Cluster-Weighted Models with Changepoints

CLUSTER-WEIGHTED MODELS WITH CHANGEPOINTS

BY

CAMERON ROOPNARINE, B.Math

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2023)	McMaster University
(Mathematics & Statistics)	Hamilton, Ontario, Canada

TITLE:	Cluster-Weighted Models with Changepoints
AUTHOR:	Cameron Roopnarine B.Math, (Statistics) University of Waterloo, Waterloo, Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	ix, 56

To my friends and family.

Abstract

A flexible family of mixture models known as cluster-weighted models (CWMs) arise when the joint distribution of a response variable and a set of covariates can be modelled by a weighted combination of several component distributions. We introduce an extension to CWMs where changepoints are present. Similar to the finite mixture of regressions (FMR) with changepoints, CWMs with changepoints are more flexible than standard CWMs if we believe that changepoints are present within the data. We consider changepoints within the linear Gaussian CWM, where both the marginal and conditional densities are assumed to be Gaussian. Furthermore, we consider changepoints within the Poisson and Binomial CWM. Model parameter estimation and performance of some information criteria are investigated through simulation studies and two real-world datasets.

Acknowledgements

I am deeply grateful to my supervisor, Dr. Paul McNicholas, for his invaluable guidance, support, and for giving me the opportunity to conduct research alongside him at McMaster University.

Furthermore, I would like to thank Dr. Pratheepa Jeganathan, Dr. Katherine Davies, and Dr. Paul McNicholas for being on my examination committee.

Lastly, I extend my sincere thanks to my friends and family for their support throughout this journey.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Changepoints, also known as breakpoints, refer to locations in the data where there is a significant change in the relationship between the dependent and independent variables. That is, changepoints denote the points at which the linear or non-linear relationship between the variables undergoes a sudden shift. Identifying changepoints in an ordinary regression framework is crucial as they indicate potential shifts in the underlying process being studied.

In a regular regression setting, changepoints have been widely studied before, and are commonly known as piecewise or segmented linear regression problems. Within a model-based clustering framework, changepoints can also be applied for models that assume one dependent and many independent variables. In this thesis, we introduce changepoints within a general conditional mixture model called a cluster-weighted model (CWM).

In Chapter 2, we give a brief summary of the history behind clustering and state some finite mixture models, as well as the definitions for such models. We outline the implementation and estimation of the CWM with changepoints in Chapter 3. Then, we perform some simulations to assess the performance of our models in Chapter 4. In Chapter 5, we apply our model to a couple of real-world datasets. Lastly, we give some concluding remarks and possible future work in Chapter 6.

Chapter 2

Background

2.1 History of Clustering

For some context of clustering, we first state the formulation of clustering given by Tiedeman (1955), which is summarized in McNicholas (2016). Suppose there are G groups, where each group is generated by a Gaussian density function. Then, remove the type identification of each group, and we have a mixture of unknown densities. Finally, we want to reconstruct the G Gaussian density functions of types. The procedure given by Tiedeman (1955) is what we know as model-based clustering today, and Wolfe (1965) developed one of the first computer programs for computing the maximum likelihood estimates (for the means and covariances) from a mixture of multivariate Gaussian distributions with a maximum of five variables and six components. Wolfe (1963) gives another definition of clustering in terms of similarity:

> A type is a set of objects which are more similar to each other than they are to objects not members of the set.

However, the criterion of a "similarity" measure is often difficult to define, which is the reason a "cluster" in general cannot be precisely defined (Estivill-Castro, 2002). Building upon the definitions given by Tiedeman (1955) and Wolfe (1963), McNicholas (2016) defines a cluster in the context of model-based clustering as follows:

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

There are a couple of things to note about this definition. First, the component should be unimodal. Otherwise, either the wrong mixture distribution is being used, or not enough components are being estimated. Second, the mixture model should be appropriate in the sense that the model has adequate "flexibility, or parameterization, to fit the data" (McNicholas, 2016). We proceed with model-based clustering in light of McNicholas (2016).

2.2 Varieties of Mixture Models

Finite mixture models are a powerful device for clustering and classification when assuming that each mixture component represents a group (or cluster) in the original data (Titterington *et al.*, 1985). Broadly speaking, there are two classes of mixture models: unconditional and conditional mixture models.

Unconditional mixture models, also known as finite mixture of distributions (FMD), assume that there are no exogenous variables that explain the means and variances of each group. A couple of examples of unconditional mixture models include: k-means (MacQueen, 1967) and finite mixture models (McLachlan *et al.*, 2019).

Conditional mixture models allow for simultaneous classification of observations as regressions, while also estimating the means and variances of the dependent variable within each group. A couple of examples of conditional mixture models include: finite mixture of regressions (FMR; DeSarbo and Cron, 1988) and finite mixtures of generalized linear models (FGLM; Wedel and DeSarbo, 1995). Similar to regular regression models, both the FMR and FGLM do not make any distributional assumptions on the covariates for clustering; that is, both models assume "assignment independence", which may be inadequate for many applications (Punzo and McNicholas, 2017). Hence, finite mixture of regressions with concomitant variables (FMRC) were introduced to account for random covariates within a FMR framework (Dayton and Macready, 1988; Wedel, 2002). Within a FMRC model, the weights of the mixture model depend on the concomitant variables (i.e., random covariates), where the weights are usually modelled by a multinomial logistic distribution. Young (2014) introduced changepoints within the FMR framework (i.e., under the assumption that the response is Gaussian and the covariates are non-random quantities), which is implemented in the function `segregmixEM` from the `mixtools` package in the R software (Benaglia *et al.*, 2009; R Core Team, 2023).

Another conditional mixture model is the cluster-weighted model (CWM; Gershenfeld, 1997). The original paper by Gershenfeld (1997) is formulated under Gaussian and linearity assumptions. However, CWMs have been shown to be flexible and can take on many distributional assumptions. For example, Ingrassia *et al.* (2012) introduced a CWM under the assumption that the response follows a Student-t distribution, which was shown to provide a better fit for noisy data. Extending the work from Ingrassia *et al.* (2012), Ingrassia *et al.* (2014) introduced a family of twelve mixture models with random covariates that are nested within the Student-t CWM.

Ingrassia *et al.* (2015) presented CWMs where the covariates have mixed data types (e.g., both Gaussian and Poisson) assuming that the response belongs to the exponential family, which is implemented in the `flexCWM` package in R (Mazza *et al.*, 2018). Similar to the purpose of the Student-t CWM, Punzo and McNicholas (2017) provided a contaminated Gaussian CWM; that is, a CWM that assumes the data has outliers. Furthermore, Punzo (2014) introduced a polynomial Gaussian CWM in the case where the covariates are univariate. In our thesis, we introduce the segmented CWM, henceforth abbreviated as sCWM (i.e., a cluster-weighted model with change-points), under the assumption that our response follows either a Gaussian, Poisson, or Binomial distribution.

2.3 Finite Mixture Model (FMM)

Suppose we have data of the form $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, where each observation is a realization of a p-dimensional random vector $\boldsymbol{X}$ with joint probability function $p(\boldsymbol{x})$. The G-component finite mixture model (FMM) is defined by

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

where $f_g(\boldsymbol{x}; \boldsymbol{\theta}_g)$ is the probability density of $\boldsymbol{x}$ (parameterized by $\boldsymbol{\theta}_g$), π_g is the g^{th} mixing proportion (with constraints $\sum_{g=1}^{G} \pi_g = 1$ and $\pi_g > 0$ for all $g = 1, \dots, G$), and $\boldsymbol{\vartheta} = \{\pi_g, \boldsymbol{\theta}_g \mid g = 1, \dots, G\}$ is the set of all parameters in the model. In many clustering applications, we set $f_g(\boldsymbol{x}; \boldsymbol{\theta}_g) = f(\boldsymbol{x}; \boldsymbol{\theta}_g)$ for all $g = 1, \dots, G$; that is, the observations within each cluster arise from the same distribution.

2.4 Finite Mixture of Regressions (FMR)

Let y_i and $\boldsymbol{x}_i$ denote the realizations of Y_i and $\boldsymbol{X}_i$ for $i = 1, \dots, N$, respectively. The G-component finite mixture of regressions (FMR) model is defined by

$$f(y; \boldsymbol{x}, \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \varphi(y; \boldsymbol{x}'\boldsymbol{\beta}_g, \sigma_g^2)\pi_g, \tag{2.2}$$

where $\varphi(\,\cdot\,; \boldsymbol{x}'\boldsymbol{\beta}_g, \sigma_g^2)$ is the Gaussian probability density function with mean $\boldsymbol{x}'\boldsymbol{\beta}_g$ and variance σ_g^2, π_g is the g^{th} mixing proportion, and $\boldsymbol{\vartheta} = \{\boldsymbol{\beta}_g, \sigma_g^2, \pi_g \mid g = 1, \dots, G\}$ is the set of all parameters in the model.

2.5 Cluster-Weighted Model (CWM)

Suppose we have data of the form $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N)$, where each observation is a realization of $(\boldsymbol{X}, Y)$ defined on Ω with joint probability function $p(\boldsymbol{x}, y)$. We say that $\boldsymbol{X}$ is the p-dimensional covariate vector and Y is the response variable. Now, suppose that Ω can be partitioned into G disjoint groups, $\Omega_1, \Omega_2, \dots, \Omega_G$; that is, $\Omega = \bigcup_{g=1}^{G} \Omega_g$ where $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$. The G-component cluster-weighted model (CWM) is defined by

$$p(\boldsymbol{x}, y; \boldsymbol{\vartheta}) = \sum_{g=1}^{G} p(y \mid \boldsymbol{x}, \Omega_g)p(\boldsymbol{x} \mid \Omega_g)\pi_g = \sum_{g=1}^{G} p(y \mid \boldsymbol{x}; \boldsymbol{\xi}_g)p(\boldsymbol{x}; \boldsymbol{\alpha}_g)\pi_g, \tag{2.3}$$

where $p(y \mid \boldsymbol{x}, \Omega_g) = p(y \mid \boldsymbol{x}; \boldsymbol{\xi}_g)$ is the conditional density of Y given $\boldsymbol{X} = \boldsymbol{x}$ in Ω_g (parameterized by $\boldsymbol{\xi}_g$), $p(\boldsymbol{x} \mid \Omega_g) = p(\boldsymbol{x}; \boldsymbol{\alpha}_g)$ is the probability density of $\boldsymbol{X}$ in Ω_g (parameterized by $\boldsymbol{\alpha}_g$), $\pi_g = p(\Omega_g)$ is the g^{th} mixing proportion, and $\boldsymbol{\vartheta} = \{\boldsymbol{\xi}_g, \boldsymbol{\alpha}_g, \pi_g \mid g = 1, \dots, G\}$ is the set of all parameters in the model.

For this thesis, we assume that the marginals are Gaussian, i.e., $\boldsymbol{X} \mid \Omega_g \sim$

$\mathcal{N}_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, so that $p(\boldsymbol{x} \mid \Omega_g) = \varphi_p(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, where φ_p is the probability density of a p-variate Gaussian distribution. Hence, we may write (2.3) as

$$p(\boldsymbol{x}, y; \boldsymbol{\vartheta}) = \sum_{g=1}^{G} p(y \mid \boldsymbol{x}; \boldsymbol{\xi}_g)\varphi_p(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\pi_g, \tag{2.4}$$

and we refer to (2.4) as the generalized linear Gaussian CWM (GLGCWM). We will now specify the types of CWMs we will use within this thesis.

2.5.1 Linear Gaussian CWM

If $Y \mid \boldsymbol{x}, \Omega_g \sim \mathcal{N}(\eta(\boldsymbol{x}; \boldsymbol{\beta}_g), \sigma^2_{Y_g})$, then $p(y \mid \boldsymbol{x}, \Omega_g) = \varphi(y; \eta(\boldsymbol{x}; \boldsymbol{\beta}_g), \sigma^2_{Y_g})$, where $\boldsymbol{\beta}_g = (\beta_{0g}, \boldsymbol{\beta}'_{1g})'$, and $\eta(\boldsymbol{x}; \boldsymbol{\beta}_g) = \beta_{0g} + \boldsymbol{\beta}'_{1g}\boldsymbol{x}$. Hence, we may write (2.4) as

$$p(\boldsymbol{x}, y; \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \varphi(y; \eta(\boldsymbol{x}; \boldsymbol{\beta}_g), \sigma^2_{Y_g})\varphi_p(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\pi_g, \tag{2.5}$$

where φ (φ_p) is the probability density of a univariate (p-variate) Gaussian distribution. We refer to (2.5) as the linear Gaussian CWM. Here, $\boldsymbol{\xi}_g = \{\boldsymbol{\beta}_g, \sigma^2_{Y_g} | g = 1, \dots, G\}$ and $\boldsymbol{\alpha}_g = \{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \mid g = 1, \dots, G\}$, so $\boldsymbol{\vartheta} = \{\boldsymbol{\beta}_g, \sigma^2_{Y_g}, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \pi_g \mid g = 1, \dots, G\}$ is the set of all parameters for the linear Gaussian CWM.

2.5.2 Poisson CWM

If Y takes values in $\{0, 1, 2, \dots\}$ and $Y \mid \boldsymbol{x}, \Omega_g \sim \text{Poisson}(\eta(\boldsymbol{x}; \boldsymbol{\beta}_g))$, then

$$p(y \mid \boldsymbol{x}; \boldsymbol{\beta}_g) = \frac{\exp(-\eta(\boldsymbol{x}; \boldsymbol{\beta}_g))(\eta(\boldsymbol{x}; \boldsymbol{\beta}_g))^y}{y!}, \text{ for } y = 0, 1, \dots, \tag{2.6}$$

where $\eta(\boldsymbol{x}; \boldsymbol{\beta}_g) = \exp(\beta_{0g} + \boldsymbol{\beta}'_{1g}\boldsymbol{x})$. If we assume (2.6) in our GLGCWM, we have the Poisson CWM. Here, $\boldsymbol{\xi}_g = \{\boldsymbol{\beta}_g \mid g = 1, \dots, G\}$ and $\boldsymbol{\alpha}_g = \{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \mid g = 1, \dots, G\}$, so

$\boldsymbol{\vartheta} = \{\boldsymbol{\beta}_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \pi_g \mid g = 1, \dots, G\}$ is the set of all parameters for the Poisson CWM.

2.5.3 Binomial CWM

If Y takes values in $\{0, 1, \dots, M\}$, where M is a positive integer, and $Y \mid \boldsymbol{x}, \Omega_g \sim \text{Binomial}(M, \eta(\boldsymbol{x}; \boldsymbol{\beta}_g))$, then

$$p(y \mid \boldsymbol{x}; \boldsymbol{\beta}_g) = \binom{M}{y} \big(\eta(\boldsymbol{x}; \boldsymbol{\beta}_g)\big)^{y} \big(1 - \eta(\boldsymbol{x}; \boldsymbol{\beta}_g)\big)^{M-y}, \text{ for } y = 0, 1, \dots, M, \tag{2.7}$$

where

$$\eta(\boldsymbol{x}; \boldsymbol{\beta}_g) = \frac{\exp(\beta_{0g} + \boldsymbol{\beta}'_{1g}\boldsymbol{x})}{1 + \exp(\beta_{0g} + \boldsymbol{\beta}'_{1g}\boldsymbol{x})} := \text{expit}(\beta_{0g} + \boldsymbol{\beta}'_{1g}\boldsymbol{x}).$$

If we assume (2.7) in our GLGCWM, we have the Binomial CWM. Here, $\boldsymbol{\xi}_g = \{\boldsymbol{\beta}_g \mid g = 1, \dots, G\}$ and $\boldsymbol{\alpha}_g = \{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \mid g = 1, \dots, G\}$, so $\boldsymbol{\vartheta} = \{\boldsymbol{\beta}_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \pi_g \mid g = 1, \dots, G\}$ is the set of all parameters for the Binomial CWM.

2.6 Performance

2.6.1 Overall Model Performance

The Bayesian information criterion (BIC; Schwarz, 1978) has been shown to not underestimate the number of components and consistently estimates the number of components under certain regularity conditions in a mixture model setting (Leroux, 1992; Keribin, 2000). For a model with parameters $\boldsymbol{\vartheta}$, the BIC is given by

$$\text{BIC} = 2\ell(\hat{\boldsymbol{\vartheta}}) - \nu \log N, \tag{2.8}$$

where $\ell(\hat{\boldsymbol{\vartheta}})$ is the maximized observed-data log-likelihood, $\hat{\boldsymbol{\vartheta}}$ is the final estimate of $\boldsymbol{\vartheta}$, ν is the number of free parameters, and N is the number of observations. Note that we select our final model by maximizing (2.8). If we do not know the number of changepoints beforehand, Muggeo (2008) also suggests using the BIC to select the best model (in a non-mixture setting). The BIC can also be useful for comparing a regular CWM to an sCWM to determine which model is best for the data. However, we note that the BIC for a FMR and CWM cannot be compared in most cases (see Ingrassia *et al.*, 2012).

2.6.2 Clustering Performance

The adjusted Rand index (ARI; Hubert and Arabie, 1985) can be used to assess clustering performance for a simulation setting as we know the true classifications. Let $\mathscr{P}$ be the predicted partition and $\mathscr{T}$ be the true partition of classifications. Define a as the number of pairs of observations that are in the same clusters in both $\mathscr{P}$ and $\mathscr{T}$, b as the number of pairs of observations that are in different clusters in both $\mathscr{P}$ and $\mathscr{T}$, c as the number of pairs of observations that are in the same clusters in $\mathscr{P}$ and different clusters in $\mathscr{T}$, d as the number of pairs of observations that are in different clusters in $\mathscr{P}$ and the same clusters in $\mathscr{T}$. The original Rand index (RI; Rand, 1971) is given by

$$\text{RI} = \frac{a+b}{a+b+c+d}.$$

In non-mathematical terms, the RI quantifies the proportion of agreement between the predicted and true classifications. The RI takes values between 0 and 1 (perfect assignment).

One issue with the RI is that under random assignment the RI has a positive expected value, so the ARI was introduced to account for this issue. The ARI is given by

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]}. \tag{2.9}$$

Hence, the ARI has an expected value of 0 under random assignment, a value of 1 under perfect assignment, and a value of less than 0 under worse than random assignment. For further details about the ARI, see Steinley (2004). In R, we calculate the ARI with the `adjustedRandIndex` function from the `mclust` package. Note that we classify each group after the algorithm converges into G groups via the maximum *a posteriori* (MAP) rule, which assumes a value of 1 if $\max_{g=1,\ldots,G}\{z_{ig}^{(k)}\}$ occurs at component g, and 0 otherwise.

Chapter 3

Methodology

3.1 CWM with Changepoints

To account for changepoints in a GLGCWM, we employ a simple data augmentation technique. Assume that there are c_{jg} changepoints for covariate $j = 1, \dots, p$ in group $g = 1, \dots, G$. We define the augmented covariate vector for observation $i = 1, \dots, N$ and covariate $j = 1, \dots, p$ by

$$\boldsymbol{x}_{ij}(\boldsymbol{\psi}_{jg}) = \left(x_{ij}, (x_{ij} - \psi_{j1g})_+, \dots, (x_{ij} - \psi_{jc_{jg}g})_+\right)',$$

where $(z - \psi)_+ = (z - \psi)\, \mathbb{I}\{z > \psi\}$ and $\mathbb{I}\{\,\cdot\,\}$ is the indicator function, and $\boldsymbol{\psi}_{jg} = (\psi_{j1g}, \dots, \psi_{jc_{jg}g})'$ is the vector of c_{jg} distinct changepoints for covariate j in group g. It is worth noting that when there are no changepoints for a given group g, the augmented covariate vector is equivalent to x_{ij}. We can then define the augmented vector of all covariates for observation $i = 1, \dots, N$ by

$$\boldsymbol{x}_i(\boldsymbol{\psi}_g) = \left(1, \boldsymbol{x}_{i1}(\boldsymbol{\psi}_{1g})', \dots, \boldsymbol{x}_{ip}(\boldsymbol{\psi}_{pg})'\right)', \tag{3.10}$$

where $\boldsymbol{\psi}_g = (\boldsymbol{\psi}'_{1g}, \dots, \boldsymbol{\psi}'_{pg})'$ is the vector of all changepoints in group g, and 1 is added at the beginning of $\boldsymbol{x}_i(\boldsymbol{\psi}_g)$ to account for the intercept regression coefficient. Let $k = 1, \dots, c_{jg}$ be the index for the known number of changepoints, $k^\star = \{0\} \cup k$, and $j^\star = \{0\} \cup j$. Now, the regression coefficients become

$$\boldsymbol{\beta}_g = (\beta_{00g}, \beta_{10g}, \dots, \beta_{1c_1g}, \dots, \beta_{p0g}, \dots, \beta_{pc_pg})',$$

where $\beta_{j^\star k^\star g}$ is the $k^{\star\text{th}}$ regression coefficient for the $j^{\star\text{th}}$ covariate in group g, where $k^\star = 0$ are the regression coefficients for the regular x_{ij}'s and $j^\star = 0$ are the regression coefficients for the intercepts. If we combine all these vectors, we can define the augmented design matrix for group $g = 1, \dots, G$ by

$$\mathbf{X}(\boldsymbol{\psi}_g) = (\boldsymbol{x}_1(\boldsymbol{\psi}_g), \dots, \boldsymbol{x}_N(\boldsymbol{\psi}_g)). \tag{3.11}$$

Using this formulation, we can see that the changepoints are accounted for within the link function of a given CWM. Dropping subscripts i for $\boldsymbol{x}_i(\boldsymbol{\psi}_g)$, for the linear Gaussian CWM, we have $\eta(\boldsymbol{x}; \boldsymbol{\beta}_g) = \boldsymbol{x}(\boldsymbol{\psi}_g)'\boldsymbol{\beta}_g$; for the Binomial CWM, we have

$$\eta(\boldsymbol{x}; \boldsymbol{\beta}_g) = \frac{\exp(\boldsymbol{x}(\boldsymbol{\psi}_g)'\boldsymbol{\beta}_g)}{1 + \exp(\boldsymbol{x}(\boldsymbol{\psi}_g)'\boldsymbol{\beta}_g)};$$

for the Poisson CWM, we have $\eta(\boldsymbol{x}; \boldsymbol{\beta}_g) = \exp(\boldsymbol{x}(\boldsymbol{\psi}_g)'\boldsymbol{\beta}_g)$. Now, $\boldsymbol{\vartheta} := \boldsymbol{\vartheta} \cup \{\boldsymbol{\psi}_g \mid g = 1, \dots, G\}$ and $\boldsymbol{\beta}_g$ will have additional terms corresponding to $\psi_{jc_{jg}g}$ for $j = 1, \dots, p$ and $g = 1, \dots, G$ for all the CWMs defined earlier. Furthermore, the regular CWM is nested within the sCWM since $\beta_{jkg} = 0$ for all $k = 1, \dots, c_{jg}$ in an sCWM yields a regular CWM.

3.2 Identifiability

Before estimating the parameters of an sCWM, it is important to discuss its identifiability. It is known that finite mixture models of the same family are invariant under $G!$ permutations, which is also known as the "label-switching" problem (McLachlan and Peel, 2000, pp. 26–28). The sCWM also suffers the same issue since the joint probability distribution (2.4) is invariant under $G!$ permutations. Furthermore, since we specify the number of changepoints c_{jg} for covariate $j = 1, \dots, p$ in group $g = 1, \dots, G$, we see that we are specifying which component the changepoint occurs at; that is, we have an identifiability issue whenever $c_{jg} \neq c_{jg'}$ for all $g = g'$ since we do not know which group the changepoint will occur at. Hence, to address both of these issues, we follow a similar approach to Aitkin and Rubin (1985):

$$\pi_1 \leq \pi_2 \leq \cdots \leq \pi_G.$$

Therefore, if we specify $c_{j1} = 2$, $c_{j2} = 1$, and $c_{j3} = 0$ for a CWM with three groups, we can identify that the smallest component will have two changepoints, the second-smallest component will have one changepoint, and the largest component will have zero changepoints. The drawbacks of this approach is that we will have many permutations to test for c_{jg}. It is important to note that if at least two of the mixing proportions are equal, the segmented CWM will still be unidentifiable. To address this issue, it is possible to also order other parameters. For this thesis, we will not focus on cases where at least two of the mixing proportions are equal, as the models can become quite cumbersome to test.

3.3 Estimation

For a regular CWM without changepoints, maximum likelihood estimation is usually performed using the expectation-maximization (EM) algorithm from Dempster *et al.* (1977). However, in the case of a CWM with changepoints, we cannot directly maximize both the changepoints and the other parameters in one maximization step of the algorithm. Therefore, we perform parameter estimation using an expectation-conditional maximization (ECM) algorithm from Meng and Rubin (1993). The ECM algorithm iterates between three steps, including one E-step (expectation) and two CM-steps (conditional maximization), until convergence is achieved.

Given a random sample $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N)$ of $(\boldsymbol{X}, Y)$, the observed-data likelihood function is given by

$$\mathcal{L}(\boldsymbol{\vartheta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i, y_i; \boldsymbol{\vartheta}) = \prod_{i=1}^{N} \sum_{g=1}^{G} p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\xi}_g) p(\boldsymbol{x}_i; \boldsymbol{\alpha}_g) \pi_g,$$

and the observed-data log-likelihood function is given by

$$\ell(\boldsymbol{\vartheta}) = \sum_{i=1}^{N} \log\left(\sum_{g=1}^{G} p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\xi}_g) p(\boldsymbol{x}_i; \boldsymbol{\alpha}_g) \pi_g \right).$$

Let z_{ig} be an indicator variable to denote component membership, i.e.,

$$z_{ig} = \begin{cases} 1, & \text{if } (\boldsymbol{x}_i, y_i) \text{ belongs to component } g, \\ 0, & \text{otherwise,} \end{cases} \qquad \text{for } i = 1, \dots, N;\ g = 1, \dots, G.$$

Hence, the complete-data likelihood is given by

$$\mathcal{L}_{\mathrm{c}}(\boldsymbol{\vartheta}) = \prod_{i=1}^{N} \prod_{g=1}^{G} \left(p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\xi}_g) p(\boldsymbol{x}_i; \boldsymbol{\alpha}_g) \pi_g\right)^{z_{ig}},$$

and the complete-data log-likelihood is given by

$$\begin{aligned}\ell_{\mathrm{c}}(\boldsymbol{\vartheta}) &= \log\left(\prod_{i=1}^{N}\prod_{g=1}^{G}(p(y_i \mid \boldsymbol{x}_i;\boldsymbol{\xi}_g)p(\boldsymbol{x}_i;\boldsymbol{\alpha}_g)\pi_g)^{z_{ig}}\right)\\ &= \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\log\pi_g + \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\log(p(y_i \mid \boldsymbol{x}_i;\boldsymbol{\xi}_g))\\ &\quad+\sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\log(p(\boldsymbol{x}_i;\boldsymbol{\alpha}_g)). \end{aligned} \tag{3.12}$$

Our objective is to maximize $\mathcal{L}(\boldsymbol{\vartheta})$ using $\mathcal{L}_{\mathrm{c}}(\boldsymbol{\vartheta})$ using our ECM algorithm.

3.3.1 E-step

The E-step on the $(k+1)^{\text{th}}$ iteration, $k = 0, 1, ...$, requires calculating the expectation of the complete-data log-likelihood given the observed data and the current estimate of $\boldsymbol{\vartheta}$ at the $(k+1)^{\text{th}}$ iteration, denoted by $\boldsymbol{\vartheta}^{(k)}$. Since $\ell_{\mathrm{c}}(\boldsymbol{\vartheta})$ is linear with respect to z_{ig}, we can simply calculate the current conditional expectation of Z_{ig} given the observed data, where Z_{ig} is the random variable corresponding to z_{ig}. Therefore, for $i = 1, \dots, N$ and $g = 1, \dots, G$, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\vartheta}^{(k)}}\left[Z_{ig} \,\middle|\, (\boldsymbol{x}_i, y_i)\right] &:= z_{ig}^{(k)} \\ &= \frac{\pi_g^{(k)}p(y_i \mid \boldsymbol{x}_i;\boldsymbol{\xi}_g^{(k)})p(\boldsymbol{x}_i;\boldsymbol{\alpha}_g^{(k)})}{p(\boldsymbol{x}_i, y_i;\boldsymbol{\vartheta}^{(k)})},\end{aligned} \tag{3.13}$$

which is the posterior probability that $(\boldsymbol{x}_i, y_i)$ belongs to group g using the current estimate $\boldsymbol{\vartheta}^{(k)}$ for $\boldsymbol{\vartheta}$. For the CM-steps, on the $(k+1)^{\text{th}}$ iteration, $k = 0, 1, ...$, we want to maximize the conditional expectation of the complete-data log-likelihood given the observed data. Hence, we replace each z_{ig} in (3.12) by their expectations, i.e., $z_{ig}^{(k)}$, to get

$$
\begin{aligned}
\mathcal{Q}(\boldsymbol{\vartheta};\boldsymbol{\vartheta}^{(k)}) &= \mathbb{E}_{\boldsymbol{\vartheta}^{(k)}}\left[\ell_{\mathrm{c}}(\boldsymbol{\vartheta}) \,\middle|\, (\boldsymbol{x}_i, y_i)\right] \\
&= \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}^{(k)} \log \pi_g + \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}^{(k)} \log(p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\xi}_g)) \\
&\quad + \sum_{i=1}^{N}\sum_{g=1}^{G} z_{ig}^{(k)} \log(p(\boldsymbol{x}_i; \boldsymbol{\alpha}_g)).
\end{aligned}
$$

To perform our CM-steps, we partition $\boldsymbol{\vartheta}$ as $\{\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2\}$, where

$$
\begin{aligned}
\boldsymbol{\vartheta}_1 &= \{\boldsymbol{\psi}_g \mid g = 1, \dots, G\}, \\
\boldsymbol{\vartheta}_2 &= \{\boldsymbol{\xi}_g, \boldsymbol{\alpha}_g, \pi_g \mid g = 1, \dots, G\}.
\end{aligned}
$$

3.3.2 CM-step 1

In the first CM-step, we estimate the changepoints; that is, calculate

$$
\boldsymbol{\vartheta}_1^{(k+1)} = \arg\max_{\boldsymbol{\vartheta}_1} \mathcal{Q}(\boldsymbol{\vartheta};\boldsymbol{\vartheta}^{(k)}), \tag{3.14}
$$

where $\boldsymbol{\vartheta}_2$ is fixed at $\boldsymbol{\vartheta}_2^{(k)}$. There is no closed-form expression for (3.14), but using a first-order Taylor expansion around pre-specified values of changepoints, we can obtain an approximation, which is implemented in the R package `segmented` (Muggeo, 2008).

For the sake of simplicity, consider a simple linear regression with one covariate x_i and one changepoint ψ for observation $i = 1, \dots, N$:

$$
f(x_i; \psi) = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \psi)_+.
$$

Now, the linear approximation to $f(x_i;\psi)$ at $\psi=\tilde{\psi}^{(0)}$ is given by

$$\begin{aligned} f(x_i;\psi) &\approx f(\tilde{\psi})+f'(\tilde{\psi}^{(0)})(x_i-\tilde{\psi}^{(0)}) \\ &= \beta_0+\beta_1 x_i+\beta_2(x_i-\tilde{\psi}^{(0)})_+ - \beta_2(\psi-\tilde{\psi}^{(0)})\,\mathbb{I}\{x_i>\tilde{\psi}^{(0)}\} \\ &= \beta_0+\beta_1 x_i+\beta_2(x_i-\tilde{\psi}^{(0)})_+ - \alpha\,\mathbb{I}\{x_i>\tilde{\psi}^{(0)}\}, \end{aligned}$$

where $\alpha=\beta_2(\psi-\tilde{\psi}^{(0)})$ is described as a re-parameterization of ψ, which accounts for the changepoint estimation (Muggeo, 2003). Note that $-\mathbb{I}\{x_i>\tilde{\psi}^{(0)}\}$ is the first derivative of $(x_i-\psi)_+$ with respect to ψ evaluated at $\psi=\tilde{\psi}^{(0)}$. Therefore, we iteratively fit

$$\beta_0+\beta_1 x_i+\beta_2(x_i-\tilde{\psi}^{(t)})_+ - \alpha\,\mathbb{I}\{x_i>\tilde{\psi}^{(t)}\}, \text{ for } t=0,1,\dots.$$

The initial changepoint $\tilde{\psi}^{(0)}$ is the 0.5 quantile of $\boldsymbol{x}=(x_1,\dots,x_N)'$. Then, update the changepoint via

$$\tilde{\psi}^{(t+1)}=\tilde{\psi}^{(t)}+\frac{\hat{\alpha}}{\hat{\beta}_2}, \text{ for } t=0,1,\dots.$$

Muggeo (2003) refers to $\hat{\alpha}$ as the "gap" measurement since it measures the difference between two regressions (i.e., before and after the estimate of $\tilde{\psi}^{(t)}$). The algorithm converges when $\hat{\alpha}\approx 0$.

For the general case, given initial changepoints $\tilde{\psi}^{(0)}_{j1g},\dots,\tilde{\psi}^{(0)}_{jc_{jg}g}$, we iteratively fit

$$\beta_{j0g}x_{ij}+\sum_{k=1}^{c_{jg}}(\beta_{jkg}(x_{ij}-\tilde{\psi}^{(t)}_{jkg})_+ - \alpha_{jkg}\,\mathbb{I}\{x_{ij}>\tilde{\psi}^{(t)}_{jkg}\}), \tag{3.15}$$

for $i=1,\dots,N$, $j=1,\dots,p$, $t=0,1,\dots$, and $g=1,\dots,G$. We partition the data into G groups via the MAP, and then we iteratively fit linear models in (3.15) for each group g. Finally, we update the changepoints via

$$\tilde{\psi}_{jkg}^{(t+1)} = \tilde{\psi}_{jkg}^{(t)} + \frac{\hat{\alpha}_{jkg}}{\hat{\beta}_{jkg}},$$

and stop when $|\hat{\alpha}_{jkg}| < 1 \times 10^{-5}$ for all j, k, g, which is the default convergence criterion in `segmented`. While it is possible to iteratively fit linear models in (3.15) for each group g with weights $z_{ig}^{(k)}$, we instead partitioned the data as above since it yielded more sensible results.

We note that `segmented` estimates the changepoints within the quantiles α and $1-\alpha$ of a given covariate, with a default value of $\alpha = \max(0.5, 1/N)$. Also, the initial breakpoints at $t = 0$ are the

$$\frac{1}{c_{jg}+1}, \frac{2}{c_{jg}+1} \ldots, \frac{c_{jg}}{c_{jg}+1}$$

quantiles of a given covariate (in a given group), which is the default starting values in `segmented`. To obtain the estimates of the coefficients, we utilize R's functions `lm` (for a Gaussian response) and `glm` (for a Poisson and Binomial response). It is worth noting that the algorithm depends on the existence of a changepoint and the initial value $\tilde{\psi}^{(0)}$. It is possible that even if a changepoint exists, the algorithm will fail due to the nature of the data (Muggeo, 2003). For example, the algorithm can fail when the coefficient is small for the intercept in a Poisson model or the data's sample size is small (Muggeo, 2003). Also, the `segmented` package states that they implemented a bootstrap restarting algorithm given in Wood (2001) to escape possible local optima of the objective function. If the estimation of the changepoints fails, we simply set $\boldsymbol{\vartheta}_1^{(k+1)} = \boldsymbol{\vartheta}_1^{(k)}$ (assuming changepoints were estimated in the previous iteration) and proceed with the next CM-step. If no changepoints were estimated in the previous iteration, we proceed without the data augmentation (i.e., the next CM-step is equivalent to the maximization step of a regular CWM).

3.3.3 CM-step 2

In the second CM-step, we calculate

$$\boldsymbol{\vartheta}_2^{(k+1)} = \arg\max_{\boldsymbol{\vartheta}_2} \mathcal{Q}(\boldsymbol{\vartheta}; \boldsymbol{\vartheta}^{(k)}),$$

where $\boldsymbol{\vartheta}_1$ is fixed at $\boldsymbol{\vartheta}_1^{(k)}$. We can write

$$\mathcal{Q}(\boldsymbol{\vartheta}; \boldsymbol{\vartheta}^{(k)}) = \mathcal{Q}_1(\boldsymbol{\pi}; \boldsymbol{\vartheta}^{(k)}) + \mathcal{Q}_2(\boldsymbol{\xi}; \boldsymbol{\vartheta}^{(k)}) + \mathcal{Q}_3(\boldsymbol{\alpha}; \boldsymbol{\vartheta}^{(k)}), \tag{3.16}$$

Note that we can maximize each term of (3.16) separately since the cross-derivatives are zero. The second CM-step is similar to the existing literature for estimating regular CWMs (see e.g., Dang *et al.*, 2017; Ingrassia *et al.*, 2015; Mazza *et al.*, 2018). The major difference is that the calculation of the parameters related to Y have a different link function.

Mixing proportions

Maximizing $\mathcal{Q}_1$ with respect to $\boldsymbol{\pi} = (\pi_1, \dots, \pi_G)'$ subject to the constraints on these parameters, yields

$$\pi_g^{(k+1)} = \frac{1}{N} \sum_{i=1}^{N} z_{ig}^{(k)}, \text{ for } g = 1, \dots, G.$$

Parameters related to Y

Maximizing $\mathcal{Q}_2$ with respect to $\boldsymbol{\xi}$ is equivalent to maximizing

$$\mathcal{Q}_{2g}(\boldsymbol{\xi}_g;\boldsymbol{\vartheta}^{(k)}) = \sum_{i=1}^{N} z_{ig}^{(k)} \log(p(y_i \mid \boldsymbol{x}_i;\boldsymbol{\xi}_g)) \tag{3.17}$$

with respect to $\boldsymbol{\xi}_g$ for $g = 1, \dots, G$. For a CWM with a response variable being a member of the exponential family, maximizing (3.17) is equivalent to the maximization problem of a generalized linear model (GLM) using the complete data with weights $z_{ig}^{(k)}$ for each observation $(\boldsymbol{x}_i, y_i)$. More details can be found in Wedel and DeSarbo (1995) and McLachlan and Peel (2000, pp. 147–148). Note that for the estimation of $\boldsymbol{\xi}_g$, the link function utilizes the augmented covariate vector $\boldsymbol{x}(\boldsymbol{\psi}_g)$, which leads to different estimates for $\boldsymbol{\beta}_g$ compared to a regular CWM. For example, consider the linear case: instead of regressing y on $\boldsymbol{x}$, we regress y on $\boldsymbol{x}(\boldsymbol{\psi}_g)$.

Parameters related to $\boldsymbol{X}$

Maximizing $\mathcal{Q}_3$ with respect to $\boldsymbol{\alpha}$ depends on the specification of $\boldsymbol{\Sigma}_g$, where many parsimonious structures for $\boldsymbol{\Sigma}_g$ can be found in Table A.9. Assuming an unconstrained (VVV) model, the updates for $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ for $g = 1, \dots, G$ are

$$\begin{aligned}
\boldsymbol{\mu}_g^{(k+1)} &= \frac{\sum_{i=1}^{N} z_{ig}^{(k)} \boldsymbol{x}_i}{\sum_{i=1}^{N} z_{ig}^{(k)}}, \\
\boldsymbol{\Sigma}_g^{(k+1)} &= \frac{\sum_{i=1}^{N} z_{ig}^{(k)} (\boldsymbol{x}_i - \boldsymbol{\mu}_g^{(k+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_g^{(k+1)})'}{\sum_{i=1}^{N} z_{ig}^{(k)}}.
\end{aligned}$$

For the constrained models, the M-step varies based on the decomposition of the covariance matrix $\boldsymbol{\Sigma}_g$; the updates are given in Punzo and McNicholas (2016), which were derived from Celeux and Govaert (1995). The focus on this thesis will be on low-dimensional datasets, so we will not perform any estimation with the constrained model. In R, the updates are calculated via the `mstep` function from the `mclust` package (Scrucca *et al.*, 2016).

3.4 Initialization

The EM algorithm (and its variants) are known to depend heavily on the starting values (Baudry and Celeux, 2015). The most common way to start an EM algorithm is to provide the initial values of z_{ig} in (3.13) for the first E-step of the algorithm (McLachlan and Peel, 2000). We proceed with a multiple random soft initialization strategy; that is, we generate G positive values in $\boldsymbol{z}_i^{(0)} = (z_{i1}^{(0)}, \dots, z_{iG}^{(0)})'$ such that $\sum_{g=1}^{G} z_{ig}^{(0)} = 1$ for all $i = 1, \dots, N$ and run our ECM algorithm $r = 1, \dots, R$ times. Then, we select the model that maximizes the observed-data likelihood among the R runs. To generate these values, we utilize the `z_ig_random_soft` function from the `mixture` package in R (Pocuca *et al.*, 2022). Also, we must specify the number of changepoints c_{jg} for covariate $j = 1, \dots, p$ in group $g = 1, \dots, G$, which also implies that the number of groups G must be specified. Note that we assume that number of changepoints are known, so they are not parameters; however, the changepoints $\boldsymbol{\psi}_g$ are unknown for group $g = 1, \dots, G$, so they are parameters. After the ECM estimates are obtained, we compute the MAP classification to plot the results.

3.5 Convergence Criteria

Let $\ell^{(k+1)}$ denote the observed-data log-likelihood on the $(k+1)^{\text{th}}$ iteration, $k = 0, 1, \dots$. The stopping criterion associated with an EM algorithm is usually in terms of size of the relative change in the parameter estimates or the log-likelihood, i.e.,

$$\ell^{(k)} - \ell^{(k-1)} < \epsilon, \tag{3.18}$$

for some small positive real-valued ϵ. However, this is a "measure of lack of progress, but not of actual convergence" (Lindstrom and Bates, 1988). The stopping criterion in (3.18) has been shown to underestimate the correct value of the log-likelihood (McNicholas *et al.*, 2010). Böhning *et al.* (1994) applied the Aitken's acceleration procedure to a sequence of log-likelihood values to give an estimate of the limiting value for the log-likelihood. The Aitken's acceleration (Aitken, 1926) at iteration k is given by

$$a^{(k)} = \frac{\ell^{(k+1)} - \ell^{(k)}}{\ell^{(k)} - \ell^{(k-1)}},$$

which leads to the asymptotic estimate of the log-likelihood at iteration $k+1$, given by

$$\ell_{\infty}^{(k+1)} = \ell^{(k)} + \frac{\ell^{(k+1)} - \ell^{(k)}}{1 - a^{(k)}}. \tag{3.19}$$

From (3.19), Lindsay (1995) proposed that the algorithm can be stopped when

$$\ell_{\infty}^{(k+1)} - \ell^{(k+1)} < \epsilon, \tag{3.20}$$

for some small positive real-valued ϵ. McNicholas *et al.* (2010) proposed a similar stopping criterion that is no less strict than (3.20); that is, the algorithm can be stopped when

$$\ell_{\infty}^{(k+1)} - \ell^{(k)} \in (0, \epsilon), \tag{3.21}$$

for some small positive real-valued ϵ. We will use (3.21) with $\epsilon = 1 \times 10^{-4}$ to determine when the algorithm should be stopped. It is worth noting that the `cwm` function from the `flexCWM` package uses the stopping criterion in (3.20) with $\epsilon = 1 \times 10^{-4}$ and the `segregmixEM` function from the `mixtools` package uses the stopping criterion in (3.18) with $\epsilon = 1 \times 10^{-8}$ by default.

Chapter 4

Simulations

4.1 Linear Gaussian CWM

For our first simulation study, we consider $G = 3$ groups and one covariate (i.e., $p = 1$) for a linear Gaussian CWM with three changepoints. The purpose of this study is to demonstrate that when we have random covariates (generated via the Gaussian distribution) with changepoints under a moderately complex setting, the sCWM will outperform both the regular CWM and the FMR with changepoints models on average. Define $\boldsymbol{X}_g = (X_{g1}, X_{g2}, \ldots, X_{gN_g})'$ and $\boldsymbol{Y}_g = (Y_{g1}, Y_{g2}, \ldots, Y_{gN_g})'$ for $g = 1, 2, 3$. First, we generate our covariate vector with respect to each group by

$$\begin{aligned}
\boldsymbol{X}_1 &\sim \mathcal{N}(8, 2^2),\\
\boldsymbol{X}_2 &\sim \mathcal{N}(0, 5^2),\\
\boldsymbol{X}_3 &\sim \mathcal{N}(-5, 3^2),
\end{aligned}$$

where $N_1 = 100$, $N_2 = 175$, and $N_3 = 200$. Next, define regression coefficients $\boldsymbol{\beta}_1 = (2, 3, 12)'$, $\boldsymbol{\beta}_2 = (4, -6, 10, -10)'$, $\boldsymbol{\beta}_3 = (3, -2)'$, and changepoints $\psi_1 = 8$, $\boldsymbol{\psi}_2 = (-7, 0)'$. Finally, we generate our response vector with respect to each group by

$$\begin{aligned}
\boldsymbol{Y}_1 &\sim \mathcal{N}(\mathbf{X}(\psi_1)\boldsymbol{\beta}_1, 3^2),\\
\boldsymbol{Y}_2 &\sim \mathcal{N}(\mathbf{X}(\boldsymbol{\psi}_2)\boldsymbol{\beta}_2, 5^2),\\
\boldsymbol{Y}_3 &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}_3, 4^2).
\end{aligned}$$

Note that $\mathbf{X}(\boldsymbol{\psi}_g)$ is the augmented design matrix given in (3.11) with respect to the g^{th} group, $g = 1, 2, 3$. Therefore, for this simulation study we have $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2', \boldsymbol{X}_3')'$ and $\boldsymbol{Y} = (\boldsymbol{Y}_1', \boldsymbol{Y}_2', \boldsymbol{Y}_3')'$, which implies that our sample size is $N = 475$.

Also, define the number of changepoints for each covariate in each group by $c_{11} = 1$, $c_{21} = 2$, and $c_{31} = 0$. For this simulation study, our set of parameters is given by

$$\boldsymbol{\vartheta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \sigma_{Y_1}, \sigma_{Y_2}, \sigma_{Y_3}, \mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3, \pi_1, \pi_2, \pi_3, \psi_1, \boldsymbol{\psi}_2\},$$

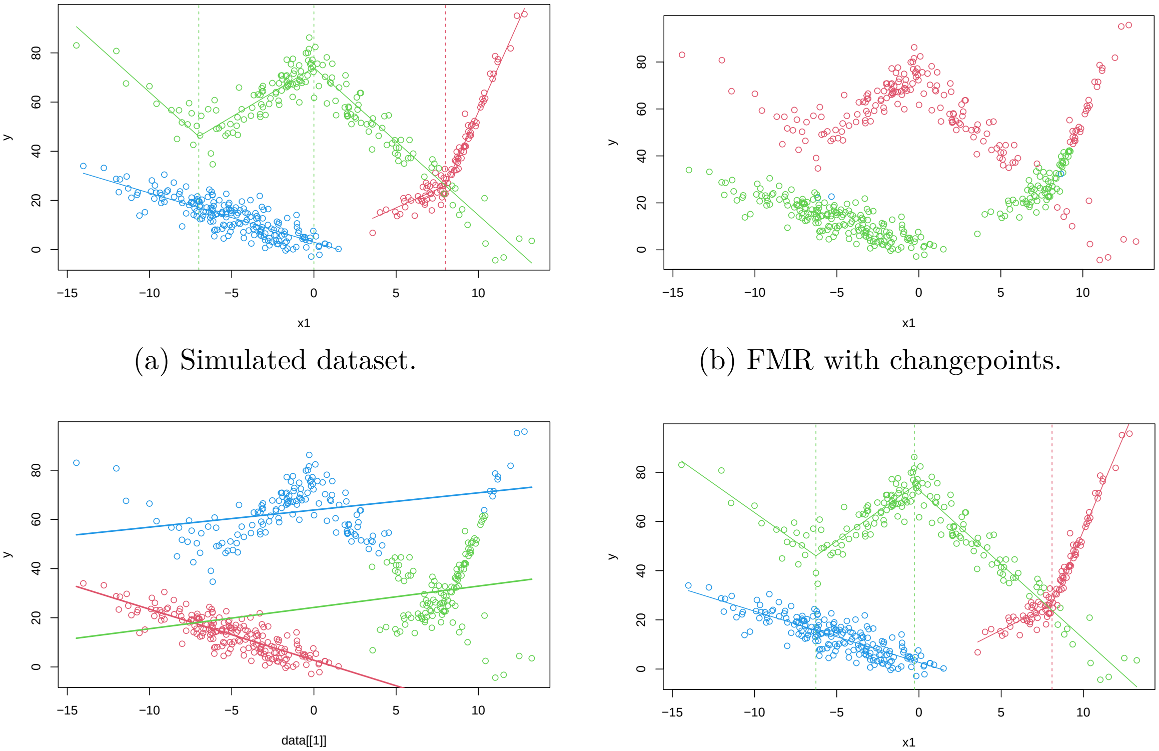
where $\boldsymbol{\beta}_1 = (\beta_{001}, \beta_{101}, \beta_{111})'$, $\boldsymbol{\beta}_2 = (\beta_{002}, \beta_{102}, \beta_{112}, \beta_{122})'$, and $\boldsymbol{\beta}_3 = (\beta_{003}, \beta_{103})'$. Note that the mixing proportions are positive and sum to one, so we only count two free parameters for the mixing proportions. Therefore, we have $\nu = 23$ free parameters for the CWM with changepoints, and also note that there are $\nu = 17$ free parameters for the regular CWM for this simulation study.

The results for one simulation can be visualized in Figure 4.1, where Figure 4.1a shows the true clusters, Figure 4.1b shows the clusters given by a finite mixture of regressions with changepoints model, Figure 4.1c shows the clusters given by a regular CWM, and Figure 4.1d shows the clusters given by a CWM with changepoints. For each plot in Figure 4.1, we colour the MAP estimates for each observation, and the dashed lines given in Figures 4.1a, 4.1c and 4.1d indicate the locations where the changepoints occur. Also, the solid lines given in Figures 4.1a, 4.1c and 4.1d are simply the fitted regression lines on each group. Note that we did not indicate the dashed/solid lines in Figure 4.1b since the locations of changepoints and regression coefficients were estimated poorly with respect to the true values.

It is clear that the FMR with changepoints model (which was estimated assuming $G = 3$ groups) does not estimate the true number of clusters correctly under this simulation study due to the fact that we have random covariates. Furthermore, the regular CWM estimates one cluster correctly (the cluster with no changepoints), but the other two clusters are not correctly estimated due to the fact that we have changepoints. Lastly, the CWM with changepoints model estimates the true number of clusters correctly and the regression segments look similar to the true values for this simulation.

For a more detailed comparison, we generated 100 datasets under this framework, where both the CWM and CWM with changepoints were estimated with $R = 5$ random soft initializations. For `segregmixEM`, we only performed one estimation for each dataset as there is no mention of a multiple random initialization strategy for this package.

A histogram of the ARIs can be visualized in Figure 4.2a for the FMR with changepoints, CWM, and sCWM, noting that the sCWM has a higher ARI in general. Furthermore, a histogram of the BICs can be visualized in Figure 4.2b, and we note that the sCWM always yielded a higher BIC. Note that we cannot compare the BIC for a FMR to an sCWM under this setting as mentioned earlier. The major



(a) Simulated dataset.

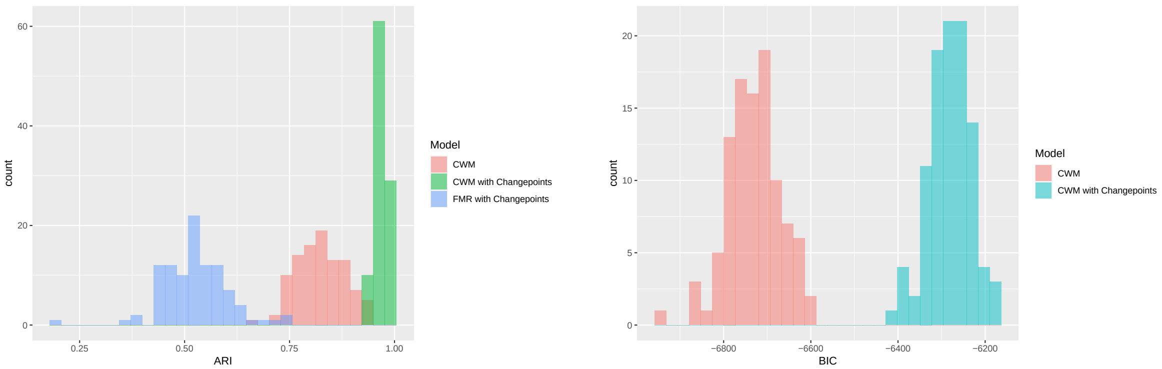
(b) FMR with changepoints.

(c) Regular CWM.

(d) CWM with changepoints.

Figure 4.1: Linear Gaussian CWM with changepoints simulation.

downside of using an sCWM is the computation time, which can be seen in Table 4.1; however, the mean computation time is faster on average compared to a FMR with changepoints. In terms of raw estimation performance, we can view the true parameter values along with their mean and standard deviations in Table 4.2. We see that the mean parameter estimates are close to the true values, but the standard deviations are large for the regression coefficients (especially the intercepts for the groups with changepoints), most likely due to the complicated structure of the simulation. Overall, our model performs well under this setting, but it is computationally expensive.



(a) Histogram of ARIs.

(b) Histogram of BICs.

Figure 4.2: Histograms of ARIs and BICs for the Gaussian simulation.

Table 4.1: Summary of computation time (in seconds).

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Regular CWM	0.740	1.021	1.166	1.230	1.373	2.324
CWM with Changepoints	23.720	30.170	36.940	40.550	48.700	91.070
FMR with Changepoints	1.137	27.057	84.069	94.650	126.399	459.751

Table 4.2: True parameter values along with mean and standard deviations of the parameter estimates for the Gaussian simulation.

Parameter	True values	Mean estimates	Standard deviations
$\boldsymbol{\beta}_1$	$(2, 3, 12)'$	$(2.061, 2.983, 11.980)'$	$(4.382, 0.748, 0.735)'$
$\boldsymbol{\beta}_2$	$(4, -6, 10, -10)'$	$(5.859, -5.771, 9.844, -10.062)'$	$(9.875, 1.169, 1.160)'$
$\boldsymbol{\beta}_3$	$(3, -2)'$	$(3.030, -1.993)'$	$(0.584, 0.096)'$
σ_{Y_1}	3	2.962	0.253
σ_{Y_2}	5	4.917	0.297
σ_{Y_3}	4	4.009	0.219
μ_1	8	7.983	0.194
μ_2	0	-0.020	0.416
μ_3	-5	-5.015	0.206
σ_1	2	1.997	0.251
σ_2	5	4.972	0.254
σ_3	3	3.009	0.154
π_1	0.211	0.211	0.005
π_2	0.368	0.368	0.005
π_3	0.421	0.421	0.001
ψ_1	8	7.994	0.142
$\boldsymbol{\psi}_2$	$(-7, 0)'$	$(-6.915, -0.023)'$	$(0.391, 0.158)'$

4.2 Poisson CWM

For our second simulation study, we consider $G = 2$ groups and one covariate (i.e., $p = 1$) for a Poisson CWM with one changepoint. The purpose of this study is to illustrate that the CWM with changepoints can be applied to a response which follows a Poisson distribution under a basic changepoint setting.

First, we generate our covariate vector with respect to each group by

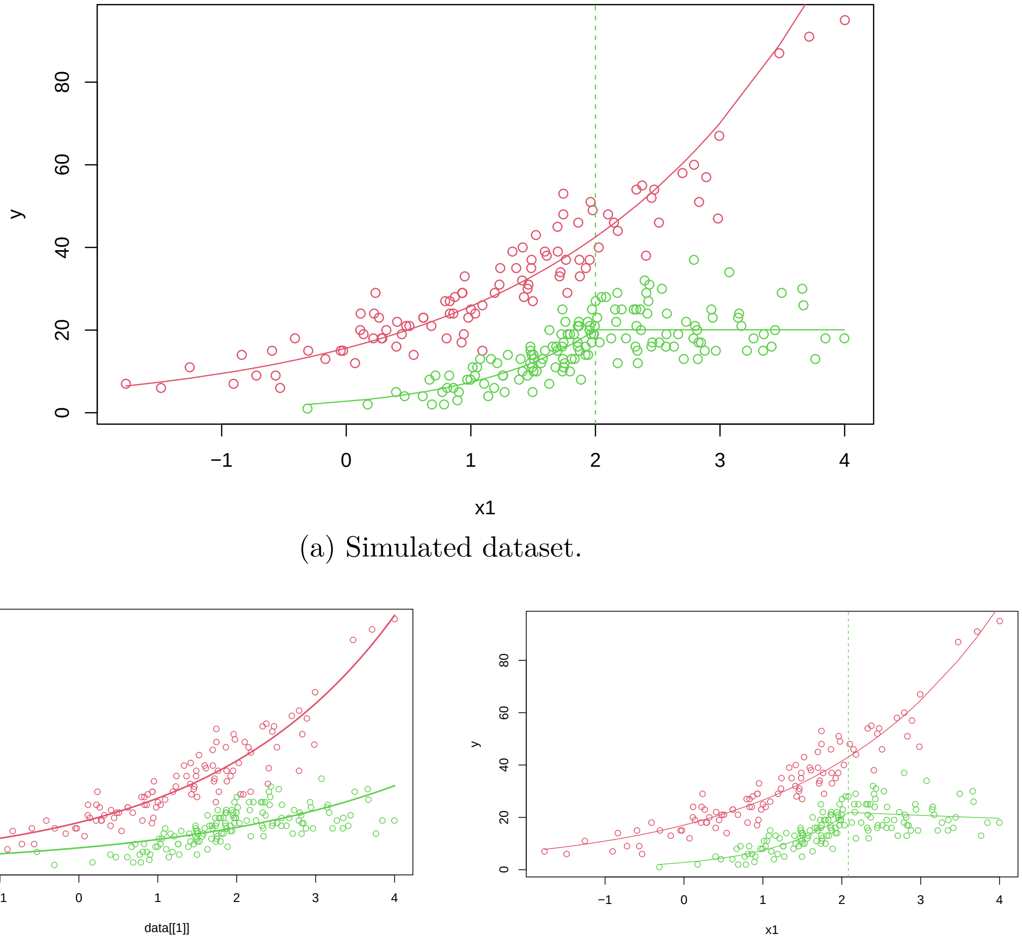
$$\boldsymbol{X}_1 \sim \mathcal{N}(1, 1.25^2),$$
$$\boldsymbol{X}_2 \sim \mathcal{N}(2, 0.8^2),$$

where $N_1 = 100$ and $N_2 = 175$. Next, define regression coefficients $\boldsymbol{\beta}_1 = (2.75, 0.5)'$, $\boldsymbol{\beta}_2 = (1, 1, -2)'$, and one changepoint $\psi_2 = 2$. Finally, we generate our response vector with respect to each group by

$$\boldsymbol{Y}_1 \sim \text{Poisson}(\exp(\mathbf{X}\boldsymbol{\beta}_1)),$$
$$\boldsymbol{Y}_2 \sim \text{Poisson}\Big(\exp(\mathbf{X}(\psi_2)\boldsymbol{\beta}_2)\Big).$$

Therefore, for this simulation study we have $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')'$ and $\boldsymbol{Y} = (\boldsymbol{Y}_1', \boldsymbol{Y}_2')'$, which implies that our sample size is $N = 275$. We can visualize this dataset in Figure 4.3, which also shows the estimation of the CWM and the sCWM under this setting.

Similar to the linear Gaussian study, we generated 100 datasets under this framework, where both the CWM and sCWM were estimated with $R = 10$ random soft initializations. The results for the ARIs and BICs can be visualized in Figure 4.4, where we note that the sCWM yielded higher BIC values for all the simulations, and the sCWM's ARIs were higher on average. Furthermore, Table 4.3 demonstrates that the parameters' true values and mean estimates for this study are fairly close with a low standard deviation, which implies that the sCWM is recovering the parameters well in this setting.

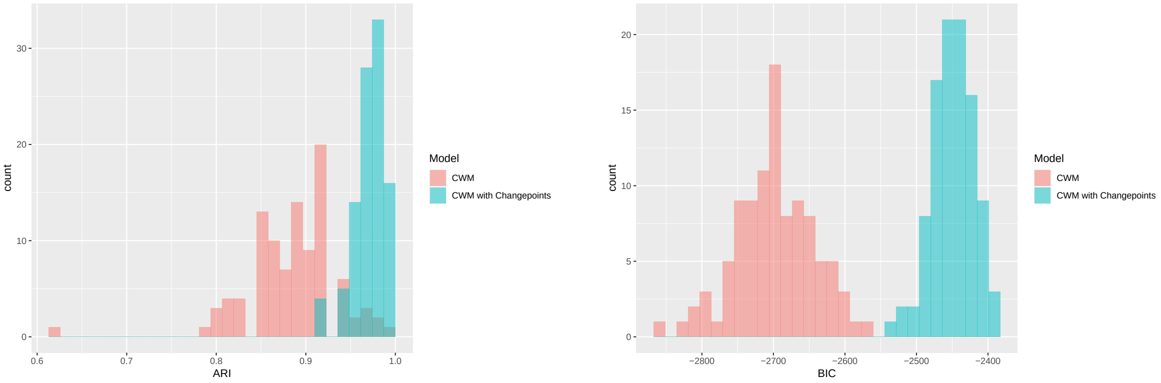


(a) Simulated dataset.

(b) Regular CWM.

(c) CWM with one changepoint.

Figure 4.3: Poisson CWM with one changepoint simulation.



(a) Histogram of ARIs.

(b) Histogram of BICs.

Figure 4.4: Histograms of ARIs and BICs for the Poisson simulation.

Table 4.3: True parameter values along with mean and standard deviations of the parameter estimates for the Poisson simulation.

Parameter	True values	Mean estimates	Standard deviations
$\boldsymbol{\beta}_1$	$(2.75, 0.5)'$	$(2.754, 0.498)'$	$(0.034, 0.016)'$
$\boldsymbol{\beta}_2$	$(1, 1, -2)'$	$(0.982, 1.012, -2.018)'$	$(0.168, 0.107, 0.137)'$
μ_1	1	0.997	0.120
μ_2	2	1.997	0.069
σ_1	1.25	1.233	0.092
σ_2	0.8	0.797	0.044
π_1	0.4	0.401	0.005
π_2	0.6	0.599	0.005
ψ_2	2	1.995	0.050

4.3 Binomial CWM

For our last simulation study, we consider $G = 2$ groups and one covariate (i.e., $p = 1$) for a Binomial CWM with one changepoint. The purpose of this study is to illustrate that the CWM with changepoints can be applied to a response which follows a Binomial distribution under a basic changepoint setting.

First, we generate our covariate vector with respect to each group by

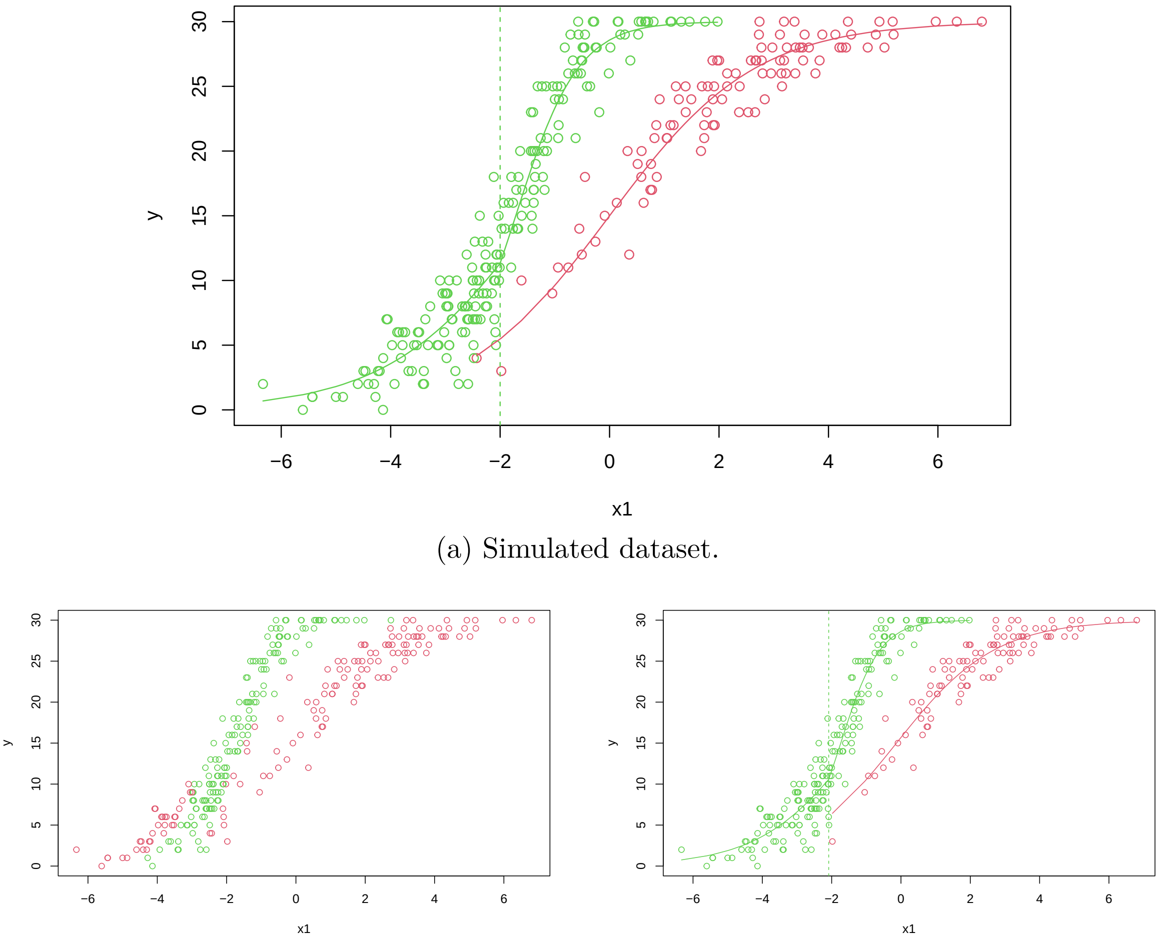
$$\boldsymbol{X}_1 \sim \mathcal{N}(2, 2^2),$$
$$\boldsymbol{X}_2 \sim \mathcal{N}(-2, 1.5^2),$$

where $N_1 = 100$ and $N_2 = 200$. Next, define regression coefficients $\boldsymbol{\beta}_1 = (0, 0.75)'$, $\boldsymbol{\beta}_2 = (1, 0.75, 1)'$, and one changepoint $\psi_2 = -2$. Finally, we generate our response vector with respect to each group by

$$\boldsymbol{Y}_1 \sim \text{Binomial}\big(\text{expit}(\mathbf{X}\boldsymbol{\beta}_1)\big),$$
$$\boldsymbol{Y}_2 \sim \text{Binomial}\Big(\text{expit}\big(\mathbf{X}(\psi_2)\boldsymbol{\beta}_2\big)\Big).$$

Therefore, for this simulation study we have $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')'$ and $\boldsymbol{Y} = (\boldsymbol{Y}_1', \boldsymbol{Y}_2')'$, which implies that our sample size is $N = 300$. We can visualize this dataset in Figure 4.5, which also shows the estimation of the CWM and the sCWM under this setting.

Furthermore, we generated 100 datasets under this framework, where both the CWM and sCWM were estimated with $R = 10$ random soft initializations. The results for the ARIs and BICs can be visualized in Figure 4.6, where we note that the sCWM yielded higher BIC and ARI values on average compared to the CWM. The parameter's true values and mean estimates are in Table 4.4, which shows that the model is recovering the parameters adequately.

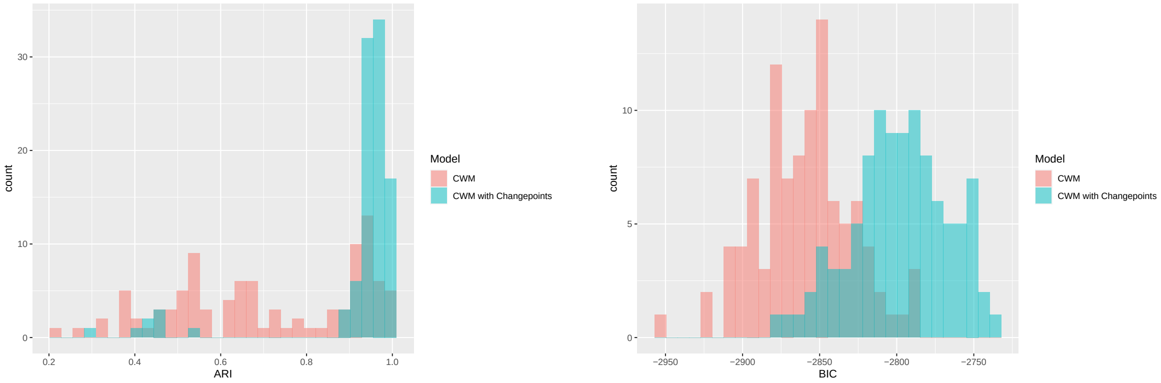


(a) Simulated dataset.

(b) Regular CWM.

(c) CWM with one changepoint.

Figure 4.5: Binomial CWM with one changepoint simulation.



(a) Histogram of ARIs.

(b) Histogram of BICs.

Figure 4.6: Histograms of ARIs and BICs for the Binomial simulation.

Table 4.4: True parameter values along with mean and standard deviations of the parameter estimates for the Binomial simulation.

Parameter	True values	Mean estimates	Standard deviations
$\boldsymbol{\beta}_1$	$(0, 0.75)'$	$(0.215, 0.811)'$	$(0.752, 0.199)'$
$\boldsymbol{\beta}_2$	$(1, 0.75, 1)'$	$(0.933, 0.730, 0.941)'$	$(0.342, 0.110, 0.210)'$
μ_1	2	1.704	0.964
μ_2	-2	-1.835	0.580
σ_1	2	1.926	0.256
σ_2	1.5	1.621	0.437
π_1	0.333	0.345	0.039
π_2	0.667	0.655	0.039
ψ_2	-2	-1.838	0.579

Chapter 5

Applications

5.1 Nitrogen Oxide

The nitrogen oxide data from Brinkman (1981) has $N = 88$ observations for the concentration of nitrogen oxide and the equivalence ratio (a measure of the air-ethanol mixture for burning ethanol in a single-cylinder car engine). This dataset has been previously analyzed in numerous finite mixture of regressions literature (e.g., Henry *et al.*, 2010; Berrettini *et al.*, 2022; Xiang *et al.*, 2019; Young, 2014). From the previous studies listed, and from the obvious two-component structure visualized in Figure 5.7, we focus on CWMs with two groups. The response variable is equivalence ratio, and the concomitant covariate is nitrogen oxide. From Figure 5.7, we see that the upper component has a slight change from the nitrogen oxide amounts about 1–2. Hence, we want to determine if a Gaussian CWM with changepoints model would fit the data better than a regular Gaussian CWM for this dataset.

We performed $R = 10$ random soft initializations, where we vary the changepoint structure such that a maximum of two changepoints can occur within the concomitant

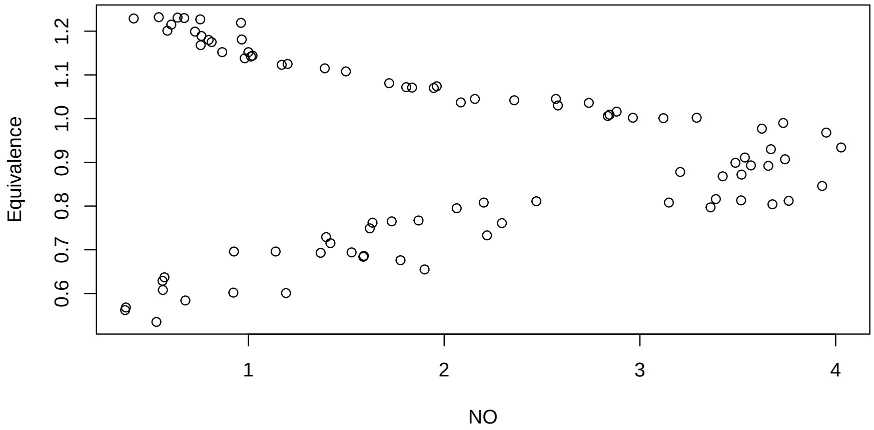


Figure 5.7: Nitrogen oxide data.

covariate. From Table 5.5, the BIC is maximized when we have a $(1, 0)$ changepoint structure, which corresponds to one changepoint in the smallest group only, and we also note that the $(0, 0)$ model (i.e., the regular CWM) has a smaller value. Therefore, it is clear that the sCWM model outperforms the regular CWM model with respect to the BIC for this dataset. The plots of the nitrogen oxide data for the best regular CWM and sCWM are given in Figure 5.8a and Figure 5.8b, respectively, and the parameter estimates for the best sCWM are given in Table 5.6. These results are fairly close to the results presented by Young (2014) under a FMR with changepoints model.

5.2 Fishing Data

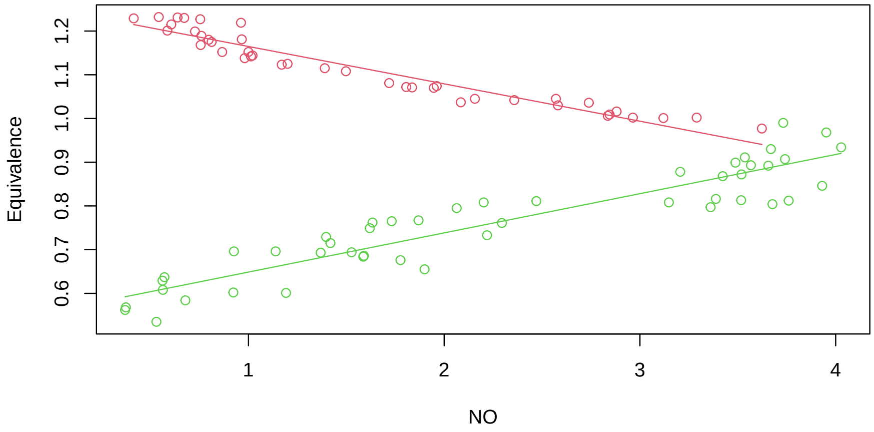
We analyze fishing data from Bailey *et al.* (2009), which was adapted from Zuur *et al.* (2013). The goal of the study from Bailey *et al.* (2009) was to determine how deep-water fish communities were impacted before and after commercial fishing began. There are $N = 147$ fishing sites, where the response variable is the total fish

Table 5.5: BIC values for nitrogen oxide data.

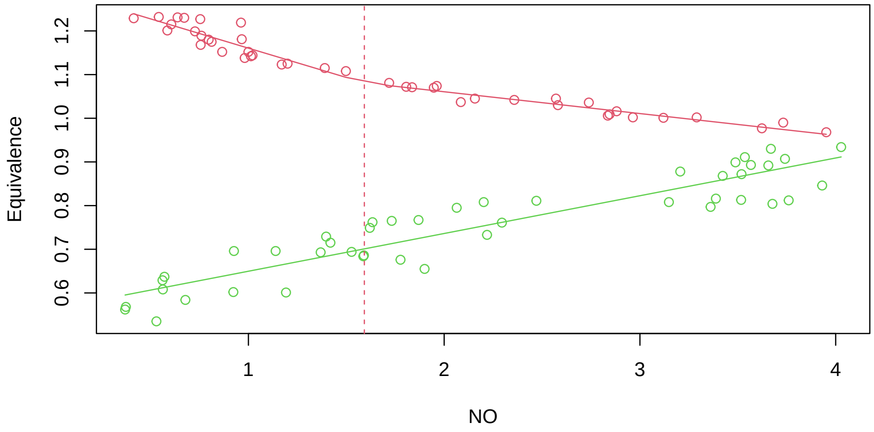
Changepoint Structure	BIC	ν
$(0, 0)$	-70.874	11
$(0, 1)$	-77.454	13
$(0, 2)$	-85.667	15
$(1, 0)$	**-51.392**	13
$(1, 1)$	-56.694	15
$(1, 2)$	-65.465	17
$(2, 0)$	-57.447	15
$(2, 1)$	-62.759	17
$(2, 2)$	-72.619	19

Table 5.6: Parameter estimates for best sCWM fitted to the nitrogen oxide data.

Parameter	Estimates
$\boldsymbol{\beta}_1$	$(1.295, -0.134, 0.084)'$
$\boldsymbol{\beta}_2$	$(0.563, 0.086)'$
σ_{Y_1}	0.016
σ_{Y_2}	0.043
μ_1	1.685
μ_2	2.214
σ_1	0.997
σ_2	1.180
π_1	0.485
π_2	0.515
ψ_1	1.592



(a) Best regular CWM for nitrogen oxide data.



(b) Best sCWM for nitrogen oxide data.

Figure 5.8: Comparison of CWM and sCWM for nitrogen oxide data.

counted per site (`totabund`) and the concomitant covariate is the trawl depth per site in metres (`meandepth`). We note that the data has been previously studied in the context of clustering in Hilbe (2014, pp. 232–235), where their goal was to determine if the data was generated from more than one mechanism. That is, Hilbe (2014) fit a negative-binomial mixture with two and three components, with response `totabund` and covariate `meandepth`. It is worth noting that when they fit a negative-binomial mixture with three components, their third component was not significant (in the context of their mixture model). If we visualize the data in Figure 5.9, we see that there is a large change with respect to the `meandepth` covariate for the amounts 1000–2000. Hence, our goal is to determine if the segmented Poisson CWM would fit the data better than an ordinary Poisson CWM.

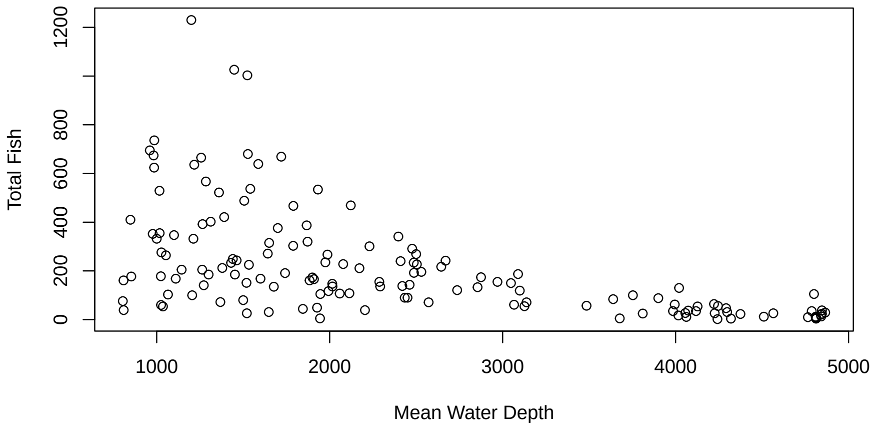


Figure 5.9: Fishing data.

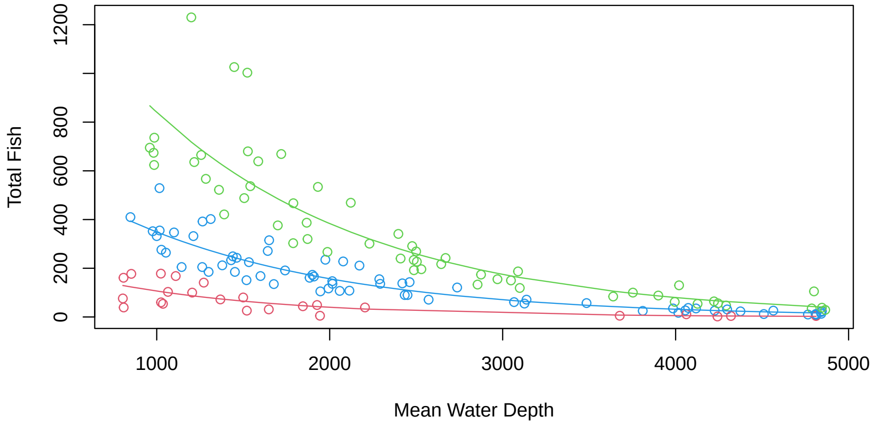
For this study, we performed $R = 10$ random soft initializations for $G = 1, 2, 3$ groups varying the changepoint structure such that a maximum of two changepoints were tested within each model. The BICs for this study are in Table 5.7 (where the best BIC is bolded for each group), and it is clear that the BICs for the changepoints model are always larger than without changepoints, so the sCWM is performing better than the regular CWM with respect to the BIC. We have abbreviated the changepoint structure as (c_{11}, c_{12}, c_{13}), with c_{11} and c_{13} being the number of changepoints in the smallest and largest components, respectively. From Table 5.7, it is worth noting that the model with two changepoints in one group was not estimated, hence the values are NA. Furthermore, the best model for $G = 2$ groups is the $(c_{11}, c_{12}) = (2, 0)$ model, which corresponds to two changepoints in the smallest component and zero changepoints in the largest component. Lastly, the best model for $G = 3$ groups is the $(c_{11}, c_{12}, c_{13}) = (1, 0, 1)$, which corresponds to one changepoint within both the smallest and largest components.

The plots of the fishing data for the best CWM and sCWM are given in Figure 5.10a and Figure 5.10b, respectively, and the parameter estimates for the best sCWM are given in Table 5.8. Although the coefficients appear to be small, we note that we are working on the exponential scale. Not shown here, we note that the coefficients for the best regular CWM are also small with respect to the covariate in each group. It is important to see that the mixing proportions appear to be highly similar, which may cause issues with identifiability.

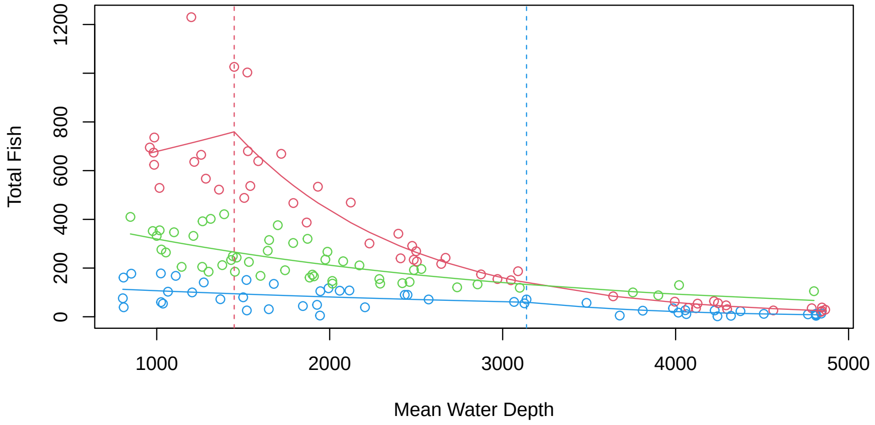
It is worth noting that the BIC values were increasing as we increased the number of groups for a regular CWM; however, it is computationally expensive to test all possible changepoint structures within an sCWM, and our goal was to determine whether an sCWM is warranted for this particular study. Therefore, we conclude that a Poisson CWM with changepoints fits the data better than a regular Poisson CWM for this study under a few groups.

Table 5.7: BIC values for fishing data.

G	Changepoint Structure	BIC	ν
1	(0)	$-19\,284.100$	4
	(1)	$\mathbf{-18\,008.250}$	6
	(2)	NA	NA
2	$(0,0)$	-8874.100	9
	$(1,0)$	-8386.777	11
	$(0,1)$	-8639.808	11
	$(1,1)$	-8477.997	13
	$(2,0)$	$\mathbf{-8295.186}$	13
	$(0,2)$	-8556.663	13
3	$(0,0,0)$	-6724.700	14
	$(1,0,0)$	-6442.377	16
	$(0,1,0)$	-6606.770	16
	$(0,0,1)$	-6498.248	16
	$(1,1,0)$	-6383.981	18
	$(0,1,1)$	-6425.173	18
	$(1,0,1)$	$\mathbf{-6357.125}$	18
	$(2,0,0)$	-6412.389	18
	$(0,2,0)$	-6546.916	18
	$(0,0,2)$	-6535.993	18



(a) Best regular CWM for fishing data for three groups.



(b) Best sCWM for fishing data for three groups.

Figure 5.10: Comparison of CWM and sCWM for fishing data.

Table 5.8: Parameter estimates for best sCWM fitted to the fishing data.

Parameter	Estimates
$\boldsymbol{\beta}_1$	$(6.270, 0.000\,250, -0.001\,25)'$
$\boldsymbol{\beta}_2$	$(6.178, -0.000\,411)'$
$\boldsymbol{\beta}_3$	$(4.945, -0.000\,273, -0.000\,911)$
μ_1	2660.126
μ_2	1923.409
μ_3	2656.064
σ_1	1304.176
σ_2	843.863
σ_3	1384.549
π_1	0.332 57
π_2	0.333 48
π_3	0.333 95
ψ_1	1447.999
ψ_3	3138.000

Chapter 6

Conclusions and Future Work

We demonstrated that Gaussian, Poisson, and Binomial CWMs with changepoints can be seen as an improvement to regular CWMs when we believe that changepoints are present within the data. From our simulation studies with changepoints, the sCWM usually outperformed the regular CWM in terms of the ARI and BIC values. For real-world datasets, we used a Gaussian and Poisson changepoint CWM for the nitrogen oxide and fishing data, respectively. Within both datasets, the best Gaussian and Poisson CWM with changepoint model outperformed the best regular CWM with respect to the BIC.

Furthermore, we note that there are major weaknesses of the CWM with changepoints model. First, due to our formulation of the model, we needed to test multiple permutations of the changepoint structure. Also, fitting the changepoint model takes a much longer time compared to a regular model. Due to these two major issues, the model will not scale well in higher dimensions with many covariates and groups, which is the reason we did not present results for such data. However, the methodology/estimation is presented for the multivariate case.

To address these two issues, we propose two solutions, but these solutions may cause more issues. For the first issue, we could ignore the identifiability "label-switching" problem and run the algorithm as usual; this would reduce the number of permutations to be estimated. However, the changepoints may not be estimated within the right component. For the second issue, Muggeo (2020) implemented a way to select the number of breakpoints within a segmented regression framework. It may be possible to replace the first CM-step in our algorithm with their algorithm for selecting the number of breakpoints, and then we would not need to test any permutations for the changepoint structure. If it is possible to implement their algorithm, then the segmented CWM may also scale well in higher dimensions. However, by selecting the number of breakpoints at each iteration, the log-likelihood may not be non-decreasing as the number of parameters may decrease at each iteration. We also note that all code was written in R. For model estimation, a faster language such as C or C++ can be used, but this would also imply rewriting the `segmented` package as used in the first CM-step. Future work can also attempt to assume different distributional assumptions on both the response and covariates. For example, a Student's t-distribution and a contaminated CWM can be explored for a segmented CWM to give a more robust inference for data with many outliers.

Appendix A

Additional Tables and Figures

Table A.9: Covariance structure and number of free parameters in $\mathbf{\Sigma}_g$.

Model	Volume	Shape	Orientation	$\mathbf{\Sigma}_g$	Number of Free Covariance Parameters
EII	Equal	Spherical	—	$\lambda\mathbf{I}$	1
VII	Variable	Spherical	—	$\lambda_g\mathbf{I}$	G
EEI	Equal	Equal	Axis-Aligned	$\lambda\mathbf{\Delta}$	p
VEI	Variable	Equal	Axis-Aligned	$\lambda_g\mathbf{\Delta}$	$G+p-1$
EVI	Equal	Variable	Axis-Aligned	$\lambda\mathbf{\Delta}_g$	$1+G(p-1)$
VVI	Variable	Variable	Axis-Aligned	$\lambda_g\mathbf{\Delta}_g$	Gp
EEE	Equal	Equal	Equal	$\lambda\mathbf{\Gamma}\mathbf{\Delta}\mathbf{\Gamma}'$	$p(p+1)/2$
VEE	Variable	Equal	Equal	$\lambda_g\mathbf{\Gamma}\mathbf{\Delta}\mathbf{\Gamma}'$	$G+p-1+p(p-1)/2$
EVE	Equal	Variable	Equal	$\lambda\mathbf{\Gamma}\mathbf{\Delta}_g\mathbf{\Gamma}'$	$1+G(p-1)+p(p-1)/2$
EEV	Equal	Equal	Variable	$\lambda\mathbf{\Gamma}_g\mathbf{\Delta}\mathbf{\Gamma}_g'$	$p+Gp(p-1)/2$
VVE	Variable	Variable	Equal	$\lambda_g\mathbf{\Gamma}\mathbf{\Delta}_g\mathbf{\Gamma}'$	$Gp+p(p-1)/2$
VEV	Variable	Equal	Variable	$\lambda_g\mathbf{\Gamma}_g\mathbf{\Delta}\mathbf{\Gamma}_g'$	$G+p-1+Gp(p-1)/2$
EVV	Equal	Variable	Variable	$\lambda\mathbf{\Gamma}_g\mathbf{\Delta}_g\mathbf{\Gamma}_g'$	$1+G(p-1)+Gp(p-1)/2$
VVV	Variable	Variable	Variable	$\lambda_g\mathbf{\Gamma}_g\mathbf{\Delta}_g\mathbf{\Gamma}_g'$	$Gp(p+1)/2$

Bibliography

Aitken, A. C. (1926). On Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, **46**, 289–305.

Aitkin, M. and Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **47**(1), 67–75.

Bailey, D. M., Collins, M., Gordon, J. D., Zuur, A. F., and Priede, I. G. (2009). Long-term changes in deep-water fish populations in the northeast atlantic: a deeper reaching effect of fisheries? *Proceedings of the Royal Society B: Biological Sciences*, **276**(1664), 1965–1969.

Baudry, J.-P. and Celeux, G. (2015). EM for mixtures: Initialization requires special care. *Statistics and Computing*, **25**(4), 713–726.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, **32**(6), 1–29.

Berrettini, M., Galimberti, G., and Ranciati, S. (2022). Semiparametric finite mixture of regression models with Bayesian P-splines. *Advances in Data Analysis and Classification*, pages 1–31.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, **46**(2), 373–388.

Brinkman, N. D. (1981). Ethanol fuel—single—cylinder engine study of efficiency and exhaust emissions. *SAE Transactions*, **90**(2), 1410–1424.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**(5), 781–793.

Dang, U. J., Punzo, A., McNicholas, P. D., Ingrassia, S., and Browne, R. P. (2017). Multivariate response and parsimony for Gaussian cluster-weighted models. *Journal of Classification*, **34**, 4–34.

Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, **83**(401), 173–178.

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

DeSarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, **5**, 249–282.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter*, **4**(1), 65–75.

Gershenfeld, N. (1997). Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, **808**(1), 18–24.

Henry, M., Kitamura, Y., and Salanié, B. (2010). Identifying finite mixtures in econometric models. Technical report, Yale University.

Hilbe, J. M. (2014). *Modeling Count Data*. Cambridge University Press, Cambridge, England.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of Classification*, **29**, 363–401.

Ingrassia, S., Minotti, S. C., and Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics & Data Analysis*, **71**, 159–182.

Ingrassia, S., Punzo, A., Vittadini, G., and Minotti, S. C. (2015). The generalized linear mixed cluster-weighted model. *Journal of Classification*, **32**, 85–113.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, **62**(1), 49–66.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, **20**(3), 1350–1360.

Lindsay, B. G. (1995). Mixture models: theory, geometry, and applications. *Institute of Mathematical Statistics*, **5**.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, **83**(404), 1014–1022.

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

Mazza, A., Punzo, A., and Ingrassia, S. (2018). flexCWM: a flexible framework for cluster-weighted models. *Journal of Statistical Software*, **86**(2), 1–30.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley, New York.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, **6**, 355–378.

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

McNicholas, P. D., Murphy, T. B., McDaid, A. F., and Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis*, **54**(3), 711–723.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**(2), 267–278.

Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, **22**, 3055–3071.

Muggeo, V. M. (2008). segmented: an R package to fit regression models with broken-line relationships. *R News*, **8**(1), 20–25.

Muggeo, V. M. (2020). Selecting number of breakpoints in segmented regression: implementation in the R package segmented. Technical report, University of Palermo.

Pocuca, N., Browne, R. P., and McNicholas, P. D. (2022). *mixture: Mixture Models for Clustering and Classification*. R package version 2.0.5.

Punzo, A. (2014). Flexible mixture modelling with the polynomial gaussian cluster-weighted model. *Statistical Modelling*, **14**(3), 257–291.

Punzo, A. and McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, **58**(6), 1506–1537.

Punzo, A. and McNicholas, P. D. (2017). Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. *Journal of Classification*, **34**, 249–293.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846–850.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, **8**(1), 289–317.

Steinley, D. (2004). Properties of the Hubert-Arable adjusted Rand index. *Psychological Methods*, **9**(3), 386–396.

Tiedeman, D. V. (1955). On the study of types. *Symposium on Pattern Analysis*.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. Applied section. Wiley, Chichester, England.

Wedel, M. (2002). Concomitant variables in finite mixture models. *Statistica Neerlandica*, **56**(3), 362–375.

Wedel, M. and DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification*, **12**, 21–55.

Wolfe, J. H. (1963). *Object Cluster Analysis of Social Areas*. Master's thesis, University of California, Berkeley.

Wolfe, J. H. (1965). A computer program for the maximum likelihood analysis of types. *Technical Bulletin*, **65**, 15.

Wood, S. N. (2001). Minimizing model fitting objectives that contain spurious local minima by bootstrap restarting. *Biometrics*, **57**(1), 240–244.

Xiang, S., Yao, W., and Yang, G. (2019). An overview of semiparametric extensions of finite mixture models. *Statistical Science*, **34**(3), pp. 391–404.

Young, D. S. (2014). Mixtures of regressions with changepoints. *Statistics and Computing*, **24**(2), 265–281.

Zuur, A., Hilbe, J., and Ieno, E. (2013). *A Beginner's Guide to GLM and GLMM with R: A Frequentist and Bayesian Perspective for Ecologists.* Highland Statistics Ltd. book series. Highland Statistics Limited, Newburgh, United Kingdom.