\widetilde{R}_l^*(t) - \widehat{R}^*(t)\right)^2}.$$

Because of the lack of a closed-form, the AUC is computed using the trapezoidal rule defined as follows:

$$\widetilde{A}_s^* = \frac{1}{2}\sum_{i=2}^{n}\left(\widetilde{\text{FP}}^*_{s_i} - \widetilde{\text{FP}}^*_{s_{i-1}}\right)\left(\widetilde{\text{TP}}^*_{s_i} + \widetilde{\text{TP}}^*_{s_{i-1}}\right). \tag{4.15}$$

The MSE is computed as

$$\widetilde{\mathrm{MSE}}_s^* = \frac{1}{D}\sum_{d=1}^{D}\left(\mathrm{TP}_{s_d}^* - \widetilde{\mathrm{TP}}_{s_d}^*\right)^2, \tag{4.16}$$

where d represents the number of cuts on the x-axis, TP_d indicates the true positive rate from the true ROC curve, for $s = 1, \ldots, S$, respectively.

4.3 Applications

To illustrate the flexibility of the newly proposed method, we applied it on publicly available case-control cancer data published by Wieand *et al.* (1989). The study was conducted at the Mayo Clinic and was concerned with the accuracy of two biomarkers for pancreatic cancer: a cancer antigen, CA 125, and a monoclonal antibody with a carbohydrate antigenic determinant, CA 19-9. The serum concentrations of the two biomarkers were collected from 51 patients without cancer but with pancreatitis and 90 patients with pancreatic cancer, both of which are measured on the continuous positive scale with higher values being more indicative of cancer.

4.3.1 Pancreas Cancer: CA-125

Recall our goal is to model the true unknown curve, i.e., the empirical curve, but with more smoothness and flexibility. First, we notice from the histogram, in Figure 4.1, that the normality assumption is not valid for both populations. The MG method detects $G_X = 3$ and $G_Y = 4$, the MT method finds $G_X = 2$ and $G_Y = 3$, and the MST method detects $G_X = G_Y = 3$ for non-diseased and diseased group, respectively. Unsurprisingly, the ROC curve of the standard binormal curve performs poorly compared to other methods, see Figure 4.2. Note the standard binormal is computed directly via (4.4) without any transformation. The MG and MST methods outperform the well-known LABROC. Figure 4.3 also confirms this statement. Table 4.1 presents the performance measures, i.e., the AUC and the MSE. When comparing the AUC value, LABROC obtained the closest value to the empirical. However, as mentioned in Remark 3.4.2, two distinct curves can have the same AUC. Therefore, the MSE provides an additional information. The MSE confirms that GM and MST methods surpass other methods with a MSE of 0.0016.

Table 4.1: Area under the curve (AUC) and mean square error (MSE) of ROC estimates for biomarker CA 125.

Biomarker	Measure	BIN	LABROC	MG	MT	MST
Pancreas	AUC (0.6966)	0.5923	0.6946	0.7082	0.7153	0.7187
CA 125	MSE	0.0529	0.0021	0.0016	0.0052	0.0016

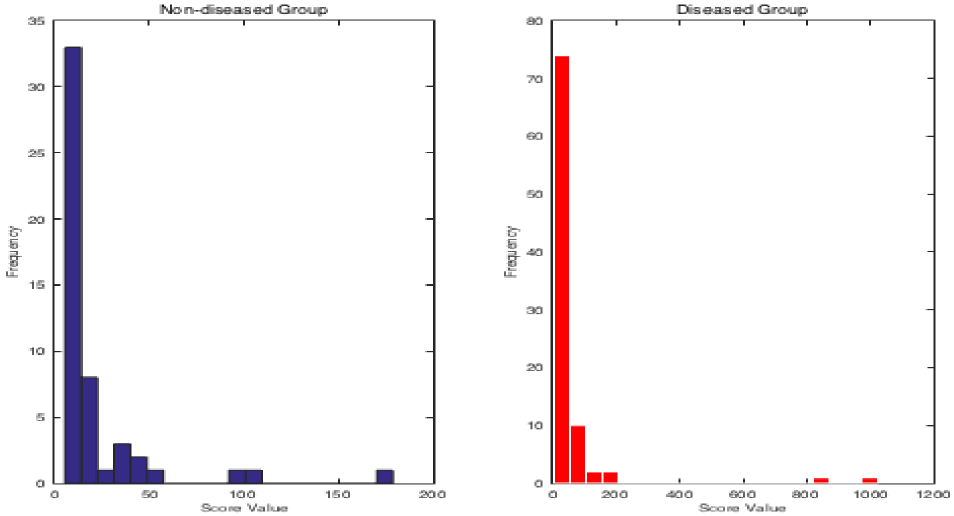


Figure 4.1: Histograms for the pancreatic cancer data using biomarker CA 125 for the non-diseased (blue) and the disease (red) group.

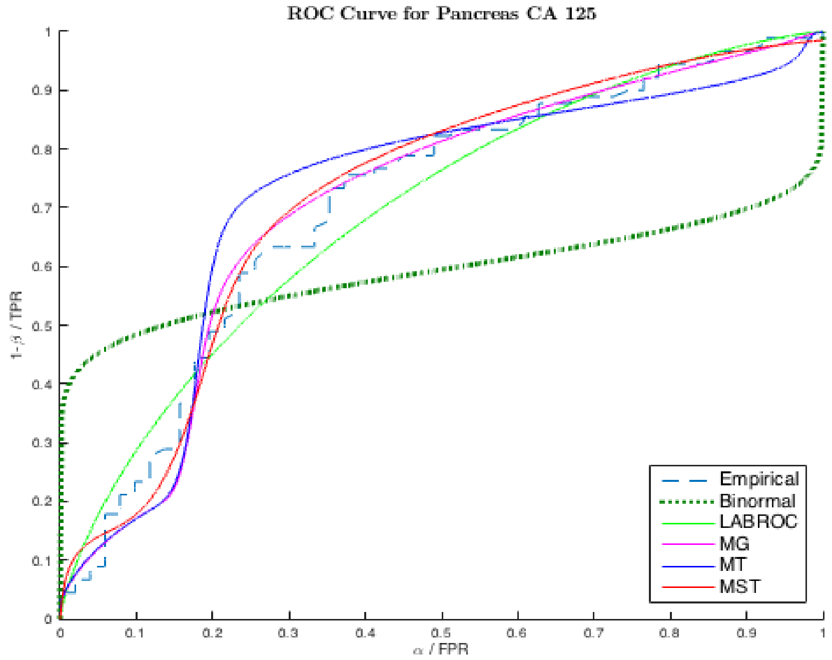


Figure 4.2: ROC curves for the pancreatic cancer data using biomarker CA 125 with different approaches.

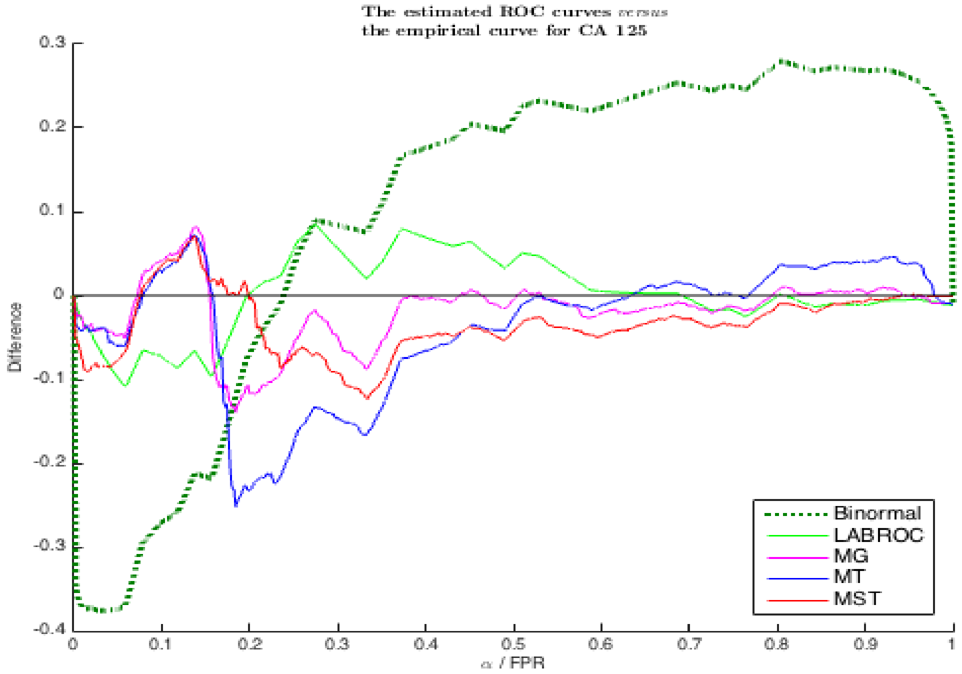


Figure 4.3: Difference between the estimated ROC curves *versus* the empirical curve for pancreatic cancer data using biomarker CA 125.

4.3.2 Pancreas Cancer: CA 19-9

Again, the standard binormal give irrelevant information, see Figures 4.4 and 4.5, caused by the unsuitable assumption of normality. The MG method detects $G_X = 3$ and $G_Y = 6$, while the MT method and the MST method find $G_X = G_Y = 3$ for non-diseased and diseased group, respectively. These data seem less skewed, but with heavier tails, which may explained the outstanding performance of MT method. Compared to the LABROC method, the MT method performs relatively well in terms of replication and closeness to the empirical curve. Quantitatively, the MT method obtains an AUC close to the empirical AUC, as reflected in a small AUC difference, i.e., $|0.8521 - 0.8570| = 0.0049$, see Table 4.2. In term of MSE, LABROC has the smallest MSE (0.0017) followed by MT (0.0028). Without any monotonic transformation, the proposed methods surpass significantly the standard binormal and perform as well as LABROC.

Table 4.2: Area under the curve (AUC) and mean square error (MSE) of ROC estimates for biomarker CA 19-9.

Biomarker	Measure	BIN	LABROC	MG	MT	MST
Pancreas	AUC (0.8521)	0.6773	0.8625	0.8585	0.8570	0.9269
CA 19-9	MSE	0.0442	0.0017	0.0031	0.0028	0.0083

4.4 Discussion

Direct approaches are easy to compute and are based on minimal assumptions. However, empirical ROC curves may suffer from large variability, notably with small sample sizes (Gönen, 2013). In contrast, indirect approaches give smoother curves.

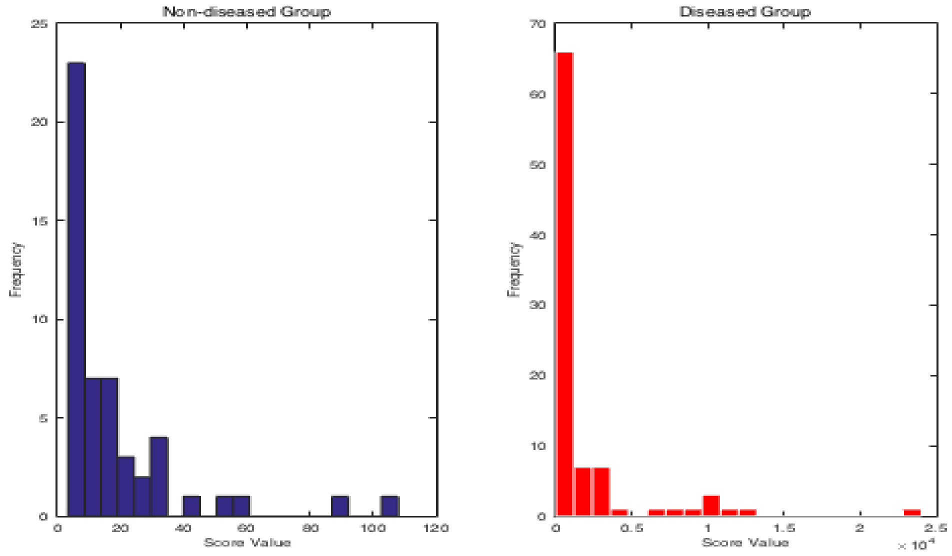


Figure 4.4: Histograms for the pancreatic cancer data using biomarker CA 19-9 for the non-diseased (blue) and the disease (red) group.

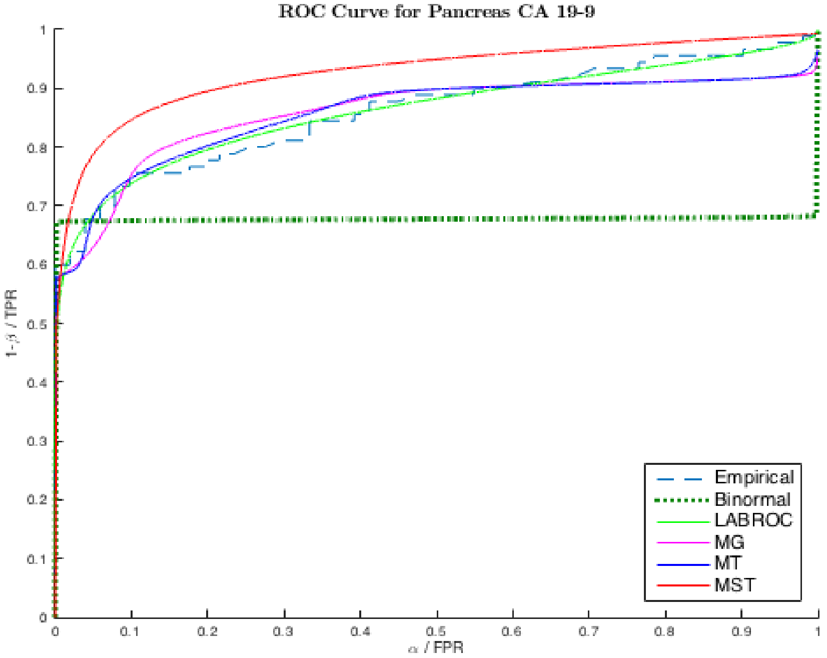


Figure 4.5: ROC curves for the pancreatic cancer data using biomarker CA 19-9 with different approaches.

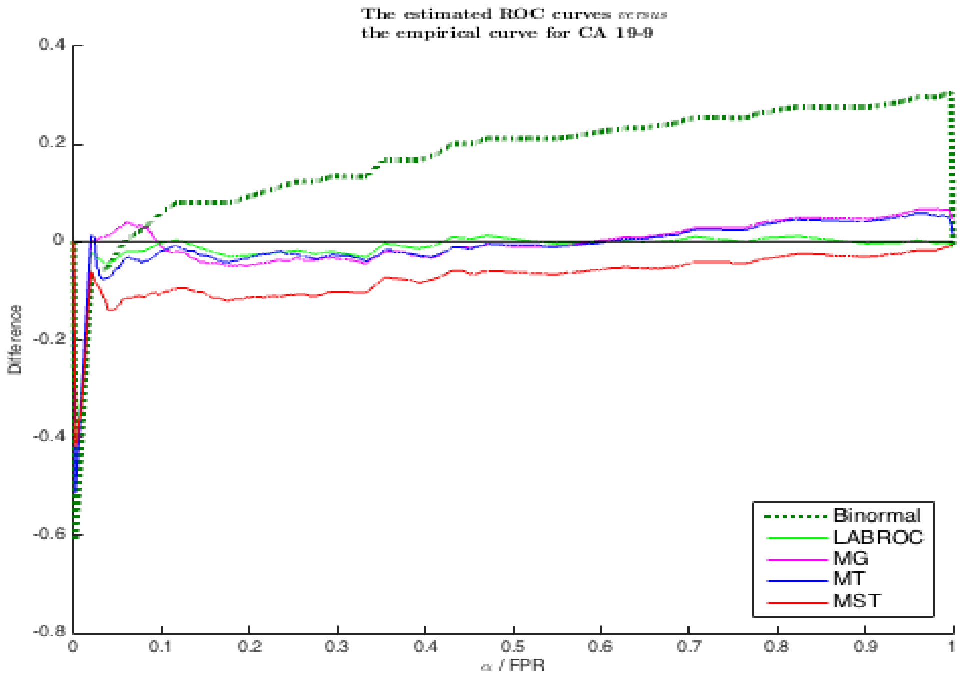


Figure 4.6: Difference between the estimated ROC curves *versus* the empirical curve for pancreatic cancer data using biomarker CA 19-9.

The binormal model has been widely used because it gives convenient maximum likelihood estimates of the ROC curve parameters. The underlying assumption for the binormal is that there exists some unknown monotone increasing function, which would simultaneously transform the observations into normally distributed random variables. Unfortunately, real data may not necessarily follow normal distributions, even after transformation.

This chapter provides a brief summary of the existing method, LABROC, and details of the proposed method for modelling the ROC curve. The proposed method utilizes Gaussian and non-Gaussian mixture distributions for both populations, in conjunction with the Monte Carlo method. The novel method is applied to two pancreatic cancer datasets illustrating its flexibility and smoothness. Depending on the dataset, the proposed methods (MG, MT, MST) either relatively outperform or equivalently perform, when compared to LABROC. Note that the LABROC method still assume that a single monotone transformation exists that would make both groups normal simultaneously. Unlike the LABROC, the proposed method does not require any monotonic transformation.

Chapter 5

Model-Based Clustering

5.1 Overview

In cluster analysis, the goal is to partition observations into meaningful homogeneous subgroups. Tyron and Bailey (1970) stated

> *"Understanding our world requires conceptualizing the similarities and differences between the entities that compose it."*

Obviously the word 'cluster' has a significant meaning in this chapter. Therefore it is entirely natural to ask the question *'What is a cluster?'*. Several authors have reflected on the definition and meaning of a cluster (Everitt *et al.*, 2011; Hennig, 2015). However the most popular definition remains the one based upon similarity (cf. Wolfe, 1963, for discussion). Recently, McNicholas (2016a) proposed a definition more specialized to finite mixture models:

> *"A cluster is a unimodal component within an appropriate finite mixture model."*

The author further defined his particular word choice for a better understanding. The reader is referred to McNicholas (2016a) for additional information.

The term ‘clustering’ is adopted in several research fields to designate methods for grouping unlabeled observations. Interest in clustering has increased astronomically due to the emergence of new domains of application such as astronomy, biology and more (Eisen *et al.*, 1998; Jang and Hendry, 2007). Consequently, many clustering approaches have been developed, which can be classified into two types: hierarchical and non-hierarchical (or partitional). Note that some authors may prefer to organize the literature differently such as ‘discriminative’ or distance-based versus ‘generative’ or model-based (Zhong and Gosh, 2003).

Hierarchical clustering finds successive clusters using previously established clusters. It can be carried out using either an agglomerative or a divisive method with the first being the most commonly utilized. In agglomerative (or bottom-up) clustering, each observation represents a cluster of its own and then successively merges clusters together until a stopping criterion is verified or all clusters are contained within a single cluster. Conversely, divisive (or top-down) clustering begins with all observations in a single cluster and then successively divide into sub-clusters until a stopping criterion is reached. Hence, the problem reduces to selecting an appropriate distance measures (Jain *et al.*, 1999). These methods present one major advantage: their construction is intuitive. Nevertheless their lack of statistical basis and their inability to back-track appear to be a limitation for their use. In addition, the time complexity is at least $\mathrm{O}(n^2)$, where n is the total number of observations.

Unlike hierarchical methods that are based on distance measures, partitional clustering is an iterative relocation algorithm. The clustering process starts with an initial partition. The quality of this partition is then improved by minimizing the within-cluster variation. If the reallocation of the observation to another cluster decreases the within-cluster variation, this observation remains in this assigned cluster. Like its counterpart, partitional clustering can be divided into four categories: prototype-based (e.g., K-means or K-medoids see MacQueen, 1967 and Hartigan, 1975), density-based (e.g., DBSCAN see Ester *et al.*, 1996), grid-based (e.g., CLIQUE see Agrawal *et al.*, 2005) and model-based clustering (Fraley and Raftery, 2002; McNicholas, 2016b). For the purpose of this thesis, the prime focal point is model-based clustering because it gives more flexibility. Compared to hierarchical clustering, model-based clustering offers better interpretability because each cluster has its own postulated model. In addition, the computational complexity is linear with respect to the number of observations.

5.2 Spatio-Temporal Data

In the last few years, spatio-temporal data has become ubiquitous thanks to the availability of cheap sensor devices and remarkable development of computer power. As a consequence, it attracted tremendous attention from researchers in diverse fields such as medicine (Gaudart *et al.*, 2006), geography (Birant and Kut, 2007; Wu *et al.*, 2015) and others, where there is a drive to understand and interpret complex spatio-temporal phenomena. A spatial framework consists of a collection of locations and a neighbour relationship, while a time series consists of a sequence of observations taken progressively in time. Combining these two notions, spatio-temporal data is a collection of times series, each referencing a location in a common spatial framework.

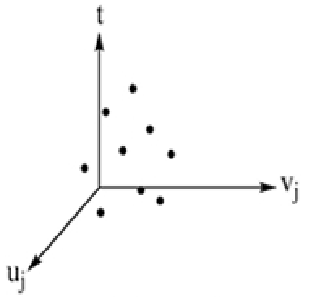
Due to its complex nature, it is pragmatic to partition these data into homogeneous subgroups by considering the temporal and spatial information. While most of the prior work on clustering treated space and time separately, it is undeniable that accounting for both could yield better results. For instance, in public health, the aim is to detect clusters that are prominent in time which may be indicative of a naturally occurring disease outbreak such as the influenza or an environmental hazard like a radiation leak. Thus the clustering method depends heavily on the type of data used. Kisilevich *et al.* (2010) provided a classification of different types of spatio-temporal data: spatio-temporal (ST) events, geo-referenced variables, geo-referenced time series, moving points and trajectories.

For a better understanding, Table 5.1, which is taken from Kisilevich *et al.* (2010), summarizes each type of data by dividing it into fixed and dynamic location. Here the word dynamic signifies that the spatial location is time-changing, for example, the GPS in the cellphone provides location and time information. For a sterling description of each type of data, the reader is referred to Kisilevich *et al.* (2010). This thesis focuses on the geo-referenced time series, i.e., a collection of time series each referencing a location. For example, each data item represents a weather station with its corresponding temperature at different times. An example of representation of collected data is presented in Table 5.2.

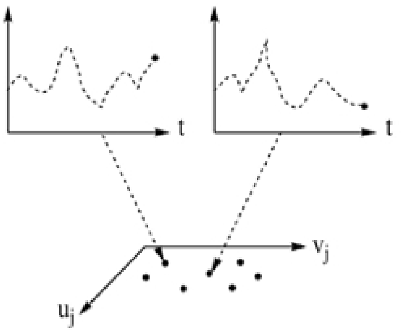
The clustering problem resumes to group the multivariate time series with spatial dependencies. In other words, clustering a set of observations requires the curvatures of the time series to be compared and associated with its corresponding location. In practice, the time series are observed only at discrete times. However, many clustering methods have been developed for continuous time series which require transformation

Table 5.1: Nomenclature of different types of spatio-temporal data.

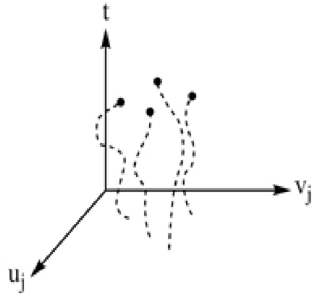
<table>
<tr><th>Time/Location</th><th>Fixed location</th><th>Dynamic location</th></tr>
<tr><td>Single snapshot</td><td>ST event



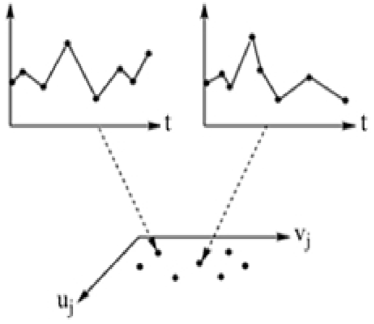
</td><td>N/A</td></tr>
<tr><td>Update snapshot</td><td>Geo-referenced variables



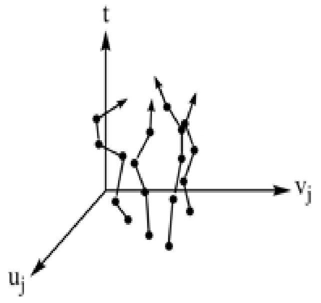
</td><td>Moving points



</td></tr>
<tr><td>Time series</td><td>Geo-referenced time series



</td><td>Trajectories



</td></tr>
</table>

Table 5.2: An illustration of data set for geo-referenced time series.

Time	Longitude	Latitude	Temperature
t_1	u_1	v_1	V_{11}
t_1	u_2	v_2	V_{12}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_1	u_J	v_J	V_{1J}
t_2	u_1	v_1	V_{21}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_T	u_J	v_J	V_{TJ}

of the observed data. Unfortunately, these transformations are not unique and the choice of method is arbitrary. For example, Appice *et al.* (2013) suggested to use trend clusters to interpolate missing data, and inverse distance weighting interpolation to estimate values at any spatial location and at any time. Meanwhile Wu *et al.* (2015) proposed employing the Bregman block average co-clustering algorithm with I-divergence which allows simultaneous study of the spatial and temporal patterns. Note that the number of clusters in these approaches is either pre-specified or distance-based. However, selecting the number of classes remains a challenging problem.

An interesting avenue to address the aforesaid problem is to apply the concept of model-based clustering. Thus, selecting the number of clusters is achievable by probability theory, and interpreting the results is easier and convenient. Samé *et al.* (2011) proposed a method in which the observed data do not require any transformation. They suggested a mixture model, where each component follows a polynomial regression mixture in which the logistic weights depend on the time. More specifically, each observation of a time series arises independently from one of the polynomial regression model specific to the component to which belongs. Unfortunately, this model considers only the univariate temporal framework.

In the upcoming chapter, a novel model is introduced extending this concept to a higher level with the addition of a spatial and an autoregressive component. The proposed model, named STM, is a mixture model, where each component is an autoregressive polynomial regression mixture in which the logistic weights depend on the spatial and temporal dimensions. Hence, the STM model is not restricted to univariate temporal framework, but can also model spatial dependencies for multivariate functional data.

5.3 Functional Data

As mentioned previously, this thesis focuses on geo-referenced time series which implies the aim of clustering is to group the multivariate time series with spatial dependencies. Subtracting the spatial dependencies, the problem is similar to functional data. Ferraty and Vieu (2006) defined functional data as a set of curves belonging to an infinite dimensional space.

The curse of dimensionality was and remains an active topic in statistics. The challenge, when dealing with functional data, emerges from the fact that the observations are supposed to come from an infinite dimensional space. However, in practice, curves are generally observed at discrete observation points. As a consequence, the first step in functional data analysis is usually the reconstruction of the functional form of the discrete observations, say $\mathbf{X}_{ij}$, of each sample path $\mathbf{X}_i(m)$ at a finite set of time grids m_{ij}, for $j = 1, \ldots, t_i$. Thus, this can be executed by predefining functional basis such as Fourier, wavelets, splines and more (Ramsay and Silverman, 2005). Consider a basis $\Psi = \{\psi_1, \ldots, \psi_L\}$, we assume that the basis expansion, for some $L \in \mathbb{N}$, is given by

$$\mathbf{X}_i(m) = \sum_{\ell=1}^{L} \boldsymbol{\alpha}_{i\ell} \psi_\ell(m), \tag{5.1}$$

where $\boldsymbol{\alpha}_{i\ell} \in \mathbb{R}$ is the basis coefficients. Although the choice of functional basis is arbitrary, it highly influences its results because, when using the same model to cluster, two different bases can obtain different partitions.

There has been a significant amount of research on functional data clustering. Jacques and Preda (2014a) classified different approaches into four categories: raw data methods, filtering methods, adaptive methods and distance-based methods. The proposed method, named TM, falls into the adaptive methods category. By adaptive methods, the authors imply any method whose functional representation of data depends on clusters, and that the dimensionality reduction and the clustering proceed simultaneously. For further details of each approach, the reader is referred to Jacques and Preda (2014a).

Within the adaptive methods, the authors divided it into two subgroups based on the choice of the probabilistic modelling: basis expansion coefficients and functional principal component analysis (FPCA) scores. For instance, James and Sugar (2003) considered Gaussian mixture distributions as the basis expansion coefficients of the curves into a natural cubic spline, i.e., $\boldsymbol{\alpha}_i \sim \mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})$. Note that the mean $\boldsymbol{\mu}_g$ is specific to each component but the variance $\boldsymbol{\Sigma}$ is common for all. Thus, we observe that the basis coefficients are considered as random variables rather than fixed variables, unlike the filtering approaches. Similarly, Jacques and Preda (2014b) applied the idea of Gaussian mixture modelling to FPCA scores. For the TM model, the basis expansion is the polynomial regression mixtures which give the possibility to change from one regression model to a different one in time and also space.

5.4 Performance Assessment

Given the knowledge of the ground truth class assignments, clustering performance evaluation is achievable. For the purpose of demonstrating the efficacy of a proposed method, a toy dataset is simulated or a dataset that has true group memberships is considered. Frequently used performance assessment measures include the Rand and adjusted Rand indices, abbreviated RI and ARI, respectively. Both indices are a measure of agreement between two partitions.

5.4.1 The Rand Index

Rand (1971) introduced a similarity function that converted the problem of comparing two partitions with possibly differing number of classes into a problem of computing pairwise label relationships. Thus, the RI is simply the proportion of pairwise agreements between two partitions and is expressed as

$$\mathrm{RI} = \frac{\text{number of pairwise agreements}}{\text{number of pairwise agreements} + \text{number of pairwise disagreements}}. \tag{5.2}$$

Note, we refer the ‘number of pairwise agreements’ as the number of pairs of observations that were correctly clustered into the same group plus the number of pairs of observations that were correctly clustered into the different groups. The ‘number of pairwise disagreements’ corresponds to the number of pairs that were incorrectly clustered.

5.4.2 The Adjusted Rand Index

To assess clustering results in this thesis, we mostly use the ARI (Hubert and Arabie, 1985). An issue with the RI is that its expected value for random classification does not take a constant value such as zero. Thus, the ARI corrects the RI for chance by considering the fact that if a classification is performed randomly, some cases will be correctly classified by chance. Therefore the expected value of an ARI under random classification is 0.

5.5 Software: MCLUST

MCLUST is a contributed R package for Gaussian mixture modelling and model-based clustering (Fraley and Raftery, 2002). Given observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, MCLUST assumes a Gaussian mixture model

$$f(\mathbf{x}_i \mid \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \mathbf{\Sigma}_g), \tag{5.3}$$

where

$$\phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \mathbf{\Sigma}_g) = \frac{1}{\sqrt{(2\pi) \mid \mathbf{\Sigma}_g \mid}} \exp\left\{\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_g)'\mathbf{\Sigma}_g^{-1}(\mathbf{x} - \boldsymbol{\mu}_g)\right\} \tag{5.4}$$

is the density of a multivariate Gaussian with mean $\boldsymbol{\mu}_g$ and covariance $\mathbf{\Sigma}_g$, and π_g is the probability of membership of group g.

Banfield and Raftery (1993), Celeux and Govaert (1995) and Fraley and Raftery (2002) exploited a model-based framework by parameterizing the following eigenvalue decomposition of the covariance matrix

$$\mathbf{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}'_g, \tag{5.5}$$

where

(i) $\mathbf{D}_g$ is the orthogonal matrix of eigenvectors which determines the orientation of the principal components of $\mathbf{\Sigma}_g$;

(ii) $\mathbf{A}_g$ is the diagonal matrix whose elements are proportional to the eigenvalues of $\mathbf{\Sigma}_g$ which determines the shape of the density contours, and $\det(\mathbf{A}_g) = 1$;

(iii) λ_g is a constant which specifies the volume of the corresponding ellipsoid.

Characteristics (i.e., orientation, shape and volume) of distributions are usually estimated from the data and can either vary between clusters or be constrained across clusters. Parameter estimation is carried out using the EM algorithm.

Chapter 6

Modelling Spatio-Temporal Data

6.1 Model

6.1.1 General Mixture Model

We introduce a novel model using a mixture, where each component is an autoregressive polynomial regression mixture in which the logistic weights depend on the spatial and temporal dimensions. In this chapter, some notations are redefined from the previous Chapters 3 and 4. Let $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_n')'$ be independently and identically distributed random variables with realizations $\mathbf{x} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$. Recall the density of a finite mixture model in Section 2.2,

$$f(\mathbf{x}_i \mid \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i \mid \boldsymbol{\vartheta}_g), \tag{6.1}$$

where G is the number of components, π_g are the mixing proportions such that

$$\sum_{g=1}^{G} \pi_g = 1 \quad \pi_g > 0,$$

and $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_G)$ is the parameter vector. The $f_g(\mathbf{x}_i \mid \boldsymbol{\vartheta}_g)$ are called component densities.

6.1.2 Autoregressive Mixture Model

Before presenting the novel model, it is interesting to break down each aspect by recalling its definition. In the proposed model, we introduce an autoregressive aspect to the univariate time series, i.e., an ordered sequence of measurements of the same variable collected over time. Usually, the measurements are made at evenly spaced times, e.g., hourly, monthly or yearly. Here, we omit the spatial dimension and only consider the temporal dimension. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{iT})'$, for $i = 1, \ldots, n$, be the collection of observable variables at time t. For this thesis, we assume there exists a strong relationship between the immediate past and current values. This suggests using $x_{i(t-1)}$ to explain x_{it} in a regression model. Using previous values of a series to forecast current values of a series is termed an autoregression.

Definition 6.1.1. *The autoregressive model of order 1, denoted by AR(1), is given by*

$$X_{it} = \omega_0 + \omega_1 x_{i(t-1)} + \varepsilon_{it}, \quad t = 2, \ldots, T \tag{6.2}$$

where ε_{it} is a white noise process with zero mean and constant variance σ^2_ε, and ω_0, and ω_1 are unknown parameters.

In the AR(1) model, the parameter ω_0 may be any fixed constant, while the parameter $\mid \omega_1 \mid < 1$. By restricting ω_1, it can be established that the AR(1) model is stationary. Note that (6.2) is

1. a white noise process if $\omega_1 = 0$,

2. a random walk if $\omega_1 = 1$.

Here, we focus on the stationary process, i.e., the mean and variance stay steady over time. Thus, $E[X_{it}] = E[X_{i(t-1)}] = 0$ and $V[X_{it}] = V[X_{i(t-1)}] = \sigma_\varepsilon^2$.

Definition 6.1.2. *The mixture of autoregressive components can be defined by, for* $x_{i0} = 0$*,*

$$X_{it} \mid x_{i(t-1)} \sim \sum_{g=1}^{G} \omega_g \prod_{t=1}^{T} \phi(x_{it} \mid c_{gt}, \sigma_g^2), \tag{6.3}$$

where $\phi(\cdot \mid \mu, \sigma^2)$ *is the univariate Gaussian distribution with mean* μ *and variance* σ^2*, and the mixing weights* ω_g *satisfy the constraints* $\omega_g > 0$ *and* $\sum_{g=1}^{G} \omega_g = 1$*, and*

$$c_{gt} = \mu_g + a_{g1}(x_{i(t-1)} - \mu_g). \tag{6.4}$$

Note that (6.4) is the conditional mean of a stationary autoregressive model of order 1 with stationary mean μ_g and autoregressive coefficient a_{g1}.

6.1.3 Polynomial Regression Mixture Model

Another underlying aspect of this new model is polynomial regression. Polynomial regression is among the most frequently used nonlinear models because it can approximate the shape of the empirical curvilinear, when the distribution is unknown or has possibly complex nonlinear relationships. Suppose the spatio-temporal data $\mathbf{x}_i = (x_{i11}, \ldots, x_{iJT})'$, for $i = 1, \ldots, n$, where j denotes the spatial index and t the temporal index that corresponds to time locations m_t. For example, the data can be described in terms of T images each with J pixels. Thus, at each pixel location j, we have the temporal sequence $\mathbf{x}_{ij}$ of length T.

Definition 6.1.3. *The Q-order polynomial regression on the time grid $\mathbf{m} = (m_1, \dots, m_T)$ with additive noise term is given by, for $i = 1, \dots, n$ and $j = 1, \dots, J$,*

$$\mathbf{x}_{ij} = \mathbf{M}\boldsymbol{\beta} + \varepsilon_{ij}, \tag{6.5}$$

where $\mathbf{M}$ is the Vandermonde matrix, i.e.,

$$\mathbf{M} = \begin{pmatrix} 1 & m_1 & \cdots & m_1^Q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & m_T & \cdots & m_T^Q \end{pmatrix}$$

and $\boldsymbol{\beta}$ is the $(Q+1)$-vector of regression coefficients, and ε_{ij} is the error term which is assumed to be Gaussian and independent over time, i.e., $\varepsilon_{ij} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with a diagonal covariance matrix $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_T^2)$.

In this thesis, we consider the problem of curve clustering. In other words, we want to group the set of curves $\mathbf{x}_{ij}$ into G components, where each component will contain curves of the same polynomial regression model. Following this idea, the polynomial regression mixture is a useful generative model that can be used to capture curvilinearity. Hence, the polynomial regression mixture is described as

$$f(\mathbf{x}_{ij} \mid \boldsymbol{\theta}) = \sum_{g=1}^{G} \omega_g f_g(\mathbf{x}_{ij} \mid \boldsymbol{\vartheta}_g), \tag{6.6}$$

where $f_g(\mathbf{x}_{ij} \mid \boldsymbol{\vartheta}_g) = \mathcal{N}(\mathbf{M}\boldsymbol{\beta}_g, \boldsymbol{\Sigma}_g)$, $\boldsymbol{\vartheta}_g = (\boldsymbol{\beta}_g, \boldsymbol{\Sigma}_g)$ is the set of parameters of the gth component, and $\omega_g \in (0, 1]$, such that $\sum_{g=1}^{G} \omega_g = 1$, are called mixing proportions.

6.1.4 Autoregressive Polynomial Regression Mixtures to Model the Components

Suppose each observation $\mathbf{x}_i = (x_{i11}, \ldots, x_{iJT})$ consists of $J \times T$ observations over a pre-specified time grid $\mathbf{m} = (m_1, \ldots, m_T)$ and spatial grid $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_J)$. The realization of X_{ijt} corresponding to observation i at site j and time t is $x_{ijt} \in \mathbb{R}$. The spatial coordinates of site j are defined by the bivariate vector $\mathbf{s}_j = (u_j, v_j)$.

The spatio-temporal model, named STM, supposes that each site j is conditionally independent on the membership component $\mathbf{z}_i$, defined in (2.5), and the dependency in time is Markov, i.e., autoregressive. Thus, we have

$$\mathrm{P}(\mathbf{X}_i \mid \boldsymbol{\vartheta}_g, Z_{ig} = 1) = \prod_{j=1}^{J} \prod_{t=1}^{T} \mathrm{P}(X_{ijt} \mid \boldsymbol{\vartheta}_g, Z_{ig} = 1, x_{ij(t-1)}), \tag{6.7}$$

where $\mathbf{X}_i = (X_{i11}, \ldots, X_{iJT})'$. The distribution of $X_{ijt} \mid \boldsymbol{\vartheta}_g, Z_{ig} = 1, x_{ij(t-1)}$ is a K-component mixture of Q-order polynomial regression. From (6.1), the pdf of the gth component of the STM model is given by

$$f_g(\mathbf{x}_i \mid \boldsymbol{\vartheta}_g) = \prod_{j=1}^{J} \prod_{t=1}^{T} \sum_{k=1}^{K} \omega_{gjtk}(\boldsymbol{\lambda}_g) \phi(x_{ijt} \mid [\mathbf{M}_t - \mathbf{M}_{t-1}]'\boldsymbol{\beta}_{gk} + x_{ij(t-1)}, \sigma_{gk}^2), \tag{6.8}$$

where

- $\boldsymbol{\vartheta}_g = (\boldsymbol{\lambda}_g, \boldsymbol{\beta}_{gk}, \sigma_{gk}; k = 1, \ldots, K)$ is the set parameters of the gth component;

- $\omega_{gjtk}(\boldsymbol{\lambda}_g)$ are the weights of the autoregressive polynomial regression mixtures that depend on both the spatial and the time dimensions using a logistic function defined as follows

$$\omega_{gjtk}(\boldsymbol{\lambda}_g) = \frac{\exp(\lambda_{gk1}u_j + \lambda_{gk2}v_j + \lambda_{gk3}m_t + \lambda_{gk4})}{\sum_{\ell=1}^{K}\exp(\lambda_{g\ell 1}u_j + \lambda_{g\ell 2}v_j + \lambda_{g\ell 3}m_t + \lambda_{g\ell 4})} \quad \forall\ (g,j,t,k) \qquad (6.9)$$

and $\boldsymbol{\lambda}_g = (\lambda_{g11}, \dots, \lambda_{gK4})$ is the parameter;

- $\phi(\cdot \mid \mu, \sigma^2)$ is the pdf of the univariate Gaussian distribution with mean μ and σ^2;

- $\mathbf{M}_t = (1, m_t, \dots, m_t^Q)$ represents the vector of the Q degree polynomial of $\mathbf{M}_t$ and $\mathbf{M}_0$ is a null vector of size $Q+1$;

- $\boldsymbol{\beta}_{gk} = (\beta_{gk0}, \dots, \beta_{gkQ})$ is the coefficient vector of the kth regression model for the gth component.

Note we define $x_{ij0} := 0$. Because the gth component is itself a mixture, it requires a second latent variable, denoted $\mathbf{y}_{ijt} = (y_{ijt1}, \dots, y_{ijtK})$, conditional on $z_{ig} = 1$, outlined as

$$y_{ijtk} = \begin{cases} 1 & \text{if } x_{ijt} \in k\text{th polynomial regression of the } g\text{th component,} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $\mathbf{Y}_{ijt}$ can be viewed as a label assignment of the polynomial regression mixture that adjusts the observation x_{ijt} conditionally on $z_{ig} = 1$. Thus, it follows a multinomial distribution $\mathcal{M}_K(\omega_{jt1}(\boldsymbol{\lambda}_g), \dots, \omega_{jtK}(\boldsymbol{\lambda}_g))$. To ensure model identifiability, we impose that $\lambda_{g1h} = 0$ for $h = 1, \dots, 4$. The proof of the model identifiability is detailed in the following section. Note that the elements of the observation $\mathbf{x}_i$ do not necessarily emerge from the same polynomial regression.

Remark 6.1.4. *If we remove the spatial dependencies inside the weights ω_{gjtk} in* (6.9)*, i.e., $\lambda_{gk1} = \lambda_{gk2} = 0$ $\forall g, k$, the STM model can be utilized for functional data analysis which has only time dependencies. The basis expansion of* (5.1) *follows the polynomial regression model. In other words, at each time grid m_t, the observation X_{it} is assumed to evolve from one of the polynomial regression models specific to the component it belongs to.*

6.1.5 Generative Model

Equations (6.1) and (6.8) entail the following three steps of the generative model:

1. Sampling of the component membership

$$\mathbf{Z}_i \sim \mathcal{M}_G(\pi_1, \ldots, \pi_G);$$

2. Conditional sampling of the regression label assignment for all (j, t)

$$\mathbf{Y}_{ijt} \mid Z_{ig} = 1 \sim \mathcal{M}_K(\omega_{jt1}(\boldsymbol{\lambda}_g), \ldots, \omega_{jtK}(\boldsymbol{\lambda}_g));$$

3. Conditional sampling of the observation for all (j, t)

$$X_{ijt}|Z_{ig} = 1, Y_{ijtk} = 1 \sim \mathcal{N}([\mathbf{M}_t - \mathbf{M}_{t-1}]'\boldsymbol{\beta}_{gk} + x_{ij(t-1)}, \sigma^2_{gk}).$$

For a better visualization, Figure 6.1 illustrates the three steps of the generative model.

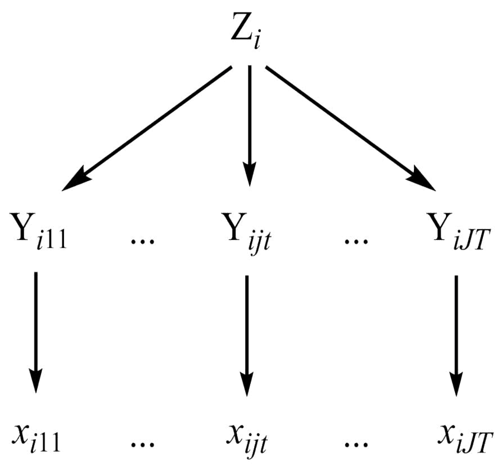


Figure 6.1: Diagram describing the three steps of the generative model.

Remark 6.1.5. *A summarization and visualization of each component can be achieved by plotting an average curve of each component using the following results:*

$$\mathbb{E}\left[X_{ijt} \mid Z_i = g\right] = \sum_{k=1}^{K} \omega_{gjtk}(\boldsymbol{\lambda}_g)\mathbf{M}_t'\boldsymbol{\beta}_{gk}, \tag{6.10}$$

because

$$\mathbb{E}\left[X_{ijt} \mid Z_i = g, Y_{ijt} = k\right] = \mathbf{M}_t'\boldsymbol{\beta}_{gk}$$

and

$$\mathbb{E}\left[X_{ijt} \mid Z_i = g, Y_{ijt} = k, X_{ij(t-1)} = x_{ij(t-1)}\right] = [\mathbf{M}_t - \mathbf{M}_{t-1}]'\boldsymbol{\beta}_{gk} + x_{ij(t-1)}.$$

Remark 6.1.6. *For our purposes, Y_{ijt} is used to model spatial and temporal dependencies although one may use it to interpret components. Conditionally on the component, the a posteriori distribution of Y_{ijt} represents the probability that the observation x_{ijt} emerges from the kth regression. The segmentation is achievable because* *conditionally to the component, we can obtain the regression with the greatest probability at time m_t for site s_j.*

6.2 Identifiability

The STM model is generically identifiable, see Definition 2.3.6, disregarding the label switching problem, if

1. $T > Q$;
2. we can construct a matrix of size $(Q+1) \times 2$ consisting of spatial coordinates of the observed sites, where the row rank equals to $(Q+1)$.

Let $\boldsymbol{\Theta}$ be the parameter space of the STM model and $A \subset \boldsymbol{\Theta}$ a subspace of measure zero such that $\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta} \setminus A$. Hence we want to prove the following condition:

$$\forall\ \mathbf{x}, \quad f(\mathbf{x} \mid \boldsymbol{\theta}) = f(\mathbf{x} \mid \widetilde{\boldsymbol{\theta}}) \implies \boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}. \tag{6.11}$$

Instead of proving the previous condition, we demonstrate an equivalent condition, i.e.,

$$\forall\ \mathbf{x},\ \forall\ (j,t) \quad f_{jt}(\mathbf{x}_{jt} \mid \boldsymbol{\theta}) = f_{jt}(\mathbf{x}_{jt} \mid \widetilde{\boldsymbol{\theta}}) \implies \boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}, \tag{6.12}$$

where $f_{jt}(\mathbf{x}_{jt} \mid \boldsymbol{\theta})$ is the marginal probability of $\mathbf{x}_{jt}$ obtained by marginalizing over $f(\mathbf{x} \mid \boldsymbol{\theta})$. Before starting the proof, recall the following important result from probability theory and statistics.

Lemma 6.2.1. *Suppose $U_1 \sim \mathcal{N}(\mu_1, \sigma^2)$ and $U_2 \mid U_1 \sim \mathcal{N}(\mu_2 - \mu_1 + u_1, \sigma^2)$, then we have $U_2 \sim \mathcal{N}(\mu_2, 2\sigma^2)$.*

Proof. See Appendix. □

From Lemma 6.2.1, the marginal probability density function of x_{jt} is defined as

$$f_{jt}(\mathbf{x}_{jt} \mid \boldsymbol{\theta}) = \sum_{g=1}^{G} \sum_{k=1}^{K} \pi_g \omega_{gjtk}(\boldsymbol{\lambda}_g) \phi(x_{jt} \mid \mathbf{M}_t' \boldsymbol{\beta}_{gk}; 2^{t-1} \sigma_{gk}^2). \tag{6.13}$$

Notice that (6.13) is a univariate Gaussian mixture with $G \times K$ components. From Proposition 2.3.5, proven by Teicher (1963), the identifiability of Gaussian mixtures entails that $\forall$ (g, k, j, t)

$$\sigma_{gk}^2 = \widetilde{\sigma}_{gk}^2, \tag{6.14}$$

$$\mathbf{M}_t' \boldsymbol{\beta}_{gk} = \mathbf{M}_t' \widetilde{\boldsymbol{\beta}}_{gk} \tag{6.15}$$

$$\pi_g \omega_{gjtk}(\boldsymbol{\lambda}_g) = \widetilde{\pi}_g \omega_{gjtk}(\widetilde{\boldsymbol{\lambda}}_g) \tag{6.16}$$

Identifiability of σ_{gk}^2

The identifiability of the parameters σ_{gk}^2 is directly demonstrated thanks to (6.14).

Identifiability of $\boldsymbol{\beta}_{gk}$

In (6.15), $\mathbf{M} = (\mathbf{M}_1, \ldots, \mathbf{M}_T)$ represents a $T \times (Q + 1)$ Vandermonde matrix. Recall that all numbers m_t are distinct. Then if $T \geq (Q + 1)$, we can construct a square sub-matrix of $\mathbf{M}$ of size $(Q+1) \times (Q+1)$. That sub-matrix is a Vandermonde matrix as well and is positive definite. Hence, we prove $\boldsymbol{\beta}_{gk} = \widetilde{\boldsymbol{\beta}_{gk}}$.

Identifiability of π_g

Recall $\omega_{gjtk}(\boldsymbol{\lambda}_g)$ is the weights of the polynomial regression mixtures, see Section 6.1.4, and, therefore, $\sum_{k=1}^{K}\omega_{gjtk}(\boldsymbol{\lambda}_g) = 1$. Equation (6.16) implies, $\forall\ (g,j,t,k)$, $\pi_g = \widetilde{\pi}_g$ which demonstrates the identifiability of π.

Identifiability of $\boldsymbol{\lambda}_g$

For all (g,j,t,k), we have $\omega_{gjtk}(\boldsymbol{\lambda}_g) = \omega_{gjtk}(\widetilde{\boldsymbol{\lambda}}_g)$, where ω_{gjtk} is a logistic function. Using the identifiability conditions proposed by Silvapulle (1981), we prove $\boldsymbol{\lambda}_g = \widetilde{\boldsymbol{\lambda}}_g$.

6.3 Parameter Estimation

As mentioned previously, the STM model requires two latent variables: the component membership $\mathbf{z}_i$ and the label assignment of polynomial regression $\mathbf{y}_{ijt}$. Applying the log-likelihood function, cf. (2.4), to (6.1) and (6.1), the complete-data log-likelihood is obtained

$$\log\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x},\mathbf{y},\mathbf{z}) = \log\mathcal{L}_1(\boldsymbol{\pi}) + \log\mathcal{L}_2(\boldsymbol{\lambda}) + \log\mathcal{L}_3(\boldsymbol{\beta},\boldsymbol{\sigma}), \tag{6.17}$$

where

$$\begin{aligned}
\log\mathcal{L}_1(\boldsymbol{\pi}) &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\log\pi_g,\\
\log\mathcal{L}_2(\boldsymbol{\lambda}) &= \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{ig}y_{ijtk}\log\omega_{gjtk}(\boldsymbol{\lambda}_g),\\
\log\mathcal{L}_3(\boldsymbol{\beta},\boldsymbol{\sigma}) &= \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{ig}y_{ijtk}\log\left\{\phi\left(x_{ijt} \mid [\mathbf{M}_t - \mathbf{M}_{t-1}]'\boldsymbol{\beta}_{gk} + x_{ij(t-1)}, \sigma^2_{gk}\right)\right\}.
\end{aligned}$$

Given an initial arbitrary parameter $\boldsymbol{\theta}^{[0]}$, the iteration $[r]$ of the EM algorithm consists of repeating the following E and M steps:

1. The E-step: Computation of the conditional probability based on the current values of $\boldsymbol{\theta}^{[r]}$, i.e., $\mathbb{E}\left[\log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) \mid \mathbf{X}, \boldsymbol{\theta}^{[r]}\right]$;

$$\eta_{ig}(\boldsymbol{\theta}^{[r]}) := \mathbb{E}\left[Z_{ig} = 1 \mid \mathbf{x}_i, \boldsymbol{\theta}_k^{[r]}\right] = \frac{\pi_g^{[r]} f_g(\mathbf{x}_i \mid \boldsymbol{\theta}_g^{[r]})}{\sum_{\ell=}^{G} \pi_\ell^{[r]} f_\ell(\mathbf{x}_i \mid \boldsymbol{\theta}_\ell^{[r]})}$$

   and

$$\begin{aligned}\tau_{igjtk}(\boldsymbol{\theta}^{[r]}) &:= \mathbb{E}\left[Y_{ijtk} = 1 \mid \mathbf{x}_i, \boldsymbol{\theta}^{[r]}, Z_{ig} = 1\right] \\ &= \frac{\omega_{jtk}^{[r]}\ \phi(x_{ijt} \mid [\mathbf{M}_t - \mathbf{M}_{t-1}]'\boldsymbol{\beta}_{gk}^{[r]} + x_{ij(t-1)}, \sigma_{gk}^{2[r]})}{\sum_{\ell=1}^{K} \omega_{jt\ell}^{[r]}\ \phi(x_{ijt} \mid [\mathbf{M}_t - \mathbf{M}_{t-1}]'\boldsymbol{\beta}_{g\ell}^{[r]} + x_{ij(t-1)}, \sigma_{g\ell}^{2[r]})}.\end{aligned}$$

   These expectations are obtained using the three steps generative model in Section 6.1.5.

2. The M-step: Maximization of the expectation of the complete-data log-likelihood, cf. (6.17);

   (i) Update the mixing weights, $\pi_g^{[r]}$, by differentiating $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{[r]})$ with respect to π_g, which leads to

$$\pi_g^{[r+1]} = \frac{n_g^{[r]}}{n},$$

   where $n_g^{[r]} = \sum_i^n \eta_{ig}(\boldsymbol{\vartheta}^{[r]})$;

   (ii) Update the parameter of regression weights, $\lambda_g^{[r]}$, by differentiating $Q(\boldsymbol{\theta} \mid$

$\boldsymbol{\theta}^{[r]}$) with respect to λ_g, which leads to

$$\boldsymbol{\lambda}_g^{[r+1]} = \arg\max_{\boldsymbol{\lambda}_g} \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{t=1}^{T} \sum_{k=1}^{K} \rho_{igjtk}(\boldsymbol{\vartheta}^{[r]}) \log \omega_{gjtk}(\boldsymbol{\lambda}_g),$$

where $\rho_{igjtk}(\boldsymbol{\vartheta}^{[r]}) = \eta_{ig}(\boldsymbol{\vartheta}^{[r]})\tau_{igjtk}(\boldsymbol{\vartheta}^{[r]})$. Recall the condition $\lambda_{g1h} = 0$ for $h = 1, \ldots, 4$ and $g = 1, \ldots, G$ from (6.9). The computation of $\boldsymbol{\lambda}_g^{[r+1]}$ is performed by Newton-Raphson algorithm (Nocedal and Wright, 2006);

(iii) Update the regression coefficients, $\beta_{gk}^{[r]}$, by differentiating $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{[r]})$ with respect to β_{gk}, which leads to

$$\boldsymbol{\beta}_{gk}^{[r+1]} = \left[\mathbf{A}'\mathbf{D}_{gk}^{[r]}\mathbf{A}\right]^{-1} \left[\mathbf{A}'\mathbf{C}_{gk}^{[r]}\right],$$

where $\mathbf{A}_t = [\mathbf{M}_t - \mathbf{M}_{t-1}]$ is the tth row of a $T \times (Q+1)$ matrix $\mathbf{A}$, $\mathbf{D}_{gk}^{[r]}$ is a $T \times T$ diagonal matrix, where the element t is $\sum_{i=1}^{n} \sum_{j=1}^{J} \rho_{igjtk}(\boldsymbol{\vartheta}^{[r]})$, and $\mathbf{C}_{gk}^{[r]}$ is the Tth vector for which the element t is $\sum_{i=1}^{n} \sum_{j=1}^{J} \rho_{igjtk}(\boldsymbol{\vartheta}^{[r]})(x_{ijt} - x_{ij(t-1)})$;

(iv) Update the variance, $\sigma_{gk}^{2[r]}$, by differentiating $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{[r]})$ with respect to σ_{gk}^2, which leads to

$$\sigma_{gk}^{2[r+1]} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{t=1}^{T} \rho_{igjtk}(\boldsymbol{\vartheta}^{[r]}) \left(x_{ijt} - x_{ij(t-1)} - \mathbf{A}_t'\boldsymbol{\beta}_{gk}^{[r+1]}\right)^2}{\text{trace}(\mathbf{D}_{gk}^{[r]})},$$

where $\text{trace}(\mathbf{D}_{gk}^{[r]}) = \sum_{g=k} d_{gg}^{[r]}$.

6.4 Initialization

The convergence toward a local optimum of the EM algorithm can highly depend on its starting point $\boldsymbol{\theta}^{[0]}$. Here, we borrow the $x$em-EM strategy described by Biernacki *et al.* (2003). The strategy consists of two phases:

- Several short runs of EM from random starting points with few iterations;
- A long run of EM from the solution maximizing the observed $\log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x})$.

A short run of EM, as defined by the authors, allows us to stop the algorithm after a few iterations or as soon as $\log \mathcal{L}(\boldsymbol{\theta}^{[r+1]}) - \log \mathcal{L}(\boldsymbol{\theta}^{[r]}) < \varepsilon$ instead of waiting for convergence. Note, for the upcoming applications, we generate 500 starting points of 20 iterations. The convergence criteria for the long run of EM is $\varepsilon = 10^{-3}$.

6.5 Competing Models

Let $\mathcal{M}$ be the set of competing candidate models. From (6.1) and (6.8), each model $\mathscr{M} \in \mathcal{M}$ is characterized by three factors: the number of components (G), the number of regressions per component (K) and the degree of polynomial regressions (Q). Hence, a model is defined by $\mathscr{M} = (G, K, Q)$ with $G, K, Q \in \mathbb{N}^*$ and the number of candidate models is $\text{card}(\mathcal{M}) = G_{\max} \times K_{\max} \times Q_{\max}$. Thus the model selection is achieved using an information criterion whose value is calculated for each model in $\mathcal{M}$.

6.6 Simulation Study

This section is devoted to an evaluation and investigation of the consistency of inference of the proposed model and the importance of modelling the spatio-temporal dependencies. Results yielded by the R package, `SpaTimeClust`, are compared to those obtained by `mclust` software for model-based clustering (Fraley *et al.*, 2012). Note that the experiment can be reproduced with the default option.

6.6.1 Simulated Data

For this simulation, observations consisted of $J = 25$ sites at $T = 10$ moments. The sample is generated with different sizes, i.e., $n = 50, 100, 200$ and 400, at various levels of overlap, i.e., 5, 10, and 15%, between classes. For each scenario, the STM model generated 100 replicates with the following settings:

$$G = 2 \qquad K = 2 \qquad Q = 1$$

$$\pi_1 = 1/3 \qquad \pi_2 = 2/3$$

$$\boldsymbol{\beta}_{11} = \boldsymbol{\beta}_{21} = (0, 0) \qquad \boldsymbol{\beta}_{12} = (10, -20) \qquad \boldsymbol{\beta}_{22} = (11, -20)$$

$$\lambda_1 = \lambda_2 \quad \text{and} \quad \begin{array}{c|cccc|} & \text{abscissa} & \text{ordinate} & \text{time} & \text{constant} \\ \lambda_{g11} & 0 & 0 & 0 & 0 \\ \lambda_{g12} & 2 & -2 & -1 & 4 \end{array}$$

$$\sigma^2_{gk} = \begin{cases} 14, \text{for } 5\% \text{ misclassification error} \\ 25, \text{for } 10\% \text{ misclassification error} \\ 37, \text{for } 15\% \text{ misclassification error.} \end{cases}$$

6.6.2 Consistency of Estimates

First, we test the consistency of estimates of the STM model. The results show that the estimates are consistent because, as n increases, the estimates rapidly converge toward zero, see Table 6.1. We notice, for example, at a 15% misclassification rate, the mean distance of $\mid \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \mid$ decreases quickly from 5.17 ($n = 50$) to 0.27 ($n = 400$).

6.6.3 Consistency of Model

The validation of the consistency of model selection is fundamental due to the unknown nature of the dataset, in practice. The BIC is computed after fitting all possible models, for each replicate, where $G_{\max} = K_{\max} = 3$ and $Q_{\max} = 2$. The model that maximizes this criterion is nominated as the 'best' model. Table 6.2 shows the statistics of the winning models. Unsurpringly, when the sample size is small and the misclassification error rate is high, G is underestimated. However, as n increases, the STM model can adequately identify the true grouping. Furthermore, the STM model is able to determine the true number of regressions (K) and the degree of polynomial regression (Q) in all scenarios.

6.6.4 Importance of Dependencies

Table 6.3 shows the misclassification error rate between the STM model and the existing MCLUST model. For the purpose of comparison, we decide to use the classic dependency for MCLUST that is not necessarily specific to spatio-temporal data. As a consequence, the misclassification error rate for MCLUST is higher than STM, which converges rapidly. These results demonstrate the importance of considering the spatio-temporal dependencies in the model.

Table 6.1: Mean (standard deviation) of the distance between the true parameters and the estimates: proportions ($\mid \hat{\boldsymbol{\pi}} - \boldsymbol{\pi} \mid$), logistic weight parameters ($\mid \hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \mid$), regression parameters ($\mid \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \mid$), and noise parameters ($\mid \hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma} \mid$).

Size	$\mid \hat{\boldsymbol{\pi}} - \boldsymbol{\pi} \mid$			$\mid \hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \mid$			$\mid \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \mid$			$\mid \hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma} \mid$		
(n)	5%	10%	15%	5%	10%	15%	5%	10%	15%	5%	10%	15%
50	0.01	0.03	0.04	0.04	0.05	0.06	0.37	1.95	5.17	0.16	1.13	5.30
	(0.01)	(0.04)	(0.05)	(0.05)	(0.06)	(0.08)	(0.03)	(3.19)	(7.41)	(0.13)	(0.49)	(8.29)
100	0.01	0.01	0.03	0.02	0.02	0.03	0.18	0.59	2.66	0.08	0.34	1.46
	(0.01)	(0.02)	(0.04)	(0.03)	(0.03)	(0.04)	(0.12)	(0.60)	(5.31)	(0.06)	(0.28)	(1.73)
200	0.00	0.01	0.01	0.01	0.01	0.01	0.09	0.21	0.55	0.04	0.18	0.57
	(0.00)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.06)	(0.15)	(0.58)	(0.03)	(0.15)	(0.65)
400	0.00	0.00	0.01	0.00	0.00	0.00	0.05	0.03	0.25	0.02	0.09	0.17
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.08)	(0.27)	(0.02)	(0.06)	(0.32)

Table 6.2: Percent chance of selecting (G, K, Q) with the STM model, where the true grouping is (2, 2, 1).

Size	Errors	G			K			Q		
(n)	(%)	1	2	3	1	2	3	0	1	2
50	5	3	97	0	0	100	0	0	99	1
	10	61	39	0	0	100	0	0	97	3
	15	97	3	0	0	100	0	0	97	3
100	5	0	100	0	0	100	0	0	100	0
	10	26	74	0	0	100	0	0	100	0
	15	83	17	0	0	100	0	0	100	0
200	5	0	100	0	0	100	0	0	100	0
	10	0	100	0	0	100	0	0	100	0
	15	37	63	0	0	100	0	0	100	0
400	5	0	100	0	0	100	0	0	100	0
	10	0	100	0	0	100	0	0	100	0
	15	11	88	0	0	100	0	0	100	0

Table 6.3: Mean (standard deviation) of misclassification errors for simulated data.

Size	5%		10%		15%	
(n)	MCLUST	STM	MCLUST	STM	MCLUST	STM
50	0.42 (0.06)	0.06 (0.03)	0.43 (0.06)	0.16 (0.03)	0.42 (0.06)	0.26 (0.11)
100	0.43 (0.05)	0.05 (0.03)	0.42 (0.05)	0.12 (0.05)	0.44 (0.05)	0.21 (0.09)
200	0.44 (0.04)	0.05 (0.01)	0.45 (0.04)	0.11 (0.02)	0.46 (0.03)	0.17 (0.04)
400	0.33 (0.02)	0.05 (0.01)	0.33 (0.02)	0.11 (0.02)	0.45 (0.02)	0.15 (0.02)

6.7 Applications

This section presents the results of experiments which aim to demonstrate the efficiency and compare the STM model to existing methods. Both applications are executed via the R package, `SpaTimeClust`, using the default options. We defined $G_{\max} = 6$, $K_{\max} = 6$, and $Q_{\max} = 4$. The geographical coordinates of the region's capital are used as the spatial coordinates of each site j. Note that the data presented in this section are available in the package. By inhibiting $\lambda_{gk1} = \lambda_{gk2} = 0 \; \forall \; g, k$, from (6.9), we end up with only the time dependencies which is similar to functional data problem. Thus, we introduce the TM model as the STM model without the spatial dependencies.

6.7.1 Influenza Data

Data

We illustrate our newly proposed model on publicly available influenza data from the website of the Réseau Sentinelles (2015). The data consist of twenty-nine incidence rates of influenza-like illness (number of cases per 100 000 inhabitants), from 1986 to 2015, from over twenty-one regions of France between the first day of Summer until the last day of Spring (approximately 52 weeks). In summary, we have $n = 29$ periods described by the incidence rates of influenza for $T = 52$ weeks at $J = 21$ regions.

Model Comparison

According to Table 6.4, the selected STM drastically outperforms MCLUST model (EII), in terms of BIC, but very closely to TM model (with $\text{BIC}_{\text{TM}} - \text{BIC}_{\text{STM}} = -3$). Hence, the proposed model is more parsimonious and fits the data better than the classical Gaussian mixture. However, the spatial dependencies within classes are not strong, which make the TM model a suitable candidate. This finding illustrates the flexibility and efficiency of the proposed model as a multivariate functional data. Table 6.5 shows that both models produce exactly the same partition because the ARI computed between both resulting partitions is equal to one. This implies that the spatial dependencies within classes are weak which confirms that the TM model is as relevant as the STM model.

Table 6.4: Selected model and information criterion for the influenza dataset.

Models		Constraints		number of	
	G	K	Q	parameters	BIC
MCLUST (EII)	7	–	–	7651	-205655
TM	3	2	4	44	-163781
STM	3	2	4	50	-163778

Table 6.5: Confusion table of ARIs computed between the three resulting partitions that were applied to the influenza data.

	MCLUST	TM	STM
MCLUST	1.00	0.35	0.35
TM	–	1.00	1.00
STM	–	–	1.00

Best Model Interpretation

Table 6.6 presents variables that can possibly explain the choice of classes of the best model according to STM. For instance, the third component groups the periods characterized by the viruses of type A only and the incidence rate curves are distinctly elongated, see Figure 6.2. Its highest peak of the observed incidence rate reaches 1530 cases per 100 000 inhabitants while the lowest one is at 533 cases per 100 000 inhabitants, in the first component. Thus, the curves of the latter are lower with all periods affected by virus B. The second component represents periods affected by the AH3N3 virus and the duration lasted about ten weeks. It is interesting to point out that the third component occurred between 1986/1987 and 1995/1996 while all the observations of the second component happened after 1996.

Table 6.6: Justification of classes using other explanatory variables for influenza (number of cases per 100 000 inhabitants).

Variables	Component (g)		
	1	2	3
Incidence rate at highest peak	533.0	847.2	1530.0
Total incidence rate	2989.1	4592.0	6663.5
Duration (week)	8.8	10.3	9.0
Number of AH1N1	7	1	2
Number of AH3N3	5	8	2
Number of B	8	1	0

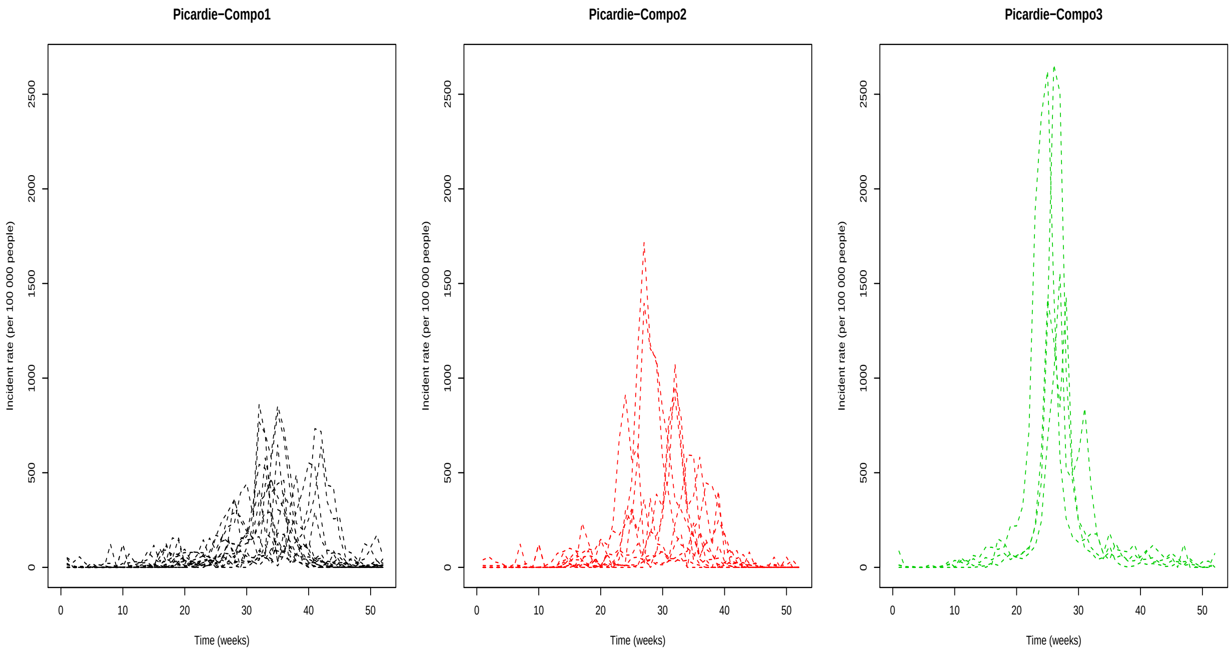


Figure 6.2: Weekly incidence rate (number of cases per 100 000 inhabitants) curves at Picardie under different components.

6.7.2 Air Pollution Data

Data

Airparif-Design Clair et Net (2010) is a non-profit organization accredited by the Ministry of Environment that promotes the air quality monitoring network in Paris and Île-de-France region, essentially the capital city region. The quantities of nitrogen dioxide (NO_2) is measured in μg/m^3, every hour, at nine Airparif stations located around the periphery of Paris for 101 days in 2014. From the previous information, we extracted: $J = 9$ stations that calculate the amount of NO_2, $T = 24$ hours per day, and $n = 101$ days. The nine stations constitute: Ivry-sur-Seine, Neuilly-sur-Seine, Pantin (RN2), Périphérice Est (abbreviated Periph-Est), place Basch (Paris

XVIeme), Porte d'Auteuil, Rue Eastmen (Paris XIIIeme), rue Flacon (Paris XVIIIeme), and Stade Lenglen. Consider the percentage of weekend days, days between April 1st and October 1st, holiday days in July and August, total rain, and average and maximum wind speed, as further explanatory variables that are extracted to reinforce the interpretation of our clustering results.

Model Comparison

From Table 6.7, we observe that the selected STM surpasses the TM and MCLUST (VEI) models regarding the BIC ($\text{BIC}_{\text{STM}} > \text{BIC}_{\text{TM}} > \text{BIC}_{\text{MCLUST}}$). The STM model is less complex compared to other models because it requires fewer parameters, facilitating the interpretation. Table 6.8 summarizes the ARI values between the pertinent model of each approach. We notice the disparity in the partitioning of classes between MCLUST and STM, i.e., ARI = 0.25, which suggests that modelling spatial and temporal helps. When spatial dependencies are accounted for, we notice the different partitioning of classes between TM and STM, i.e., ARI = 0.78.

Table 6.7: Selected model and information criterion for air pollution dataset.

Models		Constraints		number of	
	G	K	Q	parameters	BIC
MCLUST (VEI)	6	–	–	1522	-93241
TM	4	4	4	123	-79175
STM	3	4	4	110	-79006

Best Model Interpretation

Recall that the main goal is to group days presenting similarity in the curvature or characteristic. As observed in Table 6.7, the STM model divides the 101 days into

Table 6.8: Confusion table of ARIs computed between the three resulting partitions that were applied to the air pollution data.

	MCLUST	TM	STM
MCLUST	1.00	0.13	0.25
TM	–	1.00	0.78
STM	–	–	1.00

three components with a size of 58, 23 and 20. Table 6.9 compiles other relevant variables to rationalize the partitioning of classes suggested by the STM model while Figure 6.3 shows the behaviour of each component for three sites (Neuilly-sur-Seine, place Basch and Pantin). The first component highlights weekends (81.0%) and vacation periods (20.7%), as well as days with the most rain (0.154 cm) and the highest average wind speed (4.638 km/h). The low and constant quantity of NO_2 at three stations, see Figure 6.3, is largely attributable to all these factors combined together.

Table 6.9: Justification of classes using other explanatory variables for air pollution.

Variables	Component (g)		
	1	2	3
Weekend (%)	81.0	69.5	10.0
Total rain (cm)	0.154	0.147	0.111
Wind_{avg} (km/h)	4.638	2.652	3.400
Days between Apr 1 - Oct 1(%)	56.9	26.1	40.0
Spring holidays in July & Aug (%)	20.7	0.0	0.15
Wind_{max} (km/h)	11.155	11.652	13.050

Inversely, the second component contains no vacation days. This suggests that the roads are busier, which causes an elevation in the amount of NO_2. Because of the slow wind flow velocity (2.652 km/h), the quantity of NO_2 is accumulated throughout the day and reaches its maximum at 8 pm $(\underset{t=1,\ldots,24}{\arg\max}\mathbb{E}_{gjt})$, see also Table 6.10.

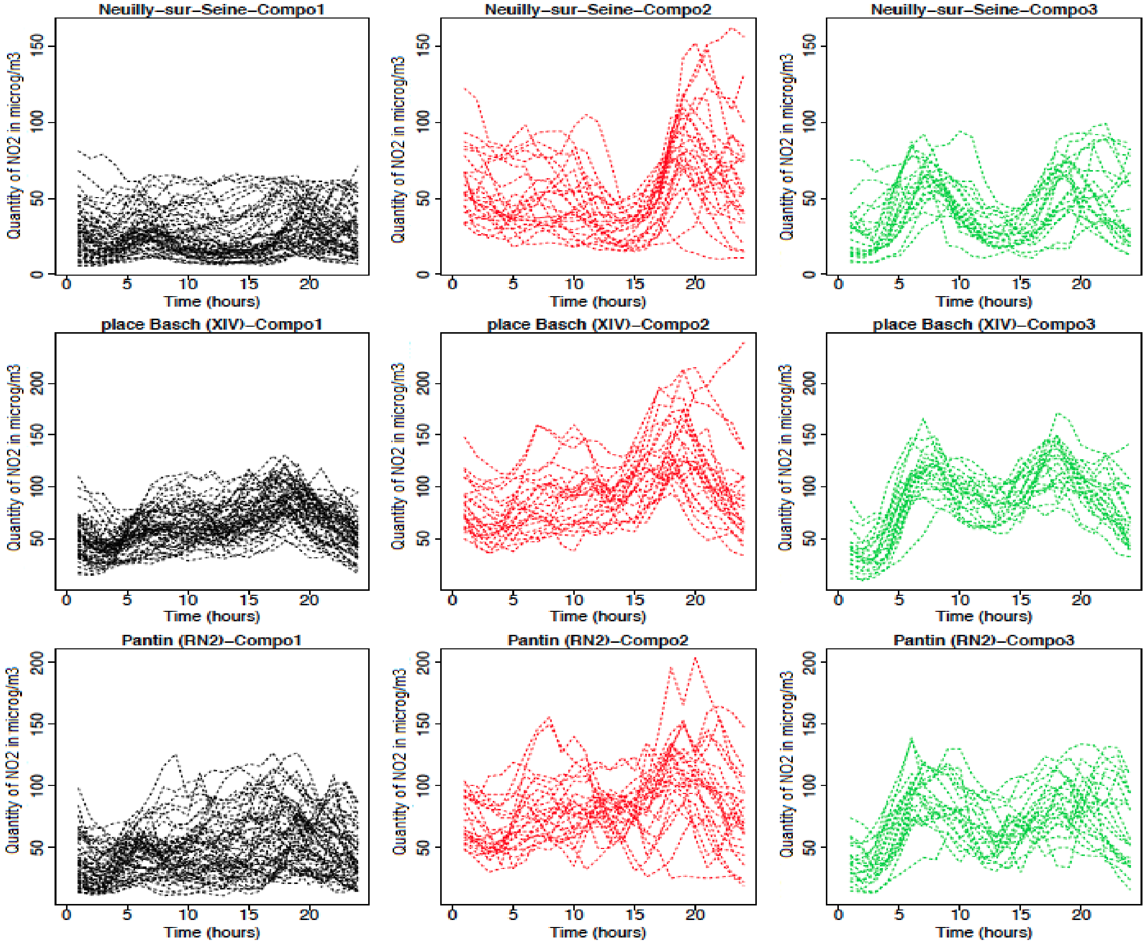


Figure 6.3: Hourly quantity of NO_2 ($\mu g/m^3$) measured at three different stations in the periphery of Paris, where each class member represents days with similar curvature or characteristics for 101 days observed.

The last component includes weekdays with little rain (0.111 cm) but strong wind (13.05 km/h). Although there was little rain, the wind carried NO_2 which explains the highest quantity of pollution is attained early in the day, i.e., 8 am. Because the third component principally describes weekdays, we observe two peaks, induced by rush hours, see Figure 6.3. This statement is witnessed in the averaged curves of the model for each component, obtained with (6.10) for $K = 4$, see Figure 6.4 (cf. green bold line). Hence, the STM model is flexible because of its ability to

Table 6.10: Class descriptions through parameters with respect to the time: $\mathbb{E}_{gjt} = \mathbb{E}\left[x_{ij1} \mid Z = g\right]$.

Site	$\mathbb{E}_{gjt}$	Component		
		$g=1$	$g=2$	$g=3$
Neuilly-sur-Seine	$\mathbb{E}_{gj1}$	23.32	63.65	32.14
	$\mathbb{E}_{gj24}$	19.83	67.87	51.91
	$\max\limits_{t=1,\ldots,24} \mathbb{E}_{gjt}$	27.27	96.93	76.05
	$\arg\max\limits_{t=1,\ldots,24} \mathbb{E}_{gjt}$	19	20	8
Ivry-sur-Seine	$\mathbb{E}_{gj1}$	24.16	59.93	32.99
	$\mathbb{E}_{gj24}$	24.33	51.35	51.33
	$\max\limits_{t=1,\ldots,24} \mathbb{E}_{gjt}$	30.76	80.26	74.71
	$\arg\max\limits_{t=1,\ldots,24} \mathbb{E}_{gjt}$	20	20	8
Pantin	$\mathbb{E}_{gj1}$	37.72	63.62	32.10
	$\mathbb{E}_{gj24}$	38.17	69.11	54.56
	$\max\limits_{t=1,\ldots,24} \mathbb{E}_{gjt}$	49.86	98.06	79.34
	$\arg\max\limits_{t=1,\ldots,24} \mathbb{E}_{gjt}$	20	20	8

replicate the curvature of the data. Figure 6.5 demonstrates the importance of spatial dependencies. We notice, notably for the first component, that the averaged curves of $x_{ijt} - x_{ij(t-1)}$ per hour for nine sites are distinct. For example, we observe that one particular station has a higher level of NO_2 than the rest. Note that the spatial information is modelled with respect to each component allowing us to obtain nine curves in each component instead of only one.

6.8 Discussion

This chapter builds on the growing trend towards clustering for spatio-temporal data by developing a flexible model. The proposed model can be regarded as an extension of ClustSeg, presented by Samé *et al.* (2011), with two additional features: an

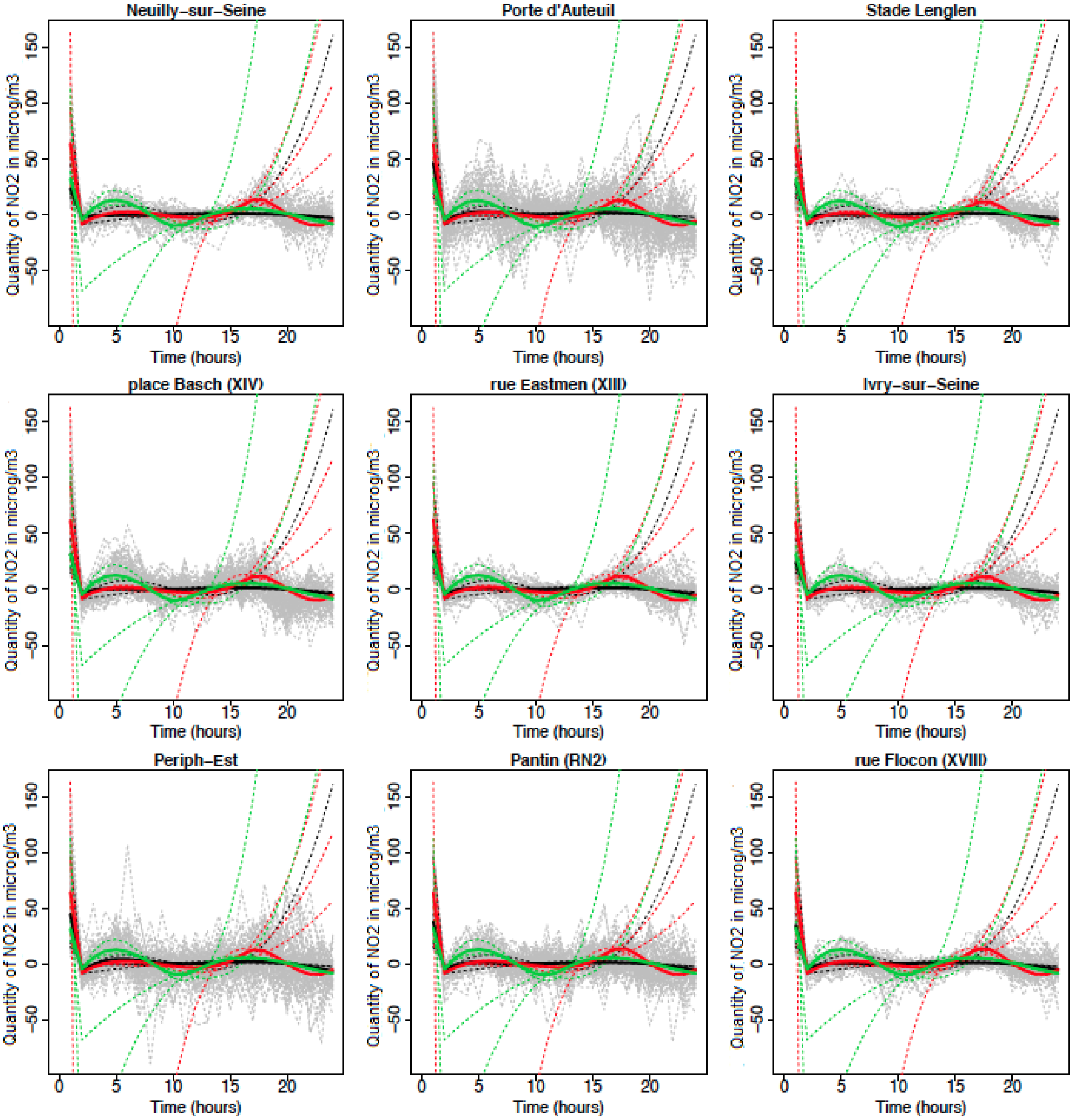


Figure 6.4: Hourly averaged curves (in bold) of the increasing quantity of NO_2 (μg/m^3), i.e., $x_{ijt} - x_{ij(t-1)}$, at nine different stations in the periphery of Paris under different components (in black for Component 1, red for Component 2 and green for Component 3). The dashed line represents each label assignment of the polynomial regression depending on the class membership. The grey curves represents the observations, i.e., 101 days. The station Périphérique Est is abbreviated as Periph-Est.

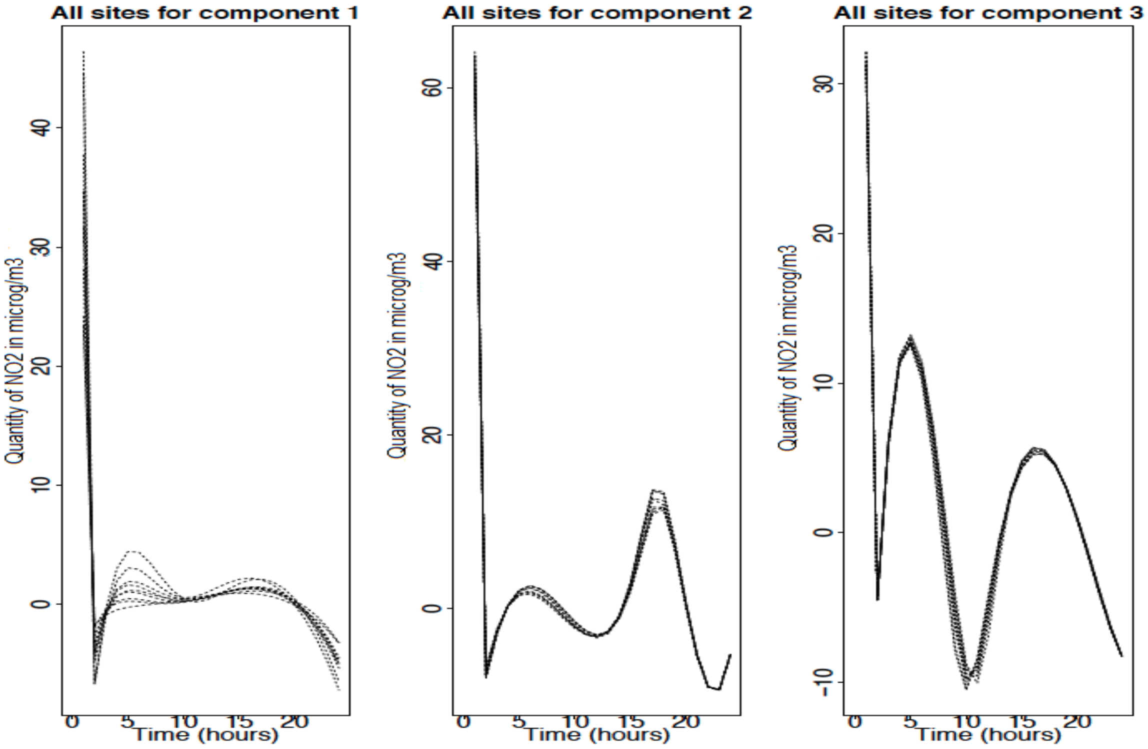


Figure 6.5: Hourly averaged curves of the increasing quantity of NO$_2$ (μg/m^3), i.e., $x_{ijt} - x_{ij(t-1)}$, for all nine stations in the periphery of Paris under their respective class member.

autoregressive regression and the possibility to operate as multivariate functional data. Hence, the TM model can also be outlined as a method to cluster multivariate functional data with an independence assumption within classes between different functions.

The STM model was illustrated through simulation study and applications. In the simulation study, we demonstrated the consistency of estimates and model along with the importance of dependencies. In the applications, we illustrated two cases: with and without spatial dependencies. When applied to the influenza dataset, our model without spatial dependencies, i.e., the TM model ($\lambda_{gk1} = \lambda_{gk2} = 0$), outperformed MCLUST. When applied to air pollution dataset, our model with the spatial dependencies, i.e., the STM model, was superior to MCLUST.

Chapter 7

Conclusions

7.1 Summary

Finite mixture models have been extensively used in the statistical literature, both as tools for modelling population heterogeneity and as a flexible method for relaxing parametric distributional assumptions. This thesis has focused on the development and implementation of two topics in modelling: receiver operating characteristics curve and model-based clustering for spatio-temporal data, using finite mixture distributions.

7.1.1 ROC curve

Most of the existing methods, especially ones focusing on the assumption of normality, require finding a single transformation that works for both the non-diseased and diseased groups, which remains the Achilles heel of the LABROC. In this thesis, we propose an alternative that explicitly recognizes the heterogeneity in both groups and at the same time allows the use of familiar parametric models. Thus, modelling the heterogeneity by means of mixture models is highly reasonable and constitutes the basic tool for cluster, latent class, and discriminant analysis, but few use it in ROC analysis.

The proposed method employs Gaussian and non-Gaussian (t and skew t distributions) mixtures for both populations, in conjunction with the Monte Carlo method. The ROC curve estimates produced by the proposed method is smooth, respects the monotonicity property and does not require any transformation of the data. These methods were then applied to data on pancreatic cancer of two biomarkers (CA 125 and CA 19-9). Each of these models performs favourably or equivalently, depending whether the data are normal, skewed or heavier tails, when compared to the existing well-establish LABROC.

7.1.2 Model-Based Clustering

Clustering is a very interesting and challenging research problem. Therefore, a wide spectrum of methodologies has been used to address these questions. Probabilistic mixture modelling is a well established model-based approach for clustering because it offers many advantages such as flexibility and selecting the number of components.

The main feature of the novel model is the incorporation of a spatial dimension and an autoregressive component. The proposed model, called STM, is a mixture model, where each component is itself a mixture, more precisely an autoregressive polynomial regression mixture in which the logistic weights depend on the spatial and temporal dimensions. As mentioned in Remark 6.1.4, by removing the spatial dependencies inside the model specification, we end up with only the time dependencies which is similar to the functional data problem. Moreover, the STM model does not require any transformation of the data or the selection of functional basis beforehand. This new model was tested using numerical experiments with both simulated and real datasets, i.e., influenza and air pollution. These experiments demonstrated the consistency of the model and estimates. Our model yields better results since it considers both the temporal and spatial dimensions. Furthermore, our model has proved to be multifunctional and versatile because it can be used as a functional data model shown in the example of influenza, or as a spatio-temporal model like in the air pollution example. In both cases, the interpretation of the results is very straightforward.

7.2 Future Work

7.2.1 Modelling the ROC Curve with No-Gold Standard

In evaluation of diagnostic accuracy of binary tests, the knowledge of true case status or a gold standard from a perfect test, i.e., zero error rates, is required. However, in practice, such information is not always available because it may be difficult or even impossible to determine the true status, and even the available reference test against which new tests are compared is subject to errors. For example, the diagnosis of Alzheimer's disease is made based on certain symptoms, but the diagnosis is not definitive until the brain tissue has been examined after death. Hence, the absence of a gold standard adds more complexity to the evaluation of new tests. Thus, we can incorporate Gaussian mixture models into the concept of no-gold standard. Let

$\mathbf{S}_i = (S_1, \ldots, S_n)$ be a vector of scores for each patient. The marginal distribution of $\mathbf{S}_i$ is defined as

$$P(\mathbf{S}_i) = \sum_{d=0}^{1} P(\mathbf{S}_i \mid D_i = d) P(D_i = d), \tag{7.1}$$

where D_i is the latent variable for the unknown disease status of the ith patient and $D_i = 1$, if patient i belongs to the disease group. Then, the joint distribution of $\mathbf{S}_i$, $P(\mathbf{S}_i \mid D_i = d)$, follows a Gaussian or non-Gaussian mixture.

7.2.2 Model Selection

The problem of selecting a statistical model with the correct complexity is fundamental in statistical modelling. Often one has to tradeoff between accurately fitting the data and the ability of the model to generalize. Sometimes BIC can cause an underfitting problem. There is work to be done on the alternatives to the BIC such as the Akaike information criterion, the integrated completed likelihood or the cross-validation, for model selection.

7.2.3 Parsimonious Spatio-Temporal Mixture Models

A better tradeoff between the bias and the variance of the estimates can be obtained by imposing constraints on the parameter space. Thus, parsimonious models are often defined for mixture models by imposing equality constraints on the parameters among components (Banfield and Raftery, 1993; Celeux and Govaert, 1991; McNicholas and Murphy, 2008), as briefly outlined in Section 5.5 for MCLUST. We propose four types of constraints:

(i) equality on the mixture weights among components;

(ii) equality of the regression weights among components;

(iii) equality of the dispersion parameters within components;

(iv) equality of the dispersion parameters among components.

The full range of possible constraints also provides a class of twelve different parsimonious spatio-temporal mixture models, described in Table 7.1. The set of the parsimonious constraints are defined by $H \in \mathcal{F}$, where $\mathcal{F} = \{\text{U,C}\}^4 \setminus \text{UUUC} \cup \text{UCUC} \cup \text{CUUC} \cup \text{CCUC}$. The four elements of H respectively indicate the presence (C) or the absence (U) of equality constraints for the mixture weights, the regression weights, the dispersion within and among components.

Table 7.1: Twelve parsimonious versions of the spatio-temporal mixture models: U indicates unconstrained parameters and C indicates an equality constraint.

π_g	$\boldsymbol{\lambda}_g$	σ^2_{gk}		Number of
		Within component	Among component	parameters
U	U	U	U	$GK(Q+6)-3G-1$
U	U	C	U	$GK(Q+5)-2G-1$
U	U	C	C	$GK(Q+5)-3G$
U	C	U	U	$GK(Q+2)+G+4K-5$
U	C	C	U	$GK(Q+1)+2G+4K-5$
U	C	C	C	$GK(Q+1)+G+4K-4$
C	U	U	U	$GK(Q+6)-4G$
C	U	C	U	$GK(Q+5)-3G$
C	U	C	C	$GK(Q+5)-4G+1$
C	C	U	U	$GK(Q+2)+4(K-1)$
C	C	C	U	$GK(Q+1)+G+4(K-1)$
C	C	C	C	$GK(Q+1)+4(K-1)+1$

For each model, the number of parameters depends neither on the number of sites nor the size of the time grid. Thus, the proposed modelling implies a reasonable number of parameters even if J and T are large. However, one should expect the computation time to increase since we are dealing with more models.

7.2.4 Generalization of the STM Model

As shown and discussed in Chapter 6, the STM model is flexible and easy to handle, i.e., putting a constraint on λ_g (see Remark 6.1.4). We can generalize the STM model by replacing the polynomial regression with other functional models. For instance, Fourier basis are excellent for describing data that are periodic (e.g., annual weather data), the spline and wavelet bases are terrific at fitting highly curvy data. Thus, the generalized STM model will automatically determine which functional basis is the best candidate with respect to the data using model selection such as BIC. Therefore, we are not restricted to a certain types of data.

Appendix

Proof of Corollary 6.2.1

Before starting the proof, let us state a useful theorem borrowing from Casella and Berger (2002).

Theorem A.1. *If X and Y are any two random variables, then*

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]], \tag{.1}$$

and

$$Var[X] = \mathbb{E}[Var[X \mid Y]] + Var[\mathbb{E}[X \mid Y]], \tag{.2}$$

provided that the expectation exists.

The proof of this theorem can be found in Casella and Berger (2002). Recall Lemma 6.2.1.

Lemma 6.2.1. *Suppose $U_1 \sim \mathcal{N}(\mu_1, \sigma^2)$ and $U_2 \mid U_1 \sim \mathcal{N}(\mu_2 - \mu_1 + u_1, \sigma^2)$, then we have $U_2 \sim \mathcal{N}(\mu_2, 2\sigma^2)$.*

Proof. The joint probability function for two random variables U_1 and U_2 is denoted by

$$f_{U_1,U_2}(u_1, u_2) = \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\Delta\right\}, \tag{.3}$$

where

$$\begin{aligned}
\Delta &= (u_1 - \mu_1)^2 + (u_2 - \mu_2 + \mu_1 - u_1)^2 \\
&= \left(u_1^2 - 2u_1\mu_1 + \mu_1^2\right) + \left[(u_2 + \mu_1 - \mu_2)^2 - 2u_1(u_2 + \mu_1 - \mu_2) + u_1^2\right] \\
&= \underbrace{2\left(u_1 - \frac{(u_2 + 2\mu_1 - \mu_2)}{2}\right)^2}_{h(u_1 \mid u_2)} + \underbrace{\frac{(u_2 + 2\mu_1 - \mu_2)^2}{2} + \mu_1^2}_{g(u_2)}. \qquad (.4)
\end{aligned}$$

Thus, the marginal probability function of U_2 is

$$\begin{aligned}
f(u_2) &= \int f(u_2 \mid u_1) f(u_1) du_1 \\
&= \int f(u_1, u_2) du_1 \\
&= \frac{1}{2\pi\sigma^2} \int \exp\left\{\frac{-1}{2\sigma^2}\left(h(u_1 \mid u_2) + g(u_2)\right)\right\} du_1 \\
&= \frac{\exp\left\{\frac{-1}{2\sigma^2} g(u_2)\right\}}{\sqrt{2\pi\sigma^2}} \underbrace{\int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2} h(u_1 \mid u_2)\right\} du_1}_{= 1 \text{ since it follows } \mathcal{N}\left(\frac{(u_2+2\mu_1-\mu_2)}{2}, \sigma^2\right)} \\
&= \frac{\exp\left\{\frac{-1}{2\sigma^2} g(u_2)\right\}}{\sqrt{2\pi\sigma^2}}. \qquad (.5)
\end{aligned}$$

By Theorem A.1, the mean is

$$\begin{aligned}\mathbb{E}[U_2] &= \mathbb{E}[\mathbb{E}[U_2 \mid U_1]] && (.6)\\ &= \mu_2 - \mu_1 + \mathbb{E}[u_1] && (.7)\\ &= \mu_2\end{aligned}$$

and the variance is

$$\begin{aligned}\mathrm{Var}[U_2] &= \mathbb{E}[\mathrm{Var}[U_2 \mid U_1]] + \mathrm{Var}[\mathbb{E}[U_2 \mid U_1]] && (.8)\\ &= \mathbb{E}[\sigma^2] + \mathrm{Var}[\mu_2 - \mu_1 + u_1] && (.9)\\ &= 2\sigma^2.\end{aligned}$$

□

Bibliography

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (2005). Automatic subspace clustering of high dimensional data. *Data Mining and Knowledge Discovery*, **11**, 5–33.

Airparif-Design Clair et Net (2010). Airparif, association de surveillance de la qualité de l'air. *http://www.airparif.asso.fr/*. Accessed March 2016.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Andrews, J. L. and McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing*, **22**, 1021–1029.

Andrews, J. L. and McNicholas, P. D. (2015). teigen: Model-based clustering and classification with the multivariate t distribution. R package version 2.1.0.

Appice, A., Ciampi, A., Malerba, D., and Guccione, P. (2013). Using trend clusters for spatiotemporal interpolation of missing data in a sensor network. *Journal of Spatial Information Science*, **6**, 119–153.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, **12**, 171–178.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 719–725.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, **41**, 561–575.

Birant, D. and Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, **60**, 208–221.

Browne, R. P. and McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, **43**, 176–198.

Cai, T. and Moskowitz, C. S. (2004). Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. *Biostatistics*, **5**, 573–586.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury, 2nd edition.

Celeux, G. and Govaert, G. (1991). Clustering criteria for discrete data and latent class models. *Journal of Classification*, **8**, 157–176.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**, 781–793.

Dasgupta, A. P. and Raftery, A. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, **93**, 294–302.

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*, **39**, 1–38.

Dorfman, D. D. and Alf, E. J. (1968). Maximum likelihood estimation of parameters of signal detection theory - a direct solution. *Psychometrika*, **33**, 117–124.

Du, P. and Tang, L. (2009). Transformation-invariant and nonparametric monotone smooth estimation of ROC curves. *Statistics in Medicine*, **28**, 349–359.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863–14868.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. John Wiley & Sons, Inc., United Kingdom, 5th edition.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer Series in Statistics, New York.

Figueiredo, M. A. T. and Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 381–396.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611–631.

Fraley, C., Raftery, A. E., and Scrucca, L. (2012). mclust: Normal mixture modeling for model-based clustering, classification, and density estimation. R package version 4.0.

Franczak, B. C., Browne, R. P., and McNicholas, P. D. (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**, 1149–1157.

Gaudart, J., Poudiougou, B., Dicko, A., Ranque, S., Toure, O., Sagara, I., Diallo, M., Diawara, S., O. A., Diakite, M., and Doumbo, O. K. (2006). Space-time clustering of childhood malaria at the household level: A dynamic cohort in a Mali village. *BMC Public Health*, **6**, 1–13.

Goddard, M. J. and Hinberg, I. (1990). Receiver operator characteristic (ROC) curves and non-normal data: An empirical study. *Statistics in Medicine*, **9**, 325–377.

Gönen, M. (2013). Mixtures of receiver operating characteristic curves. *Academic Radiology*, **20**, 831–837.

Graham, C. and Talay, D. (2013). *Stochastic Simulation and Monte Carlo Methods: Mathematical Foundations of Stochastic Simulation*. Springer-Verlag Berlin Heidelberg.

Hall, P. G. and Hyndman, R. J. (2003). Improved methods for bandwidth selection when estimating ROC curves. *Statistics & Probability Letters*, **64**, 181–189.

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc., New York.

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, **64**, 53–62.

Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, **24**, 25–40.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**, 193–218.

Jacques, J. and Preda, C. (2014a). Functional data clustering: A survey. *Advances in Data Analysis Classification*, **8**, 231–255.

Jacques, J. and Preda, C. (2014b). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, **71**, 92–106.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, **31**, 264–323.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, **98**, 397–408.

Jang, W. and Hendry, M. (2007). Cluster analysis of massive datasets in astronomy. *Statistics and Computing*, **17**, 253–262.

Keane, M. and Wolpin, K. (1997). The career decisions of young men. *Journal of Political Economy*, **105**, 473–522.

Kisilevich, S., Mansmann, F., Nanni, M., and Rinzivillo, S. (2010). Spatio-temporal clustering: A survey. *Data Mining and Knowledge Discovery Handbook*, pages 1–22.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge University Press.

Lee, S. X. and McLachlan, G. J. (2013a). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, **24**, 181–202.

Lee, S. X. and McLachlan, G. J. (2013b). On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification*, **7**, 241–266.

Lin, T. I., Lee, J. C., and Yen, S. Y. (2007). Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, **17**, 909–927.

Lloyd, C. J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, **93**, 1356–1364.

López-de Ullibarri, I., Cao, R., Cadarso-Suárez, C., and Lado, M. J. (2008). Nonparametric estimation of conditional ROC curves: Application to discrimination tasks in computerized detection of early breast cancer. *Computational Statistics and Data Analysis*, **52**, 2623–2631.

Lusted, L. B. (1971). Signal detectability and medical decision-making. *Science*, **171**, 1217–1219.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, California. University of California Press.

Madison, J. and Vermunt, J. (2004). *Latent Class Models*. Handbook of quantitative methodology for the social sciences, Newbury Park, CA.

McLachlan, G. J. and Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

McLachlan, G. J. and Peel, D. (1998). *Robust Cluster Analysis via Mixtures of Multivariate t-Distributions*, volume 1451 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley, New York.

McNicholas, P. D. (2016a). *Mixture Model-Based Classification*. Boca Raton FL: Chapman & Hall/CRC Press.

McNicholas, P. D. (2016b). Model-based clustering. *Journal of Classification*, **33**.

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**, 285–296.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267–278.

Metz, C. E. (1989). Some pratical issues of experimental design and data analysis in radiological ROC studies. *Investigative Radiology*, **24**, 234–245.

Metz, C. E., Herman, B. A., and Shen, J.-H. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, **17**, 1033–1053.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, **8**, 343–366.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer series in operations research and financial engineering. Springer, New York, NY, 2nd edition.

Pearson, K. (1894). Contributions to the theory of mathematical evolution. *Philosophical Transaction of the Royal Society of London*, **185**, 71–110.

Peng, L. and Zhou, X.-H. (2004). Local linear smoothing of receiver operating characteristic (ROC) curves. *Journal of Statistical Planning and Inference*, **118**, 129–143.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T. I., Maier, L. M., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D. A., De Jager, P. L., and Mesirow, J. P. (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 8519–8524.

Qiu, P. and Le, C. (2001). ROC curve estimation based on local smoothing. *Journal of Statistical Computation and Simulation*, **70**, 55–69.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics, New York.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.

Ren, H., Zhou, X.-H., and Liang, H. (2004). A flexible method for estimating the ROC curve. *Journal of Applied Statistics*, **31**, 773–784.

Réseau Sentinelles, I. (2015). *http://www.sentiweb.fr*. Accessed March 2016.

Samé, A., Chamroukhi, F., Govert, G., and Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis Classification*, **5**, 301–321.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Silvapulle, M. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **43**, 310–313.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**, 795–809.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin*, **99**, 100–117.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, **34**, 1265–1269.

Teicher, H. (1967). Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, **38**, 1300–1302.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Tyron, R. C. and Bailey, D. E. (1970). *Cluster Analysis*. McGraw-Hill, New York.

Vardi, Y., Sheep, L. A., and Kaufman, L. A. (1985). A statistical model for positron emission tomography. *Journal of the American Statistical Association*, **80**, 8–37.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York.

Vrbik, I. and McNicholas, P. D. (2012). A fully analytic approach for multivariate skew-t mixture models. *Statistics and Probability Letters*, **82**, 1169–1174.

Wang, K., McLachlan, G. J., Ng, S. K., and Peel, D. (2009). EMMIX-skew: EM algorithm for mixture of multivariate skew normal/t distributions. R package version 1.0-12.

Wieand, S., Gail, M. H., James, B. R., and James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, **76**, 585–592.

Wolfe, J. H. (1963). *Object cluster analysis of social areas*. Master's thesis, University of California, Berkeley.

Wu, C. F. J. (1983). On convergence properties of the EM algorithm. *The Annals of Statistics*, **11**, 95–103.

Wu, X., Zurita-Miller, R., and Kraak, M.-J. (2015). Co-clustering geo-referred time series: Exploring spatio-temporal patterns in Dutch temperature data. *International Journal of Geographical Information Science*, **29**, 624–642.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, **39**, 209–214.

Zhong, S. and Gosh, J. (2003). A unified framework for model-based clustering. *Journal of Machine Learning Research*, **4**, 1001–1037.

Zhou, X.-H. and Harezlak, J. (2002). Comparison of bandwidth selection methods for kernel smoothing of ROC curves. *Statistics in Medicine*, **21**, 2045–2055.

Zhou, X.-H. and Lin, H. (2008). Semi-parametric maximum likelihood estimates for ROC curves of continuous-scale tests. *Statistics in Medicine*, **27**, 5271–5290.