

A spatial variant of the Gaussian mixture of regressions model

by

Marion Delport

Submitted in partial fulfilment of the requirements for the degree

MSc (Mathematical Statistics)

In the Faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

30 October 2017



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Denkleiers • Leading Minds • Dilkgopolo tša Dihalefi

Declaration

I, *Marion Delpont* declare that this dissertation, which I hereby submit for the degree *MSc (Mathematical Statistics)* at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature:

Date: 30-10-2017

Acknowledgements

I would like to thank my supervisors Dr. Frans H. J. Kanfer and Mr. Sollie M. Millard. Thank you, for your patience, enthusiasm, invaluable insight and for providing your continued support. I am also grateful to the Department of Statistics at the University of Pretoria for this research opportunity.

Thank you, to Mrs. Rona Beukes at the Department of Agriculture, Forestry and Fisheries (DAFF) and Mrs. Wiltrud Durandt at the Agricultural Research Council (ARC) who provided the objective yield database. I would also like to acknowledge the Goddard Earth Sciences Data and Information Services Center (GES DISC) for disseminating the Tropical Rainfall Measurement Mission (TRMM) database.

I would like to express my deepest gratitude to my employer, the Bureau for Food and Agricultural Policy (BFAP) and colleagues, who believe in the value of my work. I look forward to the exciting work ahead and to continue to make a difference as part of this team.

To my family and friends, I would like to express my sincerest gratitude for your never-ending support, understanding and for believing in me. A special thanks to my parents Marianne Muhl, Johan Delpont and Elma Delpont for your encouragement, love and prayers.

To my loving husband, Lourens Delpont, my deepest gratitude for all your support, patience and love and that you were always there to cheer me on. Thank you, for bringing out the best in me.

Above all, I would like to thank my heavenly Father, without who none of this would have been possible.

May the God of hope fill you with all joy and peace

in believing, so that by the power of the Holy Spirit

you may abound in hope.

Romans 15:13

Abstract

In this study the finite mixture of multivariate Gaussian distributions is discussed in detail including the derivation of maximum likelihood estimators, a discussion on identifiability of mixture components as well as a discussion on the singularities typically occurring during the estimation process. Examples demonstrate the application of the finite mixture of univariate and bivariate Gaussian distributions.

The finite mixture of multivariate Gaussian regressions is discussed including the derivation of maximum likelihood estimators. An example is used to demonstrate the application of the mixture of regressions model. Two methods of calculating the coefficient of determination for measuring model performance are introduced.

The application of finite mixtures of Gaussian distributions and regressions to image segmentation problems is examined. The traditional finite mixture models however, have a shortcoming in that commonality of location of observations (pixels) is not taken into account when clustering the data. In literature, this shortcoming is addressed by including a Markov random field prior for the mixing probabilities and the present study discusses this theoretical development. The resulting finite spatial variant mixture of Gaussian regressions model is defined and its application is demonstrated in a simulated example. It was found that the spatial variant mixture of Gaussian regressions delivered accurate spatial clustering results and simultaneously accurately estimated the component model parameters.

This study contributes an application of the spatial variant mixture of Gaussian regressions model in the agricultural context: maize yields in the Free State are modelled as a function of precipitation, type of maize and season; GPS coordinates linked to the observations provide the location information. A simple linear regression and traditional mixture of Gaussian regressions model were fitted for comparative purposes and the latter identified three distinct clusters without accounting for location information. It was found that the application of the spatial variant mixture of regressions model resulted in spatially distinct and informative clusters, especially with respect to the type of maize covariate. However, the estimated component regression models for this data set were quite similar. The investigated data set was not perfectly suited for the spatial variant mixture of regressions model application and possible solutions were proposed to improve the model results in future studies. A key learning from the present study is that the effectiveness of the spatial variant mixture of regressions model is dependent on the clear and distinguishable spatial dependencies in the underlying data set when it is applied to map-type data.

Contents

1	Introduction	11
2	Finite mixture of multivariate Gaussian distributions	13
2.1	The finite mixture of multivariate Gaussian distributions	13
2.2	Estimation of Θ using the EM algorithm	15
2.2.1	Deriving the log-likelihood function and the expectation step	16
2.2.2	The maximisation step	19
2.3	Identifiability of mixture distribution components	21
2.4	Example - mixture of univariate Gaussian distributions	23
2.5	Singularity problem during estimation	26
2.5.1	EM Algorithm difficulties	26
2.5.2	Unboundedness of the mixture likelihood function	27
2.5.3	Dealing with sources of unboundedness	29
2.6	Example - mixture of multivariate Gaussian distributions	29
3	Finite mixture of multivariate regressions	33
3.1	The finite mixture of multivariate Gaussian regressions	33
3.2	Estimation of Θ using the EM algorithm	35
3.2.1	Deriving the log-likelihood function and the expectation step	35
3.2.2	The maximisation step of the EM algorithm	36
3.3	Example - mixture of univariate Gaussian regressions	38
4	The spatial variant mixture of regressions model	43
4.1	Traditional mixture of regressions model applied in an image context	44
4.2	Spatial variant mixture of regressions model	45
4.2.1	Background theory	46
4.2.2	Model specification	48
4.3	Example - spatial variant mixture of Gaussian regressions	53
4.3.1	The generated data	54
4.3.2	The model to be fitted	56
4.3.3	The estimated model	58

<i>CONTENTS</i>	6
5 Application: maize yields	63
5.1 Types of crop models and this study	64
5.2 The data	65
5.3 Exploratory analysis	66
5.3.1 Analysing the data and selecting the appropriate covariates	66
5.3.2 Simple linear regression	74
5.4 Applying the spatial variant mixture of regressions model	77
5.4.1 Traditional mixture of regressions model	78
5.4.2 Spatial variant mixture of regressions model	82
6 Conclusion and future work	89
A Results for deriving the MLE's	91
A.1 Lemma 1	91
A.2 Lemma 2 - Maximum likelihood estimators of the multivariate Gaussian distribution	92
B Standard statistical results used throughout this document	95
Bibliography	96

List of Figures

2.4.1	Overlaid histograms of simulated data: component 1 and 2	24
2.4.2	Log-likelihood function value in each EM step	25
2.4.3	Mixture distribution with responsibilities for each component	26
2.5.1	Log-likelihood of Kiefer-Wolfowitz example	28
2.5.2	Log-likelihood against σ_2	28
2.6.1	3D-Scatterplot of simulated data (\mathbf{Y}_1 left; \mathbf{Y}_2 right)	30
2.6.2	Mixture distribution of two-component bivariate mixture	31
2.6.3	Graphical representation of classification result	32
3.3.1	Simulated mixture of regressions data	39
3.3.2	Simple linear regression model fit	40
3.3.3	Mixture of regressions clustering result	41
4.2.1	Example of a Markov random field [3]	47
4.2.2	Markov random field illustration for the j^{th} cluster	49
4.3.1	Designed 16×16 image	54
4.3.2	Pixel changes over time - 3 models	55
4.3.3	Illustrating the clustering results	62
5.3.1	All yield observation by type of maize	68
5.3.2	Number of farming units per years of involvement	69
5.3.3	Histograms of total rainfall by season	70
5.3.4	Scatter plot of season rainfall to yield	71
5.3.5	Box plot of observed yields before (top) and after (bottom) outliers were removed	73
5.3.6	Observed yields	73
5.3.7	Observed yields without outliers, coloured by season	74
5.3.8	Observed vs predicted yield	76
5.3.9	Observed vs predicted yields by rainfall - simple linear regression	76
5.3.10	Residuals plotted against predicted values	77
5.4.1	Map-view of yield observations coloured according to clusters	80
5.4.2	Observed yield by rainfall coloured according to clusters	81
5.4.3	Residuals against predicted values, by cluster	82

5.4.4 Estimated mixing probabilities ($K = 3$) 84

5.4.5 Residuals against predicted values- spatial variant mixture of regressions model 85

5.4.6 Map-view of clustering result - spatial variant mixture of regressions model 86

5.4.7 Observed yield by rainfall - spatial variant mixture of regressions model 87

List of Tables

2.6.1	Parameter estimates of two component bivariate mixture of Gaussian distributions . . .	30
3.3.1	Estimated mixture of regression parameters	40
4.3.1	Simulated image with three clusters over time	56
4.3.2	Estimated responsibilities	60
4.3.3	Estimated mixing probabilities	61
5.3.1	Frequency table: objective yield data base 2004 - 2017	66
5.3.2	Number of objective yield observations per season	67
5.3.3	Frequency table: objective yield survey for harvesting season 2017	67
5.3.4	Scatter plots of monthly rainfall to yield	72
5.3.5	Variable definitions	75
B.0.1	Standard statistical results	95

List of Algorithms

2.1	EM algorithm for a K -component mixture of multivariate Gaussian distributions	21
3.1	EM algorithm for a K -component mixture of multivariate Gaussian regressions model .	38
4.1	MAP EM algorithm for the K -component spatial variant mixture of Gaussian regressions model	53

Chapter 1

Introduction

The fields of econometrics, chemometrics, biology and engineering have presented problems involving heterogeneous covariate dependent populations and have been studied in literature at the classification level (identifying the homogeneous sub populations) and the inference level (estimating the corresponding models) [33]. While many other clustering methods like k -means and hierarchical clustering provide assignments of observations to respective groups, only the mixture models calculate the classification probabilities and therefore provide very useful additional information [34].

Pearson, in 1894, was first to define the mixture of Gaussian distributions and he asserted that the analytical implications for only a two-component mixture of Gaussian distributions, rendered the application of the general theory to a real numerical example, highly unlikely [41]. More than a century later, a myriad of applications (some of which are mentioned and discussed throughout this document) can be found, many times more complex and challenging than the early definition given by Pearson. In this study the finite mixture of multivariate Gaussian distributions is discussed and maximum likelihood estimators are derived to be used in the EM algorithm. Common problems associated with the finite mixture of Gaussian distributions: identifiability of components of a mixture distribution [50] as well as the singularity problem encountered during estimation [35], are discussed and accompanied by relevant literature summaries and examples.

The finite mixture of multivariate Gaussian regressions model is examined and maximum likelihood estimators are derived. The application of the two-component mixture of regressions model is demonstrated with the help of an example and two methods by which the coefficient of determination (R^2) can be calculated for a finite mixture of Gaussian regressions model are proposed.

Finite mixture models have been successfully applied in the image segmentation¹ context including problems in the field of bioinformatics [4], the image retrieval context [21], MRI image segmentation [40] and aerial and satellite image segmentation [42]. Classical finite mixture models are not only rigorous measures for clustering performance but also assign each observation to the component that most likely generated it, based on a mixing probability. However, classical finite mixture models have

¹Image segmentation is the process that groups image pixels together based on attributes such as their intensity and spatial location [5]

shortcomings in that shared location information is not taken into account when grouping data [40]. That is, in an image segmentation context, apart from pixel intensity values the pixel (observation) location should also inform which cluster each pixel belongs to. In order to deal with the location information in image segmentation problems, an approach incorporating a Markov random field was introduced [10] with applications including series of satellite images [17], segmentation of brain MR images and mammographic images [57, 49].

Consequently, the multivariate mixture of Gaussian regressions theory is discussed in the image segmentation context application. Some background on the Markov random field theory is given and the resulting spatial variant mixture of Gaussian regressions model is defined and applied to a simulated example. The ability of the spatial variant mixture of regressions model to deliver spatially explicit clusters and simultaneously accurately estimate corresponding regression parameters is demonstrated.

The present study contributes the application of the spatial variant mixture of Gaussian regressions model to the agricultural context. Maize yield data in the Free State is modelled as a function of, precipitation, type of maize and season (time). The traditional mixture of regressions model is also fitted for comparative purposes.

Chapter 2

Finite mixture of multivariate Gaussian distributions

2.1 The finite mixture of multivariate Gaussian distributions

While many other clustering methods like k -means and hierarchical clustering provide assignments of observations to respective groups, only the mixture models calculate the classification probabilities [34] and therefore provide very useful additional information (i.e. the probability of each observation to belong to the different clusters).

The finite multivariate mixture of distributions presupposes a multivariate random variable \mathbf{Y} (a $(p \times 1)$ dimensional vector), for which the moments characterising its distribution need to be estimated, given an observed sample: $\{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_N\}$. The finite mixture of distributions poses that K independent data generating distributions of the same family (e.g. Gaussian), were involved in generating the observed sample of \mathbf{Y} . The finite mixture of distributions methodology provides a model structure through which the clustering of \mathbf{Y} into the K data generating distributions, and the estimation of the moments characterising those distributions is achieved simultaneously. The following discussion will be limited to the finite mixture of Gaussian distributions.

A finite mixture of multivariate Gaussian distributions model assumes that a set of K independent Gaussian distributions characterised by the first and second moments: $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \dots, (\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K)$ describe heterogeneous subgroups in the variable or feature of interest, \mathbf{Y} [20]. Furthermore, for each observation of \mathbf{Y} a hidden random indicator variable S , taking on an integer between 1 and K , chooses one of the K distributions to generate \mathbf{Y} . The indicator variable S follows an unknown discrete probability distribution $\Pi = (\pi_1, \pi_2, \dots, \pi_K)$ where $\sum_{j=1}^K \pi_j = 1$ for all i , and is assumed to be mutually

independent over observations $i = 1, 2, \dots, N^1$. In the simplest case, we have no prior information about Π and the underlying K distributions are given below, with each \mathbf{Y} observation being generated by one of the K multivariate Gaussian distributions with probability π_j , $j = 1, 2, \dots, K$:

$$\begin{aligned} \mathbf{Y}_1 &\sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \text{ with probability } \pi_1 \\ \mathbf{Y}_2 &\sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \text{ with probability } \pi_2 \\ &\vdots \\ \mathbf{Y}_K &\sim N(\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K), \text{ with probability } \pi_K \end{aligned}$$

where $\pi_K = 1 - \sum_{j=1}^{K-1} \pi_j$. The variable \mathbf{Y} can be written in, what is called the explicit *generative representation* [26].

$$\mathbf{Y} = \Delta_1 \mathbf{Y}_1 + \Delta_2 \mathbf{Y}_2 + \dots + \Delta_K \mathbf{Y}_K \quad (2.1.1)$$

where

$$\begin{aligned} \Delta_j &= \begin{cases} 1 & \text{with probability } \pi_j \\ 0 & \text{otherwise} \end{cases} \\ \sum_{j=1}^K \Delta_j &= 1 \text{ for all } i. \end{aligned}$$

Note that the Δ_j 's $\in \{0, 1\}$ for $j = 1, 2, \dots, K$ contain the same information as is contained in S defined above and the two definitions can be used interchangeably so that $\Delta_j = I(S = j)$.

The generative representation is explicit in the sense that during a mixture model simulation the Δ_j 's are generated and dependent on the outcomes, deliver \mathbf{Y}_1 or \mathbf{Y}_2 or \dots \mathbf{Y}_K . The set of parameters $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \dots, \boldsymbol{\Sigma}_K, \pi_1, \pi_2, \dots, \pi_K\}$ is unknown. We define $g_{\Delta_j}(\Delta_j = 1) = P(\Delta_j = 1) = \pi_j$ and it is known that if $\Delta_j = 1$ then $\mathbf{Y} = \mathbf{Y}_j \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. The probability density function of \mathbf{Y} can be given by:

¹Observations in a sample are said to be independent if they are not connected in any way. In addition to being independent, observations in a sample are often assumed to be identically distributed i.e. they originate from the same probability distribution. In statistics one often assumes a sample to be independent and identically distributed, abbreviated as *i.i.d.*

$$\begin{aligned}
p(\mathbf{Y}|\Theta) &= g_{\mathbf{Y}}(\mathbf{y}|\Theta) \\
&= \sum_{j=1}^K g_{\mathbf{Y}_j}(\mathbf{y}|\theta_j) \\
&= \sum_{j=1}^K g_{\mathbf{Y}_j|\Delta_j}(\mathbf{y}|\theta_j, \Delta_j = 1) g_{\Delta_j}(\Delta_j = 1) \text{ (conditional probability)} \\
&= \sum_{j=1}^K \pi_j \cdot \phi_{\theta_j}(\mathbf{y})
\end{aligned} \tag{2.1.2}$$

where $\phi_{\theta_j}(\mathbf{y}) = g_{\mathbf{Y}_j|\Delta_j}(\mathbf{y}|\theta_j, \Delta_j = 1) = \left(\frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_j|^{\frac{1}{2}}} \right) \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_j)' \Sigma_j^{-1} (\mathbf{y} - \boldsymbol{\mu}_j)\right)$ is the multivariate Gaussian density function and $\theta_j = (\boldsymbol{\mu}_j, \Sigma_j)$.

2.2 Estimation of Θ using the EM algorithm

Various estimation methods have been developed and improved over time. In his application of a two-component mixture of Gaussian distributions, Pearson used method of moments (MM) estimators to estimate the model parameters [41]. For the two-component Gaussian mixture estimation, five parameters: $\Theta = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi\}$, need to be estimated and therefore the first five moments were required to find the solutions. This led to Pearson defining the well-known nonic equation that needed to be solved to find the parameter estimates. In 1967, Cohen developed an expedient way of calculating the MM estimates as derived by Pearson by circumventing the direct solving of the nonic equation [13]. Day then compared the method of moments, minimum χ^2 , Bayes and maximum likelihood estimators for the two component mixture of distributions model and found that the maximum likelihood estimators were superior to the alternative estimators due to either the sampling properties of the estimates or the complexity of the computation [15]. In 1978, Quandt and Ramsey introduced the moment generating function (MGF) estimator as an alternative to the sum of squares and maximum likelihood estimators which is suitable for small sample sizes [45]. However, Quandt and Ramsey compared the performance of their MGF estimator only to the MM estimator (as developed by Cohen). In his comment on Quandt and Ramsey's work, Hosmer compared the MGF estimator with the maximum likelihood (ML) estimator and concluded that the two methods could be used in conjunction: when the MGF estimator identifies several possible estimates, the ML estimator could be used to decide on the best estimate [32].

Finally, Dempster, Laird and Rubin proved that the iterative maximum likelihood estimates are of EM (Expectation-Maximisation) type which implied that when convergence is achieved, it is to a local maximum [26, 16]. The EM algorithm is widely used in literature as an expedient way of finding the maximum likelihood estimates of a K -component mixture of distributions model. The detailed EM algorithm is described in the following section.

2.2.1 Deriving the log-likelihood function and the expectation step

Since the model is defined using a hidden random indicator variable S , the likelihood function excluding the hidden random indicator variable is known as the incomplete likelihood function, given below.

$$\begin{aligned}
 L(\Theta; \mathbf{Y}) &= \prod_{i=1}^N (p(\mathbf{Y}_i | \Theta)) \\
 &= \prod_{i=1}^N \left\{ \sum_{j=1}^K \pi_j \phi_{\theta_j}(\mathbf{y}_i) \right\} \\
 &= \prod_{i=1}^N \left(\sum_{j=1}^K \pi_j \left(\frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_j|^{\frac{1}{2}}} \right) \exp \left(-\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_j)' \Sigma_j^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_j) \right) \right). \quad (2.2.1)
 \end{aligned}$$

Direct maximisation of this likelihood function has its challenges, referring to the summed terms inside the logarithm if the likelihood function is rewritten as the log-likelihood:

$$\ln L(\Theta; \mathbf{Y}) = \sum_{i=1}^N \ln \left(\sum_{j=1}^K \pi_j \phi_{\theta_j}(\mathbf{y}_i) \right)$$

and therefore an application of the EM algorithm is discussed as a pragmatic way of finding the maximum likelihood estimates.

Assume, for the moment, that the random indicator variable S is observable. We therefore have both data sets \mathbf{Y} and S available. Then the complete likelihood is given by the joint distribution over all observations in a sample of size N :

$$\begin{aligned}
L(\Theta; \mathbf{Y}, S) &= \prod_{i=1}^N \{p(\mathbf{Y}, S | \Theta)\} \\
&= \prod_{i=1}^N \{p(\mathbf{Y} | S, \Theta) \cdot p(S | \Theta)\} \quad (\text{Bayes' Rule}) \\
&= \prod_{i=1}^N \left\{ \left[\sum_{j=1}^K \Delta_{ij} \phi_{\theta_j}(\mathbf{y}) \right] \cdot \left[\sum_{j=1}^K \Delta_{ij} \pi_j \right] \right\} \\
&= \prod_{i=1}^N \left\{ \left[\sum_{j=1}^K \Delta_{ij} \left(\frac{1}{(2\pi)^{p/2} |\Sigma_j|^{1/2}} \right) \exp \left(-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_j)' \Sigma_j^{-1} (\mathbf{y} - \boldsymbol{\mu}_j) \right) \right] \cdot \left[\sum_{j=1}^K \Delta_{ij} \pi_j \right] \right\} \\
&= \left[\sum_{j=1}^K \left(\frac{1}{(2\pi)^{Np/2} |\Sigma_j|^{N/2}} \right) \exp \left(\sum_{i=1}^N -\frac{1}{2} \Delta_{ij} \cdot (\mathbf{y} - \boldsymbol{\mu}_j)' \Sigma_j^{-1} (\mathbf{y} - \boldsymbol{\mu}_j) \right) \right] \\
&\quad \cdot \prod_{i=1}^N \left[\sum_{j=1}^K (\Delta_{ij} \pi_j) \right] \\
&= \left[\sum_{j=1}^K \left(\frac{1}{(2\pi)^{Np/2} |\Sigma_j|^{N/2}} \right) \exp \left(-\frac{1}{2} \text{tr} (\Delta_{ij} \Sigma_j^{-1} (\mathbf{y}_j - \boldsymbol{\mu}_j \mathbf{1}') (\mathbf{y}_j - \boldsymbol{\mu}_j \mathbf{1}')') \right) \right] \\
&\quad \cdot \prod_{i=1}^N \left[\sum_{j=1}^K (\Delta_j \pi_j) \right]. \tag{2.2.2}
\end{aligned}$$

Equation (2.2.2) follows from *Lemma 2a* and *2b* in the Appendix, where $\mathbf{y}_j = (\mathbf{y}_1 \ \mathbf{y}_2 \ \dots \ \mathbf{y}_N)'$. Now, define the weighted average of \mathbf{y} : $\bar{\mathbf{y}} = \frac{1}{\sum_{i=1}^N \Delta_{ij}} \sum_{i=1}^N \Delta_{ij} \cdot \mathbf{y}_i$ and $\mathbf{A} = \sum_{i=1}^N \Delta_{ij} (\mathbf{y}_i - \bar{\mathbf{y}}) (\mathbf{y}_i - \bar{\mathbf{y}})' = (\mathbf{y} - \bar{\mathbf{y}} \mathbf{1}') \mathbf{D}_j (\mathbf{y} - \bar{\mathbf{y}} \mathbf{1}')'$ where \mathbf{D}_j is a $(N \times N)$ diagonal matrix with $\{\Delta_{1j}, \Delta_{2j}, \dots, \Delta_{Nj}\}$ on the main diagonal. It is shown in *Lemma 2c* that the multivariate Gaussian density function can be rewritten in terms of $\bar{\mathbf{y}}$ and \mathbf{A} :

$$\begin{aligned}
L(\Theta; \mathbf{Y}, S) &= \left[\sum_{j=1}^K \left(\frac{1}{(2\pi)^{Np/2} |\Sigma_j|^{N/2}} \right) \exp \left(-\frac{1}{2} \text{tr} \Sigma_j^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \boldsymbol{\mu}_j)' \Sigma_j^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_j) \right) \right] \\
&\quad \cdot \prod_{i=1}^N \left[\sum_{j=1}^K (\Delta_{ij} \pi_j) \right] \\
\ln L(\Theta; \mathbf{Y}, S) &= \ln \left[\sum_{j=1}^K \left(\frac{1}{(2\pi)^{Np/2} |\Sigma_j|^{N/2}} \right) \exp \left(-\frac{1}{2} \text{tr} \Sigma_j^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \boldsymbol{\mu}_j)' \Sigma_j^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_j) \right) \right] \\
&\quad + \sum_{i=1}^N \ln \left[\sum_{j=1}^K (\Delta_{ij} \pi_j) \right] \\
\ln L(\Theta; \mathbf{Y}, S) &= \sum_{j=1}^K \left[\left(-\frac{Np}{2} \cdot \ln(2\pi) - \frac{N}{2} \cdot \ln |\Sigma_j| - \frac{1}{2} \text{tr} \Sigma_j^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \boldsymbol{\mu}_j)' \Sigma_j^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_j) \right) \right] \\
&\quad + \sum_{i=1}^N \sum_{j=1}^K \Delta_{ij} \cdot \ln(\pi_j). \tag{2.2.3}
\end{aligned}$$

Now, in Equation (2.2.3) it seems like the log of the sum is equated to the sum of the logs. However, it is presumed to be known which distribution \mathbf{Y} originated from i.e. $\Delta_{il} = 1$ while $\Delta_{ij} = 0$ for all $j \neq l$, $j \in \{1, 2, \dots, K\}$ (i.e. only the l^{th} element in the sum $\sum_{j=1}^K (\cdot)$ is non-zero). In summary, the complete likelihood is given by:

$$\ln L(\Theta; \mathbf{Y}, S) = \sum_{j=1}^K [\ln \phi_{\theta_j}(\mathbf{y})] + \sum_{i=1}^N \sum_{j=1}^K \Delta_{ij} \cdot \ln(\pi_j) \tag{2.2.4}$$

where $\phi_{\theta_j}(\mathbf{y})$ is now the distribution function of the weighted observations: $\Delta_{ij} \mathbf{y}_i$. Since the Δ_{ij} 's are usually unknown, their expected value is used:

$$\begin{aligned}
E(\Delta_{ij} | \mathbf{Y}, \Theta) &= 0.P(\Delta_{ij} = 0 | \mathbf{Y}, \Theta) + 1.P(\Delta_{ij} = 1 | \mathbf{Y}, \Theta) \\
&= P(\Delta_{ij} = 1 | \mathbf{Y}, \Theta) \\
&= \gamma_{ij}. \tag{2.2.5}
\end{aligned}$$

This is known as the *responsibility* of model j for observation i [26]; or more intuitively stated, the weight observation i contributes to the estimation of the parameters associated with model j . The Q -function results from taking the expected value of the complete log-likelihood (Equation (2.2.4)) with respect to the hidden indicator random variable Δ_{ij} :

$$\begin{aligned}
Q &= E_{\Delta_{ij}} [\ln L(\Theta; \mathbf{Y}, S)] \\
&= E_{\Delta_{ij}} \left\{ \sum_{j=1}^K [\ln \phi_{\theta_j}(\mathbf{y}_i)] + \sum_{i=1}^N \sum_{j=1}^K \Delta_{ij} \cdot \ln(\pi_j) \right\} \\
&= \sum_{j=1}^K [\ln \phi_{\theta_j}(\mathbf{y}_i)] + \sum_{i=1}^N \sum_{j=1}^K [\gamma_{ij} \cdot \ln(\pi_j)] \tag{2.2.6}
\end{aligned}$$

The maximum likelihood estimates for Θ can now be obtained by maximising the Q -function.

The *expectation* and *maximisation* steps of the EM algorithm will be shown explicitly for the l^{th} component ($j = l$) in the mixture of multivariate Gaussian distributions, but follow similarly for all $j = 1, 2, \dots, K$.

First, an expression for $E(\Delta_{il} | \mathbf{Y}, \Theta)$ needs to be obtained in the expectation step of the EM algorithm, which performs a so-called soft assignment of the observations to each component of the mixture (calculating the *responsibilities*).

$$\begin{aligned}
\hat{\gamma}_{il} &= P(\Delta_{il} = 1 | \mathbf{Y}, \Theta) \\
&= \frac{P(\Delta_{il} = 1) \cdot P(\mathbf{Y} | \Delta_{il} = 1, \Theta)}{\sum_{j=1}^K P(\Delta_{ij} = 1) \cdot P(\mathbf{Y} | \Delta_{ij} = 1, \Theta)} \quad (\text{Law of total probability}) \\
&= \frac{\pi_l \cdot \phi_{\theta_l}(\mathbf{y})}{\sum_{j=1}^K \pi_j \phi_{\theta_j}(\mathbf{y})} \tag{2.2.7}
\end{aligned}$$

The expected value of the indicator variable Δ_{ij} (for each model $j = 1, 2, \dots, K$) is equal to the probability of observation \mathbf{y}_i being generated from distribution j divided by the sum of the probability of \mathbf{y}_i being generated from any of the K distributions.

2.2.2 The maximisation step

Now, to obtain the maximum likelihood estimator $\hat{\boldsymbol{\mu}}_1$, take the partial derivative of Q with respect to $\boldsymbol{\mu}_1$:

$$\begin{aligned}
\frac{\delta Q}{\delta \boldsymbol{\mu}_1} &= \frac{\delta}{\delta \boldsymbol{\mu}_1} \left\{ \sum_{j=1}^K [\ln \phi_{\theta_j}(\mathbf{y}_i)] + \sum_{i=1}^N \sum_{j=1}^K [\gamma_{ij} \cdot \ln(\pi_j)] \right\} \\
&= \frac{\delta}{\delta \boldsymbol{\mu}_1} \left\{ \sum_{j=1}^K [\ln \phi_{\theta_j}(\mathbf{y}_i)] + 0 \right\} \\
&= \frac{\delta}{\delta \boldsymbol{\mu}_1} \left\{ -\frac{Np}{2} \cdot \ln(2\pi) - \frac{N}{2} \cdot \ln |\boldsymbol{\Sigma}_1| - \frac{1}{2} \text{tr} \boldsymbol{\Sigma}_1^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_1^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_1) + 0 \right\}
\end{aligned}$$

Since it is known that $\boldsymbol{\Sigma}_1$ is a positive definite matrix, the log-likelihood is maximised for all values of $\boldsymbol{\Sigma}_1$ where $\bar{\mathbf{y}} = \boldsymbol{\mu}_1$ (*Lemma 2d*). Therefore, according to the definition of $\bar{\mathbf{y}}$, the maximum likelihood estimator for $\boldsymbol{\mu}_1$ is given by:

$$\hat{\boldsymbol{\mu}}_1 = \frac{\sum_{i=1}^N \hat{\gamma}_{il} \mathbf{y}_i}{\sum_{i=1}^N \hat{\gamma}_{il}} \quad (2.2.8)$$

To obtain the maximum likelihood estimate of $\hat{\boldsymbol{\Sigma}}_1$, take the partial derivative of Q with respect to $\boldsymbol{\Sigma}_1$:

$$\begin{aligned} \frac{\delta Q}{\delta \boldsymbol{\Sigma}_1} &= \frac{\delta}{\delta \boldsymbol{\Sigma}_1} \left\{ \sum_{j=1}^K [\ln \phi_{\boldsymbol{\theta}_j}(\mathbf{y}_i)] + \sum_{i=1}^N \sum_{j=1}^K [\gamma_{ij} \cdot \ln(\pi_j)] \right\} \\ &= \frac{\delta}{\delta \boldsymbol{\Sigma}_1} \left\{ \sum_{j=1}^K [\ln \phi_{\boldsymbol{\theta}_j}(\mathbf{y}_i)] + 0 \right\} \\ &= \frac{\delta}{\delta \boldsymbol{\Sigma}_1} \left\{ \left[-\frac{Np}{2} \cdot \ln(2\pi) - \frac{N}{2} \cdot \ln |\boldsymbol{\Sigma}_1| - \frac{1}{2} \text{tr} \boldsymbol{\Sigma}_1^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_1^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_1) + 0 \right] \right\} \end{aligned}$$

Now, the result proven in *Lemma 2d* can be applied, using *Lemma 1* to find the maximum likelihood estimator for $\boldsymbol{\Sigma}_1$:

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\mathbf{A}}{\sum_{i=1}^N \hat{\gamma}_{il}} = \frac{(\mathbf{y} - \bar{\mathbf{y}}) \mathbf{G}_j (\mathbf{y} - \bar{\mathbf{y}})'}{\sum_{i=1}^N \hat{\gamma}_{il}}$$

where \mathbf{G}_j is a $(N \times N)$ diagonal matrix with $\{\hat{\gamma}_{1j}, \hat{\gamma}_{2j}, \dots, \hat{\gamma}_{Nj}\}$ on the main diagonal ($\mathbf{G}_j = E_{\Delta_{ij}}(\mathbf{D}_j)$). The estimates can be seen as a weighted average (with respect to the responsibilities) of the usual maximum likelihood estimators for the Gaussian distribution. The probability of \mathbf{y}_i belonging to model j (π_j) is estimated by the sum of the responsibilities with respect to model j (the effective number of observations in cluster j [3]) divided by N :

$$\hat{\pi}_j = \frac{\sum_{i=1}^N \hat{\gamma}_{ij}}{N} \quad (2.2.9)$$

Using the results derived in the text above, the EM algorithm for a K -component mixture of multivariate Gaussian distributions is given:

Algorithm 2.1 EM algorithm for a K -component mixture of multivariate Gaussian distributions

1. Find relevant starting values for the parameters in $\Theta = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \dots, \boldsymbol{\Sigma}_K, \pi_1, \pi_2, \dots, \pi_K\}$

2. *Expectation step*: calculate the responsibilities

$$\hat{\gamma}_{il} = \frac{\pi_l \cdot \phi_{\boldsymbol{\theta}_l}(\mathbf{y}_i)}{\sum_{j=1}^K \pi_j \phi_{\boldsymbol{\theta}_j}(\mathbf{y}_i)}$$

3. *Maximisation step*: calculate the weighted average maximum likelihood estimates and mixing probabilities for $l = 1, 2, \dots, K$.

$$\begin{aligned} \hat{\boldsymbol{\mu}}_l &= \frac{\sum_{i=1}^N \hat{\gamma}_{il} \mathbf{y}_i}{\sum_{i=1}^N \hat{\gamma}_{il}} \\ \hat{\boldsymbol{\Sigma}}_l &= \frac{(\mathbf{y} - \bar{\mathbf{y}}) \mathbf{G}_j (\mathbf{y} - \bar{\mathbf{y}})'}{\sum_{i=1}^N \hat{\gamma}_{il}} \\ \hat{\pi}_l &= \frac{\sum_{i=1}^N \hat{\gamma}_{il}}{N} \end{aligned}$$

4. Repeat step 2. and 3. until convergence is achieved.
-

Note that convergence is achieved when the change in estimated parameters from one iteration of the algorithm to the next is negligible. The examples in Section 2.4 and 2.6 include discussions on finding the starting values for Algorithm 2.1.

In the following sections the literature on the identifiability of the mixing distributions (components) and the well-known singularity problem arising during the estimation of mixing distributions (components) will be discussed. Examples will accompany these discussions to demonstrate principles and to illustrate how the EM algorithm is applied to estimate the finite mixture of Gaussian distributions model for the univariate and multivariate case.

2.3 Identifiability of mixture distribution components

Identifiability of finite mixture models is essential for their consistent estimation [20] in that it allows for the recovery of the mixing distributions, also referred to as components, from the mixture [37, 28]. Therefore as Holzmänn et al pointed out, the assumption of identifiability “lies at the heart of most statistical theory and practice” [28, 52].

Frühwirth-Schnatter distinguishes between three types of nonidentifiability namely,

1. Nonidentifiability due to invariance to relabeling the components of the mixture distribution,
2. Nonidentifiability due to potential overfitting and
3. Nonidentifiability as a generic property of a certain class of mixture distributions [20].

The first type of nonidentifiability is also referred to as the label-switching problem [26]: all the parameters are related and only differ in terms of the sequence of the components [20]. This nonidentifiability plays a role in parameter estimation and can be easily addressed [20] by ensuring that the order of the components remains consistent throughout the estimation process.

Nonidentifiability due to potential overfitting was noted by Crawford, who showed that any mixture of $K - 1$ components defines a nonidentifiability subset in the parameter space Θ_K (mixtures with K components) where one component is empty or two components are equal [20, 14]. It is therefore important that the number of components of a mixture be estimated if this is not known *a priori*, in order to address this nonidentifiability.

The third type of nonidentifiability is referred to in the study of identifiability of finite mixtures [28] and was pioneered by Teicher in 1961 [20]. He proved that the class of mixtures of a one-parameter additively-closed family of distributions is identifiable [50]. Teicher's definition of identifiability of finite mixture distributions is widely used and built on in literature:

$\mathcal{F} = \{F(x, \theta)\}$ is a parametric family of n -dimensional CDF's having parameter space R_1^m , a subset of Euclidean m space R^m . Let G be any m -dimensional CDF defined on the parameter space R_1^m such that G assigns probability 1 to finitely many mass points $\{\theta_1, \dots, \theta_N\}$ in R_1^m , $p(\theta_i)$ being the mass at θ_i . Then under the mapping Q defined below, G is transformed into the n -dimensional CDF

$$\begin{aligned} Q(G) &= \sum_{i=1}^N p(\theta_i) F(x | \theta_i) \\ &= H(x) \end{aligned}$$

G is called a *mixing distribution* and $H(x)$ is a finite mixture on \mathcal{F} . Let \mathcal{G} be the set of all mixing distributions G as described above. Then the image $\mathcal{H} = Q(\mathcal{G})$, the set of all finite mixtures on \mathcal{F} , is the convex hull of \mathcal{F} ^a. We say that \mathcal{H} is identifiable if Q is a one-to-one function, i.e., if

$$(G_1 \neq G_2) \text{ implies } (Q(G_1) \neq Q(G_2))$$

^aThe convex hull or envelope of a set of points X in a Euclidean space is the smallest convex set that contains X .

In summary, \mathcal{F} is identifiable if the set \mathcal{H} of finite mixtures generated by \mathcal{F} is identifiable [54]. Teicher thereby established theorems and propositions by which one may establish the identifiability of mixtures for more families of distributions (beyond the normal and Poisson distributions) [50]. Furthermore, Teicher proved that the class of all finite mixtures of normal (Gaussian) and gamma distributions are identifiable [51]. He also proved that finite mixtures of the Poisson distributions are identifiable and that finite mixtures of the uniform and binomial distributions are not.

In 1968 Yakowitz and Spragins expanded on Teicher's study of identifiability of finite mixtures by modifying his results to include multivariate CDF's and they proved that a family of CDF's induce identifiable finite mixtures if and only if the said family is linearly independent in its span over the field of real numbers [55]. They proved that finite mixtures of the following families are identifiable: n -dimensional normal (Gaussian) family, n -dimensional exponential family and the union of the two,

the one-dimensional Cauchy and one-dimensional negative binomial families. The identifiability of finite mixtures on general measurable spaces was studied by Chandra [11]. Holzmann et al studied the identifiability of finite mixtures of elliptical densities [28]. Yakowitz further related the statistical subject of identifiability of finite mixtures to the field of unsupervised learning in the engineering context and pointed out that consequently unsupervised learning is possible under surprisingly lenient conditions [54].

This study is limited to the discussion and application of finite mixture of Gaussian distributions which have been proven to be identifiable in the univariate [51] and multivariate cases [55]. The estimations of mixing distributions (mixture components), based on observations from mixture distributions included in this study, can therefore be meaningfully discussed.

2.4 Example - mixture of univariate Gaussian distributions

A univariate Gaussian mixture of distributions model is fitted to simulated data originating from two univariate Gaussian distributions (white noise is added) to demonstrate that the model identifies the component distributions well even if the location of the distributions are the same.

A random sample of size $N = 200$ was generated to simulate a two component mixture of Gaussian distribution with equal component-means and different variances so that $Y = Y_j$ with probability $\pi = 0.5$; the simulated data is shown as overlaid histograms in Figure 2.4.1.

$$\begin{aligned} Y_1 &\sim N(10, 0.5) \\ Y_2 &\sim N(10, 50) \end{aligned}$$

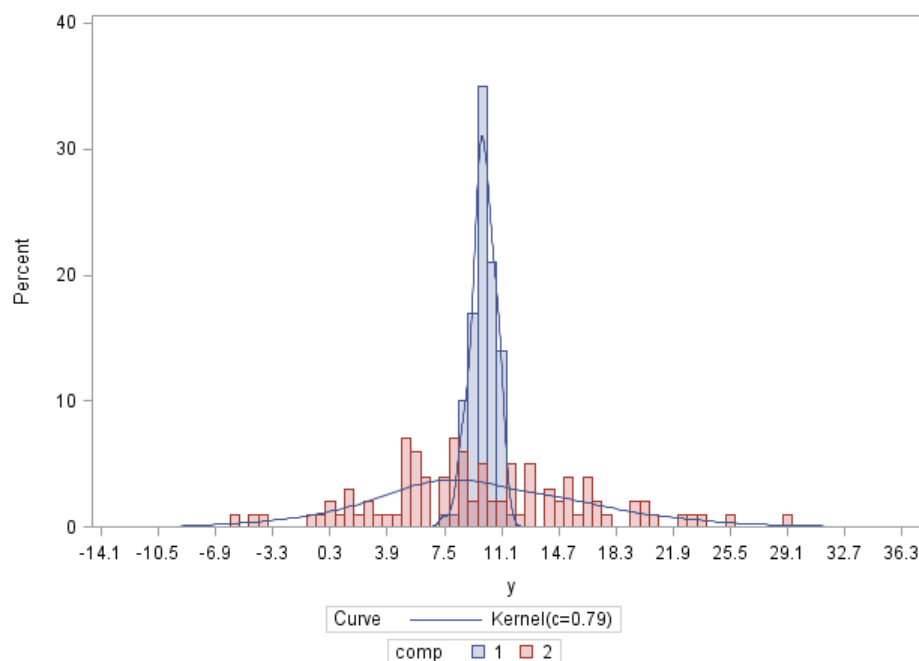


Figure 2.4.1: Overlaid histograms of simulated data: component 1 and 2

The model to be estimated is a two-component univariate mixture of Gaussian distributions:

$$Y = \pi.Y_1 + (1 - \pi).Y_2$$

The EM algorithm, as it is given for the multivariate Gaussian case in Algorithm 2.1, was used to find the maximum likelihood estimates of the parameters $\Theta = \{\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$. The starting values for the parameters were selected by choosing two random observations for the means, setting the overall sample variance as the starting value for both component variances and setting the starting values for the mixing probabilities to $\pi = \frac{1}{K} = \frac{1}{2}$. The parameter estimates are given below.

$$\hat{\pi}_1 = 0.498$$

$$\hat{\pi}_2 = 0.502$$

$$\hat{\mu}_1 = 10.15$$

$$\hat{\mu}_2 = 10.13$$

$$\hat{\sigma}_1^2 = 0.47$$

$$\hat{\sigma}_2^2 = 45.68$$

The estimated parameters are close to the known theoretical parameter values. The likelihood function was calculated in every EM algorithm step and is graphically displayed in Figure 2.4.2. It is clear that the log-likelihood of the estimated model increased with each algorithm step and the

parameter values that maximise the log-likelihood function were chosen as the parameter estimates.

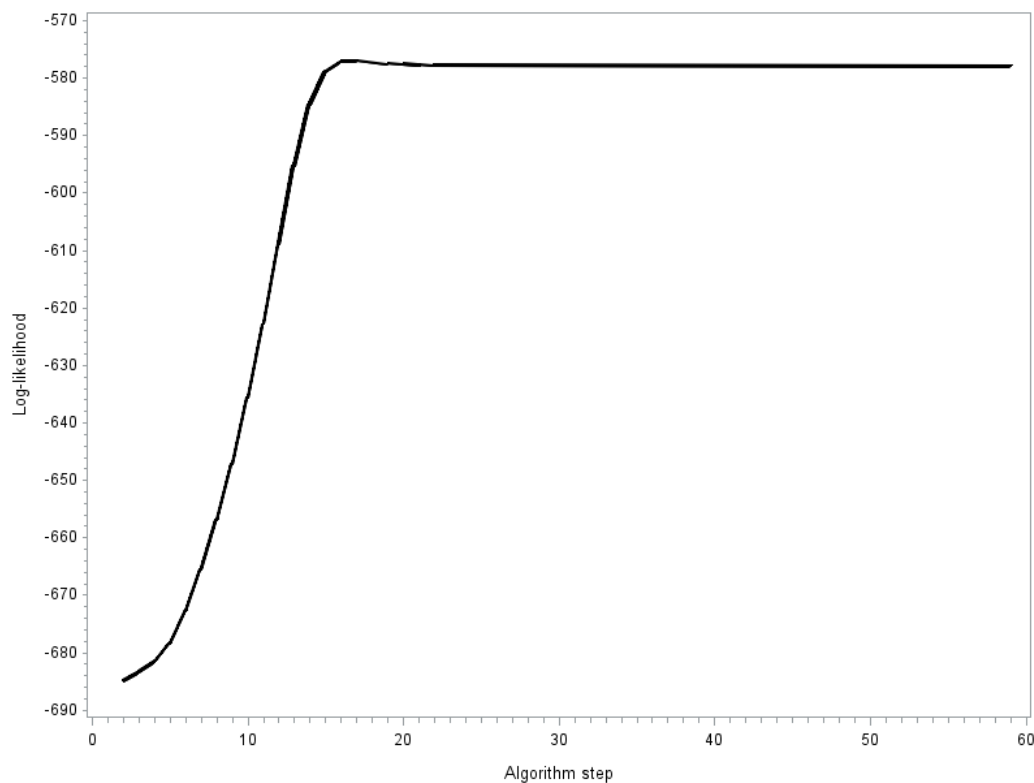


Figure 2.4.2: Log-likelihood function value in each EM step

The estimated mixture distribution with the estimated responsibilities of the respective components are graphically displayed in Figure 2.4.3. The mixture distribution for Y where the responsibility for component 1: $\gamma_1 = E(\Delta_{i1}) = E(Y = Y_1)$ is higher than for component 2 corresponds with the component 1 mixing distribution in terms of shape and location ($Y_1 \sim N(10, 0.5)$). On the other hand, where the component 2 responsibilities are higher than the component 1 responsibilities the mixture distribution resembles the component 2 distribution ($Y_2 \sim N(10, 50)$) in shape and location.

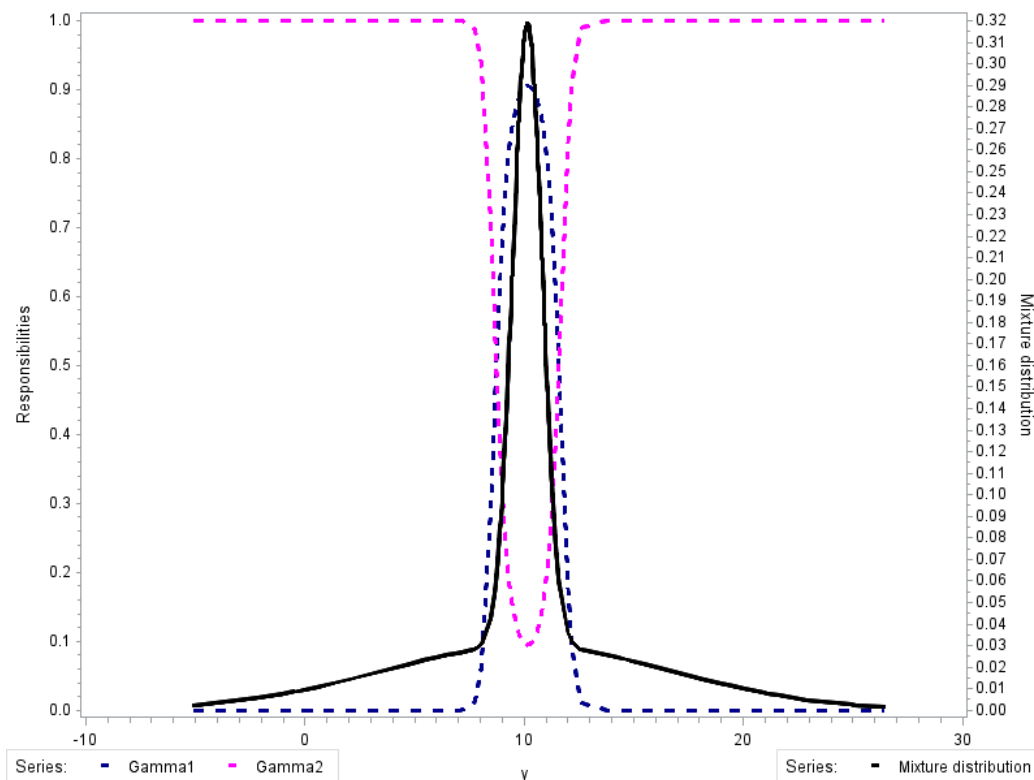


Figure 2.4.3: Mixture distribution with responsibilities for each component

Not every pair of observations chosen as the starting values for the means yield estimates for two different components. In a lot of cases, the EM algorithm converges to a single solution and gives an error as soon as one component's variance captures nearly all the variance in the sample and the other goes to zero - upon which the likelihood function becomes ill-specified. This problem is known as the singularity problem related to estimating mixtures of distributions and will be discussed in more detail in the following section.

2.5 Singularity problem during estimation

2.5.1 EM Algorithm difficulties

As mentioned in the example in Section 2.4, the EM algorithm is not always successful in estimating the parameters of a mixture of Gaussian distributions. The algorithm breaks down when the variance of a component σ_j^2 becomes numerically zero in, say, step m of the algorithm, or in the multivariate case, when Σ_j becomes singular or near-singular. This occurs when the responsibilities of the j^{th} component contain too many zeros. Consequently during step $m + 1$ of the EM algorithm, the calculation of the responsibilities through Equation (2.2.5) is no longer possible.

This problem often occurs when the EM algorithm is used to estimate a finite mixture of Gaussian distributions while overfitting the number of components [20]. This overfitting also often results from the choice of starting values i.e. two component mean and / or variance starting values are chosen to be sufficiently close to each other that the EM algorithm steps converge to the same underlying

component, leading to many zero-responsibilities for the “extra” component parameters.

Therefore, the choice of starting values and number of components in the mixture are both important when attempting to estimate the components using an EM algorithm. Regarding starting values, Hastie, Tibshirani and Friedman [26] suggest that any random observations can be used for starting values for the means; the component variance-covariance matrices can all be set equal to the overall sample variance-covariance matrix and the initial mixing probabilities to be uniformly assigned i.e., for a $K = 2$ component mixture $\pi_1 = \pi_2 = \frac{1}{K} = \frac{1}{2}$. Regarding the number of components, the probability density function of the data (to which a mixture model needs to be fitted) needs to be estimated using numerical techniques (e.g. kernel estimation) and diagnostic or exploratory analysis on this estimated probability density function can be performed to attempt to identify the number of components in the mixture.

2.5.2 Unboundedness of the mixture likelihood function

As first noted by Kiefer and Wolfowitz in 1956, the likelihood function of a univariate mixture of Gaussian distributions is unbounded and has many local spurious modes [35]. In 1969, Day noted that the unboundedness of the likelihood function is also applicable to the multivariate Gaussian mixtures where each observation \mathbf{y} gives rise to a singularity² on the boundary of the parameter space [15]. The Kiefer Wolfowitz example is revisited to demonstrate the singularity problem.

The mixture of two Gaussian distributions is considered:

$$Y = \pi Y_1 + (1 - \pi) Y_2$$

where π is fixed and it is assumed that $Y_1 \sim N(\mu, 1)$ and $Y_2 \sim N(\mu, \sigma_2^2)$, therefore μ and σ_2^2 are unknown.

$$\begin{aligned} p(\mathbf{y}|\mu, \sigma_2^2) &= \prod_{i=1}^N \{\pi \phi_{\theta_1}(y) + (1 - \pi) \phi_{\theta_2}(y)\} \\ &= \prod_{i=2}^N \left\{ \pi \left(\frac{1}{2\pi(1)} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{2(1)}(y_i - \mu)^2\right) + (1 - \pi) \left(\frac{1}{2\pi\sigma_2^2} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma_2^2}(y_i - \mu)^2\right) \right\}. \end{aligned}$$

Now, it is clear that whenever $\mu = y_i$:

$$p(\mathbf{y}|\mu = y_i, \sigma_2^2) = \prod_{i=2}^N \left\{ \pi \left(\frac{1}{2\pi(1)} \right)^{\frac{1}{2}} (1) + (1 - \pi) \left(\frac{1}{2\pi\sigma_2^2} \right)^{\frac{1}{2}} (1) \right\}.$$

Then as σ_2^2 tends to zero, the mixture likelihood function is dominated by a term proportional to $\frac{1}{\sigma_2^2}$, leading to:

²A region where values tend to infinity.

$$\lim_{\sigma_2^2 \rightarrow 0} p(\mathbf{y} | \mu = y_i, \sigma_2^2) = \infty.$$

In Figure 2.5.1 the log likelihood of the $N = 20$ observation Kiefer-Wolfowitz example (where $\mu = 0$ and $\sigma_2^2 = 4$ and $\pi = 0.2$) was calculated for a range of possible values for μ and σ_2^2 . A local maximum is observed for the theoretical parameter values however, it is seen that for $\mu = y_i; i = 1, 2, \dots, N$ various local maxima's occur.

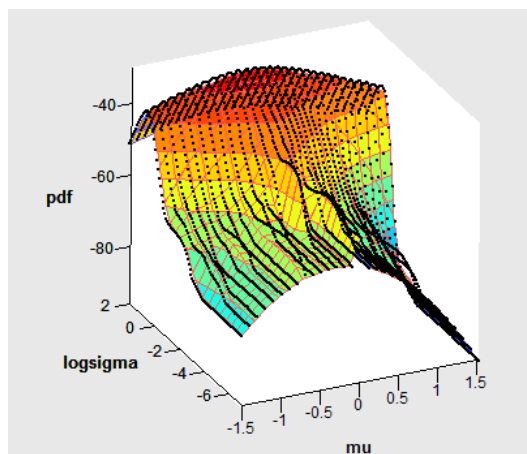


Figure 2.5.1: Log-likelihood of Kiefer-Wolfowitz example

Since these spurious local maxima exist, there does not exist a global maximiser of the likelihood function on the unconstrained parameter space [43]. Consequently, other estimation approaches were investigated or various levels of constraints were applied to the component variance ratios in order to enable maximum likelihood estimation.

In Figure 2.5.2 μ is kept constant at $\mu = 0$ and $\mu = y_{10} = -0.22627$ and the resulting log-likelihood is plotted against σ_2 . It is seen that as σ_2 goes to zero, the likelihood increases, more so with $\mu = y_{10}$.

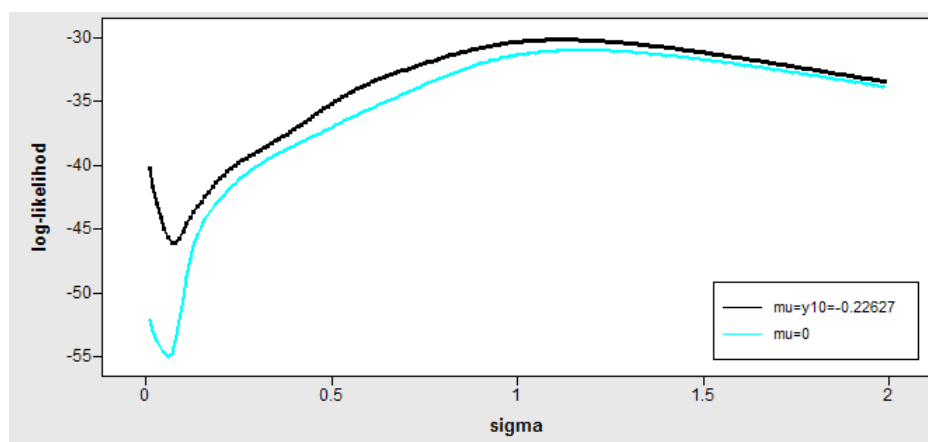


Figure 2.5.2: Log-likelihood against σ_2

2.5.3 Dealing with sources of unboundedness

Frühwirth-Schnatter points out that it is the modeller's complete ignorance about ratio between variances of the various components, that leads to the unboundedness of the mixture likelihood function [20]. Throughout literature, various ways of imposing a condition on the variance structure of the mixture were investigated. Hathaway concluded that constraints on the relative size of the standard deviations in the univariate Gaussian mixture model result in a bounded likelihood function [27]. In 1978 Quandt and Ramsey suggested that the constraint $\sigma_1 = c\sigma_2$ with c known is sufficient to make the likelihood function bounded [45]. Phillips then proposed a less restrictive constraint:

$$\min_{i,j} \left(\frac{\sigma_i}{\sigma_j} \right) \geq c > 0. \quad (2.5.1)$$

Hathaway [27] showed that constraints like the one above, ensure a global maximiser of the likelihood function for a univariate Gaussian mixture. Phillips generalised this result to apply to mixture of regressions [43] and showed that searching for the maximiser of the likelihood surface on the constrained parameter space (called the constrained MLE) is a well-posed strategy and generally produces reasonable estimates.

Hartley asserts that Quandt and Ramsey might have over exaggerated the importance of the computational obstacles to maximum likelihood estimation [25]. In his simulation studies Hosmer found that the iterative maximum likelihood estimates (e.g. EM algorithm) will not converge to the parameter values associated with the singularities if the sample size is big enough ($n < 300$) and if the components are separated well enough (satisfying $\frac{|\mu_2 - \mu_1|}{\min(\sigma_1, \sigma_2)} \geq 3$) [31]. Caudill and Acharya showed through Monte-carlo simulations that the incidence of singularities for the two component normal regression mixture problem may be significantly lower than for the mixture of univariate normal distributions [9]. Due to the multiple spurious local maximas of the likelihood function, Hartley suggests that the statistician should experiment with different starting values and where multiple solutions arise, select the solution that maximises the likelihood function [25].

2.6 Example - mixture of multivariate Gaussian distributions

This example demonstrates the application of a multivariate mixture of Gaussian distributions and shows that the components with equal means can be identified.

A random sample of $N = 200$ observations was drawn from either of two bivariate Gaussian distributions, based on the probability $\pi = 0.6$, so that $\mathbf{Y} = (1 - \pi)\mathbf{Y}_1 + \pi\mathbf{Y}_2$.

$$\begin{aligned} \mathbf{Y}_1 &\sim N \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \\ \mathbf{Y}_2 &\sim N \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} \right) \end{aligned}$$

Figure 2.6.1 below displays the sample observations from the two Gaussian distributions. Note that the shape of the two distributions look similar but when one takes the axes' scales into account,

they clearly differ according to the parameters selected above.

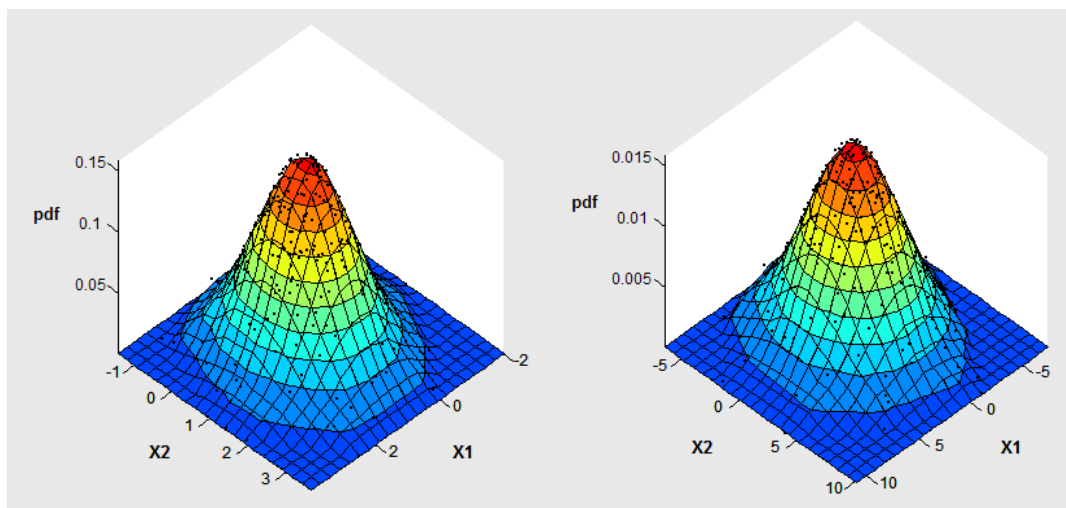


Figure 2.6.1: 3D-Scatterplot of simulated data (\mathbf{Y}_1 left; \mathbf{Y}_2 right)

A two-component mixture of bivariate Gaussian distributions is estimated:

$$\mathbf{Y} = \pi \cdot \mathbf{Y}_1 + (1 - \pi) \cdot \mathbf{Y}_2$$

where $\mathbf{Y}_1 \sim N\left(\begin{bmatrix} \mu_{11} \\ \mu_{12} \end{bmatrix}, \begin{bmatrix} \sigma_{11}^{(1)} & \sigma_{12}^{(1)} \\ \sigma_{21}^{(1)} & \sigma_{22}^{(1)} \end{bmatrix}\right)$ and $\mathbf{Y}_2 \sim N\left(\begin{bmatrix} \mu_{21} \\ \mu_{22} \end{bmatrix}, \begin{bmatrix} \sigma_{11}^{(2)} & \sigma_{12}^{(2)} \\ \sigma_{21}^{(2)} & \sigma_{22}^{(2)} \end{bmatrix}\right)$.

The EM algorithm in Algorithm 2.1, was used to find the maximum likelihood estimates of the parameters $\Theta = \{\pi, \mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, \sigma_{11}^{(1)}, \sigma_{12}^{(1)}, \sigma_{21}^{(1)}, \sigma_{22}^{(1)}, \sigma_{11}^{(2)}, \sigma_{12}^{(2)}, \sigma_{21}^{(2)}, \sigma_{22}^{(2)}\}$. The parameter estimates after convergence of the algorithm are given in Table 2.6.1. The estimated parameters are relatively close to the theoretical parameter values.

Table 2.6.1: Parameter estimates of two component bivariate mixture of Gaussian distributions

	Component 1 (Y_1)	Component 2 (Y_2)
$\hat{\pi}$	0.41	0.59
$\hat{\mu}_{j1}$	1.14	0.44
$\hat{\mu}_{j2}$	0.79	1.03
$\begin{bmatrix} \hat{\sigma}_{11}^{(j)} & \hat{\sigma}_{12}^{(j)} \\ \hat{\sigma}_{21}^{(j)} & \hat{\sigma}_{22}^{(j)} \end{bmatrix}$	$\begin{bmatrix} 0.95 & 0.03 \\ 0.03 & 1.19 \end{bmatrix}$	$\begin{bmatrix} 10.44 & 0.34 \\ 0.34 & 9.09 \end{bmatrix}$

The estimated mixture of bivariate Gaussian distributions is graphically displayed in Figure 2.6.2 (viewed from two different sides) and the distribution very clearly does not resemble a simple Gaussian distribution but a combination of the two component Gaussian distributions - note the bulge at the base.

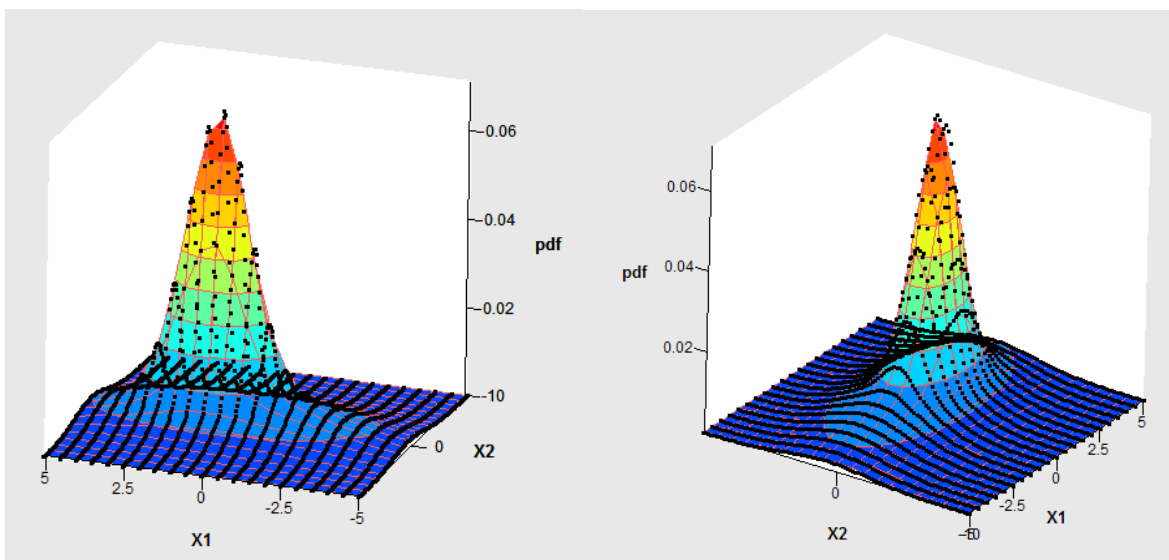


Figure 2.6.2: Mixture distribution of two-component bivariate mixture

The original observations are “hard-clustered” as being generated from distribution \mathbf{Y}_1 or \mathbf{Y}_2 respectively i.e. if the estimated responsibility for observation i is higher for component 1, the observation is clustered in cluster \mathbf{Y}_1 and otherwise in cluster \mathbf{Y}_2 . The classification result is graphically displayed in Figure 2.6.3. It is clear that the observations in the center are attributed to \mathbf{Y}_1 and the observations in the outer regions are attributed to \mathbf{Y}_2 . This is consistent with how the original sample was generated and therefore one can conclude that clustering was achieved successfully.

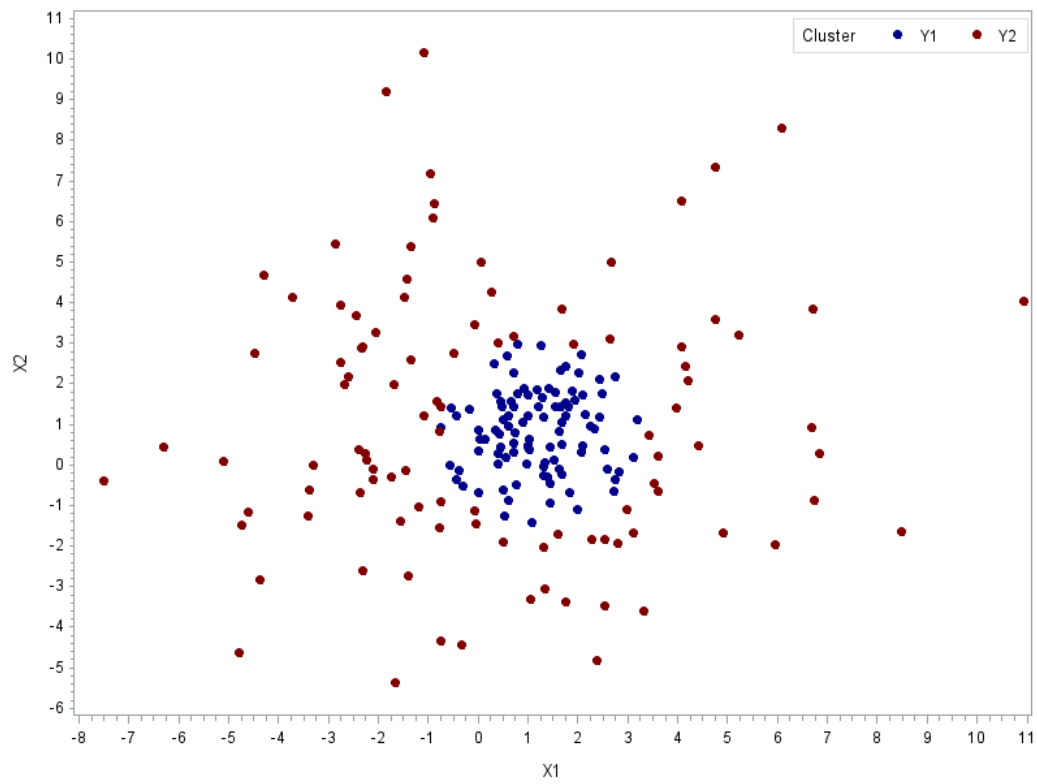


Figure 2.6.3: Graphical representation of classification result

Chapter 3

Finite mixture of multivariate regressions

The multivariate mixture of regressions involves relating a random $(p \times 1)$ dimensional random vector \mathbf{Y} , to a set of explanatory variables or covariates $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{q-1})$ through a regression type model. The conditional mean of \mathbf{Y} is assumed to be a function of $(p \times q)$ dimensional matrix $\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{X}_1 & \mathbf{X}_2 & \dots & \mathbf{X}_{(q-1)} \end{bmatrix}$ so that

$$E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is a $(q \times 1)$ vector of regression coefficients. In some instances it is inadequate to assume that $\boldsymbol{\beta}$ is constant over all observations of \mathbf{Y} and the finite mixture of regressions model is a way of dealing with this heterogeneity.

3.1 The finite mixture of multivariate Gaussian regressions

Instead of fitting a single regression type model to the available data set, there is some evidence that the population contains multiple underlying regression type models or relationships often referred to as regimes. This study will consider linear regression models where the error term is assumed to follow an Gaussian distribution. A finite mixture of regressions model assumes that a set of K independent regression models, characterised by $(\boldsymbol{\beta}_1, \boldsymbol{\Sigma}_1), (\boldsymbol{\beta}_2, \boldsymbol{\Sigma}_2), \dots, (\boldsymbol{\beta}_K, \boldsymbol{\Sigma}_K)$ describe heterogeneous subgroups in the model of interest. Moreover, for each pair of observations (\mathbf{Y}, \mathbf{X}) a hidden random indicator variable S can take on integers 1 to K and thereby chooses one of the K models to generate \mathbf{Y} . The indicator variable S follows an unknown discrete probability distribution $\Pi = (\pi_1, \pi_2, \dots, \pi_K)$ where $\sum_{j=1}^K \pi_j = 1$ and is assumed to be mutually independent over each observation $i = 1, 2, \dots, N$. In the simplest case, we have no prior information about Π and the resulting K regression models are given below, with each \mathbf{Y} being generated by one of the K models with a certain probability:

$$\begin{aligned}
\mathbf{Y}_1 &= \mathbf{X}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1, \text{ with probability } \pi_1 \\
\mathbf{Y}_2 &= \mathbf{X}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2, \text{ with probability } \pi_2 \\
&\vdots \\
\mathbf{Y}_K &= \mathbf{X}\boldsymbol{\beta}_K + \boldsymbol{\varepsilon}_K, \text{ with probability } \pi_K
\end{aligned}$$

where $\pi_K = 1 - \sum_{j=1}^{K-1} \pi_j$ and $\boldsymbol{\varepsilon}_j \sim N(0, \boldsymbol{\Sigma}_j)$. Therefore \mathbf{Y}_j can be modelled by a Gaussian distribution: $\mathbf{Y}_j \sim N(\mathbf{X}\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_j)$ for $j = 1, 2, \dots, K$.

In economics, a mixture of regressions model is also described as a switching regression system which is equivalent to assuming the presence of a so-called structural change [45]. In the specification above, the investigator is assumed to be ignorant of what moves the system from one structural form to another (also described as an unobservable latent variable) as opposed to situations where the structural change may depend deterministically on some observable variables. In other words, nature is assumed to generate each \mathbf{Y} from \mathbf{X} by regime j with probability π_j [36, 44].

Another way to denote this notion is to maintain, that for each model $j = 1, 2, \dots, K$ an indicator variable Δ_j exists which takes on the value 1 with probability π_j and is otherwise zero, essentially capturing the same information contained in S such that $I(S = j) = \Delta_j$. Define the Δ_j 's as adhering to the constraint: $\sum_{j=1}^K \Delta_j = 1$, implying that for each observation \mathbf{Y} only one of the K regression models can be chosen as the data generating process. This notation allows us to write the data generating process for each observation of \mathbf{Y} in a generative form:

$$\mathbf{Y} = \Delta_1 \mathbf{Y}_1 + \Delta_2 \mathbf{Y}_2 + \dots + \Delta_K \mathbf{Y}_K \quad (3.1.1)$$

with the set of parameters $\boldsymbol{\Theta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_K, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \dots, \boldsymbol{\Sigma}_K, \pi_1, \pi_2, \dots, \pi_K\}$ unknown. It is known that $g_{\Delta_j}(\Delta_j) = P(\Delta_j = 1) = \pi_j$ then, similar to Equation (2.1.2), the probability density function of \mathbf{Y} ¹ is given by

$$\begin{aligned}
p(\mathbf{Y}|\boldsymbol{\Theta}) &= g_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\Theta}) \\
&= \sum_{j=1}^K \pi_j \phi_{\boldsymbol{\theta}_j}(\mathbf{y})
\end{aligned} \quad (3.1.2)$$

where $\phi_{\boldsymbol{\theta}_j}(\mathbf{y})$ is the p -variate Gaussian density function with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}_j$ and $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_j)$. It is important to note here, that the estimation of a finite mixture of regressions model with Gaussian error terms is equivalent to estimating parameters of a finite mixture of Gaussian distributions with $\boldsymbol{\mu}_j = \mathbf{X}\boldsymbol{\beta}_j$ and therefore the results derived in Section 2.2 follow in a similar fashion [36, 20].

¹Upper case \mathbf{Y} indicates the random variable, while lower case \mathbf{y} denotes the observed value. Upper case \mathbf{X} denotes the matrix of observed values or random variables.

3.2 Estimation of Θ using the EM algorithm

3.2.1 Deriving the log-likelihood function and the expectation step

The model is defined using a hidden random indicator variable S therefore the likelihood function excluding this hidden random variable is regarded as the incomplete likelihood function given below (similar to Equation (2.2.1)).

$$L(\Theta; \mathbf{Y}) = \prod_{i=1}^N \left(\sum_{j=1}^K \pi_j \left(\frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_j|^{\frac{1}{2}}} \right) \exp \left(-\frac{1}{2} (\mathbf{y}_i - \mathbf{X}_i \beta_j)' \Sigma_j^{-1} (\mathbf{y}_i - \mathbf{X}_i \beta_j) \right) \right).$$

When the likelihood is rewritten into the log-likelihood form, direct maximisation proves to be challenging due to the sum of the terms inside the logarithm. Therefore the EM algorithm is used to simplify the maximum likelihood calculations [16].

Assume, for the moment, that the random indicator variable S is observed: the Δ_j 's in $\mathbf{Y} = \Delta_1 \mathbf{Y}_1 + \Delta_2 \mathbf{Y}_2 + \dots + \Delta_K \mathbf{Y}_K$ are known with $P(\Delta_{ij} = 1) = \pi_j$. We therefore have both data sets \mathbf{Y} and S available. Then the complete likelihood is derived, just as in the previous chapter, using the *Lemmas* in the Appendix, to get the following:

$$\begin{aligned} \ln L(\Theta; \mathbf{Y}, S) &= \sum_{j=1}^K \left[\left(-\frac{Np}{2} \cdot \ln(2\pi) - \frac{N}{2} \cdot \ln |\Sigma_j| - \frac{1}{2} \text{tr} \Sigma_j^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \mathbf{X}_j \beta_j)' \Sigma_j^{-1} (\bar{\mathbf{y}} - \mathbf{X}_j \beta_j) \right) \right] \\ &\quad + \sum_{i=1}^N \sum_{j=1}^K \Delta_{ij} \cdot \ln(\pi_j) \end{aligned}$$

where $\mathbf{y} = (\mathbf{y}_1 \ \mathbf{y}_2 \ \dots \ \mathbf{y}_N)$, $\bar{\mathbf{y}} = \frac{\sum_{i=1}^N \Delta_{ij} \mathbf{y}_i}{\sum_{i=1}^N \Delta_{ij}}$ and $\mathbf{A} = \sum_{i=1}^N \Delta_{ij} (\mathbf{y}_i - \bar{\mathbf{y}}) (\mathbf{y}_i - \bar{\mathbf{y}})' = (\mathbf{y} - \bar{\mathbf{y}} \mathbf{1}') \mathbf{D}_j (\mathbf{y} - \bar{\mathbf{y}} \mathbf{1})'$ where \mathbf{D}_j is a $(N \times N)$ diagonal matrix with $\{\Delta_{1j}, \Delta_{2j}, \dots, \Delta_{Nj}\}$ on the main diagonal.

In reality Δ_{ij} 's are unknown and therefore we use the expected value and the fact that each Δ_{ij} can take on either 1 or 0 for $j = 1, 2, \dots, K$. This is known as the *expectation step* of the EM algorithm.

$$\begin{aligned} E(\Delta_{ij} | \mathbf{Y}, \Theta) &= 0.P(\Delta_{ij} = 0 | \mathbf{Y}, \Theta) + 1.P(\Delta_{ij} = 1 | \mathbf{Y}, \Theta) \\ &= \gamma_{ij}. \end{aligned}$$

As in the mixtures of Gaussian distributions case, the expected value of Δ_{ij} is called the *responsibility* of model j for observation \mathbf{y}_i which can be described as the weight observation \mathbf{y}_i contributes towards the estimation of the parameters of model j . The Q -function results from taking the expected value of the complete log-likelihood with respect to the hidden indicator random variable Δ_{ij} :

$$\begin{aligned}
Q &= E_{\Delta_{ij}} [\ln L(\Theta; \mathbf{Y}, S)] \\
&= E_{\Delta_{ij}} \left\{ \sum_{j=1}^K \left[\left(-\frac{Np}{2} \cdot \ln(2\pi) - \frac{N}{2} \cdot \ln |\Sigma_j| - \frac{1}{2} \text{tr} \Sigma_j^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \mathbf{X}\beta_j)' \Sigma_j^{-1} (\bar{\mathbf{y}} - \mathbf{X}\beta_j) \right) \right] \right. \\
&\quad \left. + \sum_{i=1}^N \sum_{j=1}^K \gamma_{ij} \cdot \ln(\pi_j) \right\}
\end{aligned}$$

where $\bar{\mathbf{y}} = \frac{\sum_{i=1}^N \gamma_{ij} \mathbf{y}_i}{\sum_{i=1}^N \gamma_{ij}}$ and $\mathbf{A} = \sum_{i=1}^N \gamma_{ij} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' = (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}') \mathbf{G}_j (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}')'$ where \mathbf{G}_j is a $(N \times N)$ diagonal matrix with $\{\gamma_{1j}, \gamma_{2j}, \dots, \gamma_{Nj}\}$ on the main diagonal. The maximum likelihood estimates for Θ can now be obtained by maximising the Q -function.

First, an expression for $E(\Delta_{ij} | \mathbf{Y}, \Theta)$ for $j = 1, 2, \dots, K$ needs to be obtained in order to perform the necessary calculation for the expectation step of the EM algorithm. The results will be shown for the l^{th} component with the remaining components following in similar fashion:

$$\begin{aligned}
\hat{\gamma}_{il} &= P(\Delta_{il} = 1 | \mathbf{Y}, \Theta) \\
&= \frac{P(\Delta_{il} = 1) \cdot P(\mathbf{Y} | \Delta_{il} = 1, \Theta)}{\sum_{j=1}^K P(\Delta_{ij} = 1) \cdot P(\mathbf{Y} | \Delta_{ij} = 1, \Theta)} \quad (\text{Law of total probability}) \\
&= \frac{\pi_l \cdot \phi_{\theta_l}(\mathbf{y})}{\sum_{j=1}^K \pi_j \phi_{\theta_j}(\mathbf{y})}.
\end{aligned} \tag{3.2.1}$$

Intuitively, the expected value of the hidden indicator random variable for the l^{th} component model (Δ_{il}) is equal to the probability of observation \mathbf{y} being generated by model l divided by the sum of the probability of \mathbf{y} being generated by any of the $j = 1, 2, \dots, K$ models.

3.2.2 The maximisation step of the EM algorithm

The equations for the maximisation step for the EM Algorithm are given for the l^{th} component in the mixture of regressions model, and follow in a similar fashion for all K components. To obtain the maximum likelihood estimator for $\hat{\beta}_1$, take the partial derivative of Q with respect to β_1 :

$$\begin{aligned}
\frac{\delta Q}{\delta \beta_1} &= \frac{\delta}{\delta \beta_1} \left\{ \sum_{j=1}^K \left[-\frac{Np}{2} \cdot \ln(2\pi) - \frac{N}{2} \cdot \ln |\Sigma_j| - \frac{1}{2} \text{tr} \Sigma_j^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \mathbf{X}\beta_j)' \Sigma_j^{-1} (\bar{\mathbf{y}} - \mathbf{X}\beta_j) \right] + 0 \right\} \\
&= \frac{\delta}{\delta \beta_1} \left(-\frac{Np}{2} \cdot \ln(2\pi) - \frac{N}{2} \cdot \ln |\Sigma_1| - \frac{1}{2} \text{tr} \Sigma_1^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \mathbf{X}\beta_1)' \Sigma_1^{-1} (\bar{\mathbf{y}} - \mathbf{X}\beta_1) \right)
\end{aligned}$$

Since it is known that Σ_1 is a positive definite matrix, the log-likelihood is maximised for all values of Σ_1 where $\bar{\mathbf{y}} = \mathbf{X}\beta_1$ (*Lemma 2d*). Therefore the maximum likelihood estimator for β_1 is given below:

$$\begin{aligned}
\bar{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}}_l \\
\frac{\sum_{i=1}^N \hat{\gamma}_{ij} \mathbf{y}_i}{\sum_{i=1}^N \hat{\gamma}_{ij}} &= \mathbf{X}\hat{\boldsymbol{\beta}}_l \\
\frac{\sum_{i=1}^N \hat{\gamma}_{ij} \mathbf{y}_i' \mathbf{X}}{\sum_{i=1}^N \hat{\gamma}_{ij}} &= \mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}}_l \\
\hat{\boldsymbol{\beta}}_l &= \frac{\left(\sum_{i=1}^N \hat{\gamma}_{ij} \mathbf{y}_i' \mathbf{X} \right) (\mathbf{X}' \mathbf{X})^{-1}}{\sum_{i=1}^N \hat{\gamma}_{il}} \tag{3.2.2}
\end{aligned}$$

To find the maximum likelihood estimator $\hat{\boldsymbol{\Sigma}}_1$, take the partial derivative of Q with respect to $\boldsymbol{\Sigma}_1$:

$$\begin{aligned}
\frac{\delta Q}{\delta \boldsymbol{\Sigma}_1} &= \frac{\delta}{\delta \boldsymbol{\Sigma}_1} \left\{ \sum_{j=1}^K \left[-\frac{Np}{2} \cdot \ln(2\pi) - \frac{N}{2} \cdot \ln |\boldsymbol{\Sigma}_j| - \frac{1}{2} \text{tr} \boldsymbol{\Sigma}_j^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \mathbf{X}_j \boldsymbol{\beta}_j)' \boldsymbol{\Sigma}_j^{-1} (\bar{\mathbf{y}} - \mathbf{X}_j \boldsymbol{\beta}_j) + 0 \right] \right\} \\
&= \frac{\delta}{\delta \boldsymbol{\Sigma}_1} \left(-\frac{Np}{2} \cdot \ln(2\pi) - \frac{N}{2} \cdot \ln |\boldsymbol{\Sigma}_1| - \frac{1}{2} \text{tr} \boldsymbol{\Sigma}_1^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \mathbf{X}_1 \boldsymbol{\beta}_1)' \boldsymbol{\Sigma}_1^{-1} (\bar{\mathbf{y}} - \mathbf{X}_1 \boldsymbol{\beta}_1) \right).
\end{aligned}$$

Now, we can apply the result proven in *Lemma 2d*, using *Lemma 1* to find the maximum likelihood estimator of $\boldsymbol{\Sigma}_1$:

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_1 &= \frac{\mathbf{A}}{\sum_{i=1}^N \hat{\gamma}_{il}} \\
&= \frac{(\mathbf{y} - \bar{\mathbf{y}}) \mathbf{G}_1 (\mathbf{y} - \bar{\mathbf{y}})'}{\sum_{i=1}^N \hat{\gamma}_{il}} \\
&= \frac{(\mathbf{y} - \mathbf{X} \boldsymbol{\beta}_l) \mathbf{G}_1 (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}_l)'}{\sum_{i=1}^N \hat{\gamma}_{il}}.
\end{aligned}$$

Similar to Equation (2.2.8); where \mathbf{G}_1 is a $(N \times N)$ diagonal matrix, containing the estimated responsibilities $\hat{\gamma}_{il}$ for $i = 1, 2, \dots, N$ on the main diagonal.

The estimates can be seen as a weighted average (by estimated responsibilities) of the usual maximum likelihood estimates for a simple linear regression model. The probability of \mathbf{y} belonging to model j (π_j) is estimated by the sum of the responsibilities of model j for all the observations divided by N :

$$\hat{\pi}_j = \frac{\sum_{i=1}^N \hat{\gamma}_{ij}}{N}. \tag{3.2.3}$$

The EM algorithm for estimating the components of the mixture of multivariate regressions model is then given below. The definition of convergence remains the same as in Algorithm 2.1 and the example in Section 3.3 demonstrates how the starting values for this algorithm are typically found.

Algorithm 3.1 EM algorithm for a K -component mixture of multivariate Gaussian regressions model

1. Find relevant starting values for the parameters in $\Theta = \{\beta_1, \beta_2, \dots, \beta_K, \Sigma_1, \Sigma_2, \dots, \Sigma_K, \pi_1, \pi_2, \dots, \pi_K\}$
2. *Expectation step*: calculate the responsibilities

$$\hat{\gamma}_{il} = \frac{\pi_l \cdot \phi_{\theta_l}(\mathbf{y}_i)}{\sum_{j=1}^K \pi_j \phi_{\theta_j}(\mathbf{y}_i)}$$

3. *Maximisation step*: calculate the weighted average maximum likelihood estimates and mixing probabilities for $l = 1, 2, \dots, K$.

$$\begin{aligned} \hat{\beta}_l &= \frac{\left(\sum_{i=1}^N \hat{\gamma}_{il} \mathbf{y}_i \mathbf{X}' \right) (\mathbf{X}' \mathbf{X})^{-1}}{\sum_{i=1}^N \hat{\gamma}_{il}} \\ \hat{\Sigma}_l &= \frac{\left(\mathbf{y} - \mathbf{X} \hat{\beta}_l \right) \mathbf{G}_1 \left(\mathbf{y} - \mathbf{X} \hat{\beta}_l \right)'}{\sum_{i=1}^N \hat{\gamma}_{il}} \\ \hat{\pi}_l &= \frac{\sum_{i=1}^N \hat{\gamma}_{il}}{N} \end{aligned}$$

4. Repeat step 2. and 3. until convergence is achieved.
-

3.3 Example - mixture of univariate Gaussian regressions

The application of a finite mixture of Gaussian regressions model will be demonstrated based on a simulated example and two methods of calculating the coefficient of determination for the finite mixture of regressions model will be discussed.

A data set with a total of $N = 500$ observations was generated consisting of Y_1 and Y_2 as a function of X ; where X is a fixed set of random $N(0, 1)$ observations, and

$$\begin{aligned} Y_1 &= \alpha_1 + \beta_1 X + \varepsilon_1 \\ Y_2 &= \alpha_2 + \beta_2 X + \varepsilon_2. \end{aligned}$$

The theoretical parameter values for the generated data are given by $\Theta = \{\alpha_1 = 0, \beta_1 = 1, \alpha_2 = 3, \beta_2 = 1.2, \pi = 0.6\}$. Therefore $Y = Y_1$ with probability $\pi = 0.6$ and $Y = Y_2$ with probability $1 - \pi = 0.4$. The generated data is plotted in Figure 3.3.1, coloured by component and includes the theoretical component regression lines (note that *Response* represents Y_1 and Y_2 observations).

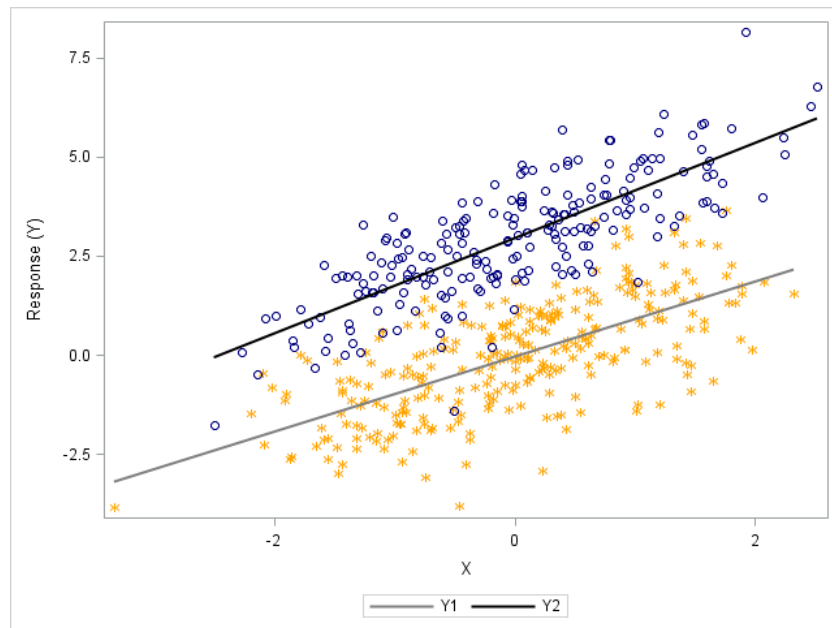


Figure 3.3.1: Simulated mixture of regressions data

Ignoring the underlying mixture of regressions, a simple linear regression is fitted to the data using the regression procedure (*proc reg*) in SAS. With an $R^2 = 0.25$, the resulting model explains 25% of the variation in Y .

$$E(Y|X) = 1.24 + 1.04X$$

Both parameters in the estimated model are significant (the null hypothesis of a zero parameter is rejected at a 1% significance level in both cases) and the overall model fits reasonably well, even though its performance is not excellent. The fitted model is illustrated in the Figure 3.3.2.

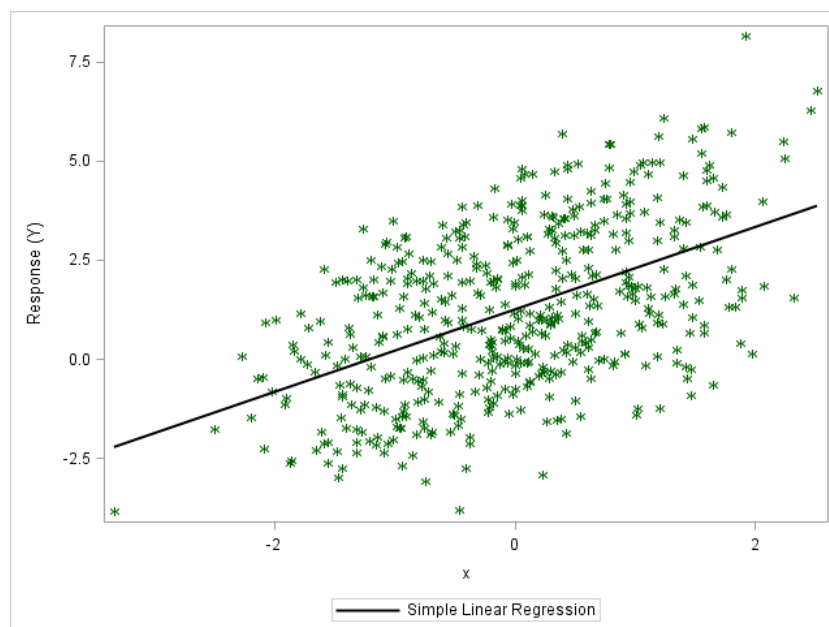


Figure 3.3.2: Simple linear regression model fit

The finite mixture models procedure (*proc fmm*) in SAS is used to fit a model where the assumption of a mixture of two regressions is incorporated, essentially estimating the exact model structure from which the data was generated. The estimated parameters, $\hat{\Theta}$, are given in the table below:

Table 3.3.1: Estimated mixture of regression parameters

Θ	$\hat{\Theta}$	Θ	$\hat{\Theta}$
α_1	0.03	α_2	2.99
β_1	0.92	β_2	1.18
π	0.59		

All the parameter estimates are significantly different from zero, apart from $\hat{\alpha}_1$ which is theoretically, equal to zero and is expected not to be different from zero. The estimates are very close to the theoretical values. The resulting component regression model estimates are given by

$$E(Y_1|X) = 0.03 + 0.92X$$

$$E(Y_2|X) = 2.99 + 1.18X.$$

The component regression models have been accurately estimated (i.e. the estimated parameters are close to the theoretical parameter values). The clustering result is investigated by hard clustering² the observations. The corresponding clustering result is shown in Figure 3.3.3 and resembles the original theoretical clustering closely, however, the slight overlap in the original clusters is not captured by the

²hard clustering is when the observations are assigned to the cluster for which the highest responsibility was estimated

hard clustering, although the magnitudes of the estimated responsibilities reflect this overlap.

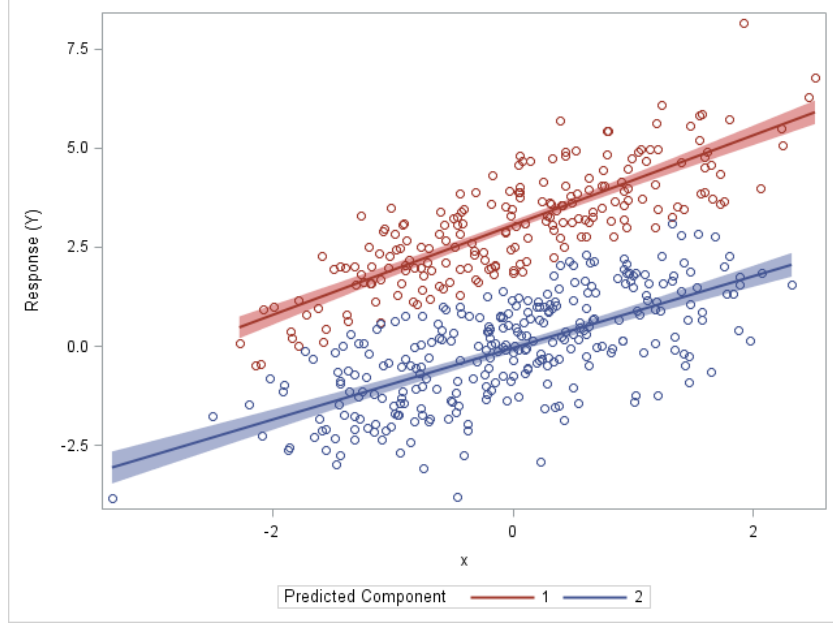


Figure 3.3.3: Mixture of regressions clustering result

Ideally, one would like to evaluate the goodness of fit for the two-component mixture of regressions model, in order to decide whether it fits better than the simple linear regression model. The coefficient of determination for the simple linear regression model is traditionally calculated as:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 0.25.$$

This study has explored two methods of calculating the coefficient of determination for the mixture of regressions model. The first method involves hard clustering of the observations; essentially splitting up the data set into two groups representing the two clusters. The estimated parameters for the respective clusters, can then be used to calculate the R^2 statistic for each cluster.

$$R_{Y_1}^2 = 1 - \frac{\sum_{i=1}^{n_1} (y_i - \hat{y}_{1i})^2}{\sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2} = 0.45$$

$$R_{Y_2}^2 = 1 - \frac{\sum_{i=1}^{n_2} (y_i - \hat{y}_{2i})^2}{\sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2} = 0.63$$

45% of the variation in cluster 1 was explained by component 1 of the mixture of regressions model and 63% of the variation in cluster 2 was explained by component 2. Both components' R^2 statistics are a significant improvement on the simple linear regressions R^2 however, $K = 2$ R^2 statistics are not strictly comparable with a single R^2 statistic in order to decide which model fits best.

The second method involves calculating a weighted R^2 statistic where each observation's residuals for component 1 and component 2 are weighted with the corresponding responsibilities.

$$R_{weighted}^2 = 1 - \frac{\sum_{i=1}^n [\hat{\gamma}_{1i} (y_i - \hat{y}_{1i})^2 + \hat{\gamma}_{2i} (y_i - \hat{y}_{2i})^2]}{\sum_{i=1}^n (y_i - \bar{y}_{weighted})^2} = 0.75$$

where $\bar{y}_{weighted} = \frac{\sum_{i=1}^n (\hat{\gamma}_{1i} y_i + \hat{\gamma}_{2i} y_i)}{n} = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$ since $\sum_{j=1}^K \hat{\gamma}_j = 1$. This weighted R^2 statistic indicates that 75% of the total variation in Y is explained by the two-component mixture of regressions model which is significantly higher than the simple linear regression model R^2 . Therefore, as expected, the mixture of regression model fits the data better and would be the chosen model based on the coefficient of determination.

Chapter 4

The spatial variant mixture of regressions model

Traditional mixture models have been widely used as a pixel labeling or clustering technique and some applications are discussed here. Finite mixture models have been successfully applied in, among other fields, bioinformatics [4], the image retrieval context [21], MRI image segmentation [40] and aerial and satellite image segmentation [42].

Some improvements and developments around the application of finite mixture models for image segmentation are listed. Gupta and Sortrakul developed a selective-sampling-Gaussian-mixture-parameter-estimation segmentation algorithm which performed well in accurately segmenting diverse images, including degraded images [24]. Permuter et al found that their Gaussian mixture models formulation showed an overall performance improvement relative to other models in terms of the classification of man-made and natural areas in aerial images [42]. In addressing the model selection problem of choosing the appropriate number of components in a Gaussian mixture model, Lu and Ip introduce an entropy regularized likelihood (ERL) learning algorithm which outperforms other competitive learning algorithms in the application of unsupervised image segmentation [39]. An unsupervised algorithm that estimates the parameters and number of components of a finite mixture model online, was developed and the objective of computational speed was achieved which made it particularly relevant for image processing [22].

Nikou et al conclude that finite mixture models are not only a rigorous measure for clustering performance but also hold the advantage of assigning each pixel to the component that most likely generated it based on the mixing probability [40]. In the following section the mixture of regressions model will be formulated as it is applied in the image segmentation context, followed by the model specification of the spatial variant mixture of regressions model.

4.1 Traditional mixture of regressions model applied in an image context

Consider spatiotemporal data: $\{Y_{il}\}_{i=1,\dots,N}^{l=1,\dots,T}$ where i denotes the spatial index and l denotes the time index corresponding to time $t = t_l$. The data consists of T images each with N pixels and therefore for each pixel we have an “observation” of a T dimensional time series which can be regarded as a multivariate dependent variable \mathbf{Y} , with an observation represented as $\mathbf{y}_i = [y_{i1} \ y_{i2} \ \cdots \ y_{iT}]$. The model, or what Blekas et al refer to as “the curve” [6], to be fitted to the data can be formulated as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & t_1 & \cdots & t_1^p \\ 1 & t_2 & \cdots & t_2^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_T & \cdots & t_T^p \end{pmatrix}$$

$$\boldsymbol{\beta}' = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_p \end{pmatrix}$$

(\mathbf{X} is known as the Vandermonde matrix) and \mathbf{e} is a T -dimensional vector assumed to follow a Gaussian distribution and to be independent over time; $\mathbf{e} \sim i.i.d. N(\mathbf{0}, \boldsymbol{\Sigma})$. The joint probability density of \mathbf{Y} can be modelled by a Gaussian distribution $N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$. In the case where the pixels $i = 1, 2, \dots, N$ need to be clustered into various “models” (i.e., subgroups of pixels exhibit heterogeneous changes of pixel intensity over time), multiple regression functions can be formulated as:

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{X}\boldsymbol{\beta}_1 + \mathbf{e}_1 \\ \mathbf{Y}_2 &= \mathbf{X}\boldsymbol{\beta}_2 + \mathbf{e}_2 \\ &\vdots \\ \mathbf{Y}_K &= \mathbf{X}\boldsymbol{\beta}_K + \mathbf{e}_K \end{aligned}$$

where $\mathbf{e}_j \sim i.i.d. N(\mathbf{0}, \boldsymbol{\Sigma}_j)$. These $j = 1, 2, \dots, K$ regression models capture the heterogeneous sources of curves or the assumption that the data came from K different data generating processes (of the same family of distributions). The generative form of \mathbf{Y} is formulated using hidden indicator variables Δ_j where $j = 1, 2, \dots, K$, each taking on either 0 or 1 and $\sum_{j=1}^K \Delta_j = 1$ (therefore only one regression model is chosen for each pixel).

$$\mathbf{Y} = \Delta_1 \mathbf{Y}_1 + \cdots + \Delta_K \mathbf{Y}_K$$

Now, similar to Equation (2.1.2) and (3.1.2). the probability density function of \mathbf{Y} is given by

$$p(\mathbf{Y}|\Theta) = \sum_{j=1}^K \pi_j p(\mathbf{Y}|\theta_j)$$

where $p(\mathbf{Y}|\theta_j) = N(\mathbf{X}\beta_j, \Sigma_j)$ and $\theta_j = (\beta_j, \Sigma_j)$; π_j for $j = 1, 2, \dots, K$ are the mixing probabilities adhering to the following constraints $\pi_j \geq 0$ for all $j = 1, 2, \dots, K$ and $\sum_{j=1}^K \pi_j = 1$. According to this model, pixel i with sequence of pixel intensities \mathbf{y}_i , will have been generated from the j^{th} regression model with probability π_j .

The maximum likelihood estimation of this model is performed expediently using the EM algorithm and is widely known in mixture model literature. The estimation algorithm performs two steps iteratively: the *expectation step* during which the responsibilities ($\gamma_{ij} = E(\Delta_{ij})$) of each pixel towards the K models are estimated (see Equation (2.2.7) and (3.2.1)).

$$\hat{\gamma}_{ij} = \frac{\pi_j p(\mathbf{Y}|\theta_j)}{\sum_{s=1}^K \pi_s p(\mathbf{Y}|\theta_s)}$$

The second step is the *maximisation step* whereby the mixture parameter estimators are determined by maximising the complete likelihood function and similar to the results derived in Section 3.2.2, the update rules for the parameters are given below.

$$\begin{aligned} \hat{\pi}_j &= \frac{\sum_{i=1}^N \hat{\gamma}_{ij}}{N} \\ \hat{\beta}_j &= \frac{\left(\sum_{i=1}^N \hat{\gamma}_{ij} \mathbf{y}_i' \mathbf{X}\right) (\mathbf{X}' \mathbf{X})^{-1}}{\sum_{i=1}^N \hat{\gamma}_{ij}} \\ \hat{\sigma}_{ju}^2 &= \frac{\sum_{i=1}^N \hat{\gamma}_{ij} \left(y_{iu} - [\mathbf{X} \hat{\beta}_j]_u\right)^2}{\sum_{i=1}^N \hat{\gamma}_{ij}} \end{aligned}$$

where $[\cdot]_u$ indicates the u^{th} component of the T -dimensional vector. $\hat{\Sigma}_j$ are assumed to be diagonal matrices with $\hat{\sigma}_{ju}^2$; $u = 1, 2, \dots, T$ on the main diagonal. $\hat{\Sigma}_j$ is assumed to be diagonal to simplify some derivations and calculations and it is important to note that the covariance structure of the π_j 's allows for spatial dependence, not necessarily the covariance structure of the dependent variable.

4.2 Spatial variant mixture of regressions model

Classical finite mixture models, in the image processing context, have shortcomings in that commonality of location is not taken into account when grouping the data [40] that is, apart from pixel intensity values, the relative pixel location should also be used in determining the cluster or segment to which each pixel belongs. To overcome this shortcoming an approach incorporating a Markov random field was proposed by Chalmond in 1989 [10]. The Markov random field incorporates the pixel location information in terms of the pixel labeling. Some studies are mentioned that investigated and improved

the Markov random field approach, in conjunction with finite mixture models, towards incorporating pixel locations in the clustering algorithm.

Caillol et al introduced fuzziness in Gaussian mixtures and modelled spatial information in both the segmentation and parameter estimation levels [8]. Sanjay-Gopal and Herbert proposed the spatial variant finite mixture model for pixel labeling and image segmentation; with their generalised EM maximum *a posteriori*¹ algorithm they bridged the gap between computationally intensive algorithms based on random fields and simpler segmentation algorithms based on mixture models [47]. Cai and Liu proposed Markov random field models for pattern recognition which provide a natural and flexible framework for modelling interactions between spatially related random variables [7]. A novel spatially constrained generative model also based on the Markov random field approach was proposed by Diplaros et al; they concluded that their EM algorithm for image segmentation is simple to implement and performs competitively in terms of speed and solution quality, compared to other Markov random field algorithms for image segmentation [18]. Zhang et al proposed a modified Gaussian mixture model algorithm - incorporating the Markov random field - that increased the robustness of the traditional Gaussian mixture model to noise in terms of image segmentation [56].

There is a vast literature that demonstrates the application of the Gaussian mixture model with Markov random fields in image segmentation problems; a few are mentioned here: Dezzani and Al-Dousari [17] applied a Markov random field based approach to investigate the rate of change of oil residue deposits in Kuwait - this analysis was based on satellite images over a given time period. Zhang et al applied the Markov random field and EM algorithm in the segmentation of brain MR images [57]. Suliga, Deklerck and Nyssen used the Markov random field approach to classify masses on mammographic images with emphasis on efficient clustering on image edges [49]. Blekas et al applied the spatial variant mixture of regressions model to spatiotemporal data (specifically hear MR images) [6].

In the following section, the spatial variant mixture of regressions model will be defined, after providing some background knowledge for the Markov random field theory.

4.2.1 Background theory

Spatial variation in terms of the pixel clustering, when using finite mixture models for image segmentation, is introduced through a Markov random field and has been applied and investigated throughout literature (as demonstrated previously). A Markov random field is defined as a set of nodes as well as a set of links which connect pairs of nodes; nodes correspond to a variable or group of variables [3]. The Markov random field is an undirected graph as illustrated in Figure 4.2.1 and is a graphical definition of a probability distribution. One can find a way to express the joint distribution $p(x)$ of all the variables or group of variables represented in the Markov random field, as a product of functions defined over sets of variables that are local to the graph [3]. The appropriate notion of locality needs to be identified (i.e. sets of variables within a locality on the graph), here the concept of *cliques* will be used.

A *clique* is a subset of nodes in a graph such that there exists a link between all pairs of nodes

¹Again, the word *posteriori* does not refer to a Bayesian methodology, but in mixture model literature responsibilities are often referred to as posterior probabilities, given the prior knowledge about the distribution Π .

in the subset [3]. The green and blue groups in Figure 4.2.1 are examples of cliques. The blue group is also a *maximal clique*, defined as a clique for which it is impossible to include other nodes from the graph without it ceasing to be a clique [3]. Bishop states that the joint distribution $p(x)$ of the Markov random field can be decomposed into a product of functions of variables in maximal cliques (cliques are by definition subsets of maximal cliques and therefore this result extends to general cliques as well).

Let C represent a clique and x_C the set of variables in that clique, then the joint distribution can be written as a product of the potential functions $\psi_C(x_C)$ over the maximal cliques of the graph:

$$p(x) = \frac{1}{Z} \prod_C \psi_C(x_C). \quad (4.2.1)$$

Z is a normalising constant given by $Z = \sum_x \prod_C \psi_C(x_C)$ which ensures that the joint distribution is a probability measure and $p(x) \geq 0$ since potential functions satisfy $\psi_C(x_C) \geq 0$. Potential functions² are strictly positive and can conveniently be expressed in terms of exponentials: $\psi_C(x_C) = \exp\{-E(x_C)\}$ where $E(x_C)$ is called an energy function³. The joint probability function $p(x)$ is defined by the product of potentials and so the total energy is obtained by adding the energies of each of the maximal cliques [3].

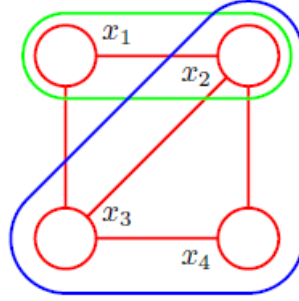


Figure 4.2.1: Example of a Markov random field [3]

A Gibbs distribution is a probability measure p on the state space of the Markov random field (all image pixels), that is represented as

$$p(x) = \frac{1}{Z} \exp\left(\frac{U(x)}{T}\right) \quad (4.2.2)$$

where Z and T are constants and U is called the *energy function* and is of the form: $U(x) = \sum_C V_C(x_C)$ where C is the set of cliques. $V_C(x)$ depends on the coordinates or pixels that lie in C [19].

Now, the Hammersley-Clifford theorem [12] states that “the set of distributions that are consistent with the set of conditional independence statements that can be read from the graph, using graph separation, are identical to the set of distributions that can be expressed as a factorization of the

²A potential function is an arbitrary non negative function over a maximal clique

³An energy function expresses the correlation between the variables in the maximal clique

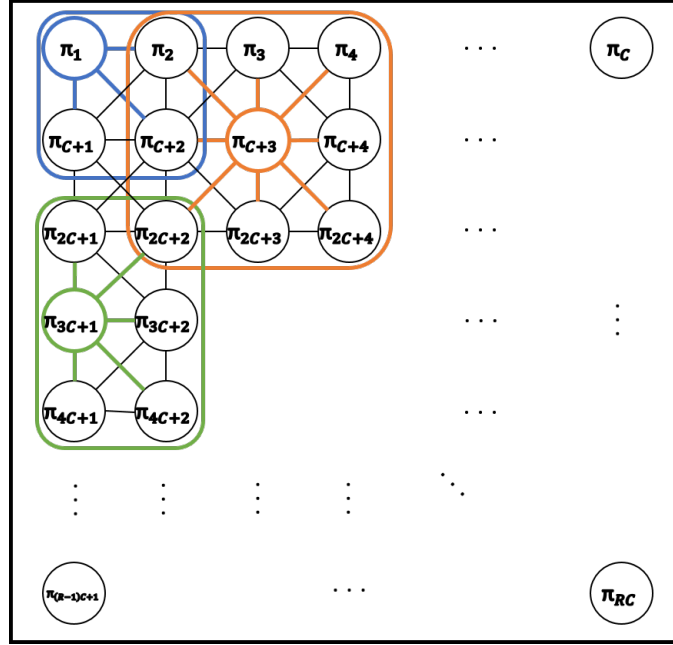
form in Equation (4.2.1) with respect to maximal cliques". In other words, the theorem states that an undirected graph G is a Gibbs random field (i.e. its density can be factorised over the cliques of the graph resulting in a Gibbs distribution) if and only if the undirected graph G describes a probability distribution that has a positive density function that satisfies one of the Markov properties - like the conditional independence properties.

This fundamental theorem is named after John Hammersley and Peter Clifford who proved the equivalence in an unpublished paper in 1971. Geoffrey Grimmett [23] and Sherman [48] also proved the theorem in 1973 with a further proof by Julian Besag in 1974 [2].

The concepts and theory discussed in this section are used in the model specification of the finite spatial variant mixture of regressions model as set out by Blekas et al [6].

4.2.2 Model specification

This model assumes the probabilities of the data labels π_{ij} to be random variables where $i = 1, 2, \dots, N$ indexes the pixels and $j = 1, 2, \dots, K$ the clusters. To handle this information a Markov random field is constructed in order to formulate the prior distribution for the data labels. Note, that this prior distribution is not related to the Bayesian context, but represents prior additional knowledge about the distribution Π , of the random indicator variable S , as mentioned in Section 2.1 and 3.1. The Markov random field construction can be described as follows: let every pixel data label (π_{ij}) for the j^{th} cluster represent a node in the Markov random field and let all the nodes be linked to their neighbours according to their position in the image. Suppose the image consists of R rows and C columns of pixels (then $R \times C = N$). Then a pixel in row r where $1 \leq r \leq R$ and column c where $1 \leq c \leq C$ is linked with the eight pixels surrounding it and the pixels on the sides and in the corners of the image are linked with less than eight neighbouring pixels. Figure 4.2.2 illustrates these linkages and demonstrates three type of cliques: π_1 's clique is marked in blue, π_{C+3} 's clique is marked in orange and π_{3C+1} 's clique is depicted in green. In a similar fashion, a clique corresponding to each pixel labels can be defined.


 Figure 4.2.2: Markov random field illustration for the j^{th} cluster

The mixture density function is given by

$$f(\mathbf{y}|\Theta) = \sum_{j=1}^K \pi_{ij} p(\mathbf{y}|\theta_j)$$

where $\Theta = \left\{ \{\pi_{ij}\}_{i=1}^N, \theta_j \right\}_{j=1}^K$, $\pi_{ij} \geq 0$. The π_{ij} 's are said to follow a Gibbs distribution with density function

$$p(\pi) = \frac{1}{Z} \exp \left(\frac{\sum_{i=1}^N \mathcal{V}_{\mathcal{N}_i}(\pi)}{T} \right).$$

Relating this back to the Gibbs distribution defined in Equation (4.2.2): the energy function is given by $U(\pi) = \sum_{i=1}^N \mathcal{V}_{\mathcal{N}_i}(\pi)$, where \mathcal{N}_i are the cliques demonstrated in Figure 4.2.2 and there are as many cliques as there are pixels (N). That means, that each clique (also called neighbourhood) includes the pixel labels of the eight linked neighbouring pixels for the j^{th} cluster. $\mathcal{V}_{\mathcal{N}_i}(\pi)$ represents the clique potential functions which are given by $\mathcal{V}_{\mathcal{N}_i}(\pi) = \sum_{m \in \mathcal{N}_i} (\pi_{ij} - \pi_{mj})^2$ according to the Gaussian-Markov random field formulation with different variances at each cluster (ξ_j) [40]. Therefore, spatial dependency is modelled by the Markov random field prior fitted to the data label mixing probabilities.

$Z = \xi_j^{-N}$ is a normalising constant and $T = 2\xi_j^2$ is the regularisation parameter. The factorised prior distribution function of the π_{ij} 's is then given by

$$p(\pi) = \prod_{j=1}^K \xi_j^{-N} \exp\left(-\frac{\sum_{i=1}^N \sum_{m \in \mathcal{N}_i} (\pi_{ij} - \pi_{mj})^2}{2\xi_j^2}\right).$$

This prior enforces smoothness of different degrees at each cluster (ξ_j) and can be estimated directly from the data. The details of this derivation are found in Nikou et al [40]. It is assumed that the errors of the labels for all spatial locations and clusters $j = 1, 2, \dots, K$ are independently identically distributed $N(0, \xi_j^2)$ random variables. For this model where a prior $p(\pi)$ is defined over the pixel labels, Bishop states that the EM algorithm can be used to find the Maximum *a Posteriori* (MAP) solutions. In this case the E-step remains the same as in the traditional maximum likelihood case and the Q -function to be maximised in the M-step is given by $Q(\Theta; \mathbf{Y}) = Q(\Theta^{old}; \mathbf{Y}) + \log p(\pi)$ [3]. The complete likelihood (Q -function) to be maximised (also called the Maximum *a Posteriori* -MAP-function) is then given by

$$\begin{aligned} Q(\Theta; \mathbf{Y}) &= \sum_{i=1}^N \sum_{j=1}^K \gamma_{ij} [\log(\pi_{ij}) + \log(p(\mathbf{y}|\theta_j))] + \sum_{j=1}^K \left[-N \log \xi_j - \frac{\sum_{i=1}^N \sum_{m \in \mathcal{N}_i} (\pi_{ij} - \pi_{mj})^2}{2\xi_j^2} \right] \\ &= \sum_{i=1}^N \sum_{j=1}^K \left\{ \gamma_{ij} [\log(\pi_{ij}) + \log(p(\mathbf{y}|\theta_j))] - \log(\xi_j) - \frac{\sum_{m \in \mathcal{N}_i} (\pi_{ij} - \pi_{mj})^2}{2\xi_j^2} \right\}. \end{aligned}$$

Estimation of the mixture parameters $(\beta_j, \Sigma_j) \in \Theta$ remains the same as in the traditional mixture of Gaussian regressions model detailed in Section 3. The estimation of the probabilities of the pixel labels π_{ij} is not as straightforward, keeping in mind that the maximisation procedure should take the following constraints into account: $0 \leq \pi_{ij} \leq 1$ for all i and j and $\sum_{j=1}^K \pi_{ij} = 1$ for all i . Blekas et al presented a projection algorithm to achieve this, which involves projecting the gradient of the MAP function onto the hyperplane of the constraints after which a line search is performed along the direction of the projected gradient to find the label parameters $\{\pi_{ij}\}$ that maximise the Q -function [5].

First take the derivative of the Q -function with respect to π_{ij} and set equal to zero:

$$\begin{aligned}
\frac{\delta}{\delta\pi_{ij}}Q(\Theta; \mathbf{Y}) &= \frac{\delta}{\delta\pi_{ij}} \sum_{i=1}^N \sum_{j=1}^K \left\{ \gamma_{ij} [\log(\pi_{ij}) + \log(p(\mathbf{y}|\theta_j))] - \log(\xi_j) - \frac{\sum_{m \in \mathcal{N}_i} (\pi_{ij} - \pi_{mj})^2}{2\xi_j^2} \right\} \\
&= \gamma_{ij} \frac{\delta}{\delta\pi_{ij}} \log(\pi_{ij}) + 0 - 0 - \frac{\delta}{\delta\pi_{ij}} \left(\frac{\sum_{m \in \mathcal{N}_i} (\pi_{ij} - \pi_{mj})^2}{2\xi_j^2} \right) \\
&= \frac{\gamma_{ij}}{\pi_{ij}} - \frac{\delta}{\delta\pi_{ij}} \left(\frac{\sum_{m \in \mathcal{N}_i} (\pi_{ij}^2 - 2\pi_{ij}\pi_{mj} + \pi_{mj}^2)}{2\xi_j^2} \right) = 0 \\
&\Rightarrow \frac{\gamma_{ij}\xi_j^2}{|\mathcal{N}_i|} - \frac{\pi_{ij}}{|\mathcal{N}_i|} \cdot \left(|\mathcal{N}_i| 2\pi_{ij} - 2 \sum_{m \in \mathcal{N}_i} \pi_{mj} \right) = 0 \\
&\Rightarrow -\pi_{ij}^2 + \pi_{ij}\tilde{\pi}_{ij} + \frac{\gamma_{ij}\xi_j^2}{|\mathcal{N}_i|} = 0 \\
&\Rightarrow \pi_{ij}^2 - \pi_{ij}\tilde{\pi}_{ij} - \frac{\gamma_{ij}\xi_j^2}{|\mathcal{N}_i|} = 0 \tag{4.2.3}
\end{aligned}$$

where $\tilde{\pi}_{ij} = \frac{1}{|\mathcal{N}_i|} \sum_{m \in \mathcal{N}_i} \pi_{mj}$; the mean of the j^{th} cluster probability in the neighbourhood of π_{ij} .

To solve $\frac{\delta}{\delta\pi_{ij}}Q(\Theta; \mathbf{Y}) = 0$, the roots to quadratic Equation (4.2.3) need to be found. Select the root with the positive sign since it yields to constraint $\pi_{ij} \geq 0$:

$$\pi_{ij} = \frac{\tilde{\pi}_{ij} + \sqrt{\tilde{\pi}_{ij}^2 + 4 \frac{\gamma_{ij}\xi_j^2}{|\mathcal{N}_i|}}}{2}. \tag{4.2.4}$$

It should be noted that the neighbourhood \mathcal{N}_i can contain updated π_{ij} and “non-updated” π_{ij} ’s. The solution in Equation (4.2.4) however is not final since it does not adhere to the constraints $0 \leq \pi_{ij} \leq 1$ for all i and j and $\sum_{j=1}^K \pi_{ij} = 1$ for all i . These constraints define a convex hull⁴; thus after calculating the updated π_{ij} using Equation (4.2.4), we project them onto the convex hull (i.e., the constraints). Blekas et al describe an efficient quadratic programming algorithm for this purpose [5].

Denote a_{ij} ($j = 1, 2, \dots, K$) the label parameter values ($\pi_{ij} \geq 0$) calculated from Equation (4.2.4). Given a vector $\mathbf{a}_i \in \mathbb{R}^K$ with $a_{ij} \geq 0$ and the hyperplane $\sum_{j=1}^K x_j = 1$, the point on the hyperplane with non negative components that is closest to \mathbf{a}_i needs to be found. This problem can be formulated as a linear constrained convex quadratic programming problem for each pixel ($i = 1, 2, \dots, N$):

$$\begin{aligned}
\min_x \sum_{j=1}^K (x_j - a_j)^2 \quad \text{subject to} \quad & \sum_{j=1}^K x_j = 1 \\
& \text{and } x_j \geq 0 \quad \forall j = 1, 2, \dots, K.
\end{aligned}$$

The quadratic programming problem formulated above can be solved using a few approaches; here a Lagrange multiplier method is used. See the Lagrange function below:

⁴a convex hull of a set of points X in a Euclidean space is the smallest convex set that contains X

$$L(x, \lambda_0, \lambda_j) = \frac{1}{2} \sum_{j=1}^K (x_j - a_j)^2 - \lambda_0 \left(\sum_{j=1}^K x_j - 1 \right) - \sum_{j=1}^K \lambda_j x_j$$

where λ_0 is the multiplier for the equality and λ_j ($j = 1, 2, \dots, K$) are the multipliers for the inequality constraints.

First order necessary conditions imply

$$x_j = a_j + \lambda_0 + \lambda_j. \quad (4.2.5)$$

Combining Equation (4.2.5) with the constraint $\sum_{j=1}^K x_j = 1$ we get

$$\begin{aligned} \sum_{j=1}^K a_j + \sum_{j=1}^K \lambda_0 + \sum_{j=1}^K \lambda_j &= 1 \\ K\lambda_0 &= 1 - \sum_{j=1}^K a_j - \sum_{j=1}^K \lambda_j \\ \lambda_0 &= \frac{1}{K} - \frac{1}{K} \sum_{j=1}^K a_j - \frac{1}{K} \sum_{j=1}^K \lambda_j, \end{aligned} \quad (4.2.6)$$

and substituting Equation (4.2.6) back into Equation (4.2.5), we get

$$x_j = \frac{1}{K} + a_j - \frac{1}{K} \sum_{j=1}^K a_j + \lambda_j - \frac{1}{K} \sum_{j=1}^K \lambda_j. \quad (4.2.7)$$

Note that $b_j = \frac{1}{K} + a_j - \frac{1}{K} \sum_{j=1}^K a_j$ is a projection of a_j onto hyperplane $\sum_{j=1}^K x_j = 1$.

The λ_j 's must be chosen so as to satisfy the inequality constraints. Kuhn-Tucker conditions state that at the minimiser x^* , $\lambda_j \geq 0$, $\lambda_j > 0$ if $x_j^* = 0$ (active constraint), $\lambda_j x_j^* = 0$ [5]. Given this information, Blekas et al [5] present a very efficient iterative strategy for calculating the λ_j 's for the problem:

1. Let x denote the vector at the current iteration.
2. Initially, set $x_j = b_j \forall j$.
3. In general there exist m negative components in x_j .
4. Define corresponding set of indices $S = \{j, \text{ where } x_j < 0\}$ (finding the active set of constraints for the current vector x).

- (a) for all $j \notin S$: $\lambda_j = 0$
- (b) for all $j \in S$: $x_j = x_j^* = 0$. Then the corresponding λ_j is calculated by solving an $m \times m$ linear system that force the inequalities to be satisfied as equalities: $x_j + \lambda_j - \frac{1}{K} \sum_{j=1}^K \lambda_j = 0$, leading to ⁵

$$\lambda_j = \frac{1}{m-k} \sum_{k \in S} x_k - x_j$$

- (c) calculate the updated x_j values for $j \notin S$ using the new vector $\boldsymbol{\lambda}$ via Equation (4.2.7).

5. These steps are repeated until a feasible point is obtained (i.e., $x_j \geq 0 \forall j$ as desired).

It is important to note, that once a x_j becomes zero, it remains so.

Finally, the Maximum *a Posteriori* (MAP) EM algorithm for the model formulation is given below:

Algorithm 4.1 MAP EM algorithm for the K -component spatial variant mixture of Gaussian regressions model

1. Choose appropriate starting values for the parameters, $\boldsymbol{\Theta}$.
2. *Expectation step*: calculate the responsibilities.

$$\hat{\gamma}_{ij} = \frac{\pi_{ij} \cdot p(\mathbf{y}_i | \boldsymbol{\theta}_j)}{\sum_{s=1}^K \pi_{is} p(\mathbf{y}_i | \boldsymbol{\theta}_s)}$$

3. *Maximisation step*: calculate the maximum likelihood estimates:

$$\hat{\boldsymbol{\beta}}_j = \frac{\left(\sum_{i=1}^N \hat{\gamma}_{ij} \mathbf{y}_i' \mathbf{X} \right) (\mathbf{X}' \mathbf{X})^{-1}}{\sum_{i=1}^N \hat{\gamma}_{ij}}$$

$$\hat{\sigma}_{ju}^2 = \frac{\sum_{i=1}^N \hat{\gamma}_{ij} \left(y_{iu} - [\mathbf{X} \hat{\boldsymbol{\beta}}_j]_u \right)^2}{\sum_{i=1}^N \hat{\gamma}_{ij}}$$

$$\hat{\xi}_j^2 = \frac{1}{N} \sum_{i=1}^N \sum_{m \in \mathcal{N}_i} (\hat{\pi}_{ij} - \hat{\pi}_{mj})^2$$

Calculating the mixing probability estimates involves another iterative algorithm as set out in the discussion above.

4. Repeat step 2 and 3 until convergence is achieved.
-

4.3 Example - spatial variant mixture of Gaussian regressions

To illustrate how the spatial variant mixture of Gaussian regressions model discussed above works,

⁵Note, that the k in this formula is a counter and does not refer to the number of mixture components: K

an example was designed by generating spatiotemporal data. A 16×16 pixel image was designed and divided into 8 sections which were randomly assigned to $K = 3$ clusters. For each cluster, the pixel intensity was set to increase and / or decrease along a pre-defined polynomial function for ten time steps ($t = 1, 2, \dots, 10$) to simulate “pixel intensity change over time”. The spatial variant mixture of regressions model was then used to identify and estimate the component regression models.

4.3.1 The generated data

A 16×16 pixel image was designed to consist of 3 clusters illustrated in white, blue and red in Figure 4.3.1.

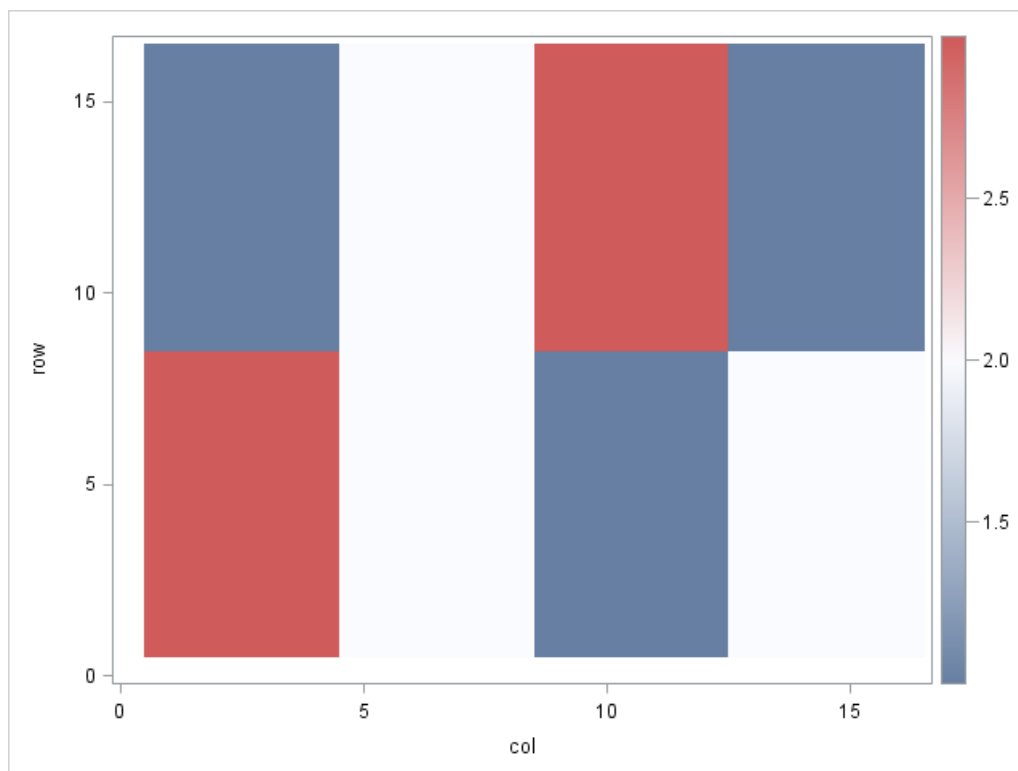


Figure 4.3.1: Designed 16×16 image

The pixel intensities were then adjusted over 10 time steps along predefined second-degree polynomial regression functions given below. The pixels in the white area in Figure 4.3.1 were set to follow the function \mathbf{Y}_1 , the blue area was set to follow \mathbf{Y}_2 and the red area was set to follow the function \mathbf{Y}_3 .

$$\mathbf{Y}_1 = 0.85t - 0.10t^2 \quad (4.3.1)$$

$$\mathbf{Y}_2 = -0.5t + 0.06t^2$$

$$\mathbf{Y}_3 = -0.05t + 0.02t^2$$

The change of the pixel intensities over time is illustrated in Figure 4.3.2 below and the colouring of the graph matches that of Figure 4.3.1 (i.e. the pixels in the red areas in Figure 4.3.1 were set to change over time according to the red line in Figure 4.3.2).

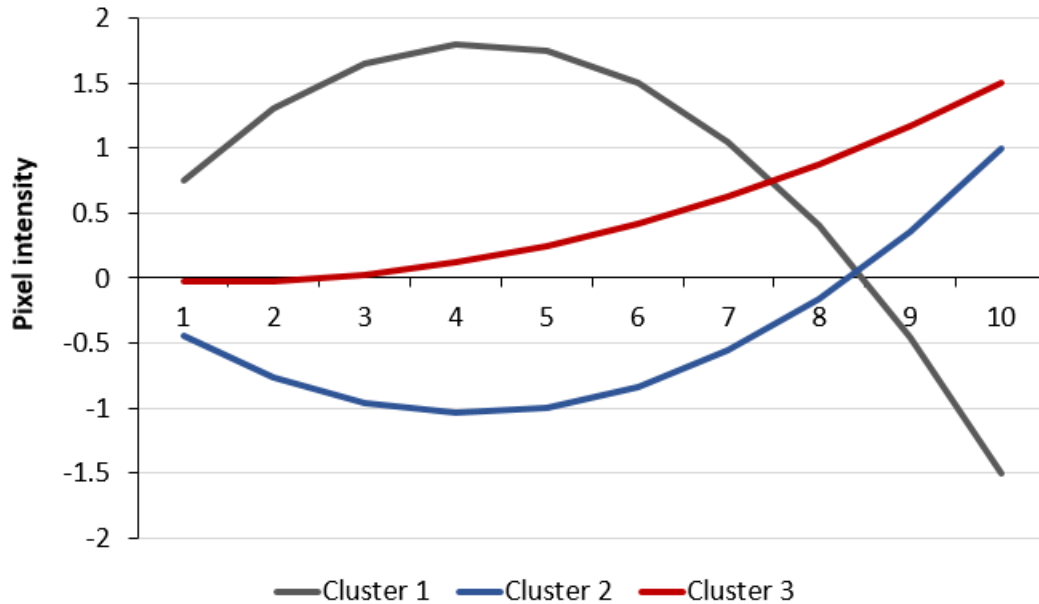
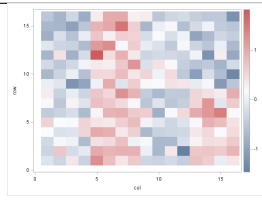
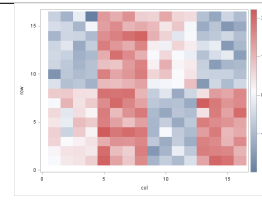
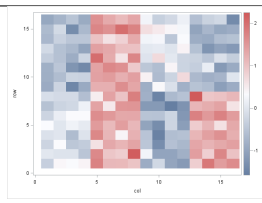
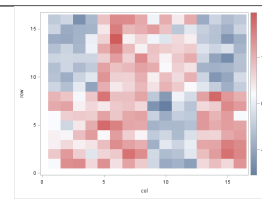
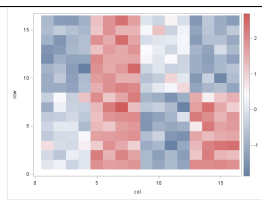
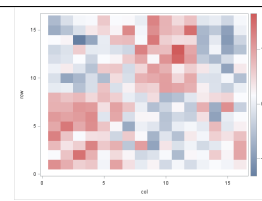
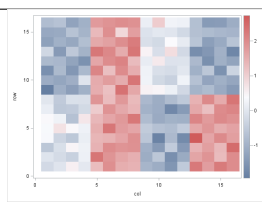
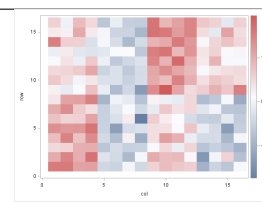
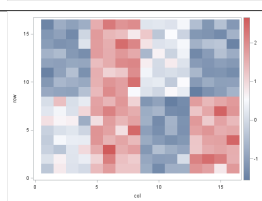
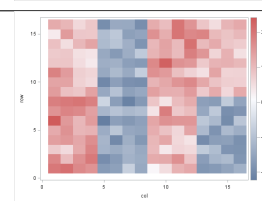


Figure 4.3.2: Pixel changes over time - 3 models

Gaussian white noise was added to each time-step and the resulting spatiotemporal data from time $t = 1$ to time $t = 10$ is shown in Table 4.3.1. It is clear that the simulated image was divided into $K = 3$ groups with the regions as indicated in Figure 4.3.1 that change according to the functions for \mathbf{Y}_1 , \mathbf{Y}_2 and \mathbf{Y}_3 . The three clusters of different pixel intensity changes over time are clearly distinguishable despite the Gaussian white noise. In the next section the spatial variant mixture of Gaussian regressions model will be fitted to the simulated spatiotemporal data.

Table 4.3.1: Simulated image with three clusters over time

t	Simulated image	t	Simulated image
1		6	
2		7	
3		8	
4		9	
5		10	

4.3.2 The model to be fitted

The model to be fitted to the data is a spatial variant of the finite mixture of Gaussian regressions model. It is known that the spatiotemporal data contains $K = 3$ clusters and that for each pixel a sequence of pixel intensities for $t = 1, 2, \dots, 10$ is observed. Therefore the generative form of the model to be fitted is given by $\mathbf{Y} = \Delta_1 \mathbf{Y}_1 + \Delta_2 \mathbf{Y}_2 + (1 - \Delta_1 - \Delta_2) \mathbf{Y}_3$ where

$$\mathbf{Y}_1 = \beta_{11} \mathbf{t} + \beta_{12} \mathbf{t}^2 + \mathbf{e}_1$$

$$\mathbf{Y}_2 = \beta_{21} \mathbf{t} + \beta_{22} \mathbf{t}^2 + \mathbf{e}_2$$

$$\mathbf{Y}_3 = \beta_{31} \mathbf{t} + \beta_{32} \mathbf{t}^2 + \mathbf{e}_3.$$

\mathbf{Y} , \mathbf{Y}_1 , \mathbf{Y}_2 , \mathbf{Y}_3 are (10×1) dimensional vectors of pixel intensity sequences for each observation and \mathbf{t} is a (10×1) vector containing time steps, $\mathbf{e}_j \sim i.i.d. N(0, \boldsymbol{\Sigma}_j)$ and $P(\Delta_j = 1) = \pi_j$. The model can be rewritten in matrix notation:

$$\mathbf{Y}_j = \mathbf{X}\boldsymbol{\beta}_j + \mathbf{e}_j$$

where

$$\mathbf{X} = \begin{bmatrix} t_1 & t_1^2 \\ t_2 & t_2^2 \\ t_3 & t_3^2 \\ t_4 & t_4^2 \\ t_5 & t_5^2 \\ t_6 & t_6^2 \\ t_7 & t_7^2 \\ t_8 & t_8^2 \\ t_9 & t_9^2 \\ t_{10} & t_{10}^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 4 \\ 3 & 9 \\ 4 & 16 \\ 5 & 25 \\ 6 & 36 \\ 7 & 49 \\ 8 & 64 \\ 9 & 81 \\ 10 & 100 \end{bmatrix}$$

$$\boldsymbol{\beta}_j = \begin{bmatrix} \beta_{j1} \\ \beta_{j2} \end{bmatrix}$$

and $\mathbf{Y}_j \sim i.i.d. N(\mathbf{X}\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_j)$ for $j = 1, 2, 3$. Thus far, the model describes the traditional finite mixture of multivariate Gaussian regressions model with $p = 2$ explanatory variables, $K = 3$ clusters and $N = 256$ observations of $T = 10$ dimensional vectors, where each observation (in this case, sequence of pixel intensities at each pixel) is classified into group j with probability π_j .

The model incorporates spatial location information, also called spatial dependency information, by estimating $K = 3$ mixing probabilities for each pixel with the help of a Markov random field. This model specification captures the notion that neighbouring pixels are likely to be classified into the same cluster and allows for varying levels of smoothness (or variance) between clusters. Therefore the spatial variant generative form of the model is given by $\mathbf{Y} = \Delta_{i1}\mathbf{Y}_1 + \Delta_{i2}\mathbf{Y}_2 + (1 - \Delta_{i1} - \Delta_{i2})\mathbf{Y}_3$, $i = 1, 2, \dots, N$ where $P(\Delta_{ij} = 1) = \pi_{ij}$.

Let the mixing probabilities for the j^{th} cluster be random variables representing the nodes of a Markov random field. The links between nodes are defined such that the neighbouring pixels of pixel i are linked (usually the 8 adjacent pixels - note that for i in the corners and along the sides of the image less than 8 pixels will be linked). Therefore, the i^{th} clique of the Markov random field \mathcal{N}_i contains the 8 (or less) adjacent pixels to pixel i . Following the theory discussed in Section 4.2.1, the mixing probabilities follow a Gibbs distribution.

$$\pi_{ij} \sim \prod_{j=1}^K \xi_j^{-N} \exp \left(\frac{\sum_{i=1}^N \sum_{m \in \mathcal{N}_i} (\pi_{ij} - \pi_{mj})^2}{2\xi_j^2} \right)$$

It is assumed that the errors of the mixing probabilities for clusters $j = 1, 2, \dots, K$ are i.i.d. $N(0, \xi_j^2)$ distributed. The Maximum *a Posteriori* EM algorithm as detailed in Algorithm 4.1 was used to estimate the model.

4.3.3 The estimated model

The parameters $\Theta_j = \{\beta_j, \Sigma_j, \pi_{ij}, \xi_j^2\}$ for $j = 1, 2, 3$ were estimated using starting values, selected as follows:

- For β_j , $K = 3$ random observations of \mathbf{Y} were selected and the corresponding starting values were calculated: $\beta_j = (\mathbf{X}'\mathbf{X})^{-1} \cdot (\mathbf{X}'\mathbf{Y})$.
- It is assumed that Σ_j is a diagonal matrix with $\frac{1}{K} \sum_{i=1}^N (\mathbf{Y} - \bar{\mathbf{Y}})^2$ on the main diagonal as starting values for $j = 1, 2, 3$.
- $\pi_{ij} = \frac{1}{K} = \frac{1}{3}$ for $i = 1, 2, \dots, N$ and $j = 1, 2, 3$.
- $\xi_j^2 = \frac{1}{K} = \frac{1}{3}$ for $j = 1, 2, 3$.

After convergence of the EM algorithm, the estimated regression models are given by

$$\begin{aligned} \mathbf{Y}_1 &= 0.83\mathbf{t} - 0.10\mathbf{t}^2 \\ \mathbf{Y}_2 &= -0.47\mathbf{t} + 0.06\mathbf{t}^2 \\ \mathbf{Y}_3 &= -0.08\mathbf{t} + 0.025\mathbf{t}^2, \end{aligned} \tag{4.3.2}$$

where the variances Σ_j are assumed to be diagonal matrices with the following vectors on the main diagonals:

$$\begin{bmatrix} 0.010 & 0.000 & 0.005 \\ 0.010 & 0.007 & 0.009 \\ 0.008 & 0.002 & 0.002 \\ 0.010 & 0.006 & 0.005 \\ 0.012 & 0.003 & 0.009 \\ 0.017 & 0.004 & 0.003 \\ 0.007 & 0.010 & 0.019 \\ 0.008 & 0.002 & 0.004 \\ 0.006 & 0.002 & 0.003 \\ 0.010 & 0.015 & 0.004 \end{bmatrix}$$

and the estimated variances of the mixing probabilities are given by

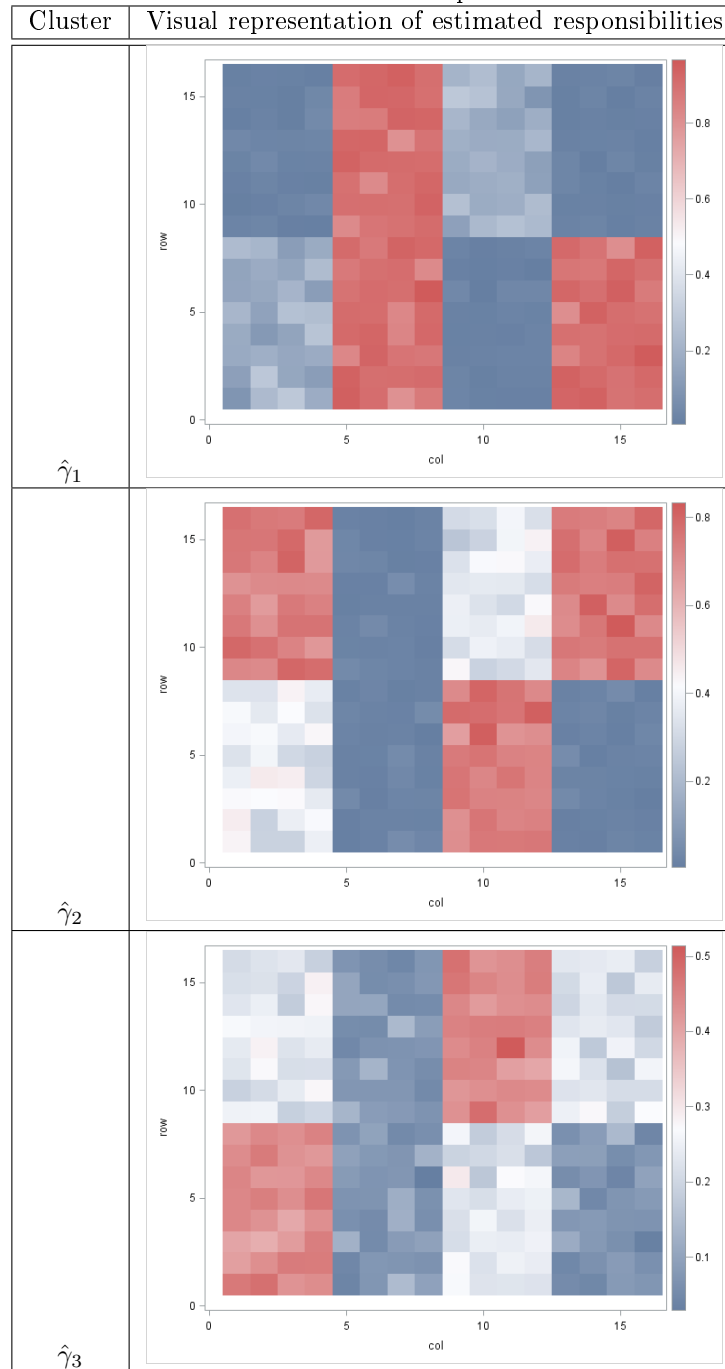
$$\begin{aligned} \xi_1^2 &= 0.0000931 \\ \xi_2^2 &= 0.0000592 \\ \xi_3^2 &= 9.533 \times e^{-6}. \end{aligned}$$

Note that label switching has been accounted for so that \mathbf{Y}_1 in Equation (4.3.2) corresponds to \mathbf{Y}_1 in Equation (4.3.1) etc. The estimated parameters are reasonably close to the theoretical parameters and it can be inferred that the model specification and algorithm accurately estimated the three different models according to which pixel intensities changed over time.

The estimated responsibilities ($\gamma_{ij} = E(\Delta_{ij})$, $i = 1, 2, \dots, N$ and $j = 1, 2, 3$) and mixing probabilities (π_{ij}) are best illustrated graphically as set out in Table 4.3.2 and 4.3.3 and discussed below.

A given pixel is assigned to the cluster for which the estimated responsibility is the largest therefore the estimated responsibilities indicate how well the underlying clusters were identified. The estimates for γ_1 are highest (red) for the areas that were indicated white in Figure 4.3.1, indicating that the correct areas were identified as cluster 1. Similarly, the estimates for γ_2 and γ_3 are highest for the areas that were indicated as blue and red in Figure 4.3.1 respectively. It is interesting to note that the estimates of γ_2 in the areas known to be cluster 3 and the estimates of γ_3 in the areas known to be cluster 2 are not as close to 0 as the estimates of γ_1 in areas known to be cluster 2 and 3. This observation indicates that pixels in cluster 2 and cluster 3 contribute to the estimation of the parameters associated with both cluster 2 and 3 to some extent (as is typical for a fuzzy clustering method).

Table 4.3.2: Estimated responsibilities



Even though the mixing probabilities don't differ a lot across clusters (see the scale in Table 4.3.3) - the colouring in the visual representations does show clearly that the estimated mixing probabilities for clusters 1 ($\hat{\pi}_1$), 2 ($\hat{\pi}_2$) and 3 ($\hat{\pi}_3$) are highest in the areas known to be clusters 1, 2 and 3 respectively. Therefore the probability of belonging to cluster j is higher in locations known to belong to cluster j , for $j = 1, 2, 3$.

Table 4.3.3: Estimated mixing probabilities

Cluster	Visual representation of the estimated mixing probabilities
$\hat{\pi}_1$	
$\hat{\pi}_2$	
$\hat{\pi}_3$	

The estimated responsibilities and mixing probabilities indicate that the model specification and applied algorithm identified the clusters correctly.

Another typical useful illustration of the clustering results of finite mixture models is to plot the dependent variable against the independent variables and colouring the observations by cluster. This example however, deals with multivariate dependent (sequence of pixel changes over time) and

independent variables (time steps $t = t_1, t_2, \dots, t_{10}$) therefore, this representation of the clustering results is adjusted as follows. In Figure 4.3.3 the estimated mean pixel change over time is shown in thick dotted lines for cluster 1 (black), cluster 2 (blue) and cluster 3 (red). Additionally, five observations from each cluster (pixel intensity sequences that were classified according to the maximum estimated responsibilities) were randomly selected, plotted on the same graph and coloured according to the classified cluster. This graph aims to demonstrate that the clusters are not only spatially correctly classified (see Tables 4.3.2 and 4.3.3) but are also clearly distinguishable in a “dependent vs independent variable” plot. For example, the observations that were classified to be in cluster 1 (black) roughly follow the same pattern as the estimated conditional mean for cluster 1 and are clearly “grouped”.

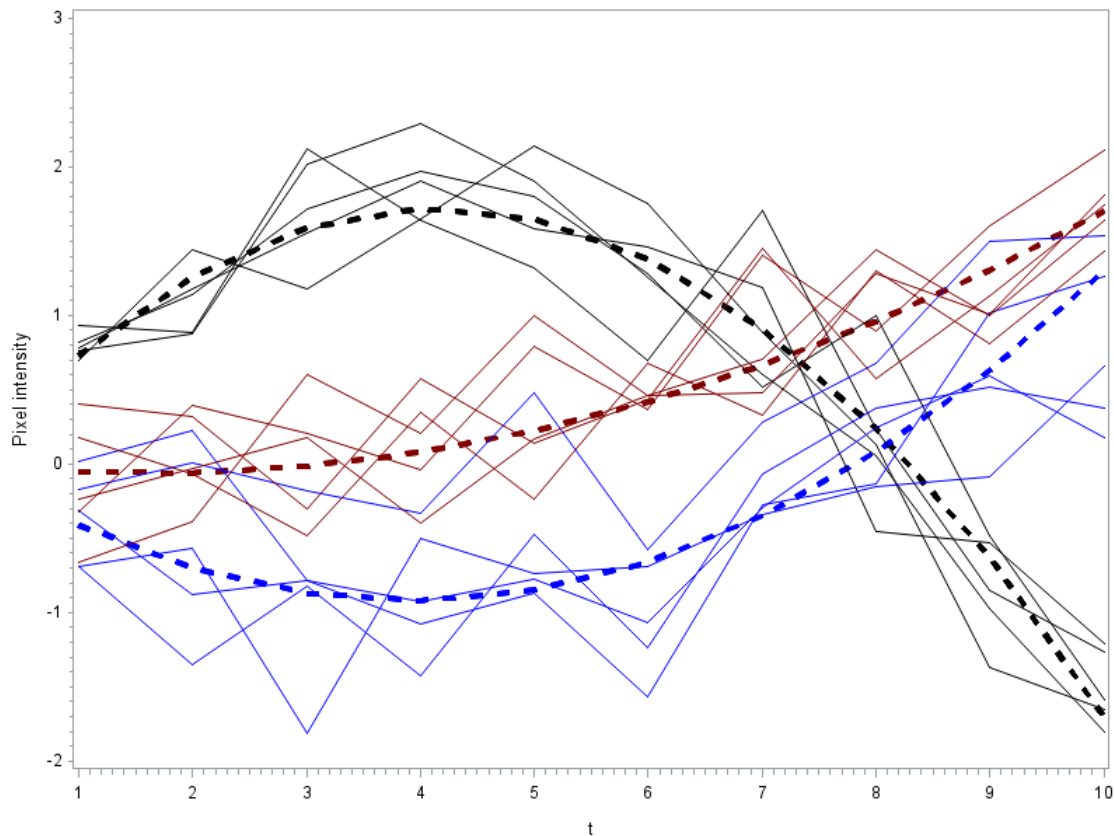


Figure 4.3.3: Illustrating the clustering results

This example therefore demonstrated that the finite spatial variant mixture of regressions model (with Gaussian error terms) simultaneously identifies underlying clusters correctly and estimates the regression models accurately, for spatiotemporal data. In the following chapter this model will be applied to maize yield data in the Free State, South Africa.

Chapter 5

Application: maize yields

Agricultural production is significantly affected by environmental factors (i.e., weather, temperature and precipitation) which influence crop growth and development and cause large intra-seasonal yield variability; furthermore, crop agronomic management (i.e., fertilizer application, planting irrigation, tillage etc.) can offset weather related yield losses [1]. This study argues that agronomic management practices and microclimates¹ can cause some farmers to mitigate climate variability more efficiently than others. Neither management practices nor microclimates can be directly observed given available yield and gridded spatial data. The present study therefore aims to identify groupings or patterns in maize yield data caused by these unobservable variables.

Environmental factors including the quality and type of natural resources and climatic conditions vary spatially; therefore, these factors clearly introduce a spatially dependent aspect to the problem. Neighbouring farmers, or farms within the same proximity, are more likely to participate in interactions (e.g., study groups or learning events) where management practices are discussed or advice is exchanged, than farmers separated by large distances. Similarly, unobservable microclimates or soil-plant-atmosphere continuums, naturally occur within a spatial proximity. The spatial orientation of agronomic management practices, microclimates and gridded precipitation data (covariate) suggest that the spatial variant mixture of regressions model is a suitable model to identify spatially specific clusters of maize yields (where nearest-neighbour yield observations are more likely to be found in the same cluster) and estimate their respective regression functions.

The aim of this study is to estimate maize yields in the Free State province as a function of type of maize produced (yellow or white), precipitation and season (harvesting year) while simultaneously identifying spatially explicit clusters of similar yield functions by fitting the spatial variant mixture of regressions model. The maize yield data, type of maize and the season in which it was produced was obtained from the Department of Agriculture Forestry and Fisheries' objective yield survey database. Precipitation data was obtained from the Tropical Rainfall Measurement Mission (TRMM) database, disseminated by the Goddard Earth Sciences Data and Information Services Center (GES DISC). It is important to note that large intra-annual variability of seasonal precipitation, especially in rain-scarce regions, has caused crop model errors and has previously led to difficulties in including precipitation

¹The weather in a particular small area, especially when this is different from the weather in the surrounding area. [30]

in statistical crop models [53, 46]; in this regard this study has been no different.

The estimation of agricultural yields is also known as crop modelling and various types of crop models are briefly discussed to contextualise this study. The data used in this study is discussed in more detail and followed by exploratory analysis. Finally, a simple linear regression model and traditional mixture of Gaussian regressions model are fitted to the data in order to facilitate comparisons with the fitted spatial variant mixture of Gaussian regressions model. The EM algorithm converged to a solution with three distinct spatial clusters; however, it is important to note that the clusters differ only slightly in terms of parameter estimates and further investigation is warranted.

5.1 Types of crop models and this study

Different types of crop yield forecasting models have been developed in literature and can be categorised as follows:

- Crop Simulation Models (CSM) are computerised representations that simulate of crop growth, development and yield through mathematical equations as functions of soil conditions, weather and management practices [29]. Their strength lies in the ability of CSM's to extrapolate the temporal patterns of crop growth and yield beyond a single experimental site. CSM's are useful in gaining scientific insight into crop physiological processes and evaluating the impact of agronomic practices. Applications listed by Hogenboom include impacts of global warming, crop response to sowing dates and spacing, characterisations of production environments and regional targeting of technologies. Extensive input data on cultivar, management practices and soil conditions, unavailable in many parts of the world, is required for CSM's [38].
- Mechanistic Models use fundamental mechanisms of plant and soil processes to simulate specific outcomes, and are often used for academic purposes to gain a better understanding of specific plant-related processes and interactions, rather than for problem solving purposes [1].
- Functional Models to simulate complex processes through simplified approaches and are often simplified versions or parts of the mechanistic models (e.g., a plant's photosynthesis process whereby solar radiation is converted to energy can be depicted by a functional model: energy is a function of, among other factors, solar radiation).
- Statistical Models for crop yield forecasting are simple and less parameter-intensive; in essence a simple statistical model is built using historical yield data and several agro-meteorological parameters (e.g. temperature and rainfall). Statistical models however are limited in the information they can provide beyond the range of values for which the model is parameterised; they do not take the timing of stresses occurring during crop growth into account and are also incapable of giving farmers important agronomic advice [1]. Because of variations in soils, landscapes and weather, beyond what is included in the population information used to estimate the statistical information, the results of statistical models generally cannot be extrapolated in space and time.

The spatial variant mixture of regressions model to be fitted in this study, is classified as a statistical model. Advantages of statistical crop forecasting models include their limited reliance on field cali-

bration data and the transparent assessment of model performance (e.g. coefficient of determination and confidence intervals around model coefficients and predictions); note that similar statistics can be combined with process-based models (i.e., CSM, mechanistic and functional models) but in practice rarely are [38]. Serious shortcomings of statistical crop forecasting models include problems with collinearity (e.g. temperature and precipitation) and assumptions of stationarity in the case of time-series models [38]. Lobell asserts that statistical crop forecasting models based on temporal and or or spatial variation in crop yields are widely used to investigate the impact of climate changes (recent and future) on crop yields.

5.2 The data

It is well established that in modelling, simulating or forecasting maize yields various input variables including weather data (precipitation, temperature, solar radiation and wind speed), crop data (variety and growth attributes), soil data (thickness of soil, pH, sand and clay percentages etc.), and crop management data (planting date, seed rate, irrigation, fertiliser application etc.) are typically used depending on the complexity of the model [1]. The data sets used in this particular statistical crop model are restricted to precipitation and information accompanying the objective yield survey data and are discussed below.

Objective yield surveys for maize are conducted annually in the South African maize production area in three provinces (Free State, North West and Mpumalanga) by the National Crop Statistics Consortium (NCSC) comprising of the Agricultural Research Council (ARC), SiQ and Geoterraimage. These surveys are aimed at estimating yield by taking in-field measurements during April in Mpumalanga and May in the Free State and North West each year, when the maize has reached physiological maturity. Seven hundred sampling locations (farming units) within the Free State, North West and Mpumalanga provinces are allocated proportional to the total area of cultivation for white maize and yellow maize under dryland or irrigated cultivation. The crop-specific total area of cultivation is determined by the Producer Independent Crop Estimates System (PICES) developed by the NCSC and these figures are released in February each year. Enumerators (ARC) visit each location and follow a predefined sampling methodology designed to ensure randomness in cob selection and eventual yield estimation. Upon arrival at the farming unit location, the average estimated yield (based on five random sampling points) is recorded per identified location (farming unit) with the field GPS coordinates linked to it.

The seven hundred locations to be visited are identified by SiQ from their database of farming units throughout the three provinces. This database is renewed on average every 5 years. Annually, these farming units are contacted telephonically to determine the area planted per crop type, for the given season. The hectares covered by the farming units database are then expanded proportionally to match the total crop type area in each province (as determined by PICES) from which the seven hundred locations are randomly selected. SiQ's methodology selects the seven hundred locations within each province with a probability proportional to size, making it a self-weighting sample (i.e., large farming units will likely have more than one sampling location included in the seven hundred).

The objective yield survey database from harvesting seasons 2004 to 2017 was obtained from the

Department of Agriculture, Forestry and Fisheries (DAFF) containing average maize yields ($Y_{estyield}$)² for yellow and white maize (X_{type}) under dryland or irrigation production (X_{DryIrr}) over the Free State, Mpumalanga and North West provinces ($X_{Province}$), as well as a code (X_{Code}) indicating the farming unit from which the yield observation was recorded and GPS coordinates linked to the field where the sample was taken (X_{Lat} , X_{Long}). The number of realised samples differs each year depending on the availability of farmers and their willingness to co-operate.

The second data set obtained was a precipitation database with a high spatial and temporal resolution. The Tropical Rainfall Measurement Mission (TRMM) database was used, which is disseminated by the Goddard Earth Sciences Data and Information Services Center (GES DISC). The 3B42 product was used, containing a 3-hourly surface precipitation estimate at a spatial resolution of 0.25 degrees (approximately 25 km grid) from 1 July 2000 to 30 June 2017. This database and product has previously been used in agricultural modelling applications [46, 53]. The total rainfall from October to March for each season in the South African summer grain production region was calculated ($X_{SeasonRain}$) and the timing of the rainfall (monthly rainfall) was also processed as possible explanatory variable.

Exploratory analysis was performed on the data and key results are presented in the following section.

5.3 Exploratory analysis

5.3.1 Analysing the data and selecting the appropriate covariates

A frequency table provides an overview of the objective yield database, showing the number of yield observations in the total database for different combinations of white maize, yellow maize, dryland and irrigation production per province (Table 5.3.1).

Table 5.3.1: Frequency table: objective yield data base 2004 - 2017

Province	Production type	White Maize	Yellow Maize	Total
Free State	Dryland	2079	1174	3253
	Irrigation	71	133	204
	Total	2150	1307	3457
Mpumalanga	Dryland	582	975	1557
	Irrigation	115	80	195
	Total	697	1055	1752
North West	Dryland	1942	504	2446
	Irrigation	120	79	199
	Total	2062	583	2645
Grand total		4909	2945	7854

The number of yield observations captured differs each season as illustrated in Table 5.3.2. Note that the grand total in Table 5.3.1 differs from that in Table 5.3.2: the Province indicator was blank for 18 yield observations. These observations were discarded.

²Note that the yield observations are objective yield estimates (as per the explanation above) and that throughout this study “estimated yields” will be used interchangeably with “observed yields” therefore, the yields estimated from the models to be fitted will be referred to as “predicted yields”.

Table 5.3.2: Number of objective yield observations per season

Season (harvesting year)	Number of observations	Mean	Standard deviation
2004	286	3.08	1.86
2005	244	4.02	2.14
2006	438	4.01	2.26
2007	562	3.08	2.46
2008	675	4.41	1.86
2009	748	4.56	2.02
2010	638	4.82	2.24
2011	666	5.05	2.57
2012	582	4.43	2.66
2013	578	4.08	2.75
2014	568	5.43	2.19
2015	604	4.03	3.01
2016	635	3.54	2.73
2017	648	6.59	2.52
Grand total	7872		

The frequency table in Table 5.3.1 is repeated for the 2017 harvesting season in order to further understand the typical composition of the annual objective yield surveys. It is clear that very few irrigation maize yield observations are recorded annually, and since maize production under irrigation is not strictly comparable with dryland production the irrigated yield observations will be excluded from the analysis.

Table 5.3.3: Frequency table: objective yield survey for harvesting season 2017

Province	Production type	White Maize	Yellow Maize	Total
Free State	Dryland	225	74	299
	Irrigation	10	7	17
	Total	235	81	316
Mpumalanga	Dryland	40	94	134
	Irrigation	8	8	16
	Total	48	102	150
North West	Dryland	133	32	165
	Irrigation	9	8	17
	Total	142	40	182
Grand total		425	223	648

The objective yield observation locations (latitudes and longitudes) for all seasons are plotted and white and yellow maize yields are indicated by colour in Figure 5.3.1. This figure can be interpreted like a map with the north-western region representing the North West province, the north-eastern region representing Mpumalanga and the center and southern area coinciding with the Free State. It seems that in the eastern part of the Free State as well as in Mpumalanga yellow maize is typically planted, more so than white maize, and in the North West province and western part of the Free State white maize is preferred.

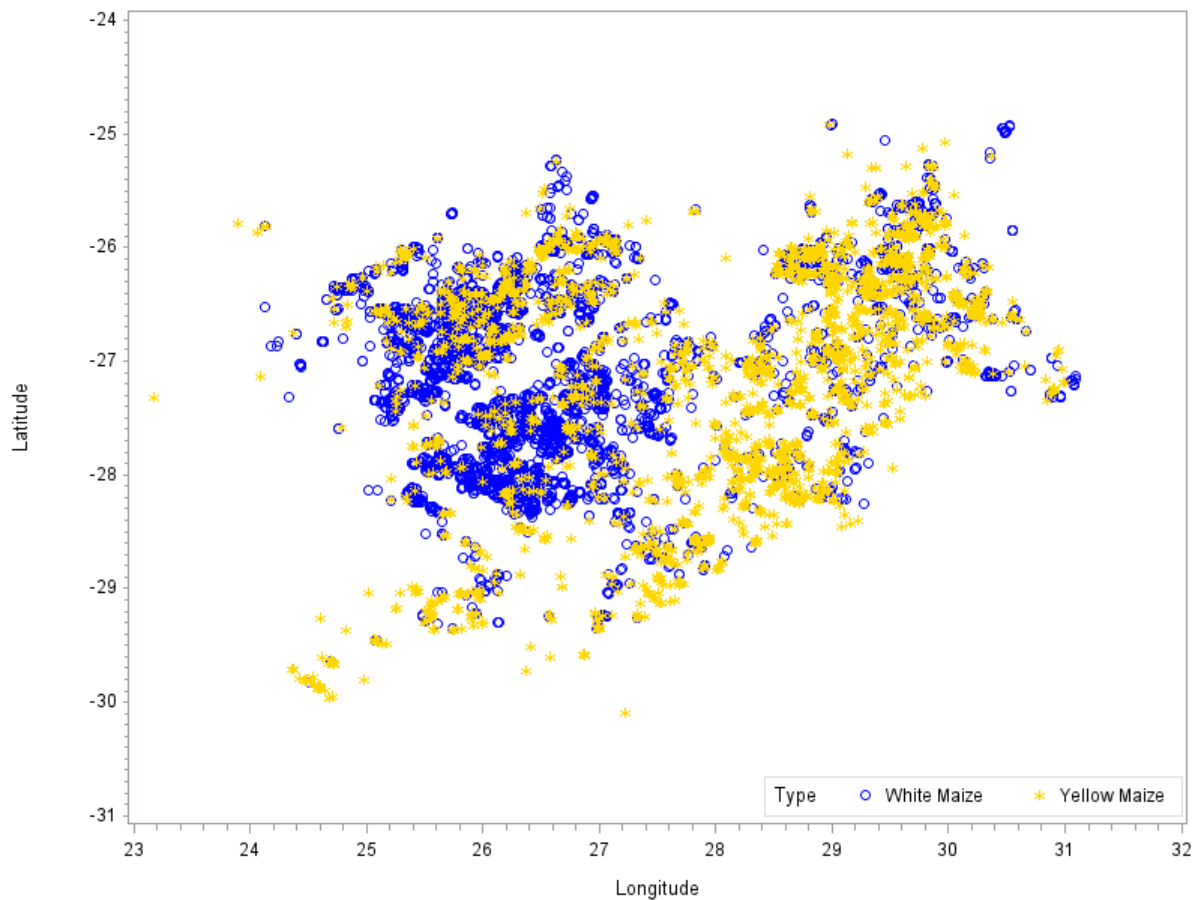


Figure 5.3.1: All yield observation by type of maize

Figure 5.3.1 indicates that the Free State has an interesting spatial combination of white and yellow maize yield observations. Thus, it was decided to restrict this analysis to the dryland yield observations in the Free State province. Therefore, a total of 3253 yield observations of which 2079 are white maize and 1174 are yellow maize yields, will be used. The following exploratory analysis relates to the Free State dryland yield observations.

The farming units are referred to as codes and Figure 5.3.2 shows the number of farming units that were visited $x = 1, 2, 3, \dots$ times (i.e., a farming unit was visited x times over the past 14 years). 38% of the farming units in the Free State were visited twice over the past 14 years of objective yield surveys. 16% of farming units were visited three times and 10% four times while only 6% of farming units in the Free State were visited 9 times or more throughout the past 14 years of objective yield surveys.

These multiple observations per farming unit can be seen as repeated measures, in which case the model needs to account for this structure. However, it can be argued that the observations ($Y_{EstYield}$) are taken at a field level (see Section 5.2) and the locations of these observations are captured in the X_{Lat} and X_{Long} GPS coordinates. More often than not, the GPS coordinates of multiple observations taken on a given farm unit differ, after taking drift into account. Therefore, the repeated measures

structure was not incorporated into the model specification.

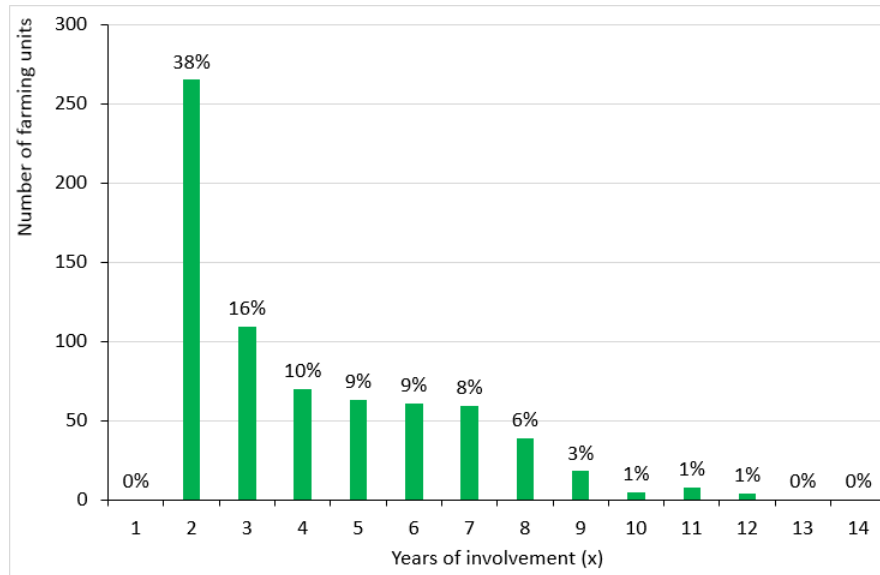


Figure 5.3.2: Number of farming units per years of involvement

The relevant season's total rainfall was linked to the yield observation by performing a spatial join³ between the gridded TRMM precipitation data and the GPS coordinates of the objective yield database (i.e., the October to March rainfall in the season in which the specific yield observation was recorded, was linked to the yield observation as variable $X_{SeasonRain}$). Then histograms of the season's total rainfall were compiled per season (i.e., year in which maize was harvested), see Figure 5.3.3. It is clear that the distribution of total rainfall differs across seasons: sometimes the total rainfall in a season at the yield observation points had a range of less than 400mm (e.g. 2011) whereas in the 2017 season the range is approximately 600mm. The shape of the season rainfall histograms also range from symmetric (e.g. 2010, 2011, 2015) to quite heavy-tailed (e.g. 2005, 2006, 2017).

³R was used to transform the precipitation data sets into a raster format, which was then joined with the GPS locations of the yield observations

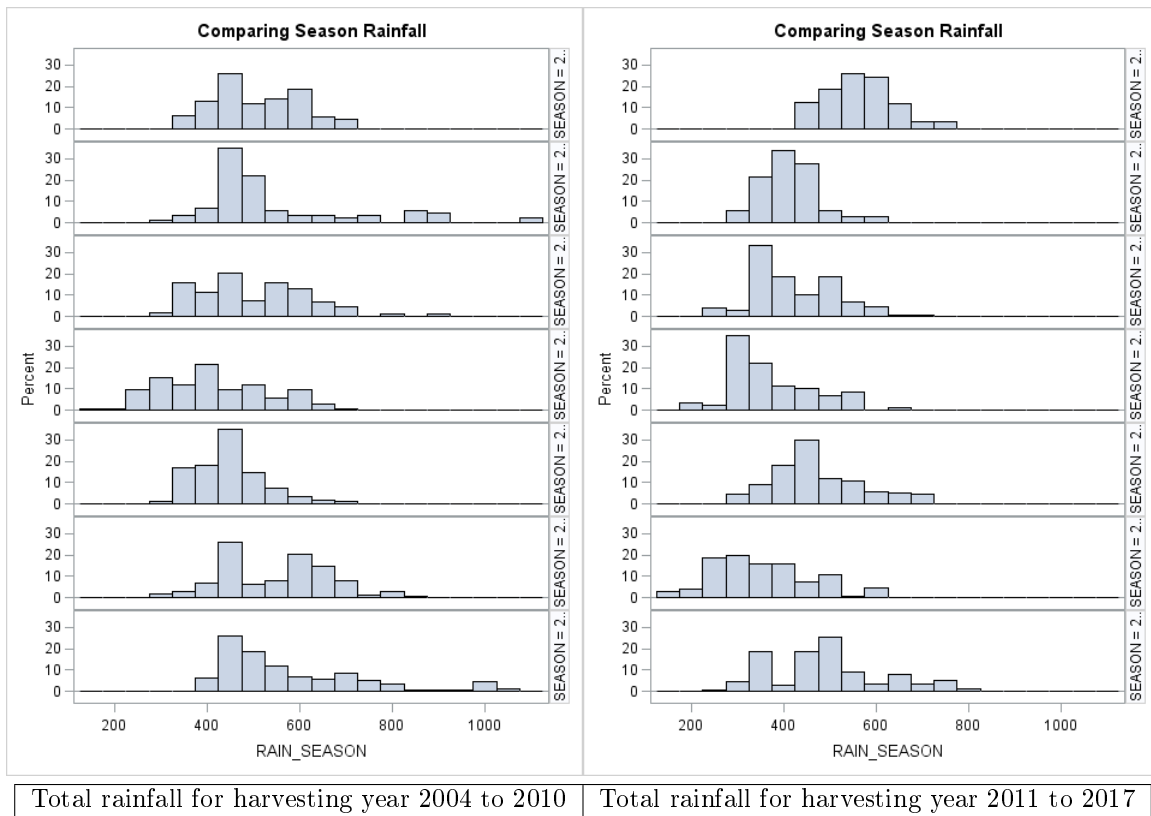


Figure 5.3.3: Histograms of total rainfall by season

Precipitation is used as the main explanatory variable and scatter plots of rainfall to observed yields were compiled to see whether the expected positive correlation is observed. Figure 5.3.4 shows the scatter plot for total season rainfall to estimated yield. A slight positive linear relationship is visible. It is interesting to observe a near-45° angled cloud of points towards the left top of the bulk of points which contrasts with the flatter cloud of points extending towards the bottom right. Also, when looking at the distinction between white and yellow maize yields, the shape of the white maize yield scatter plot differs a bit from the yellow maize yield scatter plot.

The positive relationship between the estimated yield observations and total season rainfall is not very strong, suggesting that factors besides rainfall, play a role in differentiating high from low maize yields (e.g. timing of rainfall, temperature etc.).

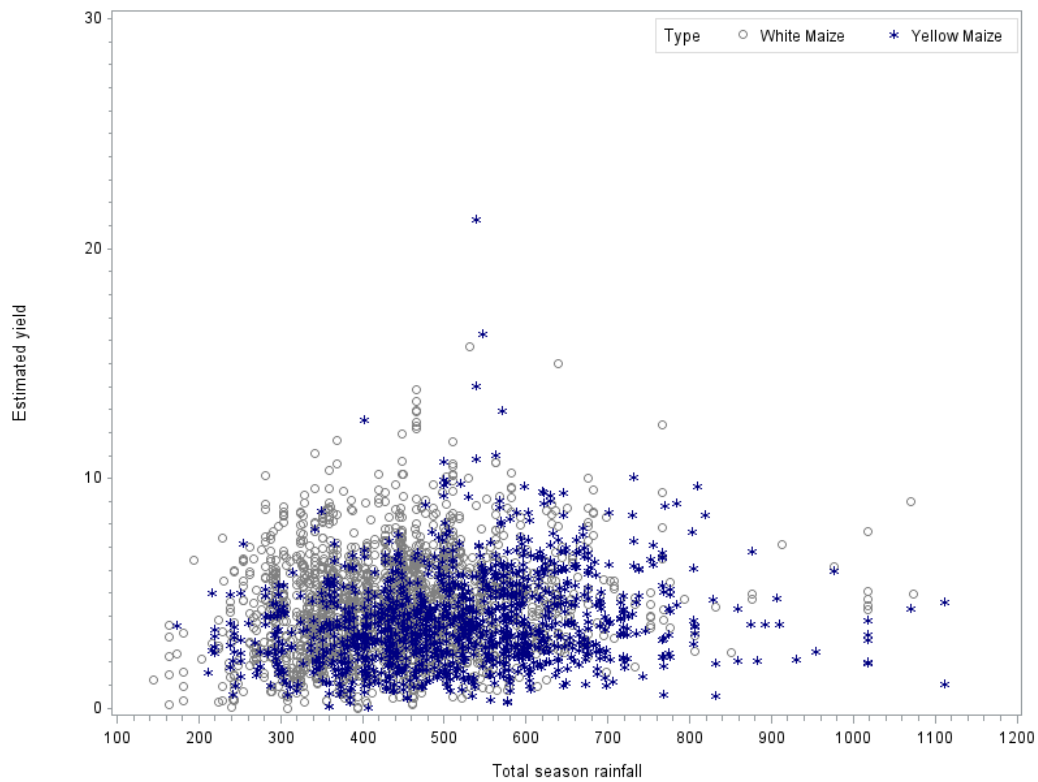
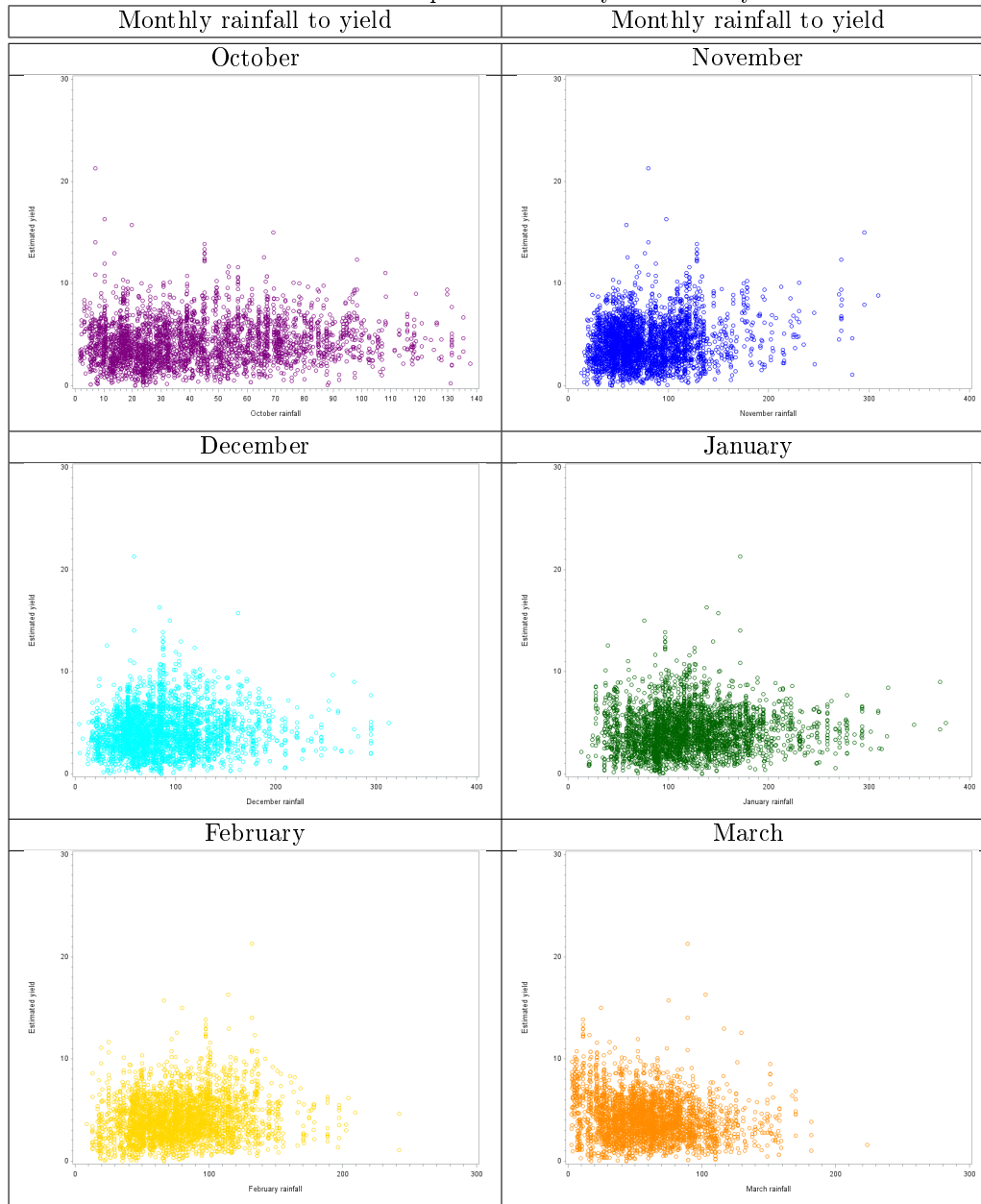


Figure 5.3.4: Scatter plot of season rainfall to yield

Even though the absolute amount of rainfall in a given season is important, the timing of that rainfall can also have a big impact on yield. Therefore, this study will look at the scatter plot of yield against the total precipitation for each month during the summer production region growing season. From Table 5.3.4 it seems that October rainfall is not clearly positively correlated with the observed yield, whereas November rainfall seems to hold a slight positive linear correlation with the observed yield. This points toward the fact that the early summer rain enables farmers to plant their summer crops earlier which might have positive impacts on yield, or that the early rain (after planting) improves the plant's production. The pattern of the December, January and February scatter plots is very similar to that in Figure 5.3.4 with a slight positive linear relationship between the monthly rainfall and yield. During February most of the maize crop in the Free State reaches flowering and fruit forming stages during which rain is crucial towards determining yield. Lastly, the March rainfall vs yield scatter plot does have similarities with the total seasonal rainfall scatter plot. However, the eastern part of the Free State does not typically receive significant amounts of rain in March, and as a result the shape of the scatter plot is contracted and does not suggest a clear positive relationship between March monthly rainfall and yield.

Once more, the exploratory analysis of precipitation data suggests that the relationship between the estimated maize yield observations and precipitation is not very strong and that a model with precipitation as its main explanatory variable will likely not explain a large proportion of variation in the estimated yield observations.

Table 5.3.4: Scatter plots of monthly rainfall to yield



Referring back to Figure 5.3.4, some extremely high yields were observed keeping in mind that these are all dryland maize yields. The first (top) box plot in Figure 5.3.5 shows the estimated yields and the box plot indicates that observed yield observations larger than 9.5 tonnes per hectare (t/ha) to be outliers, also $\mu_{estyield} + 3\sigma_{estyield} = 10.5$. Therefore, 10 t/ha was chosen as the cut-off point for outliers; Figure 5.3.6 below shows the observed yield data plotted against observation number with a reference line at a yield of 10 t/ha. Subsequently the yields greater than 10 t/ha (keeping in mind that dryland maize yields are seldom above 10 t/ha) were removed from the data set, resulting in the second (bottom) box plot in Figure 5.3.5; 40 observations were removed as outliers. The data was

sorted according to season and the estimated yields were plotted against observation number in Figure 5.3.7.

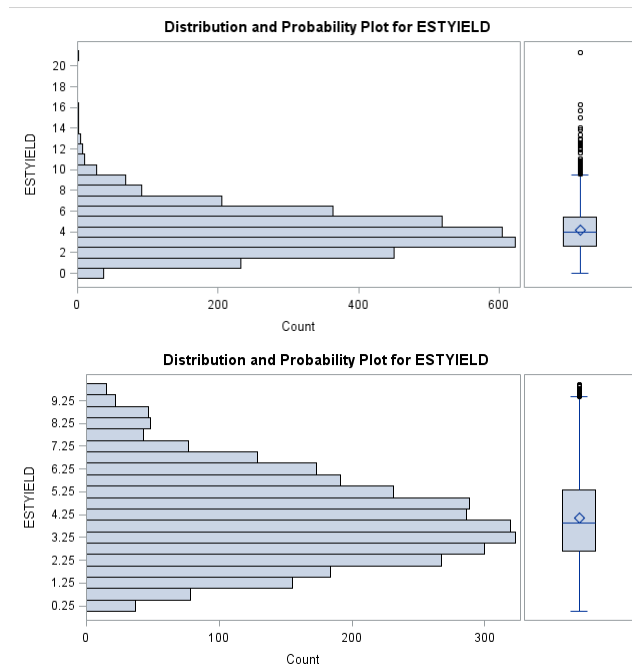


Figure 5.3.5: Box plot of observed yields before (top) and after (bottom) outliers were removed

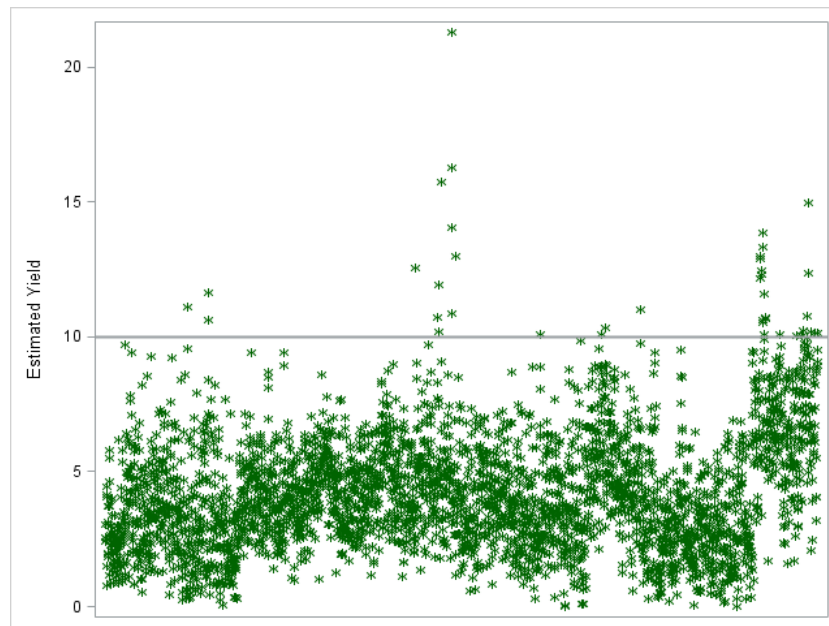


Figure 5.3.6: Observed yields

The resulting data set excluding the outliers is plotted against season in Figure 5.3.7. The data

clearly looks more randomly distributed in terms of estimated yield. The colouring clearly shows the record seasons 2014 and 2017 had higher than average yields and during the drought in 2015 and 2016 lower than average yields were realised. After removing the outliers, the data set contained 3213 observations.

Following these key exploratory analyses the variables to be included in the analysis will be selected and a simple linear regression will be fitted in the next section so that the spatial variant mixture of regressions model can be compared to the simple linear regression results.

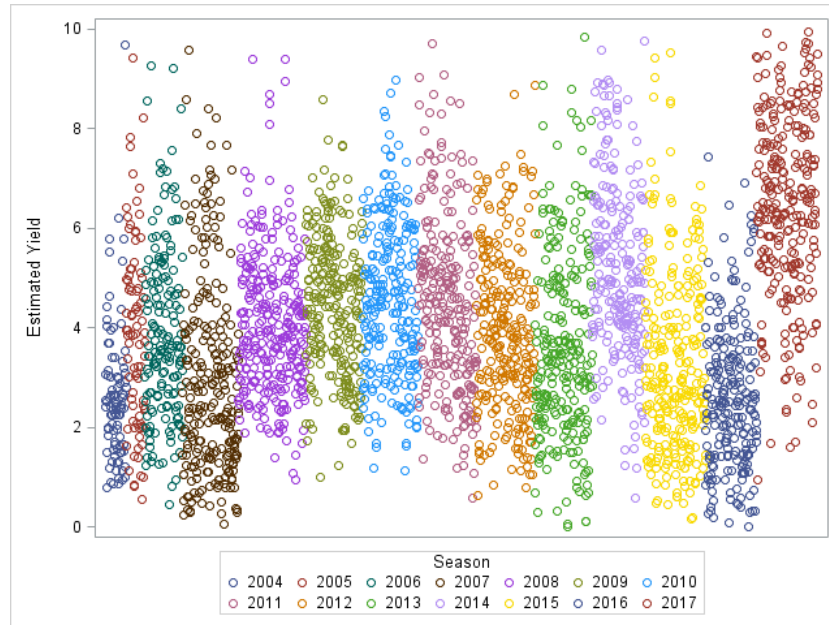


Figure 5.3.7: Observed yields without outliers, coloured by season

5.3.2 Simple linear regression

The maize yield data (without outliers) will be modelled as a function of the total season's rainfall, the season and the type of maize planted. November and February monthly rainfall regression parameters were significantly different from zero according to the t -test, however since the total season's rainfall is a function of the two the regression will run into co-linearity problems. The regression including the total season's rainfall yielded a higher R^2 , therefore total season rainfall was included in the final model instead of the November and February monthly rainfall. Note that the total season's rainfall and season variables were standardised to have a mean of 0 and standard deviation 1 to account for disparate scales. The simple linear regression model to be fitted can be described by:

$$Y_{estyield} = \beta_0 + \beta_1 X_{SeasonRain} + \beta_2 X_{Season} + \beta_3 X_{Type} + \varepsilon$$

where ε is the error term and is assumed to be independent and identically Gaussian distributed and the variable definitions are given in Table 5.3.5.

Table 5.3.5: Variable definitions

Variable	Definition
$Y_{estyield}$	Estimated yield from objective yield survey (dependent variable)
$X_{SeasonRain}$	Total rainfall for October to March in relevant season
X_{Season}	Season: harvest year e.g. 2005
X_{Type}	Maize type: for yellow maize $X_{Type} = 1$, white maize $X_{Type} = 2$

The estimated model is given by

$$\hat{Y}_{estyield} = 2.74 + 0.44.X_{SeasonRain} + 0.42.X_{Season} + 0.83.X_{Type}.$$

All the parameters are significantly different from zero with p -values of the t -test all less than 0.0001. The overall model fits reasonably well, but does not explain a large share of the total variation in the yields with the adjusted R^2 of 9.1% (i.e. 9.1% of the variation in the maize yields is explained by the model) and the $SSE = 11\,105$.

The mean observed yield $Y_{estyield}$ is plotted against the mean predicted values from the model $\hat{Y}_{estyield}$ per season in Figure 5.3.8, demonstrating to what extent the model captured the changes of yields over time. The trend and turning points of the mean observed yield (per season) have been accurately captured by the predicted values with the exception of seasons 2013, 2014 and 2015. Neither the correct turning points nor the absolute magnitude of the mean observed yields were estimated accurately during these three years. This indicates that factors beyond the covariates included in this model caused the fluctuations in average yield during this time (keep in mind that 2014 was an all-time record harvest year at the time followed by two consecutive drought years).

The model can be improved by including more covariates (for example temperature, solar radiation, soil types etc.), however this study was aimed at focusing specifically on precipitation data. Furthermore, it was previously mentioned that statistical crop models, like this regression model, do not take the full extent of detail in terms of the production process into account (for example CSM's) and the aim therefore, is not to explain all the variation but to quantify specific relationships between dependent and independent variables.

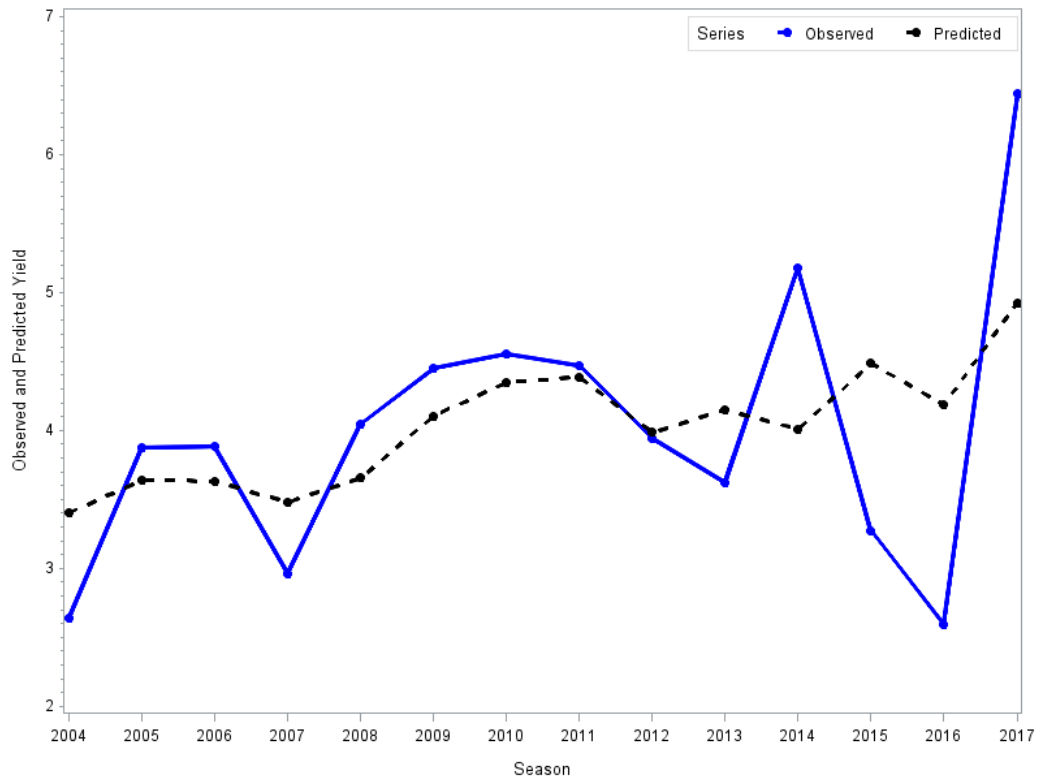


Figure 5.3.8: Observed vs predicted yield

The observed and predicted maize yields are plotted against rainfall in Figure 5.3.9 and it is clear that a lot of variation in the maize yields has gone unexplained, but the expected positive relationship between rainfall and maize yields is clearly visible. It seems also that the predicted values in Figure 5.3.9 form three distinct groups; given that the categorical covariate, type of maize, only contains two categories (white and yellow maize), these three distinct groups could be indicative of another unobservable variable causing three underlying clusters.

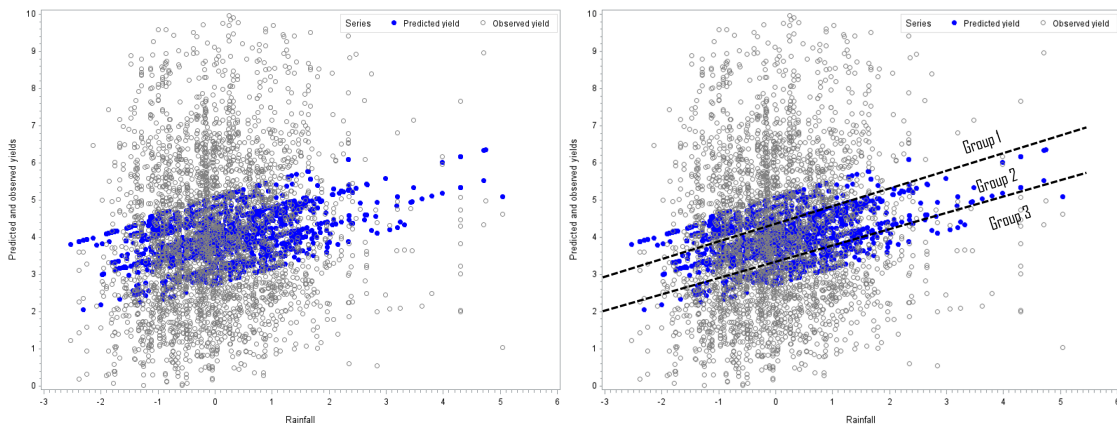


Figure 5.3.9: Observed vs predicted yields by rainfall - simple linear regression

The residuals from the regression $r = Y_{estyield} - \hat{Y}_{estyield}$ were plotted against the predicted values in Figure 5.3.10 and coloured according to the groups indicated in Figure 5.3.9. The three groups are clearly visible. Group 1 included seasons 2009-2017 while group 3 included only seasons 2004-2008. Group 3 contained yellow maize observations only, while group 1 consists of mainly white maize observations. Therefore, group 1 containing mainly white maize planted in the most recent 8 years clearly contains higher predicted yields while group 3 consisting of mainly yellow maize planted in the first 5 years has lower predicted values. Furthermore, group 2 contains average predicted yields (in every season). Overall, besides the observed and discussed grouping, the residuals vs predicted values plot does not exhibit a specific pattern.

The three groups suggested by Figure 5.3.9 can be partially explained by the covariates included in the model however, their cause will be investigated further using the traditional mixture of regression model as well as the spatial variant mixture of regressions model in following sections.

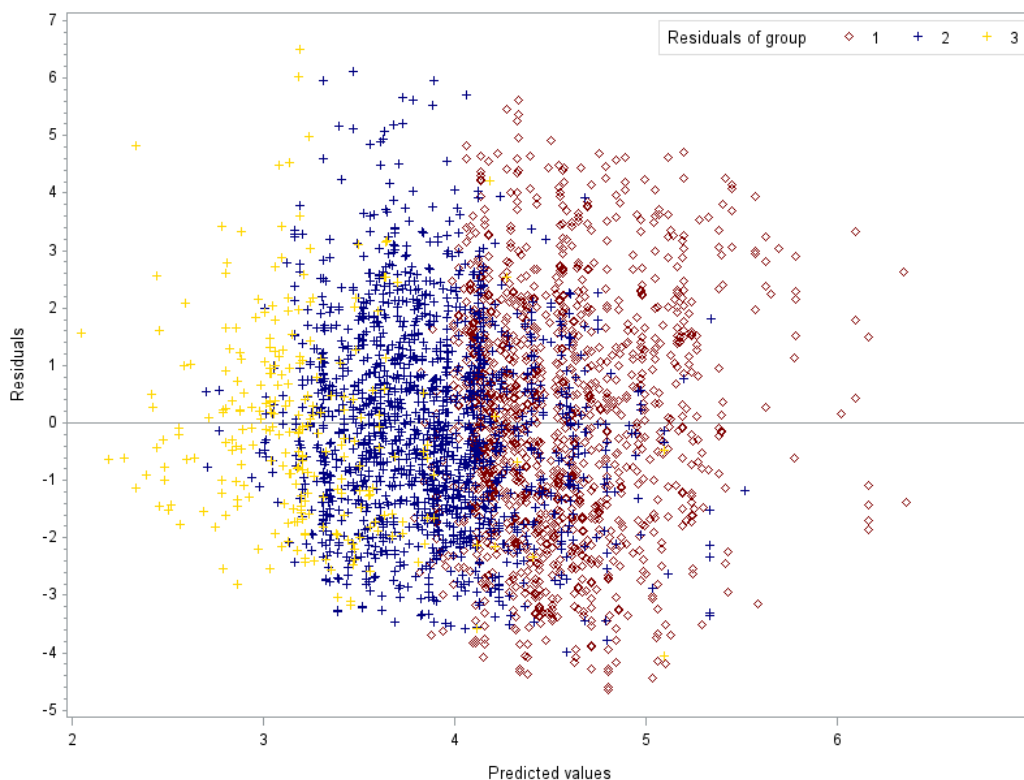


Figure 5.3.10: Residuals plotted against predicted values

5.4 Applying the spatial variant mixture of regressions model

Before applying a mixture of regressions model the number of components of the mixture needs to be determined or estimated. To this end, the finite mixture models procedure (*proc fmm*) in SAS was run to identify the 'best' number of components for a traditional mixture of regressions model fitted

to the data based on the Bayesian Information Criteria (BIC) [3]. The model with three components was identified as minimizing the BIC, as was also expected from the simple linear regression analysis and Figure 5.3.9 and 5.3.10 in the previous section. The traditional mixture of regressions model is fitted in order to facilitate comparison with the spatial variant mixture of regressions model. Since $K = 3$ components are identified as appropriate for the traditional mixture of regressions model, the same number of components are fitted for the spatial variant mixture of regressions model, so that the additional location information accounted for by the spatial variant mixture of regressions model and differences between the two model results can be studied. The results of the respective models are discussed in subsequent sections below.

5.4.1 Traditional mixture of regressions model

The traditional mixture of regressions model to be fitted can be written in the generative form with random indicator variables Δ_j : $Y = \Delta_1 Y_1 + \Delta_2 Y_2 + \Delta_3 Y_3$ where $\sum_{j=1}^3 \Delta_j = 1$ and $\Delta_{ij} \geq 0$ for all $i = 1, 2, \dots, N$ and $j = 1, 2, 3$

$$Y_j = \mathbf{X}\boldsymbol{\beta}_j + e_j$$

$$\mathbf{X} = \begin{bmatrix} 1 & X_{SeasonRain} & X_{Season} & X_{Type} \end{bmatrix}$$

$$\boldsymbol{\beta}_j = \begin{bmatrix} \beta_{j,int} \\ \beta_{j,SeasonRain} \\ \beta_{j,Season} \\ \beta_{j,Type} \end{bmatrix}$$

$e_j \sim i.i.d. N(0, \sigma_j)$ and $P(\Delta_j = 1) = \pi_j$ for $j = 1, 2, 3$; therefore Y_j can be modelled by a Gaussian distribution $N(\mathbf{X}\boldsymbol{\beta}_j, \sigma_j)$.

The estimated component regression models (after the EM algorithm converged) with the maximum likelihood parameter estimates for the traditional mixture of regressions model are given below:

$$\begin{aligned} \hat{Y}_1 &= 3.954 - 0.197.X_{SeasonRain} - 1.917.X_{Season} - 0.013.X_{Type} \\ \hat{Y}_2 &= 3.213 + 0.435.X_{SeasonRain} + 0.992.X_{Season} + 1.030.X_{Type} \\ \hat{Y}_3 &= 2.098 + 0.337.X_{SeasonRain} + 0.139.X_{Season} + 0.444.X_{Type}. \end{aligned} \tag{5.4.1}$$

The estimated error variances are given by

$$\hat{\sigma}_1 = 1.257$$

$$\hat{\sigma}_2 = 2.330$$

$$\hat{\sigma}_3 = 1.191$$

and the estimated mixing probabilities:

$$\hat{\pi}_1 = 0.134$$

$$\hat{\pi}_2 = 0.544$$

$$\hat{\pi}_3 = 0.322.$$

The estimated parameters clearly define three different sets of relationships between the dependent and independent variables. With the $SSE = 11\,166$, marginally higher than that of the simple linear regression, the extent to which yield variation is explained by the traditional mixture of regressions model is similar to that of the simple linear regression.

The estimated component regression models were used to calculate predicted yields and preliminary analysis was performed on the mean maize yields per season. The regression models for all of the clusters failed to capture the correct turning points and magnitude of yields in seasons 2013, 2014 and 2015 (as in the simple linear regression results). Cluster 1 exhibits a prominent downward trend over time, which will be investigated further, while cluster 2 seems to capture the upper “band” of historic yield trends and cluster 3 seems to capture the lower yield observation trend.

The clustering result is shown in two views: in Figure 5.4.1 the locations of the observations (latitude and longitude) were plotted like in a map and the markers coloured according to the highest responsibility for each observation; in Figure 5.4.2 the observed yields were plotted against rainfall with the markers indicating the cluster with the highest responsibility for each observation. In Figure 5.4.1, the observations found in the respective clusters seem to be uniformly distributed in terms of location (i.e., no pattern linked to the different clusters is observed). On the other hand, in Figure 5.4.2 a clear distinction between the clusters is observed: cluster 2 generally consists of higher yields and cluster 3 of lower yields, which is consistent with the preliminary analysis of mean yield per cluster. Cluster 1 groups selected observations with particularly high and low yields for a range of low to high rainfall respectively. When investigating cluster 1 in some more detail, it was observed that the cloud with low yields contains only observations from the 2015 and 2016 seasons, while the cloud of points with high yields contains observations from the 2004 season to the 2008 season. Relating this observation back to the parameter estimates in Equation (5.4.1) - cluster 1 is characterised by a negative relationship between yields and rainfall as well as season. Figure 5.3.7 shows the extreme drought years in 2015 and 2016 in contrast with the relatively high yields in 2004 to 2008; keeping in mind that cluster 1 contains the highest observations in seasons 2004 to 2008 and the lowest yield observations in seasons 2015 and 2016.

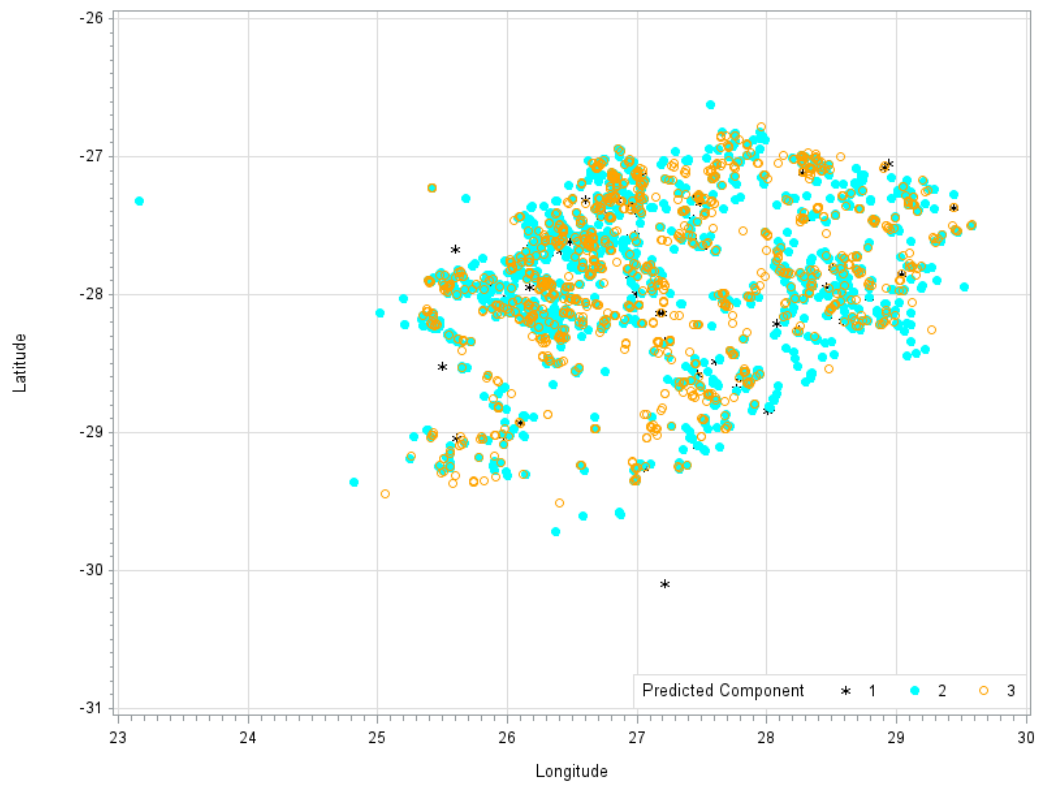


Figure 5.4.1: Map-view of yield observations coloured according to clusters

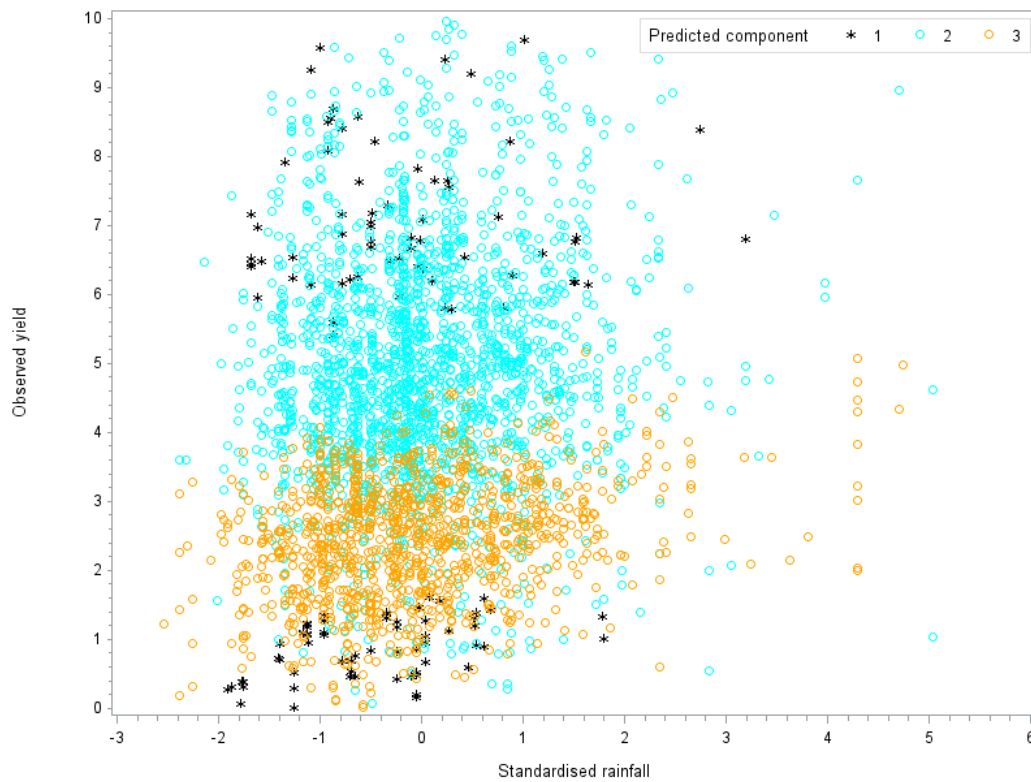


Figure 5.4.2: Observed yield by rainfall coloured according to clusters

The residuals from the traditional mixture of Gaussian regressions model are plotted against the predicted values in Figure 5.4.3, and are coloured according to the identified clusters. Clearly the clusters differ from the observed groups illustrated in Figure 5.3.10. Furthermore, it is clear that some observations in cluster 1 were overestimated (the observations in seasons 2004 to 2008) and others were underestimated (observations from the 2015 and 2016 seasons).

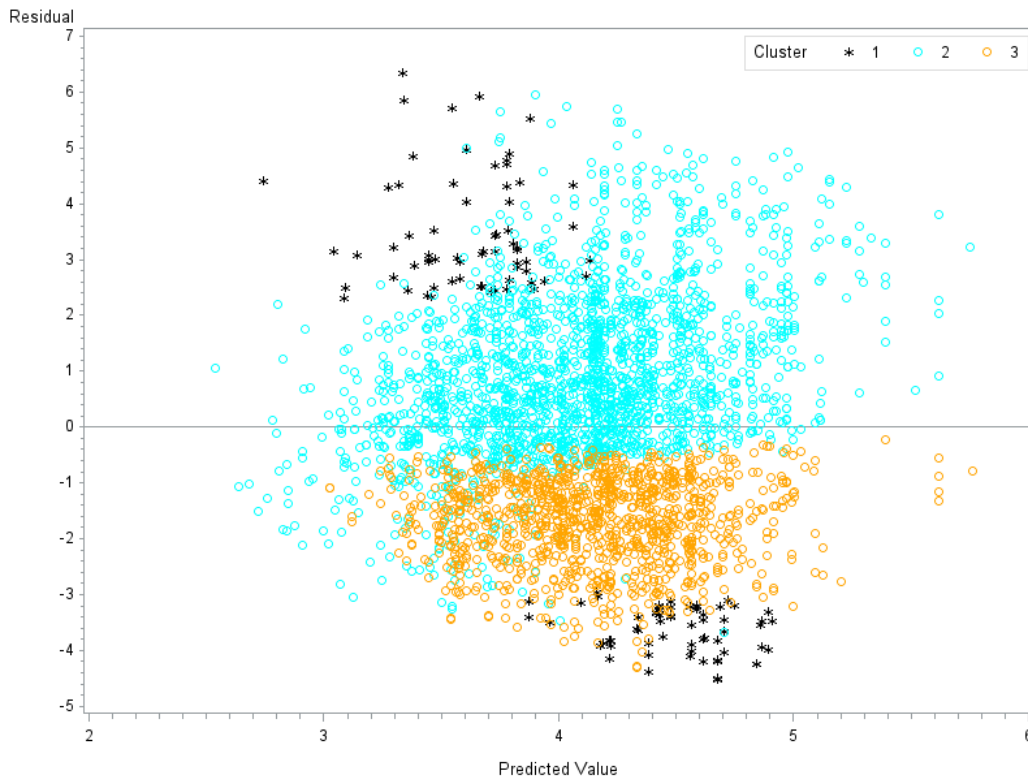


Figure 5.4.3: Residuals against predicted values, by cluster

In the following section the spatial variant mixture of regressions model will be specified and fitted to the same set of data.

5.4.2 Spatial variant mixture of regressions model

The model to be fitted can be written in the generative form as follows: $Y = \Delta_{i1}Y_1 + \Delta_{i2}Y_2 + \dots + \Delta_{iK}Y_K$ where $\sum_{j=1}^K \Delta_{ij} = 1$ and $\Delta_{ij} \geq 0$ for all $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, K$

$$Y_j = \mathbf{X}\beta_j + e_j$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & X_{SeasonRain} & X_{Season} & X_{Type} \end{bmatrix}$$

$$\beta_j = \begin{bmatrix} \beta_{j,int} \\ \beta_{j,SeasonRain} \\ \beta_{j,Season} \\ \beta_{j,Type} \end{bmatrix}$$

$e_j \sim i.i.d. N(0, \sigma_j)$ and $P(\Delta_{ij} = 1) = \pi_{ij}$ for $j = 1, 2, \dots, K$; therefore Y_j can be modelled by a Gaussian distribution $N(\mathbf{X}\beta_j, \sigma_j)$. It is important to note here that Y_j is a univariate dependent variable, unlike the example in Section 4.3 which worked with a multivariate response variable. The model describes a univariate mixture of Gaussian regressions model with $p = 4$ explanatory variables, $K = 3$ clusters and $N = 3213$ observations of $T = 1$ dimensional vectors, where the i^{th} observation (objective yield estimate) is classified into group j with probability π_{ij} .

Let the j^{th} component mixing probabilities for observations $i = 1, 2, \dots, N$ be random variables or nodes in a Markov random field with the relevant GPS coordinates for observations $i = 1, 2, \dots, N$ indicating the node positions in space. Since the yield observations are not distributed uniformly across space, the linkages between the nodes are defined such that each observation is linked to its 8 nearest neighbours, found by identifying the 8 nodes with the smallest euclidean distance between the various GPS coordinates. The i^{th} clique \mathcal{N}_i of the Markov random field is therefore defined as containing the 8 nearest neighbour nodes relative to node i . Following the theory set out in Section 4.2.1, the mixing probabilities are defined to be distributed according to a Gibbs distribution given below

$$\pi_{ij} \sim \prod_{j=1}^K \xi_j^{-N} \exp\left(-\frac{\sum_{i=1}^N \sum_{m \in \mathcal{N}_i} (\pi_{ij} - \pi_{mj})^2}{2\xi_j^2}\right).$$

It is assumed that the errors of the mixing probabilities for clusters $j = 1, 2, \dots, K$ are *i.i.d.* $N(0, \xi_j^2)$ distributed. Algorithm 4.1 was used to estimate the model specified above with parameters $\Theta_j = \{\beta_j, \sigma_j, \pi_{ij}, \xi_j^2\}$ for $j = 1, 2, \dots, K$. The starting values for the algorithm were selected as follows:

- For β_j , K random dependent and independent observation pairs (Y_k, \mathbf{X}_k) were selected and the corresponding starting values were calculated: $\beta_j = (\mathbf{X}'_k \cdot \mathbf{X}_k)^{-1} \cdot (\mathbf{X}'_k \cdot Y_k)$.
- For σ_j the starting values were taken as $\frac{1}{K} \sum_{i=1}^N (Y_i - \bar{Y})^2$ where $\bar{Y} = \frac{1}{K} \sum_{i=1}^N Y_i$ for $j = 1, 2, \dots, K$.
- $\pi_{ij} = \frac{1}{K}$ for $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, K$.
- $\xi_j^2 = \frac{1}{K}$ for $j = 1, 2, \dots, K$.

A $K = 3$ cluster spatial variant mixture of regressions model was fitted to the data with the mixing probabilities incorporating the spatial location information in the estimation as specified above. The parameter estimates and results are discussed below.

The estimated parameters (after the algorithm converged) are given in Equation (5.4.2) below.

$$\begin{aligned} \hat{Y}_1 &= 0.693 + 0.042.X_{SeasonRain} + 0.054.X_{Season} + 1.156.X_{Type} \\ \hat{Y}_2 &= 0.694 + 0.048.X_{SeasonRain} + 0.055.X_{Season} + 1.156.X_{Type} \\ \hat{Y}_3 &= 0.690 + 0.032.X_{SeasonRain} + 0.058.X_{Season} + 1.154.X_{Type}. \end{aligned} \tag{5.4.2}$$

The error variance estimates are given by

$$\hat{\sigma}_1 = 6.04$$

$$\hat{\sigma}_2 = 6.07$$

$$\hat{\sigma}_3 = 5.98$$

and the estimated variances of the mixing probabilities are given by

$$\hat{\xi}_1 = 0.0000079$$

$$\hat{\xi}_2 = 0.0000102$$

$$\hat{\xi}_3 = 0.0000091.$$

The estimated mixing probabilities are best displayed graphically, see Figure 5.4.4, where each observation $i = 1, 2, \dots, N$ is illustrated as belonging to the cluster with the largest mixing probability. It is clear that certain areas or locations have a higher probability of belonging to cluster j than others and that these areas are close together (e.g. the central Free State seems to have a higher probability to belong to cluster 3). This observation is consistent with the model definition.

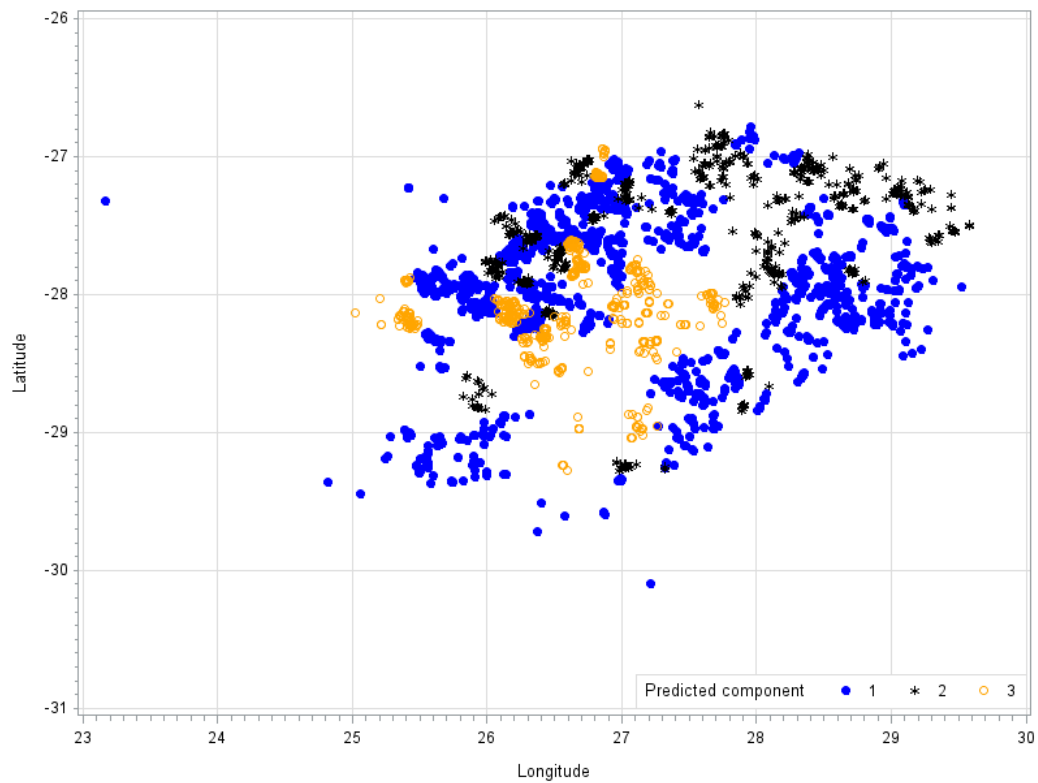


Figure 5.4.4: Estimated mixing probabilities ($K = 3$)

The estimated regression models differ from the estimated simple linear regression model, and unlike the traditional mixture of regressions model, the estimated parameters seem to be very similar across clusters. The sum of squared errors for the estimated spatial variant mixture of regressions model equals $SSE = 19348$, which is higher than that of the simple linear regression and the traditional mixture of regressions model, suggesting that the spatial variant mixture of regressions model with three clusters is not an improvement in terms of overall explanation of maize yield variance.

The estimated regression models were used to calculate predicted yields and were compared with observed yields. The component regression models do not capture the fluctuations in maize yields over time (mean yield per season) as well as the simple linear regression and traditional mixture of regressions models. This observation suggests that the season covariate is fairly constant with respect to different locations in the Free State; practically, the regions where maize was planted did not change drastically from one season to the next.

The residuals from the fitted model show that the maize yields were generally underestimated by the spatial variant mixture of regressions model, and that the residuals form two well separated groups, see Figure 5.4.5. It is also demonstrated that the two distinct groups in the residuals represent the type of maize (yellow or white maize). Observing such a distinct pattern in a residuals plot indicates that the fitted regression model is not an adequate model for the data. Due to these two distinct groups the residuals from the spatial variant mixture of regressions model do not seem to be normally distributed.

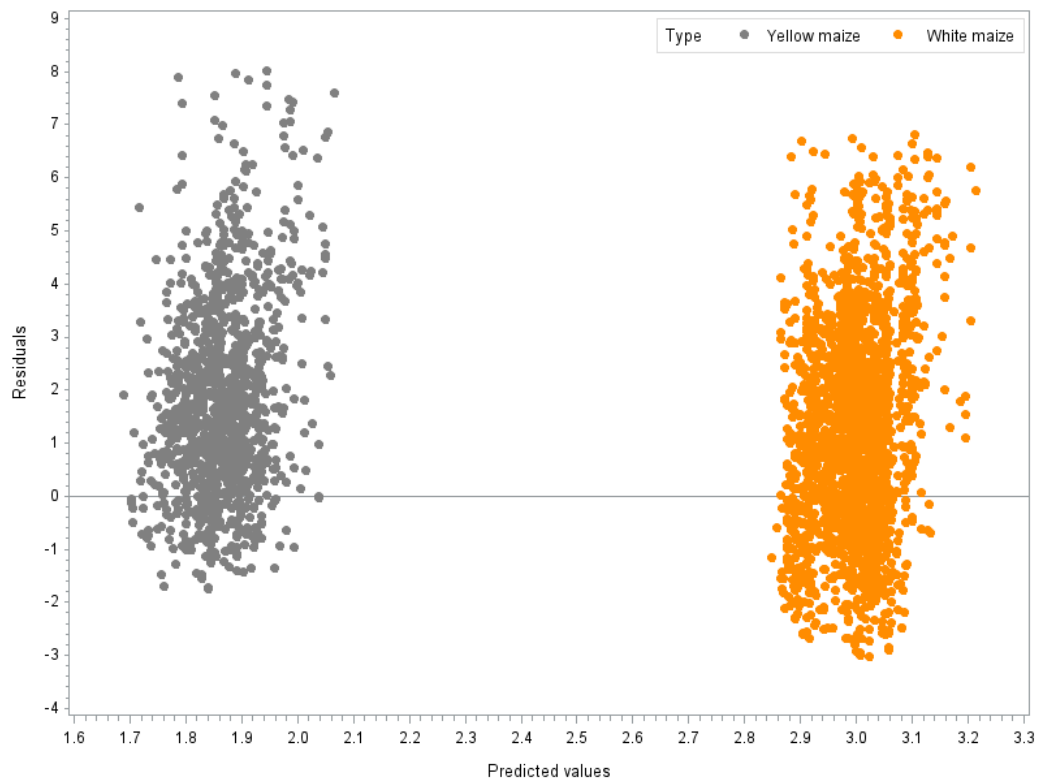


Figure 5.4.5: Residuals against predicted values- spatial variant mixture of regressions model

The clustering result for $K = 3$ clusters is shown in a map representation in Figure 5.4.6; the observations $i = 1, 2, \dots, N$ are illustrated as belonging to the cluster for which the estimated responsibility ($E(\Delta_{ij}) = \gamma_{ij}$) is the highest. The clustering result clearly incorporates the spatial dependency information suggesting that yield observations in spatial proximity will likely belong to the same cluster. The cluster assignment using the responsibilities corresponds roughly with the mixing probability estimates illustrated in Figure 5.4.4: the central and parts of the western Free State are likely to belong to cluster 3, large parts of the far-eastern and far-western Free State have the highest probability to belong to cluster 1 whereas the north-eastern parts of the Free State have the highest probability of belonging to cluster 2, together with smaller pockets throughout the western Free State. Preliminary investigation of cluster 2 shows that this cluster has the highest mean rainfall and the locations of cluster 2 in Figure 5.4.6 suggests smaller regions where a combination of environmental circumstances (like microclimates) contribute to higher rainfall. Investigation of the altitude, surrounding topography and temperatures at these locations is required to confirm this notion.

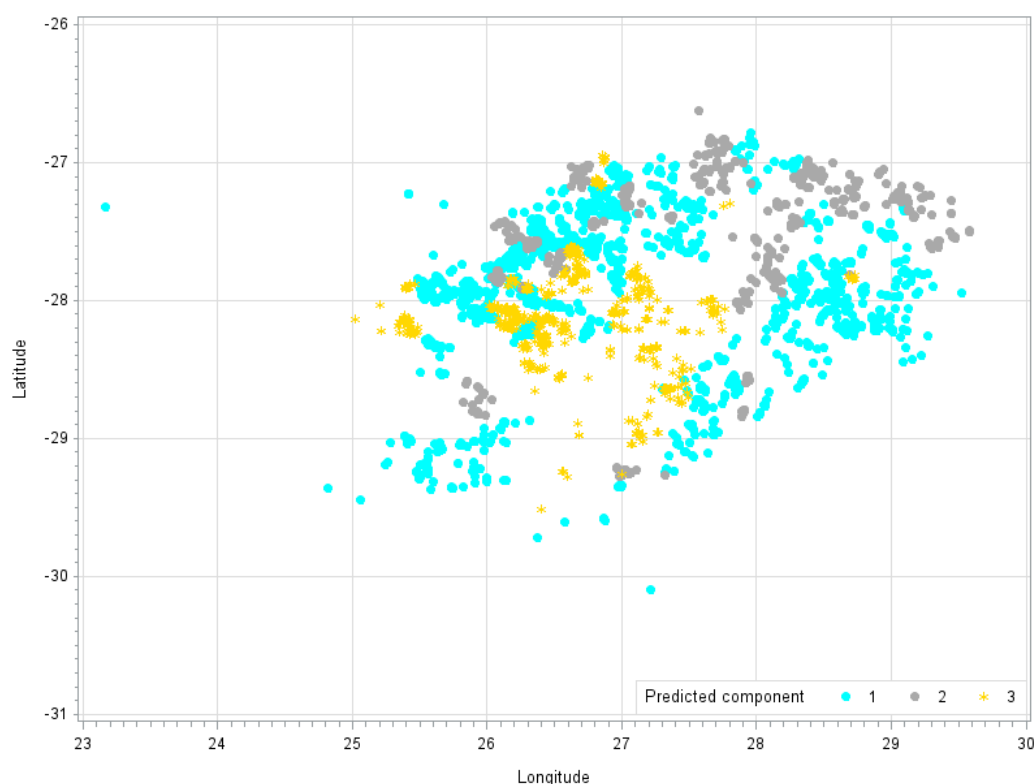


Figure 5.4.6: Map-view of clustering result - spatial variant mixture of regressions model

Further investigation showed that the yield observations included in cluster 1 had the smallest mean rainfall. The observations in all three clusters are evenly distributed across the seasons variable, but cluster 1 comprises of 61% yellow maize yield observations, and cluster 2 and 3 comprise of 55% and 34% yellow maize yield observations respectively. Cluster 1 can therefore be described as containing mostly yellow maize yield observations and cluster 3 as containing mostly white maize yield observations. Consider this observation while comparing Figure 5.4.6 and Figure 5.3.1, and it becomes

clear that the remark is consistent with the original data structure in that cluster 1 and 3 locations (Figure 5.4.6) coincide with areas where yellow and white maize are predominantly planted (Figure 5.3.1).

Figure 5.4.7 illustrates the clustering result by plotting the observed yields against standardised rainfall and colouring the markers according to the clusters. It seems like cluster 3 does not include yield observations with as high rainfall values as seen in cluster 1 and 2 however the distinction is unclear. In contrast to Figure 5.4.6, Figure 5.4.7 does not seem to clearly distinguish three definite clusters.

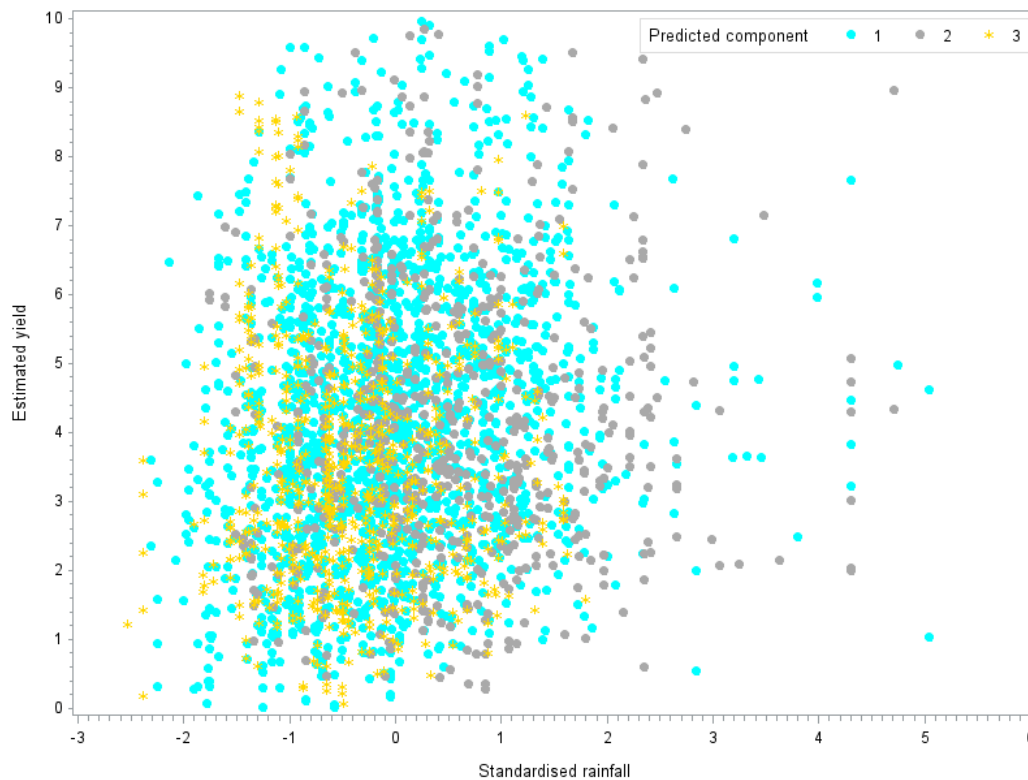


Figure 5.4.7: Observed yield by rainfall - spatial variant mixture of regressions model

After fitting the spatial variant mixture of regressions model for $K = 3$ clusters the study found that the resulting regression functions were very similar across clusters, this is evident in Figure 5.4.7. When comparing the clustering from the spatial variant mixture of regressions model in Figure 5.4.7 with that of the traditional mixture of regressions model in Figure 5.4.2 it is clear that the clustering result illustrated in the latter (clustering results for the traditional mixture of Gaussian regressions model) yields more distinct clusters. However, the opposite is true when the clustering result in Figure 5.4.6 is compared with that in Figure 5.4.1; the spatial variant mixture of regressions model provides informative spatial clustering.

Retrospectively, it is likely that this specific data set was not perfectly suited for the spatial variant mixture of regressions model, in that the latent variables: agronomic management practices and microclimates, were not clearly represented in the amount or type of data available (the fitted model

residuals in Figure 5.4.5 clearly demonstrate that the data set was not a good fit). This was also evident in the fact that the traditional mixture of regressions model clearly identified distinct regression functions across the three clusters without accounting for spatial information; whereas the spatial variant mixture of regressions model identified spatially distinct clusters with relatively similar regression functions across clusters. Ideally one would have liked to have seen spatially significant clusters and distinct regression functions in the results of the fitted spatial variant mixture of regressions model, as was demonstrated in Section 4.3. Even though this data set has a spatial underpinning as discussed in Section 5 and the spatial variant mixture of regressions model delivers useful spatial clustering, it did not deliver particularly heterogeneous regression models for the identified clusters; suggesting that not all of the covariates included in the analysis have a clear spatial dependence.

What stood out however, is that the type of maize produced was the covariate that clearly differed according to location (see Figure 5.4.6) and that this was also depicted in the resulting clusters (as discussed above). On the other hand, even though rainfall differs slightly across regions in the Free State - the distinction is not as clear or extreme as for the type of maize covariate. The season variable is not location specific either, in that the area where maize was grown did not change drastically over the years considered here. A lesson learned therefore is that the spatial variant mixture of regressions model performed well in terms of distinguishing clusters and estimating their respective regression models with respect to the variable that has a clear spatial dependence.

Chapter 6

Conclusion and future work

This study provides a detailed discussion of finite mixture of multivariate Gaussian distributions and the maximum likelihood estimators were derived for the general K -component case. It was demonstrated that this model correctly identifies and estimates the mixture components in the univariate and bivariate cases even when the location of the mixing distributions are equal and only variances differ. The study also demonstrated the theory and application of the finite mixture of multivariate Gaussian regressions model. Evaluation of model performance was briefly discussed for the finite mixture of multivariate Gaussian regressions model and two methods for calculating the coefficient of determination were put forward.

The application of mixtures of Gaussian distributions and regression models in the image segmentation context was discussed. The present study reviewed the incorporation of a Markov random field to model the pixel location information, and the resulting spatial variant mixture of Gaussian regressions model was defined. An example was designed to illustrate the application of the model. The clusters were spatially correctly identified and the corresponding regression functions were accurately estimated.

The spatial variant mixture of regressions model was applied to the agricultural context: dryland maize yield observations in the Free State were modelled as a function of maize type, precipitation and season (time). A traditional mixture of Gaussian regressions model was fitted for comparative purposes, and it clearly identified three distinct clusters without accounting for location information. The spatial variant mixture of Gaussian regressions model successfully identified underlying spatial clustering, particularly with respect to the maize type covariate. However, it was found that the estimated regression parameters were similar across clusters, suggesting that not all the covariates contained sufficient spatial information. Keep in mind, that a statistical model inherently simplifies the crop production process and that a myriad of factors and interactions determine crop yields. Retrospectively, it was found that the investigated data set was not perfectly suited for the spatial variant mixture of regressions model, consequently this study suggests that more spatially-specific covariates be included in the regression (e.g., soil type, solar radiation and temperature).

It is recommended that the following points and recommendations are investigated in future:

- Formalisation of model performance measurements for finite mixtures of regressions models is required.

- Expand the example on the performance of the spatial variant mixture of Gaussian regressions model to a formal simulation study [6] (using Monte-Carlo Markov Chain methodology), with a focus on map-type images.
- Include more spatially-specific covariates in the spatial variant mixture of Gaussian regressions model (e.g., soil type, solar radiation and temperature) to improve the model fit.
- Use a gridded approach to model spatial dependency for map-type data, as opposed to the coordinates approach used in the present study.
- The random errors in the component regression models can be assumed to be skew-Gaussian distributed (rather than $\mathbf{e}_j \sim i.i.d. N(\mathbf{0}, \Sigma_j)$) to improve model performance in real-life applications.
- Use a spatial regression model for the mixture component models rather than the simple linear regression model.
- Bayesian modelling can be used as another way of incorporating a prior and thereby to model spatial dependency.

The spatial variant mixture of Gaussian regressions model is a powerful tool to incorporate spatial dependencies in the fuzzy clustering model. It has the potential to extract spatially explicit clusters and simultaneously estimate corresponding regression models. However, the application depends on data sets that contain discernible spatial information; various applications in the agricultural production and consumption context will be investigated in future.

Appendix A

Results for deriving the MLE's

Lemma's for deriving the maximum likelihood estimators for the multivariate Gaussian distribution:

A.1 Lemma 1

Consider the function $f(\mathbf{C}) = \frac{1}{2}N \ln |\mathbf{C}| - \frac{1}{2}tr(\mathbf{C}\mathbf{D})$ where $\mathbf{C} > \mathbf{0}$ and $\mathbf{D} > \mathbf{0}$. $f(\mathbf{C})$ is at a maximum if $\mathbf{C} = N\mathbf{D}^{-1}$.

Define \mathbf{C}_0 with characteristic roots $\lambda_1, \lambda_2, \dots, \lambda_k$ such that $\mathbf{C} = \mathbf{D}^{-\frac{1}{2}}\mathbf{C}_0\mathbf{D}^{-\frac{1}{2}}$.

Proof

$$\begin{aligned} f(\mathbf{C}) &= \frac{1}{2}N \ln |\mathbf{D}^{-\frac{1}{2}}\mathbf{C}_0\mathbf{D}^{-\frac{1}{2}}| - \frac{1}{2}tr(\mathbf{D}^{-\frac{1}{2}}\mathbf{C}_0\mathbf{D}^{-\frac{1}{2}}\mathbf{D}) \\ &= \frac{1}{2}N \ln |\mathbf{D}^{-1}\mathbf{C}_0| - \frac{1}{2}tr(\mathbf{D}^{-1}\mathbf{D}\mathbf{C}_0) \\ &= \frac{1}{2}N \ln |\mathbf{D}^{-1}| + \frac{1}{2}N \ln |\mathbf{C}_0| - \frac{1}{2}tr(\mathbf{C}_0) \\ &= \frac{1}{2}N \ln |\mathbf{D}^{-1}| + \frac{1}{2} \sum_{i=1}^k \{N \ln(\lambda_i) - \lambda_i\} \end{aligned} \tag{A.1.1}$$

Let $g(\lambda) = N \ln(\lambda) - \lambda$. Find λ which maximises $g(\lambda)$ by taking the derivative and setting it equal to zero:

$$\begin{aligned} \frac{d}{d\lambda}g(\lambda) &= 0 \\ \frac{N}{\lambda} - 1 &= 0 \\ N &= \lambda \end{aligned}$$

Take the second derivative to confirm that this is indeed a maximum:

$$\begin{aligned}\frac{d^2}{d^2\lambda}g(\lambda)|_{\lambda=N} &= \frac{-N}{\lambda^2}|_{\lambda=N} \\ &= \frac{-1}{N} < 0\end{aligned}$$

therefore $g(\lambda)$ is maximised where $\lambda = N$. It follows therefore that A.1.1 is maximised where the characteristic roots of \mathbf{C}_0 are equal to N ; $\mathbf{C}_0 = \mathbf{I}.N$. Therefore, $f(\mathbf{C})$ is maximised where

$$\begin{aligned}\mathbf{C} &= \mathbf{D}^{-\frac{1}{2}}\mathbf{C}_0\mathbf{D}^{-\frac{1}{2}} \\ &= \mathbf{D}^{-\frac{1}{2}}\mathbf{I}.N\mathbf{D}^{-\frac{1}{2}} \\ &= N.\mathbf{D}^{-1}\end{aligned}$$

★

A.2 Lemma 2 - Maximum likelihood estimators of the multivariate Gaussian distribution

Suppose $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_N$, $N > p$ is a random sample of N ($p \times 1$) vector observations from a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution with $\boldsymbol{\mu} : p \times 1$. Let $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_N)$. Then $E(\mathbf{Y}) = (\boldsymbol{\mu} \boldsymbol{\mu} \dots \boldsymbol{\mu}) = \boldsymbol{\mu}\mathbf{1}'$. The observed sample is $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$.

a)

It can be shown that $\sum_{\alpha=1}^N (\mathbf{y}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_\alpha - \boldsymbol{\mu}) = \text{tr} [\boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}\mathbf{1}') (\mathbf{y} - \boldsymbol{\mu}\mathbf{1}')']$

Proof

$$\begin{aligned}\sum_{\alpha=1}^N (\mathbf{y}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_\alpha - \boldsymbol{\mu}) &= \sum_{\alpha=1}^N \text{tr} [(\mathbf{y}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_\alpha - \boldsymbol{\mu})] \quad (\text{trace of a constant}) \\ &= \text{tr} \left[\sum_{\alpha=1}^N (\mathbf{y}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_\alpha - \boldsymbol{\mu}) \right] \quad (\text{sum of a trace}) \\ &= \text{tr} [(\mathbf{y} - \boldsymbol{\mu}\mathbf{1}')' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}\mathbf{1}')] \\ &= \text{tr} [\boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}\mathbf{1}') (\mathbf{y} - \boldsymbol{\mu}\mathbf{1}')'] \quad (\text{trace multiplication})\end{aligned}$$

★

b)

The likelihood function of the multivariate Gaussian distribution is given by

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{Np/2} |\boldsymbol{\Sigma}|^{N/2}} \cdot \exp\left(-\frac{1}{2} \text{tr}(\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}\mathbf{1}')(\mathbf{y} - \boldsymbol{\mu}\mathbf{1}')')\right)$$

Proof

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{\alpha=1}^N \left[\frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{y}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_\alpha - \boldsymbol{\mu})\right) \right] \\ &= \frac{1}{(2\pi)^{Np/2} |\boldsymbol{\Sigma}|^{N/2}} \exp\left(-\frac{1}{2} \sum_{\alpha=1}^N (\mathbf{y}_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_\alpha - \boldsymbol{\mu})\right) \\ &= \frac{1}{(2\pi)^{Np/2} |\boldsymbol{\Sigma}|^{N/2}} \exp\left(-\frac{1}{2} \text{tr}(\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}\mathbf{1}')(\mathbf{y} - \boldsymbol{\mu}\mathbf{1}')')\right) \quad (\text{using result in (A.2.1)}) \end{aligned}$$

★

c)

$L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written in terms of $\bar{\mathbf{y}} = \frac{1}{N} \sum_{\alpha=1}^N \mathbf{y}_\alpha$ and $\mathbf{A} = \sum_{\alpha=1}^N (\mathbf{y}_\alpha - \bar{\mathbf{y}})(\mathbf{y}_\alpha - \bar{\mathbf{y}})' = (\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}')(\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}')'$:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{Np/2} |\boldsymbol{\Sigma}|^{N/2}} \exp\left\{-\frac{1}{2} \text{tr} \boldsymbol{\Sigma}^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu})\right\}$$

Proof

$$\begin{aligned} (\mathbf{y} - \boldsymbol{\mu}\mathbf{1}')(\mathbf{y} - \boldsymbol{\mu}\mathbf{1}')' &= \sum_{\alpha=1}^N (\mathbf{y}_\alpha - \boldsymbol{\mu})(\mathbf{y}_\alpha - \boldsymbol{\mu})' \\ &= \sum_{\alpha=1}^N ((\mathbf{y}_\alpha - \bar{\mathbf{y}}) + (\bar{\mathbf{y}} - \boldsymbol{\mu}))((\mathbf{y}_\alpha - \bar{\mathbf{y}}) + (\bar{\mathbf{y}} - \boldsymbol{\mu}))' \\ &= \mathbf{A} + N(\bar{\mathbf{y}} - \boldsymbol{\mu})(\bar{\mathbf{y}} - \boldsymbol{\mu})' \end{aligned} \quad (\text{A.2.2})$$

since the cross products are a zero matrix, e.g.:

$$\begin{aligned}
\sum_{\alpha=1}^N &= (\mathbf{y}_\alpha - \bar{\mathbf{y}}) (\bar{\mathbf{y}} - \boldsymbol{\mu})' \\
&= \left(\sum_{\alpha=1}^N \mathbf{y}_\alpha - \sum_{\alpha=1}^N \bar{\mathbf{y}} \right) (\bar{\mathbf{y}} - \boldsymbol{\mu})' \\
&= (N\bar{\mathbf{y}} - N\bar{\mathbf{y}}) (\bar{\mathbf{y}} - \boldsymbol{\mu})' \\
&= 0
\end{aligned}$$

Therefore, the result in **c**) is proven by combining the result in Equation (A.2.1) and (A.2.2):

$$\begin{aligned}
L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{Np/2} |\boldsymbol{\Sigma}|^{N/2}} \exp \left(-\frac{1}{2} \text{tr} \left(\boldsymbol{\Sigma}^{-1} (\mathbf{A} + N(\bar{\mathbf{y}} - \boldsymbol{\mu}) (\bar{\mathbf{y}} - \boldsymbol{\mu})') \right) \right) \\
&= \frac{1}{(2\pi)^{Np/2} |\boldsymbol{\Sigma}|^{N/2}} \exp \left\{ -\frac{1}{2} \text{tr} \boldsymbol{\Sigma}^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}) \right\} \quad (\text{A.2.3})
\end{aligned}$$

★

d)

The maximum likelihood estimator for $\boldsymbol{\Sigma}$ is given by $\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \mathbf{A} = \frac{1}{N} \sum_{\alpha=1}^N (\mathbf{y}_\alpha - \bar{\mathbf{y}}) (\mathbf{y}_\alpha - \bar{\mathbf{y}})'$

Proof

From Equation (A.2.3) it follows that $\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{Np}{2} \ln(2\pi) + \frac{N}{2} \ln |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \text{tr} \boldsymbol{\Sigma}^{-1} \mathbf{A} - \frac{N}{2} (\bar{\mathbf{y}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu})$. We see that since $\boldsymbol{\Sigma}^{-1}$ is positive definite, $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is maximised for any $\boldsymbol{\Sigma}$ if $\bar{\mathbf{y}} = \boldsymbol{\mu}$; $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$.

If we then want to maximise $L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in terms of $\boldsymbol{\Sigma}$, when $\bar{\mathbf{y}} = \boldsymbol{\mu}$, we need to maximise $\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{N}{2} \ln |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \text{tr} \boldsymbol{\Sigma}^{-1} \mathbf{A}$.

Applying **Lemma 1**; set $\mathbf{C} = \boldsymbol{\Sigma}^{-1}$ and $\mathbf{D} = \mathbf{A}$ then $\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is maximised where $\hat{\boldsymbol{\Sigma}}^{-1} = N\mathbf{A}^{-1}$; therefore $\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \mathbf{A} = \frac{1}{N} \sum_{\alpha=1}^N (\mathbf{y}_\alpha - \bar{\mathbf{y}}) (\mathbf{y}_\alpha - \bar{\mathbf{y}})'$

★

Appendix B

Standard statistical results used throughout this document

Table B.0.1: Standard statistical results

Description	Result
Bayes' rule	$P(A \cap B C) = P(A B, C) \cdot P(B C)$
Conditional probability	$P(A \cap B) = P(A B) \cdot P(B)$
Law of total probability	$P(A B) = \frac{P(B_i) \cdot P(A B_i)}{\sum_j P(B_j) \cdot P(A B_j)}$
Trace of a constant	$tr(c) = c$ where c is a constant
Sum of a trace	$tr(A + B) = tr(A) + tr(B)$
Trace multiplication	$tr(AB) = tr(BA)$

Bibliography

- [1] Bruno Basso, Davide Cammarano, and Elisabetta Carfagna. Review of crop yield forecasting methods and early warning systems. In *Proceedings of the First Meeting of the Scientific Advisory Committee of the Global Strategy to Improve Agricultural and Rural Statistics, FAO Headquarters, Rome, Italy*, pages 18–19, 2013.
- [2] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.
- [3] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- [4] Konstantinos Blekas, Dimitrios I. Fotiadis, and Aristidis Likas. Greedy mixture learning for multiple motif discovery in biological sequences. *Bioinformatics*, 19(5):607–617, 2003.
- [5] Konstantinos Blekas, Aristidis Likas, Nikolas P Galatsanos, and Isaac E Lagaris. A spatially constrained mixture model for image segmentation. *IEEE transactions on Neural Networks*, 16(2):494–498, 2005.
- [6] Konstantinos Blekas, Christophoros Nikou, N Galatsanos, and Nikolaos V Tsekos. A regression mixture model with spatial constraints for clustering spatiotemporal data. *International Journal on Artificial Intelligence Tools*, 17(05):1023–1041, 2008.
- [7] Jinhai Cai and Zhi-Qiang Liu. Pattern recognition using markov random field models. *Pattern Recognition*, 35(3):725–733, 2002.
- [8] Hélène Caillol, Wojciech Pieczynski, and Alain Hillion. Estimation of fuzzy gaussian mixture and unsupervised statistical image segmentation. *IEEE Transactions on Image Processing*, 6(3):425–440, 1997.
- [9] Steven B. Caudill and Ram N. Acharya. Maximum likelihood estimation of a mixture of normal regressions: starting values and singularities. *Communications in Statistics - Simulation and Computation*, 27(3):667–674, 1998.
- [10] Bernard Chalmond. An iterative gibbsian technique for reconstruction of m-ary images. *Pattern recognition*, 22(6):747–761, 1989.
- [11] Satish Chandra. On the mixtures of probability distributions. *Scandinavian Journal of Statistics*, pages 105–112, 1977.

- [12] Peter Clifford. Markov random fields in statistics. *Disorder in physical systems: A volume in honour of John M. Hammersley*, pages 19–32, 1990.
- [13] A. Clifford Cohen. Estimation in mixture of two normal distributions. *American Statistical Association and American Society for Quality*, 9(1):15–28, 1967.
- [14] Sybil L Crawford. An application of the laplace method to finite mixture distributions. *Journal of the American Statistical Association*, 89(425):259–267, 1994.
- [15] Neil E Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):463–474, 1969.
- [16] A. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal Royal Statistical Society*, 39(1):1–38, 1977.
- [17] Raymond J Dezzani and Ahmad Al-Dousari. Spatial analysis in a markov random field framework: The case of burning oil wells in kuwait. *Journal of geographical systems*, 3(4):387–409, 2001.
- [18] Aristeidis Diplaros, Nikos Vlassis, and Theo Gevers. A spatially constrained generative model and an em algorithm for image segmentation. *IEEE Transactions on Neural Networks*, 18(3):798–808, 2007.
- [19] Edward R Dougherty. *Random Processes for Image and Signal Processing*. SPIE Optical Engineering Press, 1999.
- [20] Sylvia Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.
- [21] Hayit Greenspan, Jacob Goldberger, and Arnaldo Mayer. Probabilistic space-time video modeling via piecewise gmm. *IEEE Transactions on pattern analysis and machine intelligence*, 26(3):384–396, 2004.
- [22] Nicola Greggio, Alexandre Bernardino, Cecilia Laschi, Paolo Dario, and Jose Santos-Victor. Fast estimation of gaussian mixture models for image segmentation. *Machine Vision and Applications*, 23(4):773–789, 2012.
- [23] Geoffrey R. Grimmett. A theorem about random fields. *Bulletin of the London Mathematical Society*, 5(1):81–84, 1973.
- [24] Lalit Gupta and Thotsapon Sortrakul. A gaussian-mixture-based image segmentation algorithm. *Pattern Recognition*, 31(3):315–325, 1998.
- [25] Michael J. Hartley. Estimating mixture of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(364):738–741, 1978.
- [26] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning 2nd edition*. New York: Springer, 2009.
- [27] Richard J Hathaway. A constrained em algorithm for univariate normal mixtures. *Journal of Statistical Computation and Simulation*, 23(3):211–230, 1986.

- [28] Hajo Holzmann, Axel Munk, and Tilmann Gneiting. Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics*, 33(4):753–763, 2006.
- [29] Gerrit Hoogenboom, Jeffrey W. White, and Carlos D. Messina. From genome to crop: integration through simulation modeling. *Field Crops Research*, 90(1):145–163, 2004.
- [30] Albert Sydney Hornby, Sally Wehmeier, and Michael Ashby. *Oxford advanced learner’s dictionary*, volume 1428. Oxford university press Oxford, 7th edition edition, 1995.
- [31] David W Hosmer. On mle of the parameters of a mixture of two normal distributions when the sample size is small. *Communications in Statistics-Theory and Methods*, 1(3):217–227, 1973.
- [32] David W. Hosmer. Estimating mixture of normal distributions and switching regressions: comment. *Journal of the American Statistical Association*, 73(364):741–744, 1978.
- [33] Merrilee Hurn, Ana Justel, and Christian P. Robert. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79, 2003.
- [34] Yongsung Joo, Keunbaik Lee, Joong-Hyuk Min, Seong-Taek Yun, and Trevor Park. Logistic mixture of multivariate regressions for analysis of water quality impacted by agrochemicals. *Environmetrics*, 18(5):499–514, 2007.
- [35] J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27(4):887–906, 1956.
- [36] Nicholas M. Kiefer. Discrete parameter variation: Efficient estimation of a switching regression model. *The Econometric Society*, 46(2):427–434, 1978.
- [37] Brian G. Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360, 1992.
- [38] David B. Lobell and Marshall B. Burke. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11):1443–1452, 2010.
- [39] Zhiwu Lu and Horace H.S. Ip. Image categorization with spatial mismatch kernels. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 397–404. IEEE, 2009.
- [40] Christophoros Nikou, Nikolaos P Galatsanos, and Aristidis C Likas. A class-adaptive spatially variant mixture model for image segmentation. *IEEE Transactions on Image Processing*, 16(4):1121–1130, 2007.
- [41] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, 185(1894):71–110, 1894.
- [42] Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4):695–706, 2006.

- [43] F. Phillips. A constrained approach to estimating switching regressions. *Journal of Econometrics*, 48:241–262, 1991.
- [44] Richard E. Quandt. A new approach estimating switching regressions. *Journal of the American Statistical Association*, 67(338):306–310, 1972.
- [45] Richard E. Quandt and B. Ramsey, James. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(364):730–738, December 1978.
- [46] Johanna Ramarohetra, Benjamin Sultan, Christian Baron, Thomas Gaiser, and Marielle Gosset. How satellite rainfall estimate errors may impact rainfed cereal yield simulation in west africa. *Agricultural and forest meteorology*, 180:118–131, 2013.
- [47] S. Sanjay-Gopal and Thomas J. Hebert. Bayesian pixel classification using spatially variant finite mixtures and the generalized em algorithm. *IEEE Transactions on Image Processing*, 7(7):1014–1028, 1998.
- [48] S. Sherman. Markov random fields and gibbs random fields. *Israel Journal of Mathematics*, 14(1):92–103, 1973.
- [49] Marek Suliga, Rudi Deklerck, and Edgard Nyssen. Markov random field-based clustering applied to the segmentation of masses in digital mammograms. *Computerized Medical Imaging and Graphics*, 32(6):502–512, 2008.
- [50] Henry Teicher. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1):244–248, 1961.
- [51] Henry Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34(4):1265–1269, 1963.
- [52] D.M. Titterington, A.F.M. Smith, and U.E. Makov. Statistical analysis of finite mixture models, 1985.
- [53] James Watson and Andrew Challinor. The relative importance of rainfall, temperature and yield data for a regional-scale crop model. *Agricultural and Forest Meteorology*, 170:47–57, 2013.
- [54] Sidney J. Yakowitz. Unsupervised learning and the identification of finite mixtures. *IEEE Transactions on Information Theory*, 16(3):330–338, 1970.
- [55] Sidney J. Yakowitz and John D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 38(3):659–680, 1968.
- [56] Hui Zhang, Tian Wen, Yuhui Zheng, Danhua Xu, Dingcheng Wang, Thanh Minh Nguyen, and QM Jonathan Wu. Two fast and robust modified gaussian mixture models incorporating local spatial information for image segmentation. *Journal of Signal Processing Systems*, 81(1):45–58, 2015.

- [57] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.