Master Thesis in Statistics and Machine Learning

Defining and predicting fast-selling clothing options

Sara Jesperson

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

Supervisor
Oleg Sysoev

Examiner
Linda Wänström

Abstract

This thesis aims to find a definition of fast-selling clothing options and to find a way to predict them using only a few weeks of sale data as input. The data used for this project contain daily sales and intake quantity for seasonal options, with sale start 2016-2018, provided by the department store chain Åhléns.

A definition is found to describe fast-selling clothing options as those having sold a certain percentage of their intake after a fixed number of days. An alternative definition based on cluster affiliation is proven less effective.

Two predictive models are tested, the first one being a probabilistic classifier and the second one being a k-nearest neighbor classifier, using the Euclidean distance. The probabilistic model is divided into three steps: transformation, clustering, and classification. The time series are transformed with B-splines to reduce dimensionality, where each time series is represented by a vector with its length and B-spline coefficients. As a tool to improve the quality of the predictions, the B-spline vectors are clustered with a Gaussian mixture model where every cluster is assigned one of the two labels *fast-selling* or *ordinary*, thus dividing the clusters into disjoint sets: one containing fast-selling clusters and the other containing ordinary clusters. Lastly, the time series to be predicted are assumed to be Laplace distributed around a B-spline and using the probability distributions provided by the clustering, the posterior probability for each class is used to classify the new observations.

In the transformation step, the number of knots for the B-splines are evaluated with cross-validation and the Gaussian mixture models, from the clustering step, are evaluated with the Bayesian information criterion, BIC. The predictive performance of both classifiers is evaluated with accuracy, precision, and recall. The probabilistic model outperforms the k-nearest neighbor model with considerably higher values of accuracy, precision, and recall. The performance of each model is improved by using more data to make the predictions, most prominently with the probabilistic model.

Acknowledgments

First, I would like to thank Sara Wahlman at Åhléns for this opportunity and the enthusiasm you have shown from the very start and throughout the project. I would also like to thank Erik Lundberg for answering all my questions and giving me helpful suggestions along the way.

I would like to extend my gratitude to my supervisor Oleg Sysoev for guiding me through this process and for giving me support and advice.

Lastly, I would also like to thank my opponent Alessia De Biase for a thorough read through and valuable and insightful comments.

Contents

List of Figures

List of Tables

Glossary

Hierarchy of apparel:

1. **Division** - e.g. Women's fashion
2. **Group** – e.g. Women's clothes
3. **Department** – e.g. Women's clothes private label
4. **Class** - Broadly grouping the garments according to use e.g. shirts
5. **Subclass** – A more specific grouping of the garments e.g. short-sleeved shirts
6. **Article** - A specific model of a garment
7. **Option** - Different options for the article concerning e.g. color
8. **SKU** - Stock Keeping Unit, the different sizes available for the option

Phase: A more specific time period than season, where each season consists of several phases, and each phase introducing a new range of clothing

Season: In fashion, each year is divided into two time periods called seasons: Spring/Summer and Autumn/Winter

1 Introduction

This chapter gives an overview of the previous research in the fields of apparel retail prediction and time series clustering while also presenting the objective of the thesis.

1.1 Background

Predicting sales in fashion retail offers a number of complications. As Thomassey (2014) explains, customer demand for fashion apparel is influenced by several factors that are out of the retailer's hands. Weather and the economic cycle are two of those factors affecting the consumers' buying behavior. Apparel retail is also impacted by the current fashion trends which are ever-changing, leading to new items every season and the number of products is vast since items are often offered in different colors, patterns, and materials. The lifespan of a fashion item is another obstacle since the time from designing and placing an order to sale start is long compared to the life of a product in-store, which is usually only a season. Apparel is sensitive to the season and items are selling at different rates throughout the year. Forecasting is often of interest in apparel retail since it predicts future sales based on previous sales, but because of the lack of historical sales data on the item level and the unusual lifespan of clothing items, traditional forecasting methods such as ARMA or exponential smoothing have proven to be less effective.

Over- or underestimating sales leads to a monetary loss in unsold items and lost sales caused by running out of stock too early. A stockout, where the retailer has run out of items, may also evoke negative emotions of the customer according to Kim and Lennon (2011). These emotions are stronger for apparel compared to grocery stockout. In the case of online apparel stores, the out-of-stock situation is different compared to in-store stockout since items are often shown as available even though they are sold out. When consumers experience stockouts, the perception of the store image is negatively affected and future interactions with the store might be reduced.

Most previous research in the field of fashion retail prediction has focused on forecasting the exact sales. Several studies have dealt with the lack of historical data on SKU level by basing their forecasts on class level. Using four years of sales data from an American apparel company, Frank et al. (2003) examined the results of forecasting sales for classes of apparel using both exponential smoothing and Artificial Neural Network, ANN. The results showed better predictions using the ANN but also suggest that the model might be overfitted. Sun et al. (2008) suggest using

extreme learning machines, ELM, to forecast sales for families of items, instead of ANNs. Many ANNs tune the parameters with algorithms where the gradient is used in the search for the optimal values, such as backpropagation. This is time-consuming and tends to lead to overfitted models. ELM is a neural network with a single hidden layer where the input weights and hidden biases are not tuned but instead generated randomly and thus computation time is reduced and over-tuning is prevented. Using data from a Hong Kong fashion retailer, color, size, and price were used as input to the model and the results showed improved forecasts compared to models based on backpropagation.

Some research has been done in forecasting on SKU level. Thomassey (2014) suggests a two-step approach to forecast sales in apparel on the stock keeping unit level. Historical data is used to cluster items with similar sale patterns together within a family of items. Based on the created clusters, a classification model is trained to assign cluster labels to new items based on descriptive variables. The forecast for the new item is the sales pattern of its assigned cluster adapted to its lifespan. This have been tested with k-means clustering and a decision tree for the classification (Thomassey and Fiordaliso, 2006) and a two-step clustering using self-organizing map (SOM), a type of neural network, to produce a 2D map of the input which is then clustered with k-means and classification performed by a probabilistic neural network (Thomassey and Happiette, 2007). Kumar and Patel (2010) propose another method to forecast on the item level. Using forecasts from historical data, items are clustered, and each cluster is represented by a combination of its items' forecasts weighted by their inverse variance. New items are assigned to a cluster based on a similarity measure and the forecast of the cluster is adopted by the new item. Hierarchical clustering with a 4-week moving average forecasting model produced the lowest forecasting error. A study by Goldfisher and Chan (1994) showed that by the third week of sale, successful products could be distinguished from failing products by looking at a weekly sales index. Successful products showed a higher sales index during the third week compared to the second week and failing products showed a lower sales index during the third week compared to the second week. This conclusion was reached by examining prelabeled data.

Clustering is a method for dividing data into separate clusters where the observations within a cluster are similar to each other and observations from different clusters are different from each other. The prediction of a cluster can be used as a tool in securing more robust results when the predictions of the individual observations are uncertain (Kumar and Patel, 2010). By clustering historical data, predictions of new products can be based on similar products from the past (Thomassey, 2014). Much research has been made in the field of clustering time series as stated by Aghabozorgi et al. (2015). Time series are naturally not in a format to successfully cluster with traditional algorithms and distance measures. To handle this, either the data or the algorithms must be modified, often specifically to the project at hand. The main applications within time series clustering have been to find distance measures that effectively compares the raw data representation of the time series or to find

new ways to represent the time series to reduce dimensionality. These applications are then often used with traditional clustering algorithms such as k-means and hierarchical clustering. In a study concerning time series clustering, Abraham et al. (2003) used B-splines to represent time series describing the pH evolution in different cheeses. Each time series was represented by its spline coefficients and using k-means clustering, each cluster represented a different pH evolution pattern.

1.2 Objective

There is a perceived problem at Åhléns that their biggest sales units are selling out too quickly and are often out of stock. A solution to that problem would be to alert the supply planners of predicted fast-selling options, and for them to temporary lock the restocking quantity for the smaller sales units and prioritizing restocking of fast-selling options for the bigger sales units in order to reach as many customers as possible. Thus, the objective of the thesis is to answer the following questions:

- How can a fast-selling option be defined using information from sales data?
- How can options be predicted as fast selling or not, based on a few weeks of data?
- How does the quality of the predictions change when more data is available?

This thesis will focus on fast-selling options among the seasonal options from the women's clothes department and only considers options from Åhléns' private labels. Data is also limited to only contain options with sales start from 2016 through 2018.

1.3 Outline

In chapter 2, the data source and raw data are presented and described as well as the preprocessing steps and the resulting data set. The methods used for the analysis, both for training models and evaluating them, are described in chapter 3 and the results from the analysis are presented in chapter 4. In chapter 5 the data, methods, and results are discussed and chapter 6 presents the conclusions of the thesis.

Figure 1.1 presents the process of the thesis in defining and predicting fast-selling clothing options. The flowchart broadly presents in which steps the data has been handled and how the definition of fast-selling options is reached. Two predictive classifiers are also shown: a probabilistic model, with three steps, and k-nearest neighbor, with only one step. The process starts with collecting and preprocessing data. From the resulting data set, a histogram based definition of fast-selling options is found and presented in section 4.1. The unlabeled data is passed to the probabilistic model where the first two steps are performed. This provides an alternative definition of fast-selling options based on the clustering, which can be

found in subsection 4.3.2. The preferred definition is chosen and the data is labeled according to it. The labeled data is then used for the last step of the probabilistic model as well as for the k-nearest neighbor classifier.

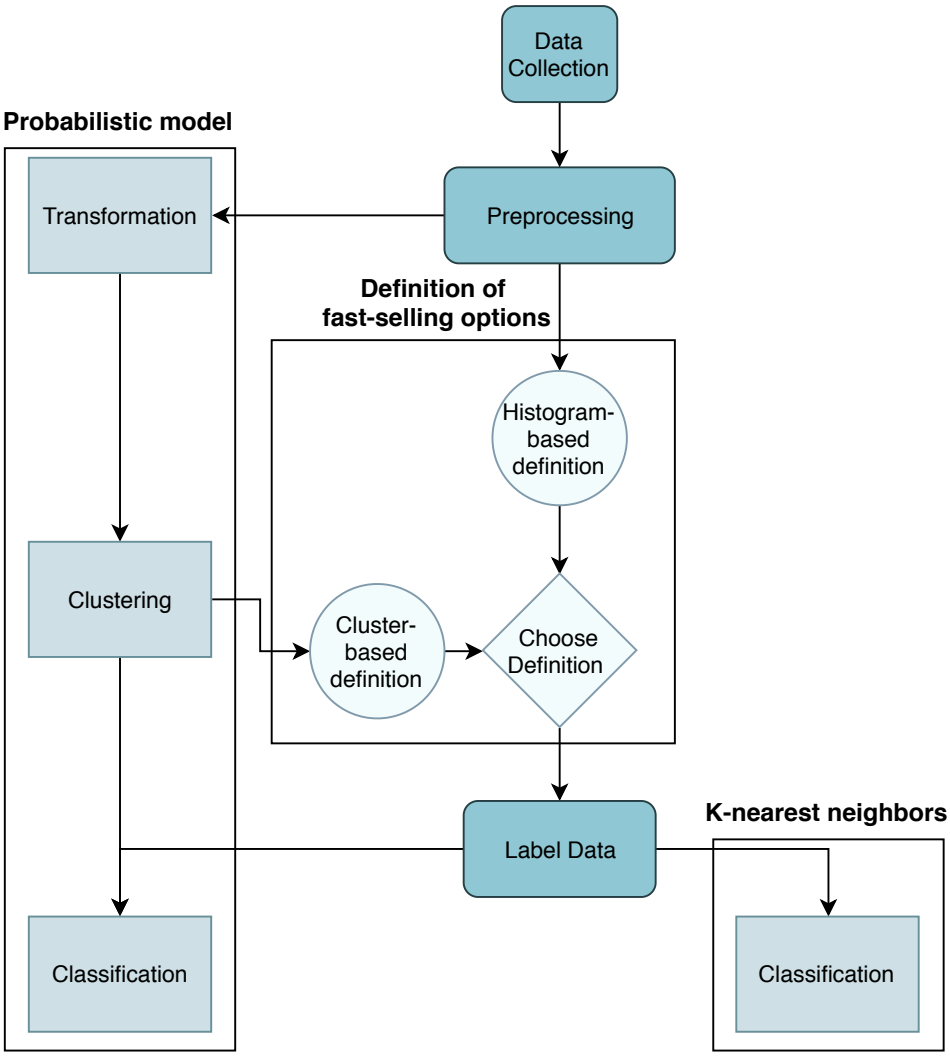

Figure 1.1: Flowchart describing the process of the analysis executed in the thesis

2 Data

This chapter introduces the data available from the commissioner, presents the different steps of the preprocessing, and describes the final data used for the analysis.

2.1 Data Sources

The data is collected from the commissioner Åhléns' database. Åhléns is a department store chain with locations all over Sweden as well as online. Among their products are seasonal options that are only for sale during a limited time period. For these options, a one-time order is placed since the production time is long and the life span in the store is relatively short. This means that no additional orders are placed and that the ordered quantity needs to be distributed wisely to the different sales units. The options are received at the warehouse and originally distributed according to a start distribution which is already defined. A quantity is left at the warehouse for restocking and later distributed to the different stores according to actual sales. Occasionally, popular seasonal items return during more than one season and additional orders are placed for the new seasons. While the option coding remains the same, the season coding is changed for both the sales and the order. Because of this, option and season are together used to identify unique observations.

2.2 Raw Data

Two data sets are extracted from the database: one containing the daily sales and the other containing the intake quantity to warehouse. The data sets are already cleaned to only contain seasonal options from the private labels in the women's clothes department and were retrieved 2019-04-01. All sales where the item is unknown were also removed before retrieval.

Table 2.1 shows the variables of the daily sales data set and a short description of them. This data set contains the sales for each option and season in the form of time series. The sales have also been grouped based on the sale type. Returns are represented by negative values which means that the variable *Quantity* can have both positive and negative values.

Table 2.1: Description of variables in daily sales

Variable	**Description**
Option	Id identifying the different clothing options
Season	Id identifying during which season the option was active
Phase start	Start date of the phase the option was active under
Date	The date of the sale
Sales type	Indicator of the type of sale: Regular, Clearance, Promotional or Personal Promotional
Quantity	The total quantity sold in all stores for the given date, option, season and sales type

Table 2.2 shows the variables for the intake quantity to warehouse data set and a short description of them. This data set contains information about how many units of each option and season were received at the warehouse. Since all seasonal options from the private labels are passing through the warehouse, the intake quantity to warehouse is used instead of the order quantity.

Table 2.2: Description of variables in intake to warehouse

Variable	**Description**
Option	Id identifying the different clothing options
Season	Id identifying during which season the option was active
Class	The garment class the option belongs to
Quantity	The total quantity received at the warehouse for each given option and season

2.3 Secondary Data

In accordance with the limits set for this thesis, the data is cleared from any options that have a sales start before 2016 or after 2018. The options with no recorded intake quantity to warehouse are also removed, as well as options with less than 50 days of regular sales or more than 350 days of regular sales.

For each unique combination of option and season, a series of preprocessing steps are performed. All dates after the first sale of type clearance are removed as well as dates more than one week before the phase start. For defining fast-selling options, the clearance is not of interest since it marks the very end of an options life in store

and sales more than a week before phase start are rare and most likely the result of a sale that has been registered incorrectly. For the remaining dates, the sales type is disregarded and the quantity sold is calculated as the sum of all sales types. The time series contain missing values, due to only sales being reported, and these days are given a quantity of sold units of zero.

Four new variables are created from the cleaned data, starting with combining option and season to into one variable *Option*, uniquely identifying each time series. From this point on, any references to option in the thesis will be to this variable combining both season and option. *Day* is created by counting the days from the first date of sales. A variable *Relative sale* describes the proportion of sold units relative to the intake quantity to warehouse of the option and by calculating the cumulative sum of the relative sale, the variable *Cumulative relative sale* is produced. After creating the new variables, options with negative cumulative sales or cumulative sales over one are removed since these values indicate mistakes in the data. An extract from the resulting data set is seen in table 2.3, showing the first five points of the time series for an option.

Table 2.3: Extract from the final data set

Option	**Day**	**Relative sale**	**Cumulative relative sale**
88127151190509-TCX537	1	0.00093	0.00093
88127151190509-TCX537	2	0.00278	0.00371
88127151190509-TCX537	3	0.00464	0.00835
88127151190509-TCX537	4	0.00093	0.00928
88127151190509-TCX537	5	0.00093	0.01020

The data is divided into two sets, where one is used to train the models and the other is used for testing the predictive performance of the models. All options with sales start before June 2018 are part of the training set and all options with sales start from June 2018 are part of the test set. This yields a training set containing 81 percent of the options and a test set containing 19 percent of the options.

2.4 Description of the Resulting Data Set

The complete resulting data set, including both training and test, contains 2 721 article options, each represented by a time series. Figure 2.1 shows the distribution of the length of the time series with the shortest being 51 days and the longest being 301 days. The distribution of length is right-skewed with the mean 131 and the median 125. Most time series have a length shorter than 200 days and the most common length of the time series is between 120 to 140 days.

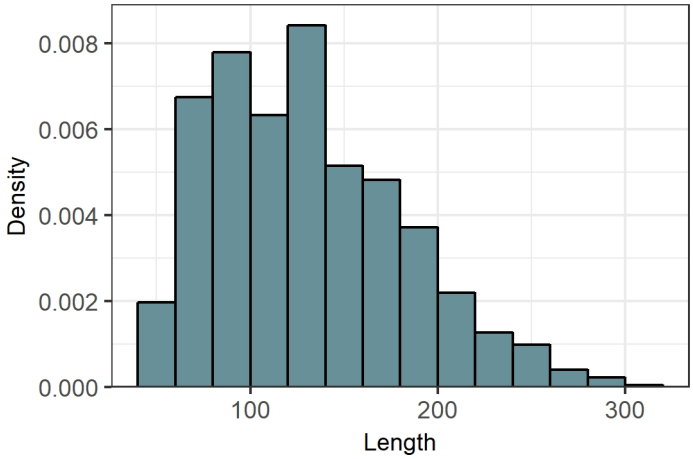

Figure 2.1: Distribution of the sales period before clearance for seasonal options 2016-2018

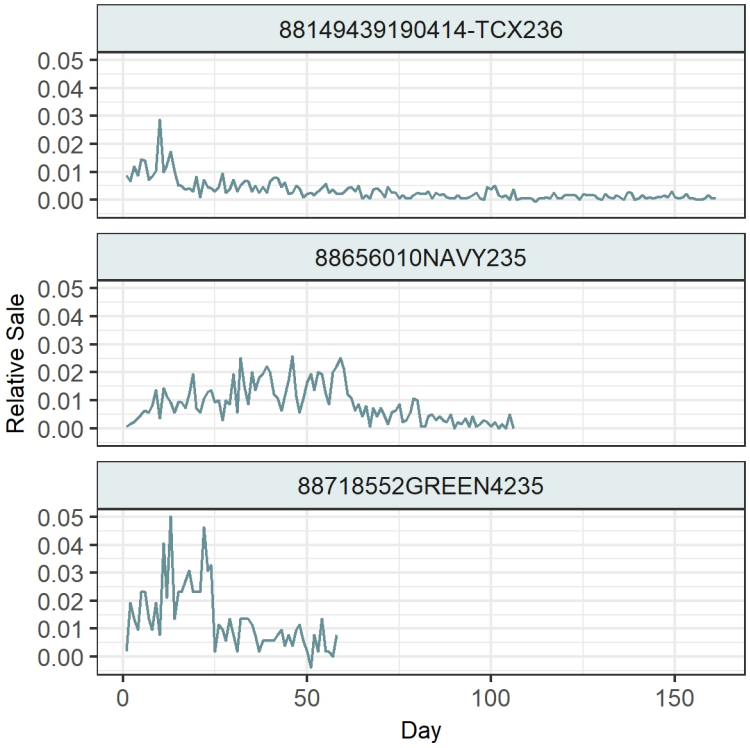

Figure 2.2: Relative sales over time for three options

Figure 2.2 presents the relative sales over time for three options. The three options show three different sales patterns and also different lengths of the time series They all exhibit a non-smooth behavior and show decreasing sales towards the end of their time periods. The top plot in figure 2.2 shows an option with slightly higher sales during the first 15 days but overall low sales during a long time. The middle

plot shows an option with the most sales around the middle of its sales period. The plot at the bottom shows the option, out of the three, with the highest peaks in sales and higher sales during the first half of its time period. The sales patterns are more easily distinguished in figure 2.3 where the cumulative sales over time for these options are presented.

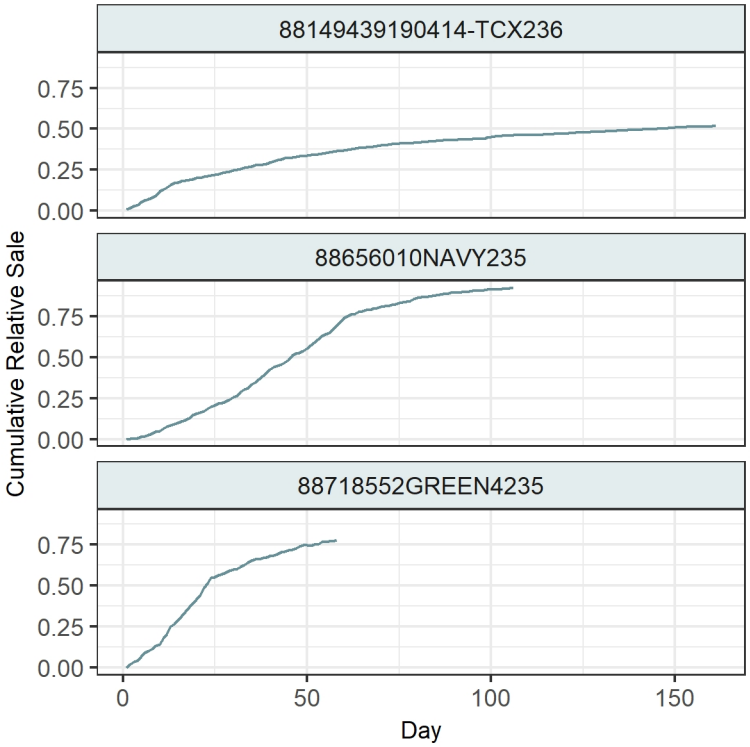

Figure 2.3: Cumulative relative sales over time for three options

Figure 2.3 shows that the option in the top plot has a slow growth throughout its lifetime and only sell just above 50 percent of its intake before clearance. The middle plot shows that the option starts out with a slower sale, only to be increased before leveling out, selling roughly 80 percent of its intake before clearance. Compared to the middle plot, the option on the bottom of figure 2.3 shows a faster pattern in the first 25 days, selling more than 50 percent of its intake in that period. The sales slow down after that resulting in the option selling just over 75 percent of its intake before clearance.

3 Methods

This chapter is aimed to present the methods used in the thesis and explain how they are applied in the analysis. Two different methods of classification are presented in this chapter. First, a probabilistic method is explained followed by the k-nearest neighbors method and lastly the evaluation methods used are presented.

3.1 Probabilistic Classification

The probabilistic method proposed in this thesis consists of three steps: transformation, clustering, and classification. The time series are transformed by B-splines to reduce dimensionality, and the B-spline coefficients and length of a time series form a new representation of that time series. The time series, in their new form, are used in a Gaussian mixture model to cluster the observations. This step is performed to increase the stability of the results of the classification, where each cluster is assigned a class label and the posterior probabilities of the different classes, given the non-transformed time series of a new observation, are used for prediction. These posterior probabilities are found by utilizing the probability distributions from the mixture model which in turn uses the simplified representation achieved with B-splines. This method is a modified and extended version of the time series clustering suggested by Abraham et al. (2003).

3.1.1 B-Splines

To transform time series, and thus simplify the input space, B-splines or basis splines are used. A time series can be represented by the coefficients of the B-splines and the position of the knots. Hastie et al. (2009) explain that a regression spline is a function $f(x)$ defined by piecewise polynomials, used to describe complex functions in a simpler way. x is one dimensional and divided by a set of knots $\xi_1, ..., \xi_R$ into contiguous intervals where each interval is represented by its own polynomial and written as

$$f(x) = \sum_{i=1}^{P} \beta_i h_{i,m}(x) \tag{3.1}$$

where β_i is the ith basis coefficient, $h_{i,m}(x)$ is the ith basis function for a spline of order m and P is the number of basis functions, decided by the order of the spline and the number of knots R. The basis functions of traditional regression splines can result in serious rounding problems when x is large. Instead of the basis functions $h_{i,m}(x)$ in equation 3.1, B-splines $B_{i,m}(x)$ can be used.

B-spines, or basis splines, use an alternative and more computationally stable basis representation for regression splines (Hastie et al., 2009). When working with B-splines, additional knots are chosen outside of the two boundary knots ξ_0 and ξ_{R+1}. These additional knots are arbitrary and often chosen to have the same values as the boundary knots, which are usually the boundaries of the domain of x. The new knot sequence τ is thus:

- $\tau_1 \leq \tau_2 \leq \ldots \leq \tau_M \leq \xi_0$
- $\tau_{r+M} = \xi_r,\ r = 1 \ldots . R$
- $\xi_{R+1} \leq \tau_{R+M+1} \leq \tau_{R+M+2} \leq \ldots \leq \tau_{R+2M}$

The B-spline basis functions are found recursively as shown in equations 3.2 and 3.3. $B_{i,m}(x)$ is the ith B-spline basis function of order m for τ where $m \leq M$ (Hastie et al., 2009).

$$B_{i,1}(x) = \begin{cases} 1 & if\ \tau_i \leq x < \tau_{i+1} \\ 0 & otherwise \end{cases},\ for\ i = 1, ..., R + 2M - 1 \tag{3.2}$$

$$B_{i,m}(x) = \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x), \quad for\ i = 1, ..., R + 2M - m \tag{3.3}$$

Smoother functions are often preferred which is achieved by increasing the order of the B-splines. For an order-4 B-spline, also known as a cubic B-spline, the human eye can no longer notice any discontinuity at the knots (Hastie et al., 2009).

Using B-splines and a fixed number of knots, equally spaced throughout the time series, an alternative representation of the time series is found for clustering. A time series can thus be represented by a vector consisting of the spline coefficients $\beta_1, \ldots, \beta_{R+m}$ and the length of the time series s.

3.1.2 Gaussian Mixture Model

As stated earlier, the prediction of a cluster can be used to strengthen the prediction of observations where the individual predictions are uncertain. The Gaussian

mixture model has the advantage of producing clusters described as probability distributions which can be used in finding posterior probabilities, used for predictions.

The assumption behind a mixture model is that the data is a set of objects from a mixture of different probability distributions: the clusters. Usually, these distributions are of the same type but with different parameters. Let $C_1, \ldots, C_K$ be the probabilistic clusters and (β, s) be the representation of the time series to cluster, then the mixture model can be expressed as

$$p\left((\beta, s)\right) = \sum_{k=1}^{K} p(C_k) p((\beta, s)\,|C_k) \tag{3.4}$$

where $p\left(C_k\right)$, usually denoted π_k, are the mixing coefficients and $p\left((\beta, s)|C_k\right)$ are the mixture components. The generative process of a mixture model has two steps: first, choosing the cluster based on the cluster probabilities and second, generating an object according to the density function of the cluster (Bishop, 2006; Tan et al., 2014). The most popular mixture model is a mixture of Gaussians where each cluster is described in terms of a Gaussian density. The clusters have the form of ellipsoids, centered around a mean vector and the shape, volume and orientation are decided by the covariance matrix of the cluster (Han et al., 2012; Scrucca et al., 2016). The K clusters follow the multinomial distribution where π_k is the mixing coefficient of cluster C_k.

$$C \sim Multinomial\left(\pi_1, \ \ldots, \ \pi_K\right) \tag{3.5}$$

The time series have been given an alternative representation with B-splines as explained in subsection 3.1.1. If β is a vector containing the spline coefficients of a time series and s is the length of that time series, cluster C_k is defined as

$$(\beta, s)\,|C_k \sim \mathcal{N}\left(\mu_k, \Sigma_k\ \right) \tag{3.6}$$

where μ_k and Σ_k are the mean vector and the covariance matrix of the kth cluster. The expectation maximization algorithm, described in subsection 3.1.2.1, is used to find the maximum likelihood estimations of the model parameters.

3.1.2.1 EM Algorithm for Gaussian Mixture Models

The Expectation Maximization (EM) algorithm is a popular alternative for solving maximum likelihood problems for models depending on latent variables, such as mixture models where the nth data observation $(\beta, s)_n$ does not contain information

from which cluster it comes. A new K-dimensional binary latent variable $\mathbf{z_n}$ is introduced where all elements of $\mathbf{z}_n$ are 0 except for the kth element which has the value 1 and thus linking the observation n to cluster k (Bishop, 2006; Hastie et al., 2009).

The EM algorithm is an iterative process comprised of two steps: Expectation (E) and Maximization (M). In the expectation step, the expected log-likelihood function for the complete data set is found, based on the posterior probabilities of the latent variables with the current parameter estimations. In the maximization step, the parameters are re-estimated to maximize the expected log-likelihood from the E-step. For each iteration of the EM algorithm, the log-likelihood is increased and the algorithm is usually stopped when the change in log-likelihood or parameters is below a threshold (Bishop, 2006).

Algorithm 1: EM algorithm for Gaussian mixture model

1. Initial guesses for μ_k^0, Σ_k^0 and π_k^0
2. E-step: for iteration j, evaluate the posterior probability that cluster k was responsible for observation $(\beta, s)_n$ with the current parameter estimations

$$p\left(z_{nk}\middle|(\beta,s)_n, \mu_k^j, \Sigma_k^j, \pi_k^j\right) = \gamma\left(z_{nk}\right) = \frac{\pi_k^j p((\beta,s)_n | \mu_k^j, \Sigma_k^j)}{\sum_{i=1}^K \pi_i^j p((\beta,s)_n | \mu_i^j, \Sigma_i^j)} \tag{3.7}$$

3. M-step: update the parameter estimations with the responsibilities from 3.7

$$\mu_k^{j+1} = \frac{1}{N_k}\sum_{n=1}^N \gamma\left(z_{nk}\right)(\beta,s)_n \tag{3.8}$$

$$\Sigma_k^{j+1} = \frac{1}{N_k}\sum_{n=1}^N \gamma\left(z_{nk}\right)((\beta,s)_n - \mu_k^{j+1})((\beta,s)_n - \mu_k^{j+1})^T \tag{3.9}$$

$$\pi_k^{j+1} = \frac{N_k}{N} \tag{3.10}$$

where

$$N_k = \sum_{n=1}^N \gamma\left(z_{nk}\right) \tag{3.11}$$

4. Iterate step 2 and 3 until convergence is reached

In algorithm 1, the EM algorithm is presented for a Gaussian mixture model using B-spline coefficients and length to represent the N time series, where $(\beta, s)_n$ is the representation of the nth time series. The responsibilities in equation 3.7 are part of the expected log-likelihood of a Gaussian mixture model.

$$E_{\mathbf{z}}\left[\ln p\left((\beta, s)\,|\pi, \mu, \Sigma\right)\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{\ln \pi_k + \ln p\left((\beta, s)_n\,|\mu_k, \Sigma_k\right)\right\} \tag{3.12}$$

By keeping the responsibilities fixed and maximizing equation 3.12, the parameter estimations in equations 3.8, 3.9 and 3.10 are found (Bishop, 2006).

3.1.3 Prediction

The last step of the probabilistic classifier makes it possible to predict the class of a new time series, without the need of transformation or for it to be complete, while still utilizing the probability distributions found in the clustering step. A probabilistic predictive model predicts the class of an observation with the help of the posterior probability for the classes (Bishop, 2006). A time series classification application of a probabilistic predictive model assumes that the dependent variable Y is modeled as a B-spline with the error terms following a suitable parametric probability distribution, such as the Laplace distribution

$$Y = \sum_{i=1}^{P}\beta_i B_{i,m}\left(x, s\right) + \varepsilon, \quad \varepsilon \sim \mathit{Laplace}\left(0, a\right) \tag{3.13}$$

where x is the time, β_i is the ith B-spline coefficient, $B_{i,m}$ is the ith basis function, P is the number of basis functions and a is the scale parameter of the error term distribution. The splines are assumed to belong to a Gaussian mixture model where each cluster belongs to a class L_v. The clusters make up disjoint sets where each set contains the clusters assigned to a specific class and a cluster can only belong to one class.

To classify a time series of length T, the class L_v which maximizes $p\left(L_v|\mathbf{x}, \mathbf{y}\right)$ is chosen, where $(\mathbf{x}, \mathbf{y})$ are the paired observations of the time series. The first step in finding the posterior probabilities of the classes is to compute the posterior probability of each cluster C_k, which is found with Bayes' theorem

$$p\left(C_k|\mathbf{x}, \mathbf{y}\right) \propto p\left(\mathbf{x}, \mathbf{y}|C_k\right) p\left(C_k\right) \tag{3.14}$$

The likelihood in equation 3.14 can be expressed in an alternative way, using the distributions of equations 3.6 and 3.13:

$$p\left(\mathbf{x}, \mathbf{y}|C_k\right) = \prod_{t=1}^{T} \int p\left(y_t|x_t, (\beta, s)\right) p\left((\beta, s)|C_k\right) d\beta_1 d\beta_2 \ldots d\beta_P ds \tag{3.15}$$

where (β, s) is a spline vector consisting of B-spline coefficients and length. For the case where the error terms are Laplace distributed, the integrals of the likelihood in equation 3.15 are not solvable analytically and must be approximated.

One way to approximate the integrals is by Monte Carlo integration which is an approximate inference method that uses numerical sampling (Bishop, 2006). Looking at the integrals in 3.15, they can be seen as the expected likelihood of (β, s) for the point (x_t, y_t), where $p\left((\beta, s)|C_k\right)$ is the probability density function. By looking at the integrals as expectations, the full likelihood can be approximated by sampling G times as follows:

For $g = 1, ..., G$

1. $(\beta, s)^{(g)}$ is sampled from $\mathcal{N}\left(\mu_k, \Sigma_k\right)$
2. $\ell_{k,t}^{(g)} = p\left(y_t \middle| x_t, (\beta, s)^{(g)}\right)$ is computed for the sampled parameters

The likelihood is then computed as the product of the approximated expectations.

$$\hat{p}\left(\mathbf{x}, \mathbf{y}|C_k\right) = \prod_{t=1}^{T} \frac{1}{G} \sum_{g=1}^{G} \ell_{k,t}^{(g)} \tag{3.16}$$

The posterior probabilities for the clusters are normalized before finally computing the posterior probability of the class L_v

$$p\left(L_v|\mathbf{x}, \mathbf{y}\right) = \sum_{C_k \in L_l} p\left(C_k|\mathbf{x}, \mathbf{y}\right) \tag{3.17}$$

3.2 K-Nearest Neighbors Classification

The idea behind the k-nearest neighbor (KNN) method is that observations of the same class will be more similar to each other than observations of different classes. Thus, the class of an observation can be determined by looking at the neighborhood of that observation. For an unknown observation x_0, the k closest training observations form its neighborhood $N_k(x_0)$. The nearest neighbors are found by

computing the distance between x_0 and all observations in the training data set. Majority voting is then used to decide the class of x_0 as

$$\hat{y}_0 = \underset{\nu}{\operatorname{argmax}} \sum_{(x_i, y_i) \in N_k(x_0)} I\left(\nu = y_i\right) \tag{3.18}$$

where ν is a class label, y_i is the class label of an observation in the neighborhood and $I(\cdot)$ is an indicator function returning 1 if the argument is true and 0 otherwise (Hastie et al. 2009, Tan et al. 2013). KNN is selected as a comparison to the more complex probabilistic model because of its simplicity and because it is intuitive.

3.3 Evaluation

Each step of the analysis is evaluated to assure the best performance and the different evaluation methods used throughout the analysis is presented in the following subsections.

3.3.1 K-Fold Cross-Validation

To evaluate the performance of the B-splines and decide the optimal number of knots, K-fold cross-validation is used. Cross-validation is a model selection method evaluating the fit of a model while at the same time controlling for overfitting. It allows each observation to be used for training the same number of times and as validation once (Han et al., 2012).

The data is divided into K folds, roughly equal in size. The model is trained K times, for $k = 1, \ldots, K$, where for each time the kth fold is used as validation and the other $K - 1$ folds are used as training (Hastie et al., 2009). Let $\kappa\left(i\right)$ be a function indicating the fold k that observation i belongs to. Then the cross-validated prediction error is

$$CV\left(\hat{f}\right) = \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, \hat{f}^{-\kappa(i)}\left(x_i\right)\right) \tag{3.19}$$

where $\hat{f}^{-k}(x)$ is the model fitted without the kth fold, $L\left(y, \hat{f}\left(x\right)\right)$ is the prediction error and n is the number of data points. When tuning parameters, such as the number of knots in a B-spline, are to be fitted, the cross-validated prediction error is

$$CV\left(\hat{f}, R\right) = \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, \hat{f}^{-\kappa(i)}\left(x_i, R\right)\right) \tag{3.20}$$

where $\hat{f}^{-k}(x, R)$ is the model trained with the kth fold left out and tuning parameter R. For a data set consisting of N time series, an overall cross-validated prediction error is computed from the CV-scores of the individual time series

$$CV^{TOT}\left(\hat{f}, R\right) = \frac{1}{N}\sum_{j=1}^{N} CV_j\left(\hat{f}, R\right) \tag{3.21}$$

and the optimal tuning parameter $\hat{R}$ is found by minimizing $CV^{TOT}\left(\hat{f}, R\right)$.

3.3.2 Bayesian Information Criterion

To determine the number of clusters of the mixture model and the best fit of the parameters, the Bayesian information criterion, BIC, is used. BIC can be used to evaluate models where the estimation of parameters is achieved by maximizing the log-likelihood and is a popular choice when to evaluate Gaussian mixture models. This criterion is a measure of how well a model fits the data, but it also penalizes the model for complexity. The general form is

$$BIC = -2\text{loglik} + \log(N) \cdot D \tag{3.22}$$

where loglik is the log-likelihood of the model, N is the sample size and D is the number of estimated parameters. BIC favors simpler models and more complex models tend to be penalized more heavily. The model selected is the one with the lowest BIC value (Hastie et al., 2009; Scrucca et al., 2016).

3.3.3 Accuracy, Precision, and Recall

To evaluate the performance of the classifiers, three measures will be used: accuracy, precision, and recall (Han et al., 2012). These three measures will be explained in this section with the help of a confusion matrix from the binary classification case, with classes positive and negative, seen in table 3.1.

Accuracy is chosen because it describes how good the model is at recognizing the correct classes overall. Accuracy is the proportion of accurately classified observations and is calculated as follows

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.23}$$

Table 3.1: Confusion matrix of the binary case to explain accuracy, precision, and recall

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Accuracy treats the classification of both classes as equally important. That is why, in addition to accuracy, the two measures precision and recall are used. They respectively measure the two types of mistakes made in binary classification, false positives and false negatives, in relation to the correctly classified positive cases. Precision is a measure of exactness, measuring the proportion of correctly classified observation among the positive predicted

$$Precision = \frac{TP}{TP + FP} \tag{3.24}$$

Recall is the proportion of correctly classified positive cases. It describes how well the classifier finds the positive observations by the following formula

$$Recall = \frac{TP}{TP + FN} \tag{3.25}$$

3.4 Technical Aspects

The R-package `mclust` is used to train and evaluate the Gaussian mixture model (Scrucca et al., 2016). The function `Mclust` tries out different numbers of clusters as well as different structures of the covariance matrices to try different models where the shape, volume, and orientation can be either equal or vary between the clusters. All the models are evaluated with BIC. As the package computes BIC differently, the resulting BIC scored are multiplied with -1 to obtain results on the form of equation 3.22.

4 Results

This chapter presents the most important results of the analysis. The result from choosing a definition of fast-selling options is presented in the first section with an alternative definition to be found in section 4.3.2. The second, third and fourth section presents the results from the different steps of the probabilistic model: B-splines, Gaussian mixture model, and prediction. The probabilistic model is evaluated at each step. Lastly, the results from a 1-nearest neighbor classifier are presented.

4.1 Definition of Fast-Selling Options

The data from the commissioner is unlabeled which means that in order to build and evaluate classification models, a definition of fast-selling options is needed. A fast-selling option should be selling larger volumes during the beginning of its sales period and to capture the options with this trait, the cumulative relative sale is of interest since it captures how much has been sold from the first sale until a given day. Histograms of the cumulative relative sale of the options are examined where each histogram presents the cumulative sale at different numbers of days. Figure 4.1 shows the distribution of the proportion of intake sold until the 50th day.

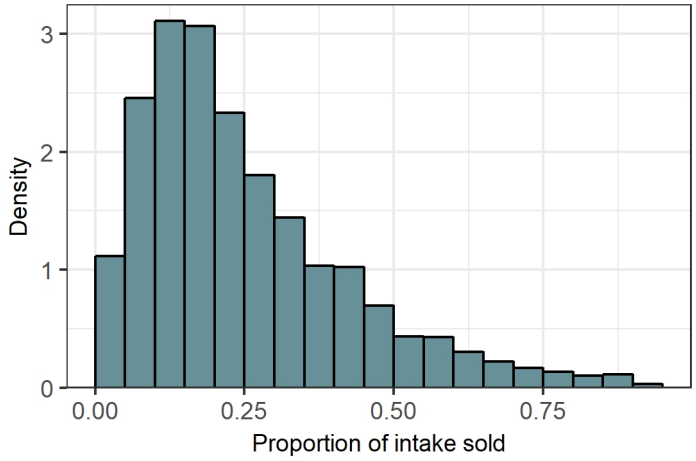

Figure 4.1: Distribution of the proportion of intake sold until the 50th day for seasonal options 2016-2018

The distribution in figure 4.1 is right skewed and unimodal with the majority of the options having sold less than 25 percent by day 50. From 50 percent of intake

sold and up, the distribution flattens and these are the options that have sold large quantities, relative to their intake, during a short period of time and is thus meeting the requirements of a fast-selling option. Figure 4.2 presents all time series, colored according to whether they sold at least 50 percent of their intake by the 50th day or not.

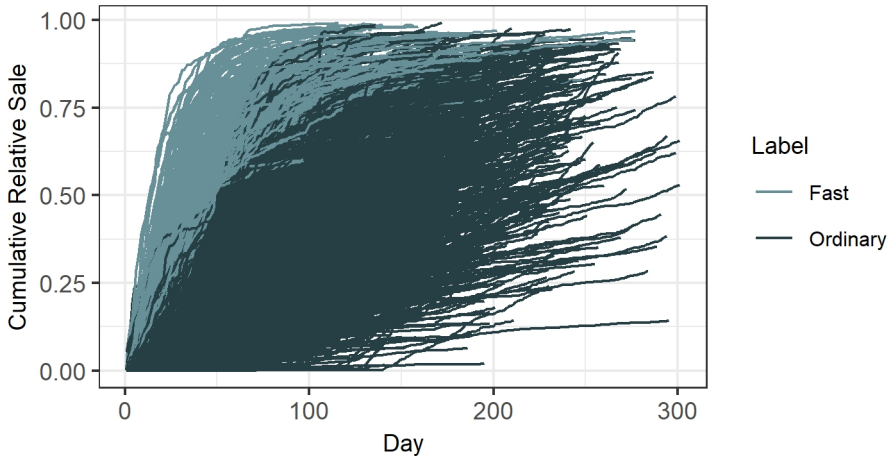

Figure 4.2: Cumulative sale over time for seasonal options 2016-2018, colored by class label

With the definition of a fast-selling option being an option having sold at least 50 percent of its intake by the 50th day, figure 4.2 shows that this definition capture the time series in the top left corner, the ones selling large percentages of its intake in the beginning, but there are some overlapping of time series of the two classes. This partition into fast-selling options and ordinary options results in 9.7 percent of the options being fast selling.

4.2 B-Splines

Cubic B-splines are fitted to the relative sale, as a function over time, for each option in the training set in order to transform the time series to reduce dimensionality. This section first presents the evaluation of the number of knots to be used in the B-splines and then evaluate the fit of the B-splines by showing the time series of three options with their fitted B-splines.

4.2.1 Evaluation of Number of Knots

Figure 4.3 presents the CV-scores for different numbers of knots where the absolute error is chosen as the prediction error. The knots are placed with equal space, depending on the number of knots and the length of each time series.

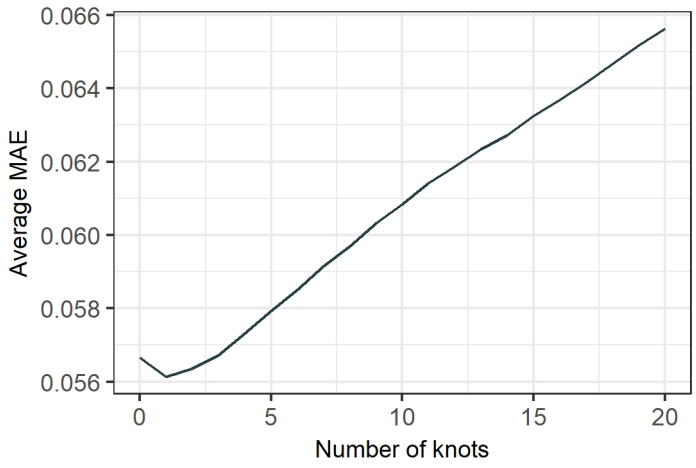

Figure 4.3: Cross-validated errors for the number of knots in B-splines

The average mean absolute error decreases from zero knots to one knot but increases after that with a growing number of knots, as seen in figure 4.3. The average MAE is very low for every number of knots and is explained by the relative sale having the domain $[-1, 1]$ but values over 0.05 or under –0.01 being very rare. The lowest CV-score is achieved with one knot and the fit of B-splines with one knot is evaluated in the next subsection.

4.2.2 Evaluation of Fit

Figure 4.4 shows the relative sale over time for three options and the fitted values of their respective B-splines. The B-splines do not follow the original time series exactly, disregarding the daily fluctuations, but instead capture the overall sales trend for each option. The options display different sale patterns with the top one having a long and slow sale and the other two with shorter and faster sales. The option on the bottom has a faster pattern in the beginning while the middle option sells more during the middle of its time series. This can be seen more clearly in figure 4.5 where the cumulative relative sale over time is presented.

Looking at figure 4.5, the B-splines seem to capture the sales trend with the lighter line, showing the original time series, and the darker line, showing the fitted values, never deviating far from each other with constant overlapping. For all three options, the most obvious deviations from the original data are when a rapid temporary growth in sales occurs but it is recovered quickly and the fitted lines exhibit patterns very close to the actual sale trends. Because the fit of the B-splines with one knot follows the sales trend, the following sections with clustering and classification will use this number of knots for the B-splines.

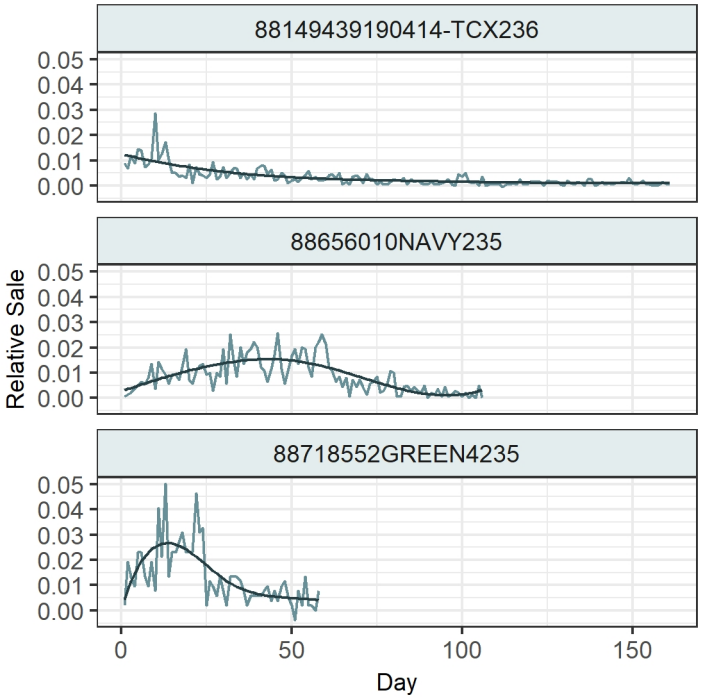

Figure 4.4: Relative sale over time for three options with fitted values from B-splines using one knot

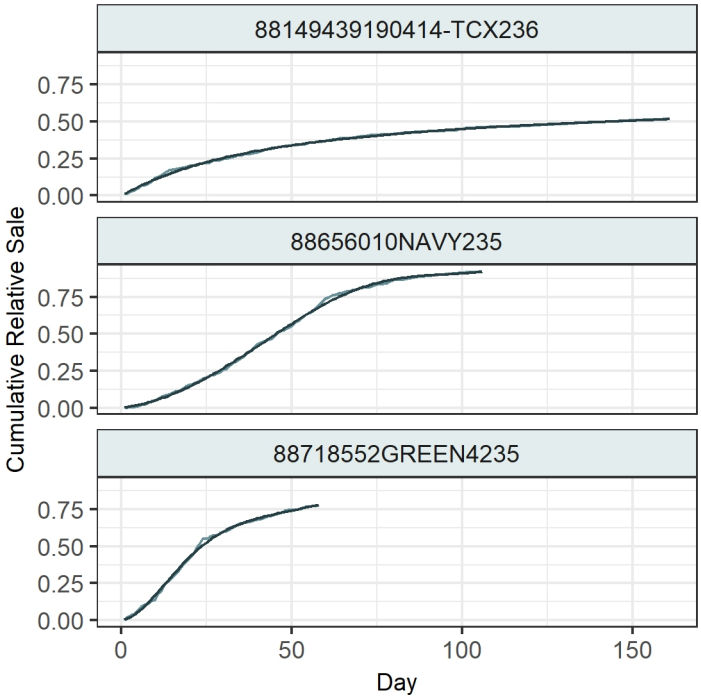

Figure 4.5: Cumulative relative sale over time for three options with fitted values from B-splines using one knot

4.3 Gaussian Mixture Model

As mentioned in section 4.3, the time series are clustered to produce more robust predictions. Using B-splines with one knot to model the relative sale, the coefficients and length of the time series form a vector, used to represent the relative sale of each option from the training data set in a Gaussian mixture model. An evaluation process of the clustering is performed ten times with different seeds. For every round, eight to twenty clusters are tested and for each number of clusters, different designs of the covariance matrices are evaluated. The parameter estimations are found with the EM algorithm and from all models tested during a run, the one with the lowest BIC is chosen as a representative and compared to the other runs. The lowest BIC of each run and the corresponding model's number of clusters are presented in table 4.1.

Table 4.1: BIC-score and number of clusters for the best model of each seed

Run	**BIC**	**Number of Clusters**
1	-71902	15
2	-71878	9
3	-71800	11
4	-71859	10
5	-71791	10
6	-71838	9
7	-71805	9
8	-71827	15
9	-71771	10
10	-71874	9

The top models, seen in table 4.1, have between 9 and 15 clusters and all BIC values are rather similar with a difference of 131 between the lowest and the highest score. The model with the lowest BIC was found with the first seed. This model has 15 clusters and will be further evaluated in the following subsection.

4.3.1 Evaluation of Clusters

For each cluster, a prototype is derived from its mean vector. These prototypes describe the mean behavior of the cluster and can be found in figure 4.6. The number after each line shows which cluster the prototype describes.

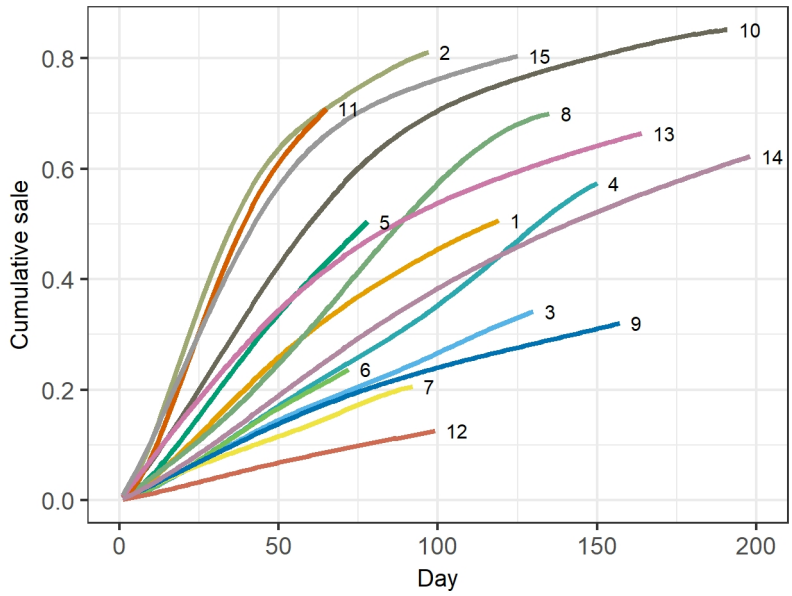

Figure 4.6: Cluster prototypes describing mean cumulative sale over time in a cluster

The cluster prototypes in figure 4.6 show that three of the clusters (2, 11 and 15) have a mean sale profile with a cumulative sale over 50 percent by the 50th day. These clusters are given the class of fast-selling since their prototypes fulfill the definition of a fast-selling option found in section 4.1. The clusters with the slowest prototypes are number 7, 9 and 12 selling less than 20 percent during the first 75 days. Noticeable is that the clusters with the slowest cluster prototypes seem to have a more constant sale over time while the clusters with the faster prototypes show a declining pattern towards the end.

Figures 4.7 and 4.8 presents the cumulative sales of the options in the training data assigned to each cluster and colored by class, as defined in section 4.1; fast-selling options are colored lighter and ordinary are colored darker. The clusters that were considered fast selling in figure 4.6 (2, 11 and 15) contain more fast-selling options than the other clusters with a majority of lighter lines. Clusters 1, 3, 4, 6, 7, 9, 12 and 14 do not contain a single fast-selling option which is supported by their prototypes, in figure 4.6, showing the slowest patterns. There are some clusters, not classed as fast selling, containing some fast-selling options. These clusters are the ones with the prototypes closest to the fast-selling clusters' prototypes.

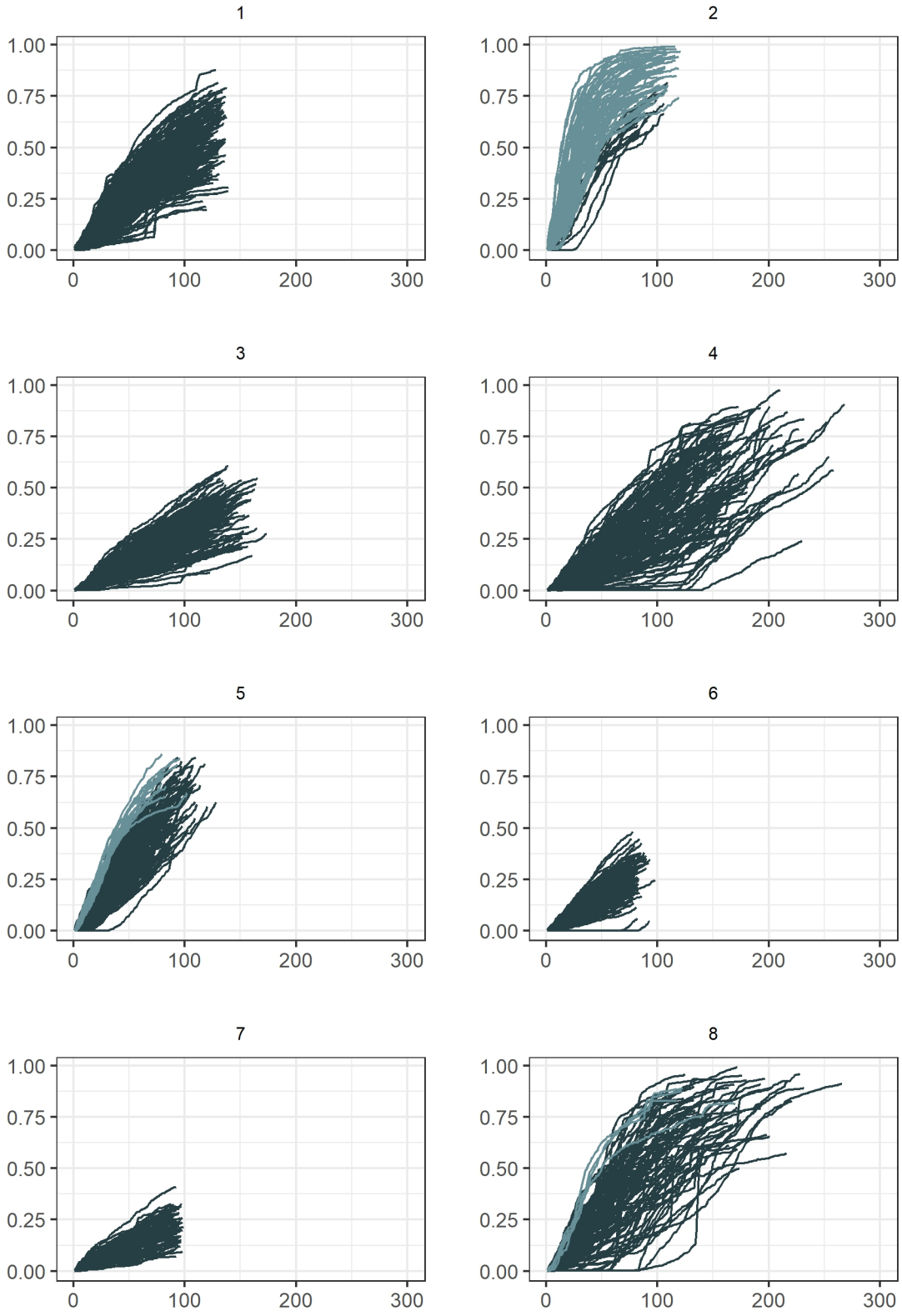

Figure 4.7: Cumulative sales over time for the options in clusters 1-8, colored by class

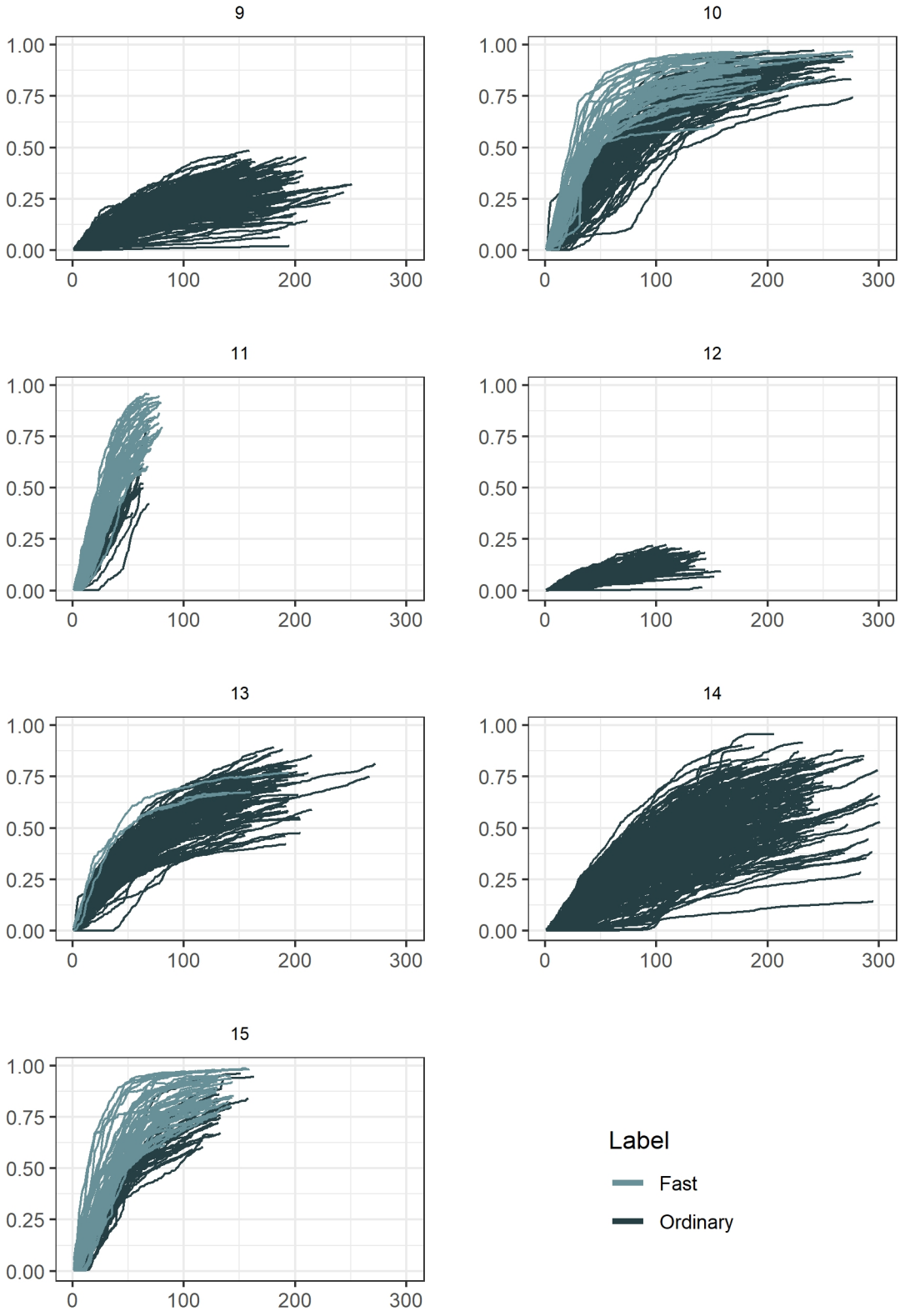

Figure 4.8: Cumulative sales over time for the options in clusters 9-15, colored by class

4.3.2 Alternative Definition of Fast-Selling Options

Besides being a part of the probabilistic classification model to improve predictions, the clustering in the previous subsection is utilized to find an alternative definition of a fast selling-option. This alternative defines an option as fast selling if it belongs to a fast-selling cluster, in this case, clusters 2, 11 and 15. All options belonging to any other cluster are considered as not fast selling by this definition. Figure 4.9 shows the cumulative sales over time for the options, colored by this alternative definition.

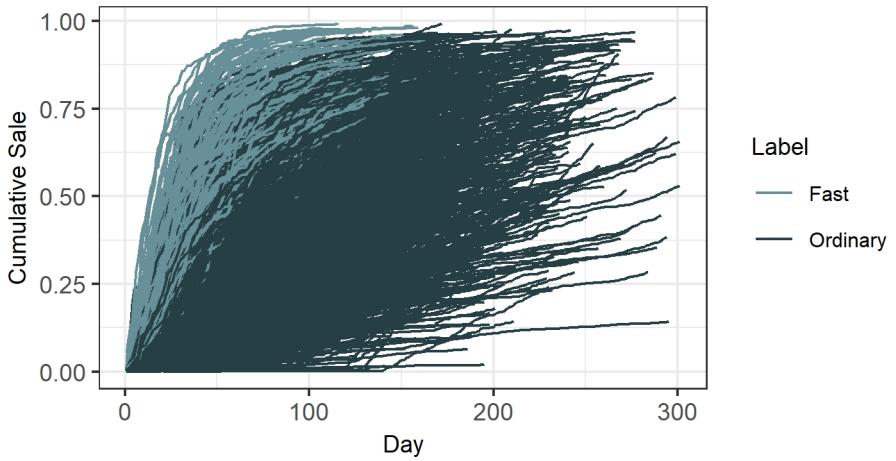

Figure 4.9: Cumulative sale over time for seasonal options 2016-2018, colored by cluster-based definition

The cluster-based definition of a fast-selling option does capture some of the options that are selling a lot in the beginning, as seen in figure 4.9. It does, however, miss some of the options that should clearly be considered as fast selling. Compared to figure 4.2, colored by the definition in section 4.1, the cluster-based definition show more overlapping of the time series of the two classes.

Table 4.2: Comparison of cluster-based definition and histogram-based definition

		Cluster-based	
		Fast	Ordinary
Histogram-based	Fast	168	64
	Ordinary	62	1908

Table 4.2 compares the two definitions. Most of the options are labeled the same way by both definitions and 126 options are labeled differently. Both definitions have approximately the same proportion of fast-selling options. The definition in section 4.1 is used to evaluate the predictions in the following two sections because it

does not miss any obvious fast-selling patterns and there is less overlapping between the time series of the different classes. Even though the clustering is not used as a definition to label the data, it is still used as a part of the probabilistic classifier in order to secure more reliable predictions.

4.4 Probabilistic Classification

This section evaluates the performance of the probabilistic classifier using the optimal number of knots for the B-splines from section 4.2 and the Gaussian mixture model with the lowest BIC from section 4.3. In the first subsection, the assumption of Laplace distributed error terms is explored and in the second subsection, the predictive performance is investigated.

4.4.1 Residual Analysis

Figure 4.10 shows four plots used to evaluate how well the residuals of the B-splines fit a Laplace distribution. The distribution of the residuals is compared to a Laplace distribution with location parameter 0 and scale parameter estimated as the mean absolute deviation from the median.

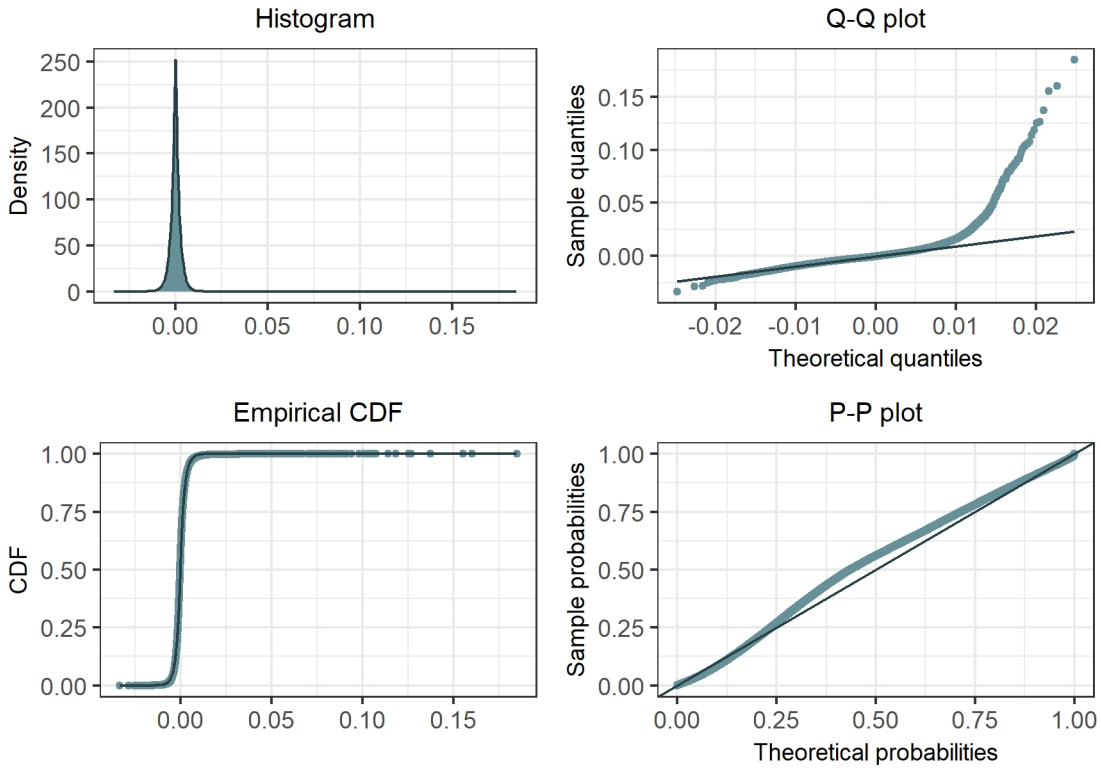

Figure 4.10: Evaluation of the residual distribution with respect to the Laplace distribution

The histogram in figure 4.10 suggests that the Laplace distribution is a good fit since it follows the theoretical PDF rather well, and the same conclusion can be drawn from the empirical CDF which follows the theoretical CDF very well. The quantile-quantile plot, as well as the probability-probability plot, disproves this conclusion by showing too heavy tails, especially the right tail.

4.4.2 Evaluation of Performance

Table 4.3 presents the performance of the probabilistic model on the test data, using different numbers of days of relative sales to make the predictions.

Table 4.3: Results from the probabilistic classification

Number of Days	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
7	93.6	48.0	37.5
14	94.0	51.2	65.6
21	96.3	68.6	75.0
28	97.5	78.8	81.3
35	98.8	88.2	93.8

Table 4.3 shows that all three measures increase with an increased number of days used for the prediction. The highest values for the three evaluation measures are found using the first 35 days of relative sales with accuracy of 98.8 percent, precision of 88.2 percent and recall of 93.8 percent.

Figure 4.11 presents the time series of the test data colored by the predictions made with the probabilistic model using different lengths of the test data. As already seen in table 4.3, the model predicts more accurately and finds more of the fast-selling options when longer time series are used. The lighter lines, showing options predicted as fast selling, are more separated from the darker lines of the ordinary options for 35 days of data than for a smaller time frame, but the probabilistic model starts to capture the fast-selling options already using only 14 days of relative sales.

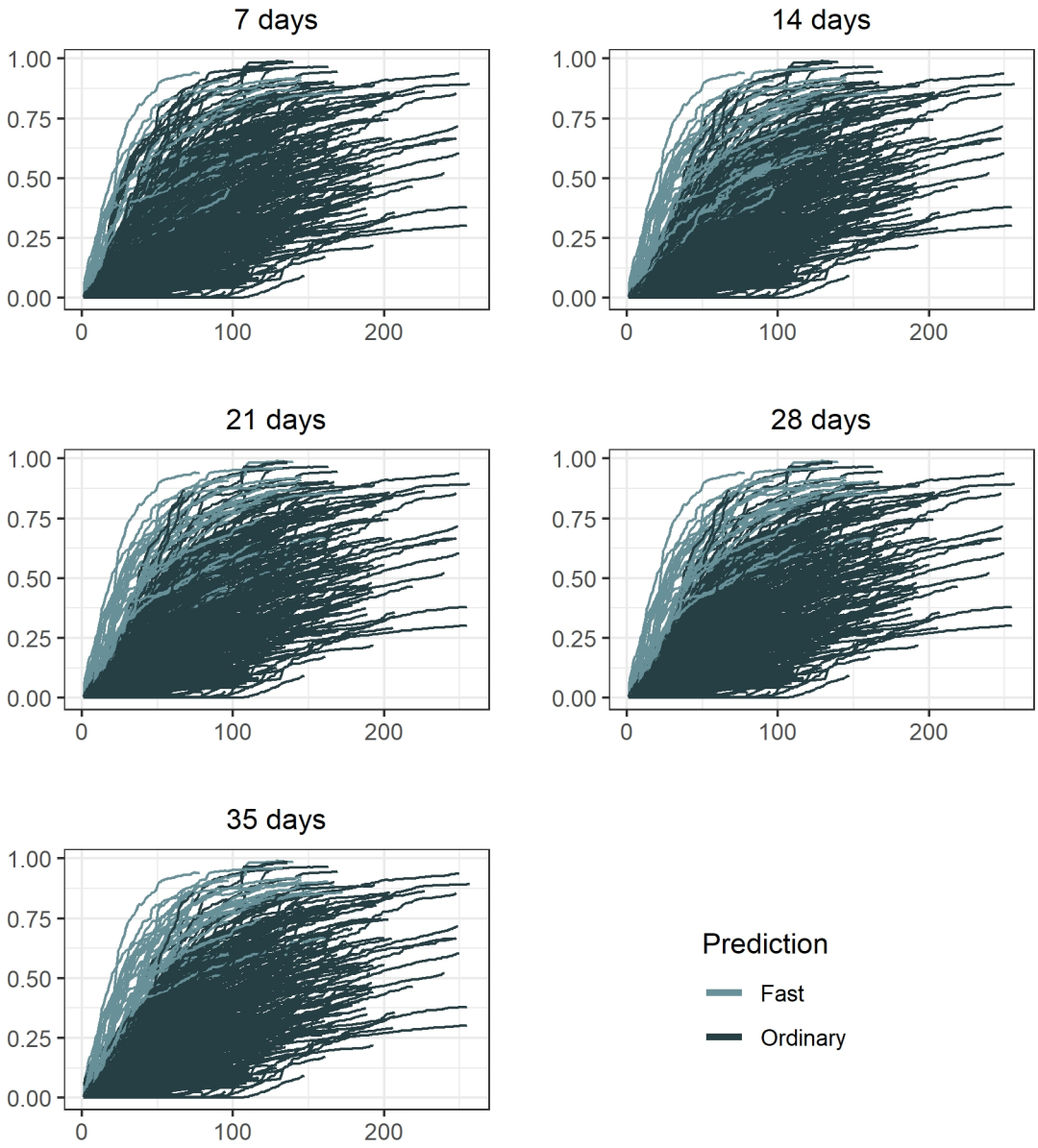

Figure 4.11: Probabilistic classification using different numbers of days for the prediction

4.5 1-Nearest Neighbor Classification

A 1-nearest neighbor classifier using the Euclidean distance is also used to predict the classes of the test data. In table 4.4, the three measures of classifications are presented for five cases, each using a different number of days of the variable relative sales to make the predictions.

Table 4.4: Results from 1-nearest neighbor classification

Number of Days	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
7	84.0	3.6	6.3
14	83.6	1.8	3.1
21	83.4	1.8	3.1
28	88.1	10.5	12.5
35	83.6	9.2	18.8

The results in table 4.4 show that the model performs the best, according to accuracy and precision, when using 28 days of sales data to make the predictions. Looking at the recall, the best performance is achieved by using 35 days of data. The 1-nearest neighbors classifier performs the worst using 14 and 21 days of data for the predictions with very low precision and recall.

Figure 4.12 presents how the test data is classified by the 1-nearest neighbor model, using different lengths of data to make the predictions. The model does not capture the fast-selling options in the test data. For any number of days, the lighter lines describing options predicted as fast selling can't be separated at all from the darker lines showing the other options.

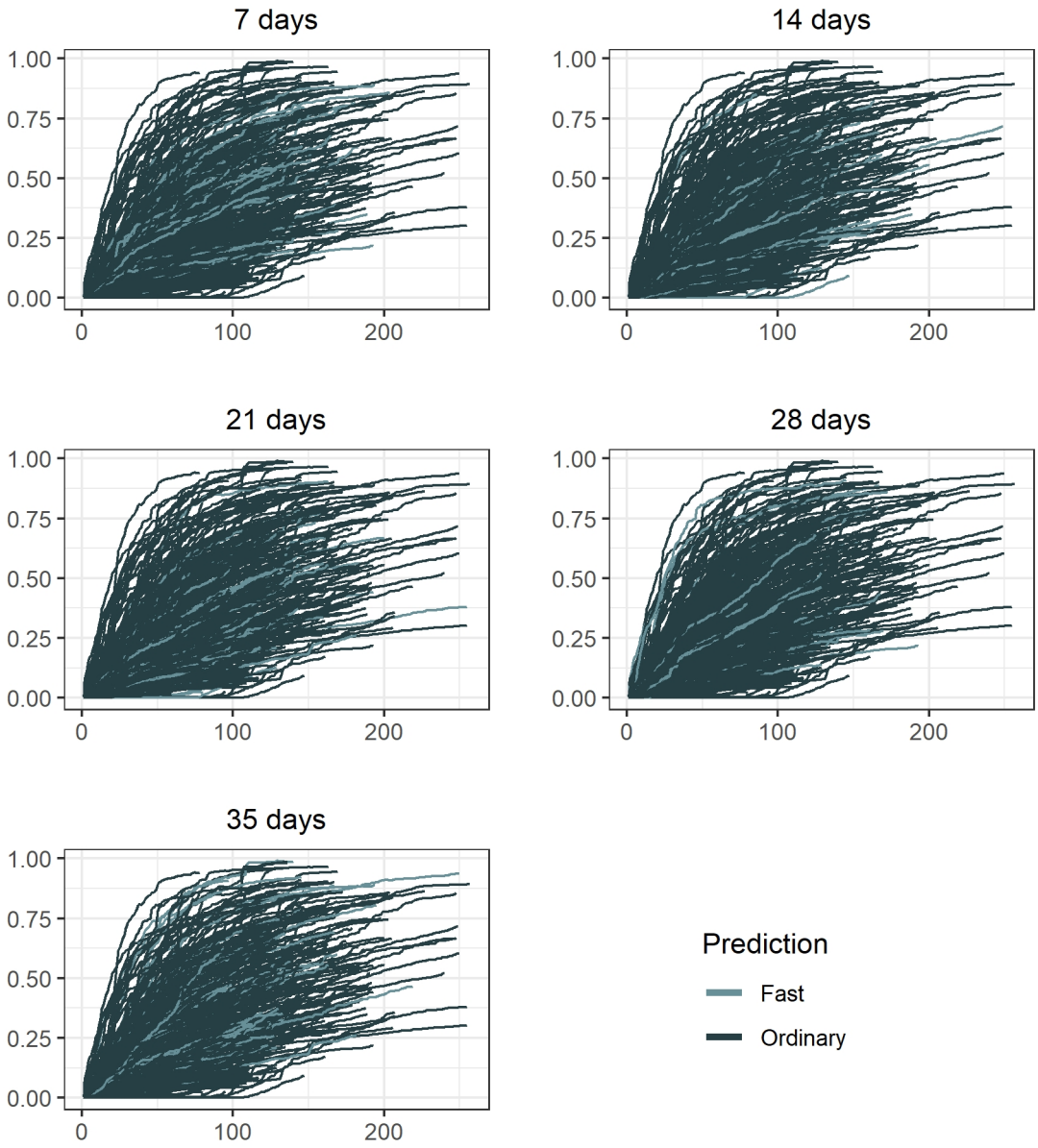

Figure 4.12: 1-nearest neighbor classification using different numbers of days for the prediction

5 Discussion

This chapter discusses the data, methods, and results of the analysis followed by thoughts for future studies and a reflection over the ethical aspects of the study.

5.1 Data

The quality of the data is considered to be good. Sales are automatically registered from the stores which improves the data quality, but there are still some mistakes made in the reporting of sales where items are registered as another option. Some of these mistakes are easily found and removed such as sales marked as unknown item and reported sales of an option long before phase start. More troubling are the sales where an option has been registered as another option, with both options being for sale at the same. Though these mistakes are regarded as unusual and dismissible.

The options for sale less than 50 days before clearance are considered too short to be useful in the analysis with not enough data to determine whether they are fast selling or not and none of the options with such short time series showed any signs of fast-selling trends. The decision to dismiss options with time series longer than 350 days, after all clearance has been removed, is motivated by the fact that these few options display highly unusual patterns and can be seen as outliers that would negatively affect the analysis. It is also improbable that a seasonal option is for sale almost a year without clearance.

The data is aggregated to the option level which gives no insight to sales at SKU level. This hides the fact that the same option very well could be considered as fast selling for some sizes and ordinary for other sizes. In the thesis, the models are aimed to find options that in general sell large quantities in all sizes, as requested by the commissioner, which makes the aggregated data appropriate. Relative sales are used instead of the quantity sold each day. This decision is made since the intake quantity to warehouse varies from option to option. By using relative sales, the percentage sold is explored which makes it possible for options with any intake quantity to be a fast seller.

In the data set collected from the commissioner, the variable *Quantity* describes the difference between the number of units of an item sold and the number of units returned for a sales type during a specific date. The objective of the thesis is to be able to predict fast-selling options in order to prevent stockouts at the

commissioner's larger sales units since customers react negatively to items being out of stock according to Kim and Lennon (2011). When a unit of an option is returned it is offered for sale again, except when it has been returned because of damage, which means that while a purchase of a unit decreases the stock level of an option, a return increases the stock level. By including returns as negative sales, the variable *Quantity* can be said to describe the change in stock level which is of interest when planning the distribution of the restocking quantity. Leaving out the returns might be misleading when looking at how many units that have actually left the stores. If an option is selling large quantities at the start but there are also many returns, the option should not be seen as fast selling because the stock levels are not changed to a degree where special care needs to be taken regarding the distribution of the restocking quantity. Additionally, returns are usually small with at most only a few units of an option returned during a day and does not greatly impact the overall sales trend.

The created variable *Day* is describing time in the form of the number of days since the first registered sale and not for how many days an option has been available for sale. This could affect the definition of a fast-selling option. An option that does not sell any units during its first weeks for sale but then starts selling very large quantities will be considered to be a fast seller because the time starts at first purchase even though it did not sell anything in the beginning. In the data, only sales are registered and thus, days without any sales before first purchase are not visible. But, this scenario is deemed rare since an option that has been in store for several weeks without any sales usually would not start selling at a very high pace. Thus, *Day* is not believed to lead to a misrepresentative definition of fast-selling options.

The partition into training and test sets is not made by random assignment but instead, a date is used. This partition is chosen because it gives the highest chance of using the options with complete time series as training data. The options in the test data have still had a chance of being for sale at least 90 days but might not have sold out yet, i.e. their time series are not guaranteed to be finished but are still long enough to establish whether they are fast selling or not.

5.2 Methods

The method for finding a definition of fast-selling options by looking at the percentage sold after a set amount of days is simple but rather effective, which is its strength. By finding this type of threshold, it is guaranteed that all fast-selling options have sold at least a decent proportion of its intake during a limited time period from their first sale. A downside is that the cumulative relative sale is considered at only one point in time. This assumes that the options reaching the threshold will behave similarly to each other before and after the chosen day. It is also a challenge to set the day and proportion of intake sold for this definition. An alternative

definition of fast-selling options is found with the clustering. This definition is not as simple and presumes that the clusters are coherent and contain exclusively fast sellers or not. The method eliminates the problems with having to choose a specific day and percentage of sold units, but with more diverse clusters, this method provides a less than satisfactory definition as seen by the results of subsection 4.3.2. The cluster-based definition might also lead to overfitting since the class variable for both training and test data would be constructed by the model.

In Abraham et al. (2003), B-spline coefficients and k-means clustering were presented as a way to group similar time series of the same length together. This thesis expands on that framework to include time series of different lengths by representing a time series as B-splines coefficients for a fixed number of knots and the length of the time series. This proves to be a sufficient representation to be used in a Gaussian mixture model for clustering, of which k-means is a special case. Using a GMM, the clusters are allowed to vary in size and shape, but they are still limited to ellipsoidal shapes. Using the B-spline representation and the Gaussian mixture model also makes it possible to build a probabilistic classifier where posterior probabilities determine the class of a new option. The error terms of the B-splines are assumed to be Laplace distributed for the probabilistic classifier in this thesis. The model itself does not rely on this assumption; any parametric probability distribution could be used as long as it is a fair assumption according to the problem at hand. Using a parametric probability distribution simplifies the computations of posterior probabilities and the likelihood can be approximated with Monte Carlo integration.

Sales are seasonal, and there is also a weekly sales pattern with higher sales during e.g. Saturdays. Even though these patterns occur, the models chosen in this thesis do not take that into consideration. This decision is made because the aim of the analysis is not to make daily predictions and explain these patterns. The overall sale trend of the options is what is of importance for finding fast-selling patterns, which motivates the choice of using spline coefficients, and the length of the time series, to represent an option. This choice also reduces the dimensionality of the data considerably, where time series of between 50 to 350 points are each represented by a vector of only six elements when using one knot.

When looking at all the time series plotted together as in figure 4.2, there is no obvious occurring partition of options into separate groups based on their sale pattern over time. This might suggest that the data should not be clustered, but the use of clustering is motivated by the clustering itself not being the aim of the thesis but instead used to strengthen predictions as seen in the studies by Thomassey and Fiordaliso (2006), Thomassey and Happiette (2007) and Kumar and Patel (2010). The parameter estimations of the evaluated Gaussian mixture models are found with the EM algorithm. A problem with this is that the EM algorithm does not always find the global maximum, it only promises to find a local maximum (Bishop, 2006). To increase the chances of finding the global maximum, the evaluation process of the clustering was run ten times, putting forward ten models to be compared with each other.

K-nearest neighbors is intuitive since fast-selling options should have a similar relative sale to each other, but it is a lazy learner and is time-consuming with a large training set. A problem with using Euclidean distance and KNN on this type of data is that there is daily fluctuation in sales and a weekly and monthly pattern that the model does not consider. This might lead to weekday of first sale and day-of-the-month for the first sale being more important for similarity rather than overall high or low sales. More sophisticated distance measures might solve this problem but would also increase the risk of even longer computation times. KNN is easy to update since more observations can be added to the training set but with a growing training set, the time to predict one option is increased.

Both methods of classification, probabilistic and KNN, have the advantage of being flexible with the number of days of relative sale for options to be predicted. The predictions improve with more data, but the commissioner can themselves decide how much data they want to use without the need to change the predictive models. An advantage of the probabilistic model is that it is not as dependent on prelabeled data as KNN. The transformation and clustering are completely executed without labels for the options with the class labels being mainly used for evaluating the clustering and the predictive performance. In deciding the clusters that are to be considered fast selling, the prototypes can be used without the knowledge if their clusters contain mainly fast-selling options. This makes the probabilistic method able to classify time series without class labels by only assigning classes to the clusters.

Three measures are used together to evaluate the classification where accuracy gives an overall measure and precision and recall each focus on the two types of mistakes. Precision and recall can be combined to a single measure, called the F_1-score, by computing the harmonic mean of the two (Tan et al., 2014). For this thesis, the F_1-score was deemed inappropriate because it treats the two measures as equally important. In this case, the recall is more important than the precision since to wrongly classifying a fast-selling option is considered worse by the commissioner than misclassifying an ordinary option. Thus, the two measures are best kept as they are and not combined. Even though recall is the most important measure in finding fast selling options, precision is important as well. There needs to be a balance between the two. In a scenario with a model where the recall is 1 but the precision is below 50 percent, all fast-selling options have been found but they are hidden since more ordinary options have been predicted as fast selling. Such a model would not be helpful in distinguishing the two classes from one another.

5.3 Results

When finding the definition of fast-selling options, as presented in section 4.1, the cumulative relative sale at different days is explored and for each day different thresholds are tested. The best definition is found by using the cumulative relative sales day

50 and the threshold 50 percent. This definition is reached by examining histograms and time series plots of the different definitions, much like the figures presented in section 4.1. The result is not perfect with the two classes overlapping at times, but it is still considered good because it captures all of the options that are selling a lot in the beginning. It is considered better than the cluster-based definition in subsection 4.3.2, which does not capture all options with fast-selling patterns and the overlapping of the two classes is more severe.

Both regular regression splines, with truncated power basis functions, and B-splines are tested to see which alternative gives better results with the clustering. For the two types of splines, both fixed number of knots and fixed knot positions are explored. The most cohesive clusters are found with B-splines, using one knot placed in the middle of each time series. The other spline alternatives produce even more mixed clusters than those seen in figures 4.7 and 4.8. The number of clusters is limited to between eight and twenty in the search for the best model in an effort to speed up the computation time. This choice is based on preliminary tests of the `Mclust` function where less than eight clusters were never chosen as the best model and more than twenty clusters gave a high risk of models with empty clusters i.e. clusters containing zero options from the training data.

In subsection 4.3.1, a relationship between the cluster prototypes and the percentage of fast-selling options in a cluster is discovered. This suggests that even though the clusters are diverse, with time series somewhat different to each other and the prototype, the prototypes are a useful description of a cluster in terms of being fast selling or not. Among the prototypes, depicted in figure 4.6, the fast-selling clusters show a declining growth in sales that is not visible for the clusters with the prototypes showing the slowest patterns. This difference is likely explained by the fact that when options reach a certain level of cumulative relative sale, it starts running out of stock in some sizes and at some locations, reducing the availability of the option for the customers and resulting in lower sales. For options that are not selling well, running out of stock is not too much of a problem and the availability is not affected, which explains the more constant growth of these prototypes. By being able to predict fast-selling options, the decline in sales, because of options not being available, might be prevented by distributing the restocking quantity wisely with this information in mind.

In the analysis of the residuals in subsection 4.4.1, the assumption of Laplace distributed error terms is questioned. The distribution of the residuals shows a very heavy right tail which is the main point against the assumption of the model. The heaviness of the right tail is caused by occasional spikes in sale for the options which are not matched by dips in sales of the same degree. These spikes can be caused by campaigns, special occasions or factors out of the commissioner's hand. More complex distributions might be able to catch this behavior but for this thesis, the Laplace distribution is found a good enough fit for the residuals since it is the distribution, among the most common parametric probability distributions, closest to the residual distribution.

The probabilistic classifier outperforms the 1-nearest neighbor classifier for all three measures. Using only seven days for the prediction with the probabilistic model still gives higher values for accuracy, precision, and recall, compared to the outcome of the 1-nearest neighbor model for any number of days, which is seen in tables 4.3 and 4.4. By looking at the results in table 4.3 and figure 4.11, the probabilistic classifier based on B-splines and a Gaussian mixture model presents a reasonable and quite effective way to predict fast-selling options. The 1-nearest neighbor classifier has high accuracy even though both precision and recall are low, at best reaching 10.5 percent and 18.8 percent respectively. This is caused by the fact that the class of fast-selling options is considerably smaller than the other class, thus the performance of the model is best explained by precision and recall. The choice of using $k = 1$ in the nearest neighbors model is motivated by larger k resulting in lower values for recall and precision. For k higher than one, the KNN does not manage to predict any of the fast-selling options in the test data correctly when using less than 28 days of relative sale.

5.4 Future Studies

For future studies, it might be interesting to look at the sales curves for each store. The sales are dependent on the item being in stock in the sales units where there is a demand for the product. Thus, the shape of the sales curve is affected by the stock level of the different sales units. An option with a curve showing a slow growth might not necessarily mean that the option could not be considered as fast selling under the right circumstances. The reason could be a situation where the sales units where the option is in demand have received too little stock while sales units where the option is not in demand have received too much stock. The sales curve might have been different if the demand from each sales unit better matched the stock level. Future studies could explore this complex connection and perhaps find fast-selling options on the store level. Another way to further improve predictions in the future could be to use descriptive variables as done in Thomassey (2014) and Sun et al. (2008) since e.g. different garment classes could show different sales patterns.

5.5 Ethical Aspects

The study does not directly raise any ethical questions. No personal data is used, only sales data and intake quantities aggregated in a way so that the purchases of a customer cannot be identified. The only concern would be that the results might be used to favor restocking of the bigger sales units, leaving customers of smaller stores with lower availability. This concern can be discarded since the online store is among the bigger sales units, which makes the products available for all consumers, regardless of geographical location.

6 Conclusions

The objective of the thesis is to find a definition of a fast-selling option and investigate how well these options can be predicted using sales data during a limited time frame.

How can a fast-selling option be defined using information from sales data?

Using sales data from options with sale start from 2016 through 2018, a criterion for fast-selling options is reached by looking at the cumulative relative sale of the 50th day after the first sale. Options that have sold 50 percent or more of their intake is considered as fast selling. This definition captures the options that sell the highest proportions of their intake during the first sales period. An alternative definition based on clustering the options with a Gaussian mixture model is also investigated but proves to be worse at finding all of the options with fast-selling patterns.

How can options be predicted as fast selling or not, based on a few weeks of data?

Two different classification methods are used to predict fast-selling options. The first one is a probabilistic classification method using B-spline coefficients and the length of a time series to represent it, and a Gaussian mixture model to cluster the options. The second method is k-nearest neighbors. The first method, using one knot for the B-splines and 15 clusters, outperformed the second, using Euclidean distance and one neighbor, in the three measures accuracy, precision, and recall but most remarkably in the latter two. Hence the probabilistic model is a more accurate way to predict whether an option will become fast selling or not based on just the first weeks of sales data.

How does the quality of the predictions change when more data is available?

The quality of the predictions for both models is improved when longer sequences of the test data are used for the predictions. The improvement is the greatest for the probabilistic model, which predicts more correctly using only seven days of data than the 1-nearest neighbor classifier does for up to five weeks of data. Recall is the measure improved the most by more data and it is also the measure used to evaluate how well the model finds the fast-selling options. It shows that with more data used, the fast-selling options are found to a higher degree.

Bibliography

Abraham, C., Cornillon, P. A., Matzner-Løber, E., and Molinari, N. (2003). Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics*, 30:1–15.

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, 53:16–38.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York.

Frank, C., Garg, A., Sztandera, L., and Raheja, A. (2003). Forecasting women's apparel sales using mathematical modeling. *International Journal of Clothing Science and Technology*, 15(2):107–125.

Goldfisher, K. and Chan, C. (1994). New product reactive forecasting. *The Journal of Business Forecasting*, 13(4):7.

Han, J., Kamber, M., and Pei, J. (2012). *Data mining : concepts and techniques*. Elsevier/Morgan Kaufmann, Waltham, Mass., 3. ed. edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer, New York, 2. ed edition.

Kim, M. and Lennon, S. J. (2011). Consumer response to online apparel stockouts. *Psychology & Marketing*, 28(2):115–144.

Kumar, M. and Patel, N. R. (2010). Using clustering to improve sales forecasts in retail merchandising. *Annals of Operations Research*, 174(1):33–46.

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):205–233.

Sun, Z.-L., Choi, T.-M., Au, K.-F., and Yu, Y. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. *Decision Support Systems*, 46(1):411–419.

Tan, P.-N., Steinbach, M., and Kumar, V. (2014). *Introduction to data mining*. Pearson, Harlow, 1. ed. edition.

Thomassey, S. (2014). Sales forecasting in apparel and fashion industry: A review. In *Intelligent fashion forecasting systems: Models and applications*, pages 9–27. Springer.

Thomassey, S. and Fiordaliso, A. (2006). A hybrid sales forecasting system based on clustering and decision trees. *Decision Support Systems*, 42(1):408–421.

Thomassey, S. and Happiette, M. (2007). A neural clustering and classification system for sales forecasting of new apparel items. *Applied Soft Computing*, 7(4):1177–1187.